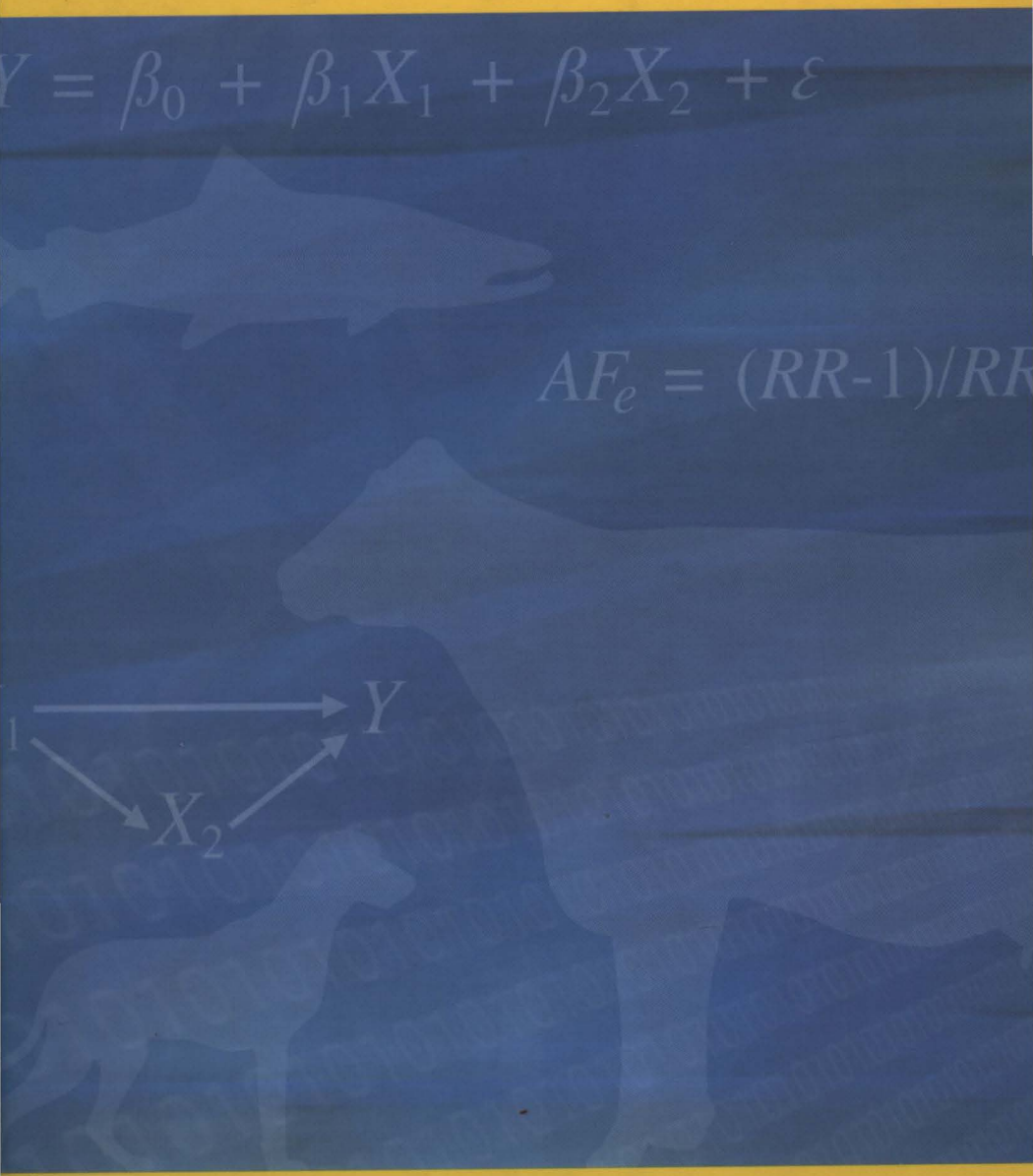
VETERINARY EPIDEMIOLOGIC RESEARCH



Ian Dohoo Wayne Martin Henrik Stryhn

VETERINARY EPIDEMIOLOGIC RESEARCH

VETERINARY EPIDEMIOLOGIC RESEARCH

Ian Dohoo
Professor of Epidemiology
Department of Health Management
University of Prince Edward Island
Charlottetown Prince Edward Island Canada

Wayne Martin
Professor of Epidemiology
Department of Population Medicine
University of Guelph
Guelph Ontario Canada

Henrik Stryhn
Associate Professor of Biostatistics
Department of Health Management
University of Prince Edward Island
Charlottetown Prince Edward Island Canada

Published by AVC Inc. • Charlottetown • Prince Edward Island • Canada

Editor/Proofreader/Compositor/Publication coordinator: S Margaret McPike
Cover designer: Gregory Mercier
Printer: Transcontinental Prince Edward Island

Printed in Canada

10 9 8 7 6 5 4 3 2 1

National Library of Canada Cataloguing in Publication
Veterinary epidemiologic research / Ian Dohoo ... [et al.].
Includes index.
ISBN 0-919013-41-4
1. Veterinary epidemiology--Research--Textbooks.
I. Dohoo, Ian Robert
SF780.9.V47 2003 636.089'44 C2003-904518-8

Dedication

This text is dedicated to all of the graduate students who have challenged and inspired us throughout our careers,
and to our families who have supported us,
especially during the writing of this text

CONTENTS

FOREWORD

Over recent decades epidemiologic thinking and methods have become central to efforts to improve animal health and protect the human population from exposure to animal pathogens.

The range of techniques used has grown enormously, and the level of understanding of the merits and limitations of the various investigational methods has increased as experience has been gained with their application. There has, however, been no single source to which one could turn for an authoritative guide to the major methods of epidemiological research, their application, and the issues which should be taken into account in using them.

With the publication of this book, that important gap in the literature has been amply filled. This is a comprehensive text for the discipline of veterinary epidemiology, written by authors who have the standing to provide wise and insightful guidance on epidemiological research methods for the novice and expert alike. They have provided both a guide to the selection and application of the various investigational and analytical techniques, and practical examples which will allow the reader to test new-found knowledge by working through the example datasets to apply the procedures to genuine case studies.

I am unaware of any other book in either veterinary or medical epidemiology which provides such a solid reference source, backed up by the understanding and wide experience of the authors. This book will become an essential reference for any epidemiologist, regardless of whether they work with human health or animal health, to be read for education and then checked on later occasions to make sure that a concept is right, a method has been correctly interpreted, or a limitation of a technique has been properly considered.

The chapters cover the fundamental principles of epidemiological research, the methods by which studies can be undertaken, and the procedures for analyzing and interpreting the data once it is gathered. The illustrative examples are real-world ones, which will be relevant to the problems being investigated by readers of the book, and the link from text examples to the teaching exercises will help users to move quickly from reading to doing, first with the example datasets and then with their own data.

I compliment the authors on assembling such a powerful tool, which will help users throughout the world to approach epidemiological investigations with enhanced confidence that they are undertaking each part of the investigation in the best possible way.

Professor Roger Morris MVSc, PhD, FAmerCE, FACVSc, FRSNZ, CNZM
Director, Massey University EpiCentre, Palmerston North, New Zealand

PREFACE

Over the past few decades, veterinary epidemiologic research has expanded both in depth and breadth of topics. As educators, in order to help new epidemiologic researchers gain the essential knowledge in the discipline, we have found a great need for a graduate-level text on the principles and methods of veterinary epidemiologic research. There are a number of excellent general veterinary epidemiology texts (*eg* Thrushfield, 1995; Martin et al, 1987; Smith, 1995; Noordhuizen et al, 1997) however, in our view, the material in these texts is insufficient for students learning how to conduct epidemiologic research or as a general reference for current epidemiologic investigators.

The primary motivation for this book came from the fact that over the years we, as teachers of graduate-level courses in epidemiologic research methods, have found it necessary to supplement available textbook material with extensive course notes in a variety of areas. For many of the study design features and analytic methods that we include in graduate courses, it has been our perspective that there are no textbooks that covered the material in a sufficiently comprehensive, yet accessible manner. Specialty textbooks on specific design or analytic methods are available, but are too detailed for students learning the subject material. Even with directed reading of selected journal papers and text material, most students needed a more concise reference for the variety of 'tools' they want in their 'tool-kit'. Although these diverse sources were comprehensive, they did not present the material in a unified framework that would be helpful both for students and researchers already in the discipline.

This text focuses on both design and analytic issues. Concerning issues of study design, we have found that existing textbooks fell into two general groups. There are a number of excellent texts, in addition to the veterinary texts mentioned above, that present the material at a level intended for use by students and health practitioners who are consumers of epidemiologic research results, but not at a level suitable for those actively involved in the design and conduct of comprehensive studies (*eg* Fletcher et al, 1996; Sackett et al, 1991; Hulley et al, 2001). On the other hand, there are a few 'high-end' reference texts that deal with the theoretical and applied bases of epidemiologic research (Breslow and Day, 1980, 1987; Kleinbaum et al, 1982, Rothman and Greenland, 1998). On the personal front, whereas we use these texts extensively, our experience is that graduate students find these texts very challenging to digest as they are learning the discipline. It is our hope that we have covered the major study design issues in Chapters 2 through 13 and have done so in a way that is comprehensible to students learning the discipline but sufficiently complete to serve as a useful reference for experienced investigators.

With respect to helping students learn the multivariable statistical methods used in epidemiology we found that, once again, the literature fell into two classes. A number of general statistics texts provide good introductory information about the more commonly used epidemiologic methods, but do not present the material with a view to their use in epidemiologic research. On the other hand, more specialised texts cover

the statistical material in great detail, but often at a level of complexity that is beyond that which can be used by many investigators that come to the discipline from a health profession background. It is our hope that in Chapters 14 through 24, we have covered the important analytical methods in a manner that is comprehensible to first-time graduate students in epidemiology and to seasoned epidemiologic investigators.

A final motivation for the preparation of this book was that the vast majority of graduate-level reference material has been written for students in human epidemiology and we felt there was a need for the material to be presented in a veterinary context. Although important, this was only a minor motivating factor as the principles of epidemiologic research are identical, regardless of whether our subjects have two legs or four (or none, since we use a number of fish health examples in this text). In fact, it is our sincere hope that students working in human epidemiology, public health research and other related disciplines will also find this text useful, even though some of the diseases discussed in the examples may not be familiar to them.

This book has grown as we have written it. While we have attempted to make it comprehensive, we realise that there are many specific topics within the realm of veterinary epidemiologic research that we have not covered (*eg* analysis of spatial data, risk analysis methodology). While important in many research projects, we felt this material fell outside what we considered to be the 'core' material required by veterinary epidemiologists, and was therefore left out to keep the book at a manageable size.

Throughout the book, but particularly in Chapters 14 through 24, we have made extensive use of examples. All of the datasets used in these examples are described in the text (Chapter 27) and are available through the book's website (http://www.upei.ca/ver). Virtually all of the examples have been worked out using the statistical program Stata™ – a program which provides a unique combination of statistical and epidemiological tools and which we use extensively in our teaching. A listing of the program files (called -do- files by Stata) used in all of the examples is provided in Chapter 28 and these are also provided on the website.

As noted above, the website is an important component of this text. Through it we provide datasets, program files, solutions to sample problems and news items relevant to the book. It is our hope that this will be a dynamic website to which we will add additional material (*eg* more problems and solution sets). In fact, we would encourage other investigators who have useful examples of sample problems to share them with us and we will post them in the relevant section of the website (with appropriate recognition of the contributor).

We hope that you find it useful in your studies and your research.

REFERENCE LIST

1. Breslow NE, Day NE. Statistical methods in cancer research Volume I - the analysis of case-control studies. IARC pub 32. Lyon, 1980.
2. Breslow NE, Day NE. Statistical methods in cancer research Volume II - the design and analysis of cohort studies. IARC pub 82. Lyon, 1987.
3. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology - the essentials. 3d ed. Philadelphia: Lippincott Williams & Wilkins, 1996.
4. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research, 2d ed. Philadelphia: Lippincott Williams & Wilkins, 2001.
5. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. London: Lifetime Learning Publications, 1982.
6. Leech FB, Sellers KC. Statistical epidemiology in veterinary science. New York: MacMillan Publishing Co., Inc., 1979.
7. Martin SW, Meek AH, Willeberg P. Veterinary epidemiology: principles and methods. Ames: Iowa State University Press, 1987.
8. Noordhuizen JPTM, Frankena K, van der Hoofd CM, Graat EAM. Application of quantitative methods in veterinary epidemiology. Wageningen: Wageningen Pers, 1997.
9. Rothman K, Greenland S. Modern epidemiology, 2d ed. Philadelphia: Lippincott-Raven, 1998.
10. Sackett DL, Haynes RB, Guyatt GG, Tugwell P. Clincal epidemiology: a basic science for clinical medicine, 2d ed. Philadelphia: Lippincott Williams & Wilkins, 1991.
11. Smith RD. Veterinary clinical epidemiology: a problem-oriented approach. 2d ed. Boca Raton: CRC Press, 1995.
12. Thrushfield MV. Veterinary epidemiology, 2d ed. Oxford: Blackwell Science Ltd., 1995.

ACKNOWLEDGEMENTS

As we completed sections of this book we called upon friends and colleagues to review individual chapters. Their input has been invaluable. They have identified many errors which we had made and contributed many useful suggestions for additions and modifications. We are deeply indebted to the following people for their contributions.

Theresa Bernardo, USA
Pat Campbell, USA
Tim Carpenter, USA
Theodore Chester, England
David Dargatz, USA
Peter Davies, New Zealand
Hubert DeLuyker, Belgium
Christian Ducrot, France
Hollis Erb, USA
Annette Ersbøll, Denmark
Christine Fourichon, France
Ian Gardner, USA
Laura Green, England
Matthias Greiner, Denmark
Yrjö Gröhn, USA
Roberto Gutierriez, USA
Larry Hammell, Canada
David Jordan, Australia
John Kaneene, USA
Ken Leslie, Canada
Ann Lindberg, Sweden
Mariam Nielen, The Netherlands
Dirk Pfeiffer, England
Paivi Rajala-Schultz, USA
Stuart Reid, Scotland
Ynte Schukken, USA
Morgan Scott, USA
Jan Sergeant, USA
Margaret Slater, USA
Scott Wells, USA
Preben Willeberg, Denmark

We believe the value of this book has been greatly enhanced by the provision of a substantial number of 'real-life' datasets, both for use in the examples and the sample problems. We are deeply indebted to the following people who, in addition to the three authors, have contributed datasets to this effort.

Contributor	Country	Dataset
Jens Agger	Denmark	scc_40
Paul Bartlett	USA	scc_40
Theresa Bernardo	USA/Canada	pig_adg
Jette Christensen	Denmark/Canada	prew_mort
Gilles Fecteau	Canada	colostrum
Lynn Ferns	Canada	nocardia
Tine Hald	Denmark	sal_outbrk
Larry Hammell	Canada	fish_morts, isa_risk, isa_test

(continued)

Contributor	Country	Dataset
Dan Hurnik	Canada	pig_farm
Greg Keefe	Canada	beef_ultra, dairy_dis
Ann Lindberg	Sweden	bvd_test
Jeanne Lofstedt	Canada	calf
Carol McClure	Canada	isa_test
Fonda Munroe	Canada	tb_real
Ane Nødtvedt	Norway/Canada	fec
Javier Sanchez	Canada	elisa_repeat, fec
Iver Thysen (published data)	Denmark	calf_pneu
Emmanuel Tillard	France	reu_cc, reu_cfs
John VanLeeuwen	Canada	dairy_dis
Håkan Vigre	Denmark	ap2, smpltype
Jeff Wichtel	USA/Canada	pgtrial

Putting this book together has been both a learning experience and a lot of fun. We are deeply indebted to Margaret McPike who has done all of the editing, proofreading, formatting and typesetting of this text. Because we chose to publish it ourselves, we had to take full responsibility for these activities and Margaret has dedicated herself to this task. All of the credit for layout of the book, and the clarity of the format, goes to Margaret.

We would also like to thank Gregory Mercier, who did the graphic design work for the book cover and website, and Debra Hannams and Kent Villard who developed the website.

1

INTRODUCTION AND CAUSAL CONCEPTS

OBJECTIVES

After reading this chapter, you should be able to:

1. Explain the history of causal thinking about disease from an epidemiological perspective.

2. Explain how observational studies and field experiments seek to estimate causal effect coefficients and how these relate to counterfactual approaches.

3. Explain the basis of component-cause models and how this concept helps to explain measures of disease association and the proportion of disease explained by a causal factor.

4. Explain the basis of causal-web models of causation.

5. Construct a logical causal diagram based on your area of research interest as an aid to guiding your study design and analyses.

6. Apply a set of causal criteria to your own research and as an aid to interpreting published literature.

1.1 INTRODUCTION

Epidemiology is largely concerned with disease prevention and therefore, with the "succession of events which result in the exposure of specific types of individual to specific types of environment" (*ie* exposures) (MacMahon and Pugh, 1970). Thus, epidemiologists strive to identify these exposures and evaluate their associations with various outcomes of interest (*eg* health, welfare, productivity) so as to improve the lives of animals and their keepers. Hence, this book is about associations: associations which are likely to be causal in nature and which, once identified, we can take advantage of to improve the health, welfare and productivity of animals and the quality and safety of foods derived from them.

Associations between exposures and outcomes exist as part of a complex web of relationships involving animals and all aspects of their environment. Thus, in striving to meet our objectives, we (epidemiologists) are constantly struggling to improve our study designs and data analyses so that they best describe this complex web. It is only by studying these associations under field conditions (*ie* in the 'real world') that we can begin to understand this web of relationships. In this regard, Meek (1993), speaking on the topic of epidemiologic research, stated:

> It is essential that all groups periodically review their mandate, achievements and direction in light of the changing needs of society . . . greater use of the naturalistic paradigm in epidemiologic research is warranted, as the scientific paradigm, while serving the 'hard' sciences well may have shortcomings when it comes to understanding multifactorial issues.

As a starting place, we believe it is useful to review briefly the history of the concept(s) of multiple interrelated causes (exposures). This will provide a sense of how we have arrived at our current concepts of disease causation and where we might need to go in the future. Because we want to identify associations which are likely to be causal, it is appropriate to review the relevant areas of the philosophy of scientific inference that relate to causal inference. Following these brief reviews, we will proceed with overviews of the key components of veterinary epidemiologic studies and discuss some current concepts of disease causation. Our objective is to provide a foundation on which a deeper understanding of epidemiologic principles and methods can be built.

1.2 A BRIEF HISTORY OF MULTIPLE CAUSATION CONCEPTS

As noted, epidemiology is based on the idea that 'causes' (exposures) and 'outcomes' (health events) are part of a complex web of relationships. Consequently, epidemiologists base their research on the idea that there are multiple causes for almost every outcome and that a single cause can have multiple effects. This perspective is not universally shared by all animal-health researchers. In this current era, when great advances are being made in understanding the genetic components of some illnesses, a significant proportion of medical and veterinary research is focused on the characteristics of only direct causal agents and how they interact with the genetic makeup of the host of

interest. As Diez-Roux (1998b) points out, while it is true that genetic abnormalities are important precursors of many diseases, in terms of maintaining health, the real questions relate to the extent that our current environmental exposures and lifestyles (we could read this as animal management) lead to genetic defects as well as the extent to which these exposures and lifestyles allow specific genetic patterns to complete a sufficient cause of disease. (The concept of 'components of sufficient cause' is discussed in section 1.7.)

From a historical perspective, it is evident that the acceptance of the concept(s) of multiple interacting causes has ebbed and flowed, depending on the dominant causal paradigm of the era. However, the roots of this concept can be traced back at least to 400 B.C. when the Greek physician Hippocrates wrote *On Airs, Waters and Places*. He stated the environmental features that should be noted in order to understand the health of populations (Buck et al, 1988). Based on this aspect of his writing, it is clear that Hippocrates had a strong multicausal concept about exposure factors in the environment being important 'causes' of disease occurrence. He carried on to discuss the importance of the inhabitant's lifestyle as a key determinant of health status, further expanding the 'web of causation.' Nonetheless, the concepts linking the state of the environment and lifestyle to the occurrence of disease seem to have been short-lived as, between 5 and 1750 A.D., humoral imbalances (events within the individual) became the major paradigm of disease causation (Schwabe, 1982).

However, from 1750 to 1885, the multifactorial nature of disease causation returned when man-created environmental filth became accepted as a central cause of disease, and the prevalent causal paradigm was that disease was due to the effects of miasmas (*ie* bad air). It was during the mid 1800s that John Snow conducted his studies on contaminated water as the cause of cholera (Frerichs, 2003). Using a combination of astute observations about the lack of spread of the disease among health workers, the geographical distribution of cholera, a series of observational studies, including natural as well as contrived (removal of the Broad Street pump handle) experiments, Snow reached the correct conclusion about the transmission of cholera (*ie* that it was spread by water contaminated by sewage effluent). It is noteworthy that he arrived at this conclusion almost 30 years before the organism (*Vibrio cholera*) was discovered, thus demonstrating an important principle: disease can be prevented without knowing the proximal causal agent.

A few years later (*ie* in the 1880s-1890s), Daniel Salmon and Frederick Kilborne determined that an insect vector (a tick: *Boophilus annulatus*) was associated with a cattle disease called 'Texas Fever' even though the direct causal agent of the disease (a parasite: *Babesia bigemina*) was not discovered until many years later (Schwabe, 1984). Their first associations were based on the similar geographical distributions between the disease and the extent of the tick's natural range; theirs was the first demonstration of the spread mechanism of the causal agent of a vector being required for the development and transmission of a parasite. Their work provided the basis for disease control, again before knowing the actual agent of the disease. Thus, in this period (mid-to-late 1800s), the study of the causes of specific disease problems

focused on multiple factors in the environment, albeit somewhat more specifically than Hippocrates had discussed earlier.

The multifactorial causal concept became submerged during the late 1800s to the mid 1900s, when the search for specific etiological agents (usually microbiological) dominated medical research. This 'golden era' of microbiology lead to a number of successes including mass-testing, immunisation, specific treatment, as well as vector control (*eg* the mosquito vector of malaria was now known) as methods of disease control. Indeed, control of many specific infectious diseases meant that by the mid 1900s, chronic, non-infectious diseases were becoming relatively more important as causes of morbidity and mortality in humans in developed countries. It was recognised early on that single agents were not likely responsible for these chronic diseases and large-scale, population-based studies examining the potential multiple causes of these diseases were initiated. For example, the Framingham Heart Study pioneered long-term surveillance and study of causes of human health beginning in 1949. Similarly, large-scale, population-based studies of animal health were also undertaken. In 1957, the British initiated a national survey of disease and wastage in the dairy industry – the survey methods were later critiqued by their author (Leech, 1971). Thus, by the early 1960s, in both human and animal-health research, there was once again a growing awareness of the complex web of causation.

By the 1970s, multiple interacting causes of diseases returned as a major paradigm of disease causation. Building on the knowledge from the microbiological revolution, the concept of the agent-host-environment causal triad appeared in an early epidemiology text (MacMahon and Pugh, 1970). In this conceptual model, a number of component causes were required to come together (either sequentially or simultaneously) in order to produce disease; later, the complex of factors that was sufficient to produce disease was known as a sufficient cause and it was assumed that most diseases had a number of sufficient causes (Rothman, 1976). In addition to multiple causes, the component cause model was not constrained to have all causal factors at the same level of organisation. A traditional veterinary example used to portray some of these concepts is yellow shanks in poultry (Martin et al, 1987). When poultry with the specific genetic defect (an individual-level factor) are fed corn (ration is usually a herd/flock level factor) they develop a discolouration of the skin and legs. If all poultry are fed corn, then the cause of the disease would be a genetic defect; however, if all birds had the genetic defect, then the cause of the disease would be deemed to be the feed. In reality, both factors are required and the disease can be prevented by removing either the genetic defect, or changing the feed, or both, depending on the specific context.

The 1970s actually appeared to be a period of peak interest in causation (Kaufman and Poole, 2000). Susser's text on causal thinking appeared in 1973 (unfortunately, it has never been reprinted) and, three years later, the concepts of necessary and sufficient causes were published by Rothman (1976), followed by a set of causal criteria by Susser (1977). Large-scale monitoring of animal diseases began in this period (Ingram et al, 1975). As an example, linking databases of veterinary schools across North America in the Veterinary Medical Data Program was initiated based on the concept of using animals as sentinels for the environment (Priester, 1975).

The 1980s seemed to be a quiet time as no major new causal concepts were brought forward. Hence (perhaps by omission), the aforementioned web of causation might have become restricted to studying individual-level directly causal factors focusing on biological malfunctioning (Krieger, 1994).

In 1990, **epigenesis** was proposed as a formal model of multivariable causation that attempted to link, explicitly, causal structures to observed risks of disease (Koopman and Weed, 1990). While this proved to be an interesting and exciting proposal, the limitations of this approach were later realised (Thompson, 1991; Martin, 1996) and the approach remained only a concept. Notwithstanding the blossoming of field-based epidemiologic research that was taking place in the mid 1990s, a paper suggesting that epidemiology had reached its limits was published in a well-known biological journal (Taubes, 1995). This article led to considerable debate within and without epidemiology and, over time, deficiencies in the arguments suggesting a limited future for epidemiology were identified.

Since the mid nineties, there has been a lot of introspective writing by epidemiologists working on human diseases with much concern over an excess focus on individuals as the units of study and analysis. We shall not review these debates in detail as excellent discussions on these topics are available elsewhere (Shy, 1997; Diez-Roux, 1998a,b; McMichael, 1999). What is apparent is that depending on the context, elements of the social, physical and biological features of the defined ecosystem should be included in each study, while the unit of concern can range from the individual, to groups (litters, pens, barns), farms/families, villages or communities, watersheds or larger ecosystems. Thus, epidemiologic research remains deeply rooted in the concept of multiple interrelated causal factors as a basis for disease and hence, for disease prevention. This conceptual basis has been supported by substantial progress in the development of epidemiologic research methodologies and these are the subject of this book.

1.3 A BRIEF HISTORY OF SCIENTIFIC INFERENCE

Epidemiology relies primarily on observational studies to identify associations between exposures and outcomes. The reasons are entirely pragmatic. First, many health-related problems cannot be studied under controlled laboratory conditions. This could be due to limitations in our ability to create 'disease' problems in experimental animals, ethical concerns about causing disease and suffering in experimental animals and the cost of studying diseases in their natural hosts under laboratory conditions. Most importantly though, if we want to understand the complex web of relationships that affects animals in their natural state, then we must study them in that natural state. This requires the use of observational studies, and inferences from these studies are based primarily on inductive reasoning.

Philosophical discussion of causal inferences appears to be limited mainly to fields where observation (in which we attempt to discern the cause) rather than experimentation (in which we try to discern or demonstrate the effect) is the chief approach to research. While the latter approach is very powerful, one cannot assume that the results of even

the best-designed experiments are infallible. Thus, because epidemiologists rely on observational studies for the majority of our research investigations, a brief review of the basis for scientific inference is in order. We pursue this review in the context that epidemiology is a pragmatic discipline, that our activities are tied to health promotion and disease prevention and, that as Schwabe (1977) indicated, the key for disease prevention is to identify causal factors that we can manipulate, regardless of the level of organisation at which they act. We will briefly present the concepts of inductive and deductive reasoning. More complete reviews on the philosophy of causal inference are available elsewhere (Rothman and Greenland, 1998; White, 2001; Aiello, 2002; Weed 2002).

Inductive reasoning is the process of making generalised inferences about (in our context) 'causation' based on repeated observations. Simply put, it is the process of drawing conclusions about the state of nature from carefully recorded and analysed observations. Francis Bacon (1620), first presented inductive reasoning as a method of making generalisations from observations to general laws of nature. As two examples, John Snow's observations during the cholera outbreaks of the mid 1800s led to a correct inference about the mechanism of the spread of the disease, while Edward Jenner's observations that milkmaids who developed cowpox didn't get smallpox, led to his conclusion that cowpox might prevent smallpox. This, in turn, led to the development of a crude vaccine which was found to be effective when tested in humans in 1796. These were both dramatic examples of the application of inductive reasoning to important health problems. In 1843, John Stuart Mill proposed a set of canons (rules) for inductive inference. Indeed, Mill's canons might have originated our concepts about the set of component causes that are necessary or sufficient to cause disease (White, 2000).

While it is easy to identify important advances in human and animal health that have been based on inductive reasoning, proponents of deductive reasoning have been critical of the philosophical basis (or lack thereof) of inductive logic. David Hume (1740) stated that there is "no logical force to inductive reasoning." He stated further that "we cannot perceive a causal connection, only a series of events." The fact that the sun comes up every day after the rooster crows, should not result in a conclusion that the rooster crowing causes the sun to rise. He noted further that many repetitions of the two events might be consistent with a hypothesis about causation but do not prove it true. Bertrand Russell (1872-1970) continued the discussion of the limitations of inductive reasoning and referred to it as "the fallacy of affirming the consequent." (In this process, we might imply that if A is present, then B occurs; so if B occurs, A must have been present.)

Deductive reasoning is the process of inferring that a general 'law of nature' exists and has application in a specific, or local, instance. The process starts with a hypothesis about a 'law of nature' and observations are then made in an attempt to either prove or refute that assumption. The greatest change in our thinking about causal inferences in the past century has been attributed to Karl Popper who stated that scientific hypotheses can never be proven or evaluated as true, but evidence might suggest they are false. This philosophy is referred to as **refutationism**. Based on Popper's philosophy, a

scientist should not collect data to try and prove a hypothesis (which Popper states is impossible anyway), but that scientists should try to disprove the theory; this can be accomplished with only one observation. Once a hypothesis has been disproven, the information gained can be used to develop a revised hypothesis, which should once again be subjected to rigorous attempts to disprove it. Popper argues that, only by disproving hypotheses do we make any scientific progress. It is partially for this reason that, when conducting statistical analyses, we usually form our hypothesis in the null (*ie* that a factor is not associated with an outcome) and, if our data are inconsistent with that hypothesis, we can accept the alternative that the factor is associated with the outcome. Thus, the current paradigm in deductive reasoning is to conjecture and then attempt to refute that conjecture.

A major benefit of using Popper's approach is that it helps narrow the scope of epidemiologic studies instead of using a data-mining 'risk-factor' identification approach. It suggests that we carefully review what is already known and then formulate a very specific hypothesis that is testable with a reasonable amount of data. In the former approach, we often generate long, multipage questionnaires, whereas, in the latter, the required information is much more constrained and highly focused on refuting the hypothesis (Petitti, 1988).

As noted, epidemiology is primarily based on inductive reasoning, but the deductive paradigm has played a large role in the development of the discipline. Epidemiologic investigations which start with a clear hypothesis are inevitably more focused and more likely to result in valid conclusions than those based on unfocused recording of observations.

Two other important concepts that relate to scientific inference are worth noting. Thomas Bayes, a Presbyterian minister and mathematician, stated that "all forms of inference are based on the validity of their premises" and that "no inference can be known with certainty" (1764). He noted that scientific observations do not exist in a vacuum, and that the information we have prior to making a series of observations will influence our interpretation of those observations. For example, numerous studies have shown that routine teat-end disinfection (after milking) can reduce the incidence of new intra-mammary infections in dairy cows. However, if a new study was conducted in which a higher rate of infections was found in cows that received teat-end disinfection, we would not automatically abandon our previous ideas about teat-end disinfection. His work has given rise to a branch of statistics known as **Bayesian** analysis, some of which will appear later in this book.

More recently, Thomas Kuhn (cited in Rothman and Greenland, 1998) reminds us that although one observation can disprove a hypothesis, the particular observation might have been anomalous and that the hypothesis could remain true in many situations. Thus, often the scientific community will come to a decision about the usefulness, if not the truth, of a particular theory. This is the role of **concensus** in scientific thinking. While hard to justify on a philosophical basis, it plays a large role in shaping our current thinking about causes of disease.

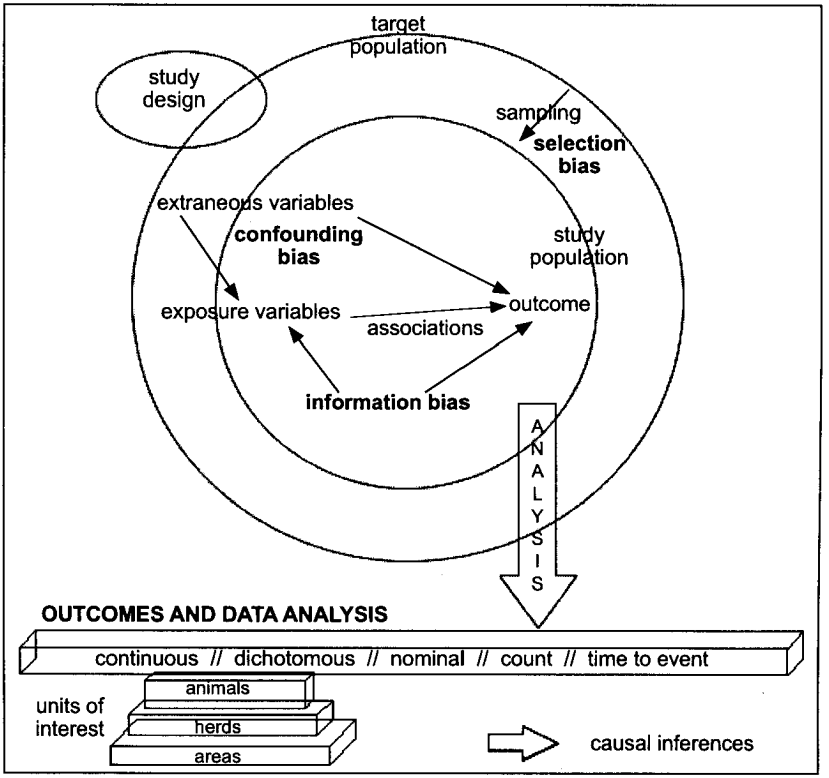
Although philosophical debates on causal inference will undoubtedly continue (Robins, 2001; White, 2001), as a summary of this section we note that . . . "all of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature. . . the tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical" (Rothman and Greenland, 1998).

While keeping these historical and philosophical bases in mind, we will now proceed to an outline of the key components of epidemiologic research.

1.4 KEY COMPONENTS OF EPIDEMIOLOGIC RESEARCH

Fig. 1.1 summarises key components of epidemiologic research. It is somewhat risky to attempt to simplify such a complex discipline and present it in a single diagram, but we believe it is beneficial for the reader to have an overview of the process of evaluating associations between exposure and outcome as a guide to the rest of the book.

Fig. 1.1 Key components of epidemiologic research



Our rationale for doing research is to identify potentially causal associations between exposures and outcomes (the centre of the diagram). In many cases, the exposures are risk factors and the outcome is a disease of interest. However, this is not the only scenario; for example, our outcome of interest might be a measure of productivity or food safety and the exposures might include certain diseases.

- Ultimately, we aim to make **causal inferences** (bottom right of diagram) and Chapter 1 discusses some important concepts of causation as they relate to epidemiologic research.
- Any study starts with an overall **study design** and the main observational study types are discussed in Chapters 7-10, with controlled trial designs being presented in Chapter 11.
- In any study, it is important to identify the **target population** and obtain a **study group** from it in a manner that does not lead to **selection bias. Sampling** is discussed in Chapter 2 and selection bias in Chapter 12.
- Once we have identified our study subjects, it is necessary to obtain data on **exposure variables, extraneous variables** and the **outcome** in a manner that does not lead to **information bias** (Chapter 12). Two important tools that are used in that process are **questionnaires** (Chapter 3) and **diagnostic and screening tests** (Chapter 5).
- In order to start the process of establishing an association between exposure and outcome, we need to settle on a **measure of disease frequency** (Chapter 4) and select a **measure of association** (Chapter 6) that fits the context. In many cases, the study design will determine the measures that are appropriate.
- **Confounding bias** is a major concern in observational studies, and the identification of factors that should be controlled as confounders is featured in Chapter 13.
- With our data in hand, we are now able to begin to model relationships with the intent of estimating causal effects of exposure (Chapter 13). Individual chapters are dedicated to the analyses appropriate for outcomes that are continuous (Chapter 14), dichotomous (Chapter 16), **nominal/ordinal** (Chapter 17), count (Chapter 18) and time-to-event data (Chapter 19). Chapter 15 presents some general guidelines on **model-building** techniques that are applicable to all types of model.
- In veterinary epidemiologic research, we often encounter **clustered** or **correlated data** and these present major challenges in their analyses. Chapter 20 introduces these while Chapters 21 and 22 focus on mixed (random effects) models for continuous and discrete outcomes. Chapter 23 presents some alternative methods of analysis for dealing with clustered data.
- Structured reviews and assessments of the literature in the form of **meta-analyses** are becoming increasingly important and are introduced in Chapter 24.
- Not all studies allow us to collect data on exposures and outcomes at the individual level and yet there is much that we can learn by **studying disease in groups** (*eg* herds). Thus, ecologic studies are introduced in Chapter 25.
- Finally, we complete the text with Chapter 26 which provides a 'road map' for investigators starting into the analysis of a complex epidemiologic dataset.

With this background, it is time to delve deeper into this discipline called epidemiology. And, at the outset it is important to stress that epidemiology is first and foremost a

biological discipline, but one which relies heavily on quantitative (statistical) methods. It is the integration of these two facets, with a clear understanding of epidemiologic principles which makes for successful epidemiologic research. As Rothman and Greenland (1998) point out:

> Being either a physician (veterinarian) or a statistician, or even both is neither a necessary nor sufficient qualification for being an epidemiologist. What is necessary is an understanding of the principles of epidemiologic research and the experience to apply them.

To help meet this goal, this book is divided roughly equally into chapters dealing with epidemiologic principles and those dealing with quantitative methods.

1.5 Seeking causes

As already noted, a major goal for epidemiologic research is to identify factors that can be manipulated to maximise health or prevent disease. In other words, we need to identify causes of health and disease. That might seem like a simple enough task, but it is, in fact, complex (that is why we wrote much of this text). Here we want to focus on what a cause is and how we might best make decisions about whether a factor is a cause. For our purposes, **a cause is any factor that produces a change in the severity or frequency of the outcome**. Some prefer to separate biological causes (those operating within individual animals) from population causes (those operating at or beyond the level of the individual). For example, a specific microorganism could be viewed as a biological cause of disease within individuals, whereas management, housing or other factors that act at the herd (or group) level – or beyond (*eg* weather) – and affect whether or not an individual is exposed to the microorganism, or affect the animal's susceptibility to the effects of exposure, would be deemed as population causes.

In searching for causes, we stress the holistic approach to health. The term holistic might suggest that we try to identify and measure every suspected causal factor for the outcome of interest. Yet, quite clearly, we cannot consider every possible factor in a single study. Rather, we place limits on the portion of the 'real world' we study and, within this, we constrain the list of factors we identify for investigation. Usually, extant knowledge and current belief are the bases for selecting factors for study. Because of this, having a concept of causation and a causal model in mind can help clarify the data needed, the key measures of disease frequency and the interpretation of associations between exposure and disease.

1.5.1 Counterfactual observations and causation

In field experiments and in observational studies, it is vital that the comparison group, comprised of the non-treated (or non-exposed) subjects, is as similar as possible to the treated (or exposed) groups with respect to factors that could affect the outcome. In this regard, the perfect comparison group would be the same treated (or exposed) individuals if they had not been treated (or exposed). This is called the **counterfactual** group. By comparing the frequency of the outcome in these two perfectly similar groups, we would obtain the 'true' causal effect of the treatment or exposure. Obviously,

this counterfactual group does not exist so we use the best practical alternatives. Randomisation helps achieve this in field experiments and statistical control and/or study-design methods attempt to achieve this in observational studies. However, neither is guaranteed to achieve the goal and so care must be taken when interpreting the results of field trials and observational studies due to potential differences in the groups that could bias the outcome.

1.5.2 Experimental versus observational evidence

Experimental evidence

Traditionally, the gold standard approach to identifying causal factors is to perform an experiment. In the ideal experiment, we randomise some animals (or other units of concern) to receive the factor and some to receive nothing, a placebo, or a standard intervention (treatment). In this context, exposure X is a proven cause of outcome Y, if in an ideal experiment X is changed and, as a result, the value or state of Y also changes. In this ideal experiment, X explicitly precedes Y temporally and all variables (known and unknown) that do not intervene between X and Y are made independent of X through the process of randomisation (this means that extraneous variables do not confound or bias the results we attribute to the exposure X). Factors that are positioned temporally or causally between X and Y are not measured and are of no concern with respect to answering the causal objective of the trial.

The measure of causation in this ideal trial is called the causal **effect coefficient** and indicates the difference in the outcome between the 'treated' and 'non-treated' groups (*ie* those with different levels of factor X). For example, if the risk of the outcome in the group receiving the treatment is denoted R_1 and the risk in the group not receiving the treatment is R_0, then we might choose to measure the effect of treatment using either an absolute measure (*ie* risk difference - *RD*) or a relative measure (*ie* risk ratio - *RR*) as shown in Chapter 6. If this difference is greater than what could be attributed to chance, then we would say that we have proved that the factor is a cause of the outcome event. A key point is that all causal-effect statements are based on **contrasts** of treatment levels; the outcome in the treated group cannot be interpreted without knowing the outcome in the untreated group. A second key feature is **exchangeability**; that is the same outcome would be observed (except for sampling error) if the assignments of treatment to study subjects had been reversed (*ie* if the treated group had been assigned to be untreated). Randomisation provides the probabilistic basis for the validity of this assumption.

Observational evidence

In observational studies, we estimate the difference in values of Y between units that happen to have different values of X. We do not control whether a subject is, or is not, exposed. Variables related to both X and Y and which do not intervene between X and Y, can be controlled analytically or through matching or restricted sampling (see Chapter 13). The appropriate measure of association (*eg* a risk ratio or regression coefficient) reflecting the difference in the value of Y between the 'exposed' and 'non-exposed' groups can be used to obtain a reasonable estimate of the causal-effect coefficient that would be obtained in the ideal experiment. The major differences between observational studies and field experiments lie in the ability to prevent selection, misclassification

and confounding bias, and dealing with the impact of unknown or unmeasured factors. Thus, by themselves, observational studies produce measures of association but cannot 'prove' causation. Nonetheless, in the ideal observational study, with total control of bias, the measure of association will estimate the causal-effect coefficient.

However, experimental evidence is deemed to provide more solid evidence of causality as, in reality, "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)."

(Attributed to Box, 1966, in Snedecor and Cochran, 1989)

Limits of experimental study evidence

Despite their advantages, performing perfect experiments is not easy even at the best of times (see Chapter 11) and, in fact, many potential causal factors of interest to epidemiologists would be difficult to study using a controlled trial format. For example, it would be impossible to perform the perfect experiment to answer the question of whether or not badgers that are infected with *M. bovis* cause tuberculosis in cattle. Laboratory studies are useful to demonstrate what can happen when animals are exposed to a specific exposure (*eg* that factor A can cause outcome B), but, if the circumstances are too contrived (very large dose, challenge by an unnatural route, limited range of cofactors), laboratory results might not be much help in deciding the issue of causation under normal, everyday conditions. For example, we could conduct an experiment in which cattle and infected badgers are maintained within a confined enclosure and assess whether or not the cattle became infected. If they did, this would demonstrate that infected badgers can cause infection in cattle, but not the extent of the problem in the field.

In field trials that are subject to non-compliance, we often have to decide how to manage the non-compliance in assessing the role of the treatment on the outcome (Heitjan, 1999) and, although any given field trial might provide more valid evidence for or against causation than any given observational study, it is not uncommon for differences in results to exist among apparently similar field trials. Hence, the ability to make perfect inferences based on field trials is illusionary and, in many instances, it is impossible to carry out experiments under conditions that even remotely resemble 'real-world' conditions.

1.6 Models of causation

Given our belief in multiple causes of an effect and multiple effects of a specific cause, epidemiologists have sought to develop conceptual models of causation; we describe the two major models in sections 1.7 and 1.8. Usually, however, the actual causal model is unknown and the statistical measures of association we use reflect, but do not explain, the number of ways in which the exposure might cause disease. Furthermore, although our main interest in a particular study might focus on one exposure factor, we need to take into account the effects of other causes of the outcome that are related to the exposure (this process is usually referred to as **control** or **controlling the effects**) if we are to learn the 'truth' about the potential causal effect of our exposure of interest.

Because our inferences about causation are based, at least in the main, on the observed difference in outcome frequency or severity between exposed and unexposed subjects, we will continue our discussion by examining the relationship between a postulated causal model and the resultant, observed, outcome frequencies. The two major conceptual models are the component-cause and the causal-web models of causation.

1.7 COMPONENT-CAUSE MODEL

The component-cause model is based on the concepts of necessary and sufficient causes (Rothman, 1976). A **necessary cause** is one without which the disease cannot occur (*ie* the factor will always be present if the disease occurs). In contrast, a **sufficient cause** always produces the disease (*ie* if the factor is present, the disease invariably follows). However, both experience and formal research have indicated that very few exposures (factors) are sufficient in and of themselves, rather different groupings of factors can combine and become sufficient causes. Thus, a **component cause** is one of a number of factors that, in combination, constitute a sufficient cause. The factors might be present concomitantly or they might follow one another in a chain of events. In turn, when there are a number of chains with one or more factors in common, we can conceptualise the web of causal chains (*ie* a causal web). This concept will be explained further under the causal-web model (section 1.8).

As an example of component causes, in Table 1.1 we portray the causal relationships of four risk factors for bovine respiratory disease (BRD). These include:

- a bacterium, namely *Mannheimia hemolytica* (Mh)
- a virus, namely the bovine respiratory syncytial virus (BRSV)
- a set of stressors such as weaning, transport, or inclement weather
- other bacteria such as *Hemophilus somnus* (Hs).

Table 1.1 Four hypothetical sufficient causes of bovine respiratory disease

	Sufficient causes			
Component causes	I	II	III	IV
Mh	+	+		
BRSV	+		+	
Stressors		+	+	+
Other organisms (eg Hs)				+

In this portrayal, there are four sufficient causes, each one containing two specific components; we assume that the four different two-factor combinations each form a sufficient cause. Hence, whenever these combinations occur in the same animal, clinical respiratory disease occurs (as mentioned, one can conceive that these factors might not need to be present concomitantly, they could be sequential exposures in a given animal). Some animals could have more than two causal factors (*eg* Mh, BRSV, Hs) but the presence of any of the two-factor combinations shown will be sufficient to produce BRD. Note that we have indicated that only some specific two-factor combinations act

as sufficient causes; Mh is a component of two of the sufficient causes, as is BRSV. Because no factor is included in all sufficient causes, there is no necessary cause in our model of BRD. Obviously, if you have not guessed by now, you should be aware that the number of causal factors and their arrangement into sufficient causes are purely for the pedagogical purposes of this example.

Now, against this backdrop of causal factors, we will assume that we plan to measure only the Mh and BRSV components (*ie* obtain nasal swabs for culture and/or blood samples for antibody titres). Nonetheless, we are aware that, although unmeasured, the other components (stressors and/or Hs) might be operating as components of one or more of the sufficient causes. In terms of the two measured factors, we observe that some cattle with BRD will have both factors, some will have only Mh and some only the BRSV components. Because of the causal effects of the other unmeasured factors (stressors and Hs), there will be some animals with BRD that have neither of these two measured factors (*eg* BRD due to sufficient cause IV).

One of the benefits of thinking about causation in this manner is that it helps us understand how the prevalence of a cofactor can impact on the strength of association between the exposure factor and the outcome of interest (Pearce, 1989). For example, assume that we are interested principally in the strength of association between infection with Mh and the occurrence of BRD (the various measures of association are explained in Chapter 6). According to our example in Table 1.1, Mh produces disease when present with BRSV, but also without BRSV when combined with 'stressors'. What might not be apparent however, is that changes in the prevalence of the virus, or of the stressors, or Hs can change the strength of association between Mh and BRD. These shared component causes that make up a sufficient cause are known as **causal complements**. To demonstrate this point, note the two populations in Examples 1.1 and 1.2.

1.7.1 The effect of the causal complement prevalence on disease risk

This example is based on the component cause model shown in Table 1.1 using three factors: Mh, BRSV and stressors. The frequency of each factor indicated above the body of the tables in Examples 1.1 and 1.2 is the same (p(stressors)=0.4 and p(Mh)=0.6) except that the frequency of BRSV is increased from 30% in Example 1.1 to 70% in Example 1.2. All three factors are distributed independently of each other; this is not likely true in the field, but it allows us to examine the effect of single factors without concerning ourselves with the biasing effects of the other factors.

If infection with Mh is our exposure factor of interest, it would be apparent that some but not all cattle with Mh develop BRD and that some cattle without Mh also develop BRD. Thus, Mh infection by itself is neither a necessary nor sufficient cause of BRD. Similarly for BRSV, some infected cattle develop BRD, some non-infected cattle also develop BRD. In order to ascertain if the occurrence of BRD is associated with Mh exposure, we need to measure and contrast the risk of BRD among the exposed (Mh+) versus the non-exposed (Mh-). In Example 1.1, these frequencies are 58% and 12%, and we can express the proportions relative to one another using a statistic called the

Example 1.1 Causal complement prevalence and disease risk - Part 1

The number and risk of BRD cases by two measured and one unknown exposures assuming joint exposure to any two factors is sufficient to cause the disease are shown below. *Mannheimia hemolytica* (Mh) is the exposure of interest (total population size is 10,000; p(stressors)=0.4; p(Mh)=0.6).

p(BRSV=0.3)

	Measured factors			
Unmeasured stressors	BRSV	Mh	Population number	Number diseased
1	1	1	720	720
1	1	0	480	480
1	0	1	1680	1680
1	0	0	1120	0
0	1	1	1080	1080
0	1	0	720	0
0	0	1	2520	0
0	0	0	1680	0

Risk of disease among the Mh+	3480/6000=0.58
Risk of disease among the Mh-	480/4000=0.12
Risk difference if Mh+	0.58-0.12=0.46
Risk ratio if Mh+	0.58/0.12=4.83

risk ratio which is 58/12=4.83. This means that the frequency of BRD is 4.83 times higher in Mh+ cattle than in Mh- cattle. We could also measure the association between Mh and BRD using a risk difference; in this instance, the *RD* is 0.46 or 46%. These measures are consistent with Mh being a cause of BRD, but do not prove the causal association. In Example 1.2, when the frequency of BRSV is increased, the relative risk for Mh+ cattle is 2.93 and the *RD* is 0.54 or 54%. Thus, we might be tempted to think that exposure to Mh+ in some sense acts differently from a causal perspective in one example to another, yet the underlying causal relationship of Mh exposure to the occurrence of BRD has not changed. The difference is due to a change in the frequency of the other components of the sufficient causes, namely BRSV. The other components that can form sufficient causes are called the **causal complement** to the exposure factor. Here with sets of two factors being sufficient causes, the causal complements of Mh are BRSV or stressors but not both (the latter cattle would have developed BRD from being stressed and having BRSV).

In general, we might note that when the prevalence of causal complements is high, measures of association between the factor of interest and the outcome that are based

Example 1.2 Causal complement prevalence and disease risk - Part 2

The number and risk of BRD cases by two measured and one unknown exposures assuming joint exposure to any two factors is sufficient to cause the disease are shown below. *Mannheimia hemolytica* (Mh) is the exposure of interest.

p(BRSV)=0.7

	Measured factors			
Unmeasured stressors	BRSV	Mh	Population number	Number diseased
1	1	1	1680	1680
1	1	0	1120	1120
1	0	1	720	720
1	0	0	480	0
0	1	1	2520	2520
0	1	0	1680	0
0	0	1	1080	0
0	0	0	720	0

Risk of disease among the Mh+	4920/6000=0.82
Risk of disease among the Mh-	1120/4000=0.28
Risk difference if Mh+	0.82-0.28=0.54
Risk ratio if Mh+	0.82/0.28=2.93

on risk differences will be increased (especially when the prevalence of exposure is low). Some, but not all, ratio or relative measures of association could have the opposite relationship with the prevalence of causal complements. In any event, although the causal mechanism remains constant, the strength of association will vary depending on the distribution of the cofactors, many of which we do not know about or remain unmeasured for practical reasons. As will be discussed, strength of association is one criterion of causation but it is not a fixed measure and we need to bear the phenomenon just discussed in mind when making causal inferences.

You might verify that the impact of BRSV on BRD as measured by the risk ratio would be the same (*RR*=3.2) in both Examples 1.1 and 1.2 even though its prevalence has changed. Although this is only one example, we could state the general rule that the strength of association for a given factor depends on the frequency of the causal complements but, providing the distribution of the other causal factors is fixed, changes in the prevalence of the factor of interest do not alter its strength of association with the outcome.

If we could measure all the cofactors including stressors and the other causal component factors, the picture would change considerably. For example, if the stressors were the only other causes of BRD, it would be obvious that, in the non-stressed animals, BRD occurred only when both Mh and BRSV were present together. This would be clear evidence of biological synergism, a feature that is detected numerically as statistical interaction (*ie* the joint effect of the two factors would be different than the sum of their individual effects – in this instance, they would have no 'individual' effect, only a joint effect). In stressed cattle, all animals exposed to Mh or BRSV would get BRD but there would be no evidence of interaction because 100% of singly, as well as jointly, exposed stressed cattle would develop BRD.

Because changes in the prevalence of the 'unknown' or 'unmeasured' factor(s) will alter the magnitude of effect for the measured exposure, we need to think of measures of association as 'population specific.' Only after several studies have found a similar magnitude of effect in different populations should we begin to think of the effect as in some sense a biological constant. Further, even if the cases have arisen from an assumed model that incorporates biological synergism, because of the distribution of the unknown causal factors, interaction (indicating synergism) might not be evident in the observed data.

1.7.2 The importance of causal factors

Using the concepts of necessary and sufficient causes, we also gain a better understanding of how much disease in the population is attributable to that exposure (or alternatively the proportion of disease that we could prevent by completely removing the exposure factor).

As explained in Chapter 6, this is called the **population attributable fraction** (AF_p). For example, if we assume that the prevalence of each of the four sufficient causes from Table 1.1 is as shown in Table 1.2, then, if we examine the amount of disease that can be attributed to each of the component causes, we see that we can explain more than 100% of the disease. Of course, we really haven't, it is simply because the components are involved in more than one sufficient cause and we are double-counting the role that each component cause plays as a cause of the disease.

Table 1.2 Hypothetical sufficient causes of bovine respiratory disease and their relationship to population attributable fraction

	Sufficient causes				
Component causes	I	II	III	IV	AF_p (%)
Mh	+	+	-		75
BRSV	+	-	+		60
Stressors	-	+	+	+	55
Hs				+	10
Prevalence of sufficient cause (%)	45	30	15	10	

Another important observation is that, when two or more factors are both essential for disease occurrence, it is difficult to attribute a specific proportion of the disease occurrence to any single causal factor. For example, in cattle that had all three factors – Mh, BRSV and stressors – it would be impossible to decide the unique importance of each factor. Our model indicates that once any two of the three were present, then BRD would occur and the presence of the third factor is of no importance causally; thus, as the saying goes 'timing is everything'. Certainly, because the frequency of cofactors can vary from subgroup to subgroup, as with relative risk measures, one should not think of AF_p as being a 'universal' measure of importance.

1.8 Causal-web model

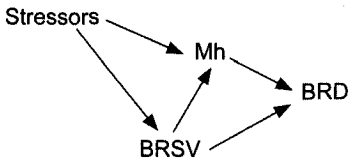
A second way of conceptualising how multiple factors can combine to cause disease is through a causal web consisting of **indirect** and **direct** causes. This concept is based on a series of interconnected causal chains or web structures; it takes the factors portrayed in the sufficient-cause approach and links them temporally. For a **direct** cause, there must be no known intervening variable between that factor and the disease (diagrammatically, the exposure is adjacent to the outcome). Direct causes are often the **proximal** causes emphasised in therapy, such as specific microorganisms or toxins. In contrast, an **indirect** cause is one in which the effects of the exposure on the outcome are mediated through one or more intervening variables. It is important to recognise that, in terms of disease control, direct causes are no more valuable than indirect causes. In fact, many large-scale control efforts are based on manipulating indirect rather than direct causes. Historically, this was also true: whether it was John Snow's work on cholera control through improved water supply, or Frederick Kilborne's efforts to prevent Texas Fever in American cattle by focusing on tick control. In both instances, disease control was possible before the actual direct causes (*Vibrio cholerae* and *Babesia bigemina*) were known, and the control programme was not focused directly on the proximal cause.

One possible web of causation of respiratory disease (BRD) based on the three factors in Examples 1.1 and 1.2 might have the structure shown in Example 1.3. The causal-web model complements the component-cause model but there is no direct equivalence between them. As we show later, causal-web diagrams are very useful to guide our analyses and interpretation of data.

The model indicates that stressors make the animal susceptible to Mh and BRSV, that BRSV increases the susceptibility to Mh and that BRSV can 'cause' BRD directly (this might be known to be true, or it might reflect the lack of knowledge about the existence of an intervening factor such as Hs which is missing from the causal model). Finally it indicates that Mh is a direct cause of BRD. If this causal model is true, it suggests that we could reduce BRD occurrence by removing an indirect cause such as stress, even though it has no direct effect on BRD. We could also control BRD by preventing the action of the direct causes Mh and BRSV (*eg* by vaccination, or prophylactic treatment with antimicrobials – we are not suggesting that you do this!). As mentioned, this model claims that stressors do not cause BRD without Mh or BRSV infection and thus

Example 1.3 Causal web

A causal-web model of BRD based on component causes from Example 1.1.



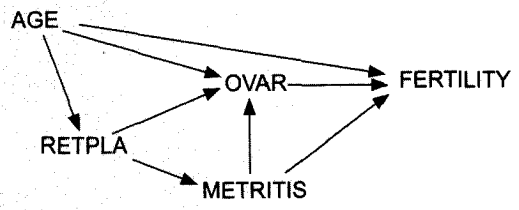
suggests a number of two- or three-factor groupings of component causes into sufficient causes. However, it does not explicitly indicate whether some of the proximal causes can produce disease in and of themselves (*ie* it is not apparent whether BRSV can cause BRD by itself or if it needs an additional unmeasured factor). From the previous examples, the outcome frequencies in BRSV-infected and non-infected cattle will depend on the distribution of the other component causes and whether, in reality, it can be a sufficient cause by itself. For now, we will discuss the relationship of the causal structure to the results of our analyses.

With a number of possible causal variables, the cause-and-effect relationships are best shown in a **causal diagram** (also called **directed acyclic graphs**, or modified **path models**). To construct a causal diagram, we begin by imposing a plausible biological causal structure on the set of variables we plan to investigate and translate this structure into graphical form that explains our hypothesised and known relationships among the variables. The causal-ordering assumption is usually based on known time-sequence and/or plausibility considerations. For example, it might be known that one variable precedes another temporally, or current knowledge and/or common sense might suggest that it is possible for one factor to cause another but not vice-versa.

1.9 Constructing a causal diagram

The easiest way to construct the causal diagram is to begin at the left with variables that are pre-determined and progress to the right, listing the variables in their causal order. The variation of these variables (those to the extreme left such as AGE in Example 1.4) is considered to be due to factors outside of the model. The remaining variables are placed in the diagram in their presumed causal order; variables to the left could 'cause' the state of variables to their right to change. If it is known or strongly believed that a variable does not cause a change in one or more variables to its right, then no causal arrow should be drawn between them. If the proposed model is correct, the analyses will not only be more informative but also more powerful than analyses that ignore the underlying structure. The only causal models to be described here are called **recursive**; that is, there are no causal feedback loops (if these are believed to exist, they can be formulated as a series of causal structures). A causal diagram of factors relating to fertility in dairy cows is shown in Example 1.4.

Example 1.4 A causal diagram of factors affecting fertility in cows



RETPLA = retained placenta
OVAR = cystic ovarian disease

Suppose the model is postulated to explain biological relationships among reproductive diseases. AGE is assumed to be a **direct cause** of retained placenta (RETPLA), cystic ovarian disease (OVAR) and FERTILITY but not METRITIS. RETPLA is the exposure variable of interest. METRITIS and OVAR are **intervening variables** between RETPLA and the outcome of interest FERTILITY. We will assume that our objective is to estimate the causal effect of RETPLA on FERTILITY based on the association between these two variables. **Note** It is the causal effect coefficient that we are interested in estimating.

The model indicates that AGE can cause changes in FERTILITY directly but also by a series of pathways involving one or more of the three reproductive diseases. It also indicates that AGE is not a direct cause of metritis. In terms of understanding relationships implied by the causal diagram, the easiest way to explain them is to think of getting (perhaps driving?) from an exposure variable (*eg* RETPLA) to a subsequent variable (*eg* FERTILITY). As we pass through other variables following the arrows, we trace out a causal path. The rule for tracing out causal pathways is that you can start backwards from any variable but once you start forward on the arrows you cannot back up. Paths which start backwards from a variable are **spurious causal** paths and reflect the impact of confounders. In displaying the relationships, if there are variables that we believe are correlated because of an unknown or unmeasured common cause, we use a line to indicate this, and you can travel in either direction between these variables. If two variables are adjacent (connected by a single direct arrow), their causal relationship is deemed to be **directly causal**. Paths which start forward from one variable and pass through intervening variables are deemed to be **indirect causal** paths (*eg* RETPLA can cause fertility changes through its effect on OVAR, but not directly). The combined effects through indirect and direct paths represent the **total effect** of the variable.

Okay, so, how does this help us? Well, in order to estimate the causal-effect coefficient, we must prevent any spurious effects, so the variables preceding an exposure factor of interest (RETPLA) that have arrows pointing toward it (*ie* from AGE) and through which FERTILITY (the outcome) can be reached on a path must be controlled. In this instance, that variable is AGE. The model also asserts that we do not control intervening variables so METRITIS and OVAR are not placed in the analytic model when estimating

the causal effect. If we assume that there are no other confounders that are missing from the model, our analyses will estimate the causal effect of RETPLA on FERTILITY. (This also assumes the statistical model is correct, but that is another story.)

We should note that if we did control for METRITIS and OVAR in this model, we would not obtain the correct estimate of causal effect. Rather, we would only obtain the direct effect of RETPLA on FERTILITY if that direct effect existed. This feature will be discussed again when regression models (*eg* Chapter 14) are described as this is a major reason why we can inadvertently break down a causal web. In the causal diagram used here, we explicitly assume there is no direct causal relationship between them (so this would be an inappropriate analysis for this reason also). RETPLA can impact on FERTILITY indirectly through the diseases METRITIS and/or OVAR; controlling these variables blocks these indirect pathways. Thus, only by excluding METRITIS and OVAR can we obtain the correct causal-effect estimate.

1.10 CAUSAL CRITERIA

Given that researchers will continue to make advances in identifying potential causes of disease using observational study techniques, a number of workers have proposed a set of causal guidelines (these seek to bring uniformity to decisions about causation (Evans, 1978)). Others suggest that we view these as a set of values and accept that different individuals might view the same facts differently (Poole, 2001). Hill (1965) proposed a list of criteria for making valid causal inferences (not all of which had to be fully met in every instance). They include: time sequence, strength of association, dose-response, plausibility, consistency, specificity, analogy and experimental evidence. Today, we might add evidence from meta-analysis to this list. Over the years, the first four of these have apparently dominated our inference-making efforts (Weed, 2000) and recently, researchers have investigated how we use these and other criteria for making inferences (Waldmann and Hagmayer, 2001). In one study, a group of 135 epidemiologists were given a variety of realistic but contrived examples and varying amounts of information about each scenario. At the end of the exercise, they had agreed on causal inferences in only 66% of the examples. This stresses the individuality of interpreting the same evidence. Because we believe a set of criteria for causal inferences is a useful aid to decision-making, we will briefly comment on Hill's list of items and give our view of their role in causal inference (Holman et al, 2001).

At the outset, we must be clear about the context for inferring causation. As Rose (1985) stated, it is important to ask whether we are trying to identify causes of disease in individuals or causes of disease in populations. Indeed, with the expansion of molecular studies, the appropriate level at which to make causal inferences, and whether such inferences are valid across different levels of organisation remains open to debate. However, clear decisions about the appropriate level to use (think back to the objectives when choosing this) will guide the study design as well as our inferences about causation.

The following set of criteria for causation can be applied at any level of organisation. The criteria are based on individual judgement, not a set of defined rules.

1.10.1 Study design and statistical issues

As will be evident after delving into study design (Chapters 7-10), some designs are less open to bias than others. For example, case-control studies are often assumed to be subject to more bias than cohort studies. However, much of this criticism is based on case-control studies using hospital or registry databases. We think it important that every study be assessed on its own merits and we need to be aware of selection, misclassification and confounding bias in all study designs.

Most often we do not make inferences about causation unless there is a statistically significant association between the exposure and the outcome (and one that is not likely to be explained by one or more of the previous biases). Certainly, if the differences observed in a well-designed study have P-values above 0.4, this would not provide any support for a causal relationship. However, beyond extremes in large P-values, statistical significance should not play a pivotal role in assessing causal relationships. Like other researchers, we suggest an effect-estimation approach based on confidence limits as opposed to a hypothesis-testing approach. Despite this, recent research indicates that P-values continue to be used frequently to guide causal inferences: P-values of 0.04 are assumed to be consistent with causal associations and P-values of 0.06 inconsistent. At the very least, this is an overemphasis of the role of assessing sampling variability vis-a-vis a causal association and is not a recommended practice.

1.10.2 Time sequence

While a cause must precede its effect, demonstrating this fact provides only weak support for causation. Further, the same factor could occur after disease in some individuals and this would not disprove causation except in these specific instances. Many times it is not clear which came first; for example, did the viral infection precede or follow respiratory disease? This becomes a greater problem when we must use surrogate measures of exposure (*eg* antibody titre to indicate recent exposure). Nonetheless, we would like to be able to demonstrate that an exposure preceded the effect or at least develop a rational argument for believing that it did – sometimes these arguments are based largely on plausibility (*ie* which time sequence is more plausible) rather than on demonstrable facts.

1.10.3 Strength of association

This is usually measured by ratio measures such as risk ratio or odds ratio but could also be measured by risk or rate differences. The belief in larger (stronger) associations being causal appears to relate to how likely it is that unknown or residual confounding might have produced this effect. However, because the strength of the association also depends on the distribution of other components of a sufficient cause, an association should not be discounted merely because it is weak. Also, when studying diseases with

very high frequency, risk ratio measures of association will tend to be weaker than with less common diseases.

1.10.4 Dose-response relationship

If we had a continuous, or ordinal, exposure variable and the risk of disease increased directly with the level of exposure, then this evidence supports causation as it tends to reduce the likelihood of confounding and is consistent with biological expectations. However, in some instances, there might be a cutpoint of exposure such that nothing happens until a threshold exposure is reached and there is no further increase in frequency at higher levels of exposure. These circumstances require considerable knowledge about the causal structures for valid inferences. Because certain physiological factors can function to stimulate production of hormones or enzymes at low doses and yet act to reduce production of these at higher levels, one should not be too dogmatic in demanding monotonic relationships.

1.10.5 Coherence or plausibility

The essence of this criterion is that if an association is biologically sensible, it is more likely causal than one that isn't. However, be careful with this line of reasoning. A number of fundamentally important causal inferences have proved to be valid although initially they were dismissed because they did not fit with the current paradigm of disease causation. As an example, when we found out that feedlot owners who vaccinated their calves on arrival subsequently had more respiratory disease in their calves than those who didn't, we didn't believe it – it didn't make sense. However, after more research and a thorough literature search in which we found the same relationship, we were convinced it was true. The problem likely related to stressing already stressed calves which made them more susceptible to a battery of infectious organisms.

Coherence requires that the observed association is explicable in terms of what we know about disease mechanisms. However, our knowledge is a dynamic state and ranges all the way from the observed association being assessed as 'reasonable' (without any biological supporting evidence) to requiring that 'all the facts be known' (a virtually nonexistent state currently). Postulating a biological mechanism to explain an association after the fact is deemed to be insufficient for causal inferences unless there is some additional evidence supporting the existence of that mechanism (Weed and Hursting, 1998).

1.10.6 Consistency

If the same association is found in different studies by different workers, this gives support to causality. This was a major factor in leading us to believe that the detrimental effects of respiratory vaccines on arrival at feedlots were indeed causal. Not only were our studies consistent but there were numerous examples in the literature indicating (or suggesting) potential negative effects of the practice. Our beliefs were further strengthened by publications from experimental work that indicated a plausible explanation for the detrimental effects.

Lack of consistency doesn't mean that we should ignore the results of the first study on a subject, but we should temper our interpretation of the results until they are repeated. This would prevent a lot of false positive scares in both human and veterinary medicine. The same approach might be applied to the results of field trials and, because there is less concern over confounding, we might not need to be as strict. Recent research has indicated that, in human medicine, once 12 studies have reached the same essential conclusion, further studies reaching the same conclusion are given little additional weight in making causal inferences (Holman et al, 2001).

Meta-analysis combines results from a number of studies on a specific exposure factor in a rigorous, well-defined manner (Weed, 2000) and consequently helps with the evolution of consistency. Evidence for or against a hypothesis can be obtained as opposed to dichotomising study results into those that support a hypothesis and those that do not. In addition, explanation of the methods used in meta-analysis tends to provide a clearer picture of the reviewer's criteria for causation than many qualitative reviews (see Chapter 24).

1.10.7 Specificity of association

Based on rigid criteria for causation such as Henle-Koch's postulates, it used to be thought that, if a factor was associated with only one disease, it was more likely causal than a factor that was associated with numerous disease outcomes. We do not believe this now and specificity, or the lack thereof, has no valid role in assessing causation – the numerous effects of smoking (heart, lungs, infant birth weight, infant intelligence) and the numerous causes for each of these outcomes should be proof enough on this point.

1.10.8 Analogy

This is not a very important criterion for assessing causation, although there are examples of its being used to good purpose. This approach tends to be used to infer relationships in cases of human diseases based on experimental results in other animal species. Today, many of us have inventive minds and explanations can be developed for almost any observation, so this criterion is not particularly useful to help differentiate between causal and non-causal associations.

1.10.9 Experimental evidence

This criterion perhaps relates partly to biological plausibility and partly to the additional control that is exerted in well-designed experiments. We tend to place more importance on experimental evidence if the same target species is used and the routes of challenge, or nature of the treatment are in line with what one might expect under field conditions. Experimental evidence from other species in more contrived settings is given less weight in our assessment of causation. Indeed, the experimental approach is just another way to test the hypothesis, so this is not really a distinct criterion for causation in its own right.

SELECTED REFERENCES/SUGGESTED READING

1. Aiello AE, Larson EL. Causal inference: the case of hygiene and health. Am J Infect Control 2002; 30: 503-511.
2. Buck C, Llopis A, Najera E, Terris M. The Challenge of Epidemiology: Issues and Selected Readings. Pan American Health Organization, Washington, 1988.
3. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analyses. Am J Pub Hlth 1998a; 88: 216-222.
4. Diez-Roux AV. On genes, individuals, society and epidemiology. Am J Epidemiol 1998b; 148: 1027-1032.
5. Evans A. Causation and disease: a chronological journey. Am J Epidemiol 1978; 108: 249-258.
6. Frerichs RR. 2003 http://www.ph.ucla.edu/epi/snow.html
7. Heitjan DF. Causal inference in a clinical trial: a comparative example. Control Clin Trials 1999; 20: 309-318.
8. Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965; 58: 295-300.
9. Holman CD, Arnold-Reed DE, deKlerk N, McComb C, English DR. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. Epidemiology 2001; 12: 246-255.
10. Ingram DG, Mitchell WR, Martin SW. eds. Animal Disease Monitoring. CC Thomas, Springfield, Illinois, 1975.
11. Kaufman JS, Poole C. Looking back on "causal thinking in the health sciences". Annu Rev Pub Hlth 2000; 21: 101-119.
12. Koopman JS, Weed DL. Epigenesis theory: a mathematical model relating causal concepts of pathogenesis in individuals to disease patterns in populations. Am J Epidemiol 1990; 132: 366-390.
13. Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med 1994; 39: 887-903.
14. Leech FB. A critique of the methods and results of the British national surveys of disease in farm animals. II. Some general remarks on population surveys of farm animal disease. Brit Vet J 1971; 127: 587-592.
15. MacMahon B, Pugh TF. Epidemiology: Principles and Methods. Little Brown, Boston, 1970.
16. Martin SW, Meek AH, Willeberg P. Veterinary Epidemiology: Principles and Methods. Iowa State Press, Ames, 1987.
17. Martin W. If multivariable modelling is the answer, what is the question? Dutch Society of Veterinary Epidemiology and Economics, Wageningen, 1996.
18. McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 1999; 149: 887-897.
19. Meek AH. Veterinary epidemiology: challenges and opportunities in research. Prev Vet Med 1993; 18: 53-60.
20. Pearce N. Analytical implications of epidemiological concepts of interaction. Int J Epidemiol 1989; 18: 976-980.
21. Petitti DB. The implications of alternative views about causal inference for the work of the practicing epidemiologist. Proc of Society for Epidemiologic Research (US). Causal inference. Chapel Hill, NC, 1985.

22. Poole C. Causal values. Epidemiol 2001; 12: 139-141.
23. Priester WA. Collecting and using veterinary clinical data. In Ingram DG, Mitchell WR, Martin SW. eds. Animal Disease Monitoring. CC Thomas, Springfield Illinois, 1975.
24. Robins JM. Data, design and background knowledge in etiologic inference. Epidemiology 2001, 12: 313-320.
25. Rose G. Sick individuals and sick populations. Int J Epidemiol 1985; 14: 32-38.
26. Rothman KJ. Causes. Amer J Epidemiol 1976; 104: 587-592.
27. Rothman KJ, Greenland S. Modern Epidemiology. 2d ed. Lippincott-Raven, Philadelphia, 1998.
28. Schwabe CW. The current epidemiological revolution in veterinary medicine. Part I. Prev Vet Med 1982; 1: 5-15.
29. Schwabe CW. The current epidemiological revolution in veterinary medicine. Part II. Prev Vet Med 1993; 18: 3-16.
30. Schwabe CW. Veterinary Medicine and Human Health. Williams and Wilkins, Baltimore 3d ed., 1984.
31. Schwabe CW, Riemann HP, Franti CE. Epidemiology in Veterinary Practice. Lea and Febiger, Philadelphia, 1977.
32. Shy C. The failure of academic epidemiology: witness for the prosecution. Am J Epidemiol 1997; 145: 479-484.
33. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Iowa State Press, Ames, Iowa, 1989.
34. Susser M. Causal Thinking in the Health Sciences: concepts and strategies of epidemiology. Oxford University Press, Toronto (out of print), 1973.
35. Susser M. Judgement and causal inference: criteria in epidemiologic studies. Am J Epidemiol 1977; 105: 1-15
36. Taubes G. Epidemiology faces its limits. Science 1995; 269: 164-169.
37. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 1991; 44: 221-232.
38. Waldmann MR, Hagmayer Y. Estimating causal strength: the role of structural knowledge and processing effort. Cognition 2001; 82: 27-58.
39. Weed DL. Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol 2000; 29: 387-390.
40. Weed DL. Environmental epidemiology: basics and proof of cause-effect. Toxicology 2002 181-182: 399-403.
41. Weed DL, Hursting SD. Biologic plausibility in causal inference: current method and practice. Am J Epidemiol 1998; 147: 415-425.
42. White PA. Causal attribution and Mill's methods of experimental inquiry: past, present and prospect. Br J Soc Psychol 2000; 39: 429-447.
43. White PA. Causal judgments about relations between multilevel variables. J. Exp Psychol Learn Mem Cogn 2001; 27: 499-513.

2

SAMPLING

OBJECTIVES

After reading this chapter, you should be able to:

1. Select a random, simple, systematic, stratified, cluster or multistage sample – given the necessary elements.

2. Recognise the advantages and disadvantages of each sampling method.

3. Select the appropriate sampling strategy for a particular situation, taking into account the requirements, advantages and disadvantages of each method.

4. List the elements that determine the sample size required to achieve a particular objective and be able to explain the effect of each upon the sample-size determination.

5. Compute required sample sizes for common analytic objectives.

6. Understand the implications of complex sampling plans on analytic procedures.

7. Select a sample appropriately to detect or rule out the presence of disease in a group of animals.

2.1 INTRODUCTION

2.1.1 Census vs sample

For the purposes of this chapter, we will assume that data are required for all individuals (animals, herds *etc*), or a subset thereof, in a population. The process of obtaining the data will be referred to as measurement.

In a census, every animal in the population is evaluated. In a sample, data are only collected from a subset of the population. Taking measurements or collecting data on a sample of the population is more convenient than collecting data on the entire population. In a census, the only source of error is the measurement itself. However, even a census can be viewed as a sample because it represents the state of the population at one point in time and hence, is a sample of possible states of the population over time. With a sample, you have both measurement and sampling error to contend with. A well-planned sample, however, can provide virtually the same information as a census, at a fraction of the cost.

2.1.2 Descriptive versus analytic studies

Samples are drawn to support both descriptive studies (often called surveys) and analytic studies (often called observational studies).

A **descriptive study** (or survey) aims to describe population attributes (frequency of disease, level of production). Surveys answer questions such as, 'What proportion of cows in the population has subclinical mastitis?' or, 'What is the average milk production of cows in Prince Edward Island (PEI)?'

An **analytic study** is done to test a hypothesis about an association between outcomes and exposure factors in the population. Analytic studies contrast groups and seek explanations for the differences among them. In epidemiology, they are used to identify associations between risk factors and disease. An analytic study might ask a question such as, 'Is barn type associated with the prevalence of subclinical mastitis?' or, 'Is subclinical mastitis associated with milk production?' Establishing an association is the first step to inferring causation, as was discussed in Chapter 1.

The distinction between descriptive and analytic studies is discussed further in Chapter 7.

2.1.3 Hierarchy of populations

There is considerable variation in the terminology used to describe various populations in a study. In this text, we will consider three: the external population, the target population and the study population. These will be discussed with reference to a study designed to quantify post-surgical mortality in dogs.

The **external population** is the population to which it might be possible to extrapolate results from a study. It is often not defined and might vary depending on the perspective

of the individual interpreting the results of the study. For example, the investigators conducting a post-surgical mortality study might have considered all dogs undergoing surgery in Canadian veterinary clinics as the external population, while someone reading the results of the study in the United States might evaluate the study assuming the external population was all dogs undergoing surgery in North America. The **external validity** relates to the capacity to extrapolate results from a study to the external population (discussed further in Chapter 12).

The **target population** is the immediate population to which the study results will be extrapolated. The animals included in the study would be derived (in some manner) from the target population. For example, if the post-surgical mortality study was to be conducted in PEI, dogs undergoing surgery in veterinary clinics/hospitals in PEI would be the target population. The **internal validity** relates to the validity of the study results for members of the target population (see Chapter 12).

The **study population** is the population of individuals (animals or groups of animals) selected to participate in the study (regardless of whether or not they actually participate). If three veterinary clinics were randomly selected as sites at which post-surgical mortality would be recorded, dogs having surgery at those three clinics would make up the study population.

One important consideration you must address when taking a sample is: does the study population truly represent the target population? If you want to quantify post-surgical mortality in dogs, you could do it at a veterinary teaching hospital; however, the types of patient seen there are much different than those at general veterinary practices and surgical management might also be different. This would make it difficult to generalise the results from such a study. Overall, it is much more important that the study population be representative of the target population if you are doing a descriptive study. Results from an analytic study (*eg* an association between an exposure and a disease) can often be extrapolated to a target population even if the study population has some characteristics that make it different from the target population.

2.1.4 Sampling frame

The **sampling frame** is defined as the list of all the **sampling units** in the target population. Sampling units are the basic elements of the population that is sampled (*eg* herds, animals). A complete list of all sampling units is required in order to draw a simple random sample, but it might not be necessary for some other sampling strategies. The sampling frame is the information about the target population that enables you to draw a sample.

2.1.5 Types of error

In a study based on a sample of observations, the variability of the outcome being measured, measurement error, and sample-to-sample variability all affect the results we obtain. Hence, when we make inferences based on the sample data, they are subject to

error. Within the context of an analytic study, there are two types of error in statistics:

Type I (α) error: You conclude that the outcomes are different when in fact they are not.

Type II (β) error: You conclude that the outcomes are not different when in fact they are.

A study was carried out to determine if an exposure had an effect on the probability of disease occurrence or not. The results shown in Table 2.1 are the possible outcomes.

Table 2.1 Types of error

		True state of nature	
		Effect present	Effect absent
Conclusion of statistical analysis	Effect present (reject null hypothesis)	Correct	Type I (α) error
	No effect (accept null hypothesis)	Type II (β) error	Correct

Statistical test results reported in medical literature are aimed at disproving the **null hypothesis** (*ie* that there is no difference among groups). If differences are found, they are reported with a P-value which expresses the probability that the observed differences could be due to chance, and not due to the presence of the factor being evaluated. P is the probability of making a Type I (α) error. When $P<0.05$, we are 'reasonably' sure that any effect detected is not due to chance.

Power is the probability that you will find a statistically significant difference when it exists and is of a certain magnitude; (*ie* power=$1-\beta$). The probability of making a Type II (β) error, or failing to detect a difference, is seldom stated because usually only positive results are reported in the literature. So-called **negative findings** (failure to find a difference) are seldom reported. There are a number of reasons why a study might find no effect of the factor being investigated.

- There truly was no effect.
- The study design was inappropriate.
- The sample size was too small (low power).
- Bad luck.

An evaluation of the power of the study will at least determine how likely you are to commit this error for a given alternative hypothesis.

2.2 NON-PROBABILITY SAMPLING

Samples that are drawn without an explicit method for determining an individual's probability of selection are known as **non-probability samples**. Whenever a sample

is drawn without a formal process for random selection, it should be considered a non-probability sample, of which there are three types: judgement, convenience, and purposive. Non-probability samples are inappropriate for descriptive studies except in the instance of initial pilot studies (and even then, use of non-probability samples might be misleading). However, non-probability sampling procedures are often used in analytic studies.

2.2.1 Judgement sample

This type of sample is chosen because, in the judgement of the investigator, it is 'representative' of the target population. This is almost impossible to justify because the criteria for inclusion and the process of selection are largely implicit, not explicit.

2.2.2 Convenience sample

A convenience sample is chosen because it is easy to obtain. For instance, nearby herds, herds with good handling facilities, herds with records that are easily accessible, volunteer herds *etc* might be selected for study. Convenience sampling is often used in analytic studies where the need to have a study population that is representative of the target population is less strict. For example, Chapter 17 will focus on the relationship between ultrasound measurements taken in beef cattle at the start of the finishing period and the final carcass grade of the animals. Even though the study was from a convenience sample of herds, the results would probably be applicable to beef cattle in general, provided they were fed and managed under reasonably comparable conditions.

2.2.3 Purposive sample

The selection of this type of sample is based on the elements possessing one or more attributes such as known exposure to a risk factor or a specific disease status. This approach is often used in observational analytic studies. If a random sample is drawn from all sampling units meeting the study criteria, then it becomes a probability sample from the subset of the target population.

2.3 PROBABILITY SAMPLING

A **probability sample** is one in which **every** element in the population has a known **non-zero** probability of being included in the sample. This approach implies that a formal process of random selection has been applied to the sampling frame. The following sections will describe how to draw different types of probability sample. Procedures for analysing data derived from the samples will be discussed in section 2.9.

2.4 SIMPLE RANDOM SAMPLE

In a **simple random sample**, every element in the target population has an equal probability of being included. A complete list of the target population is required and a formal random process is used (random is **not** the same as haphazard). Random sampling can be based on drawing numbers from a hat, using computer-generated random numbers, using a random-numbers table, flipping a coin or throwing dice.

For example, suppose you wish to draw a sample of the 5,000 small animal patients in a veterinary clinic to determine the proportion whose vaccinations are up to date. You require a sample of 500. You could draw up a list of all 5,000 patients, number each name on the list, and then randomly pick 500 numbers between 1 and 5,000. These numbers would identify the animals whose records you would examine.

2.5 SYSTEMATIC RANDOM SAMPLE

In a **systematic random sample**, a complete list of the population to be sampled is not required provided an estimate of the total number of animals is available and all of the animals (or their records) are sequentially available (*eg* cattle being run through a chute). The **sampling interval** (*j*) is computed as the study population size divided by the required sample size. The first element is chosen randomly from among the first *j* elements, then every j^{th} element after that is included in the sample. It is a practical way to select a probability sample if the population is accessible in some order, but bias might be introduced if the factor you are studying is related to the sampling interval. Consequently, a simple random sample would be preferable, but might not be feasible.

Assume once again that you want a sample of 500 patients in a veterinary clinic. You know how many you need to sample (500) and approximately how many patients there are (5,000) but generating a list of those patients would be very time consuming. However, all of their records are in a file cabinet. You need to sample every 10^{th} patient. To start, randomly pick a number between 1 and 10, then pull out every 10^{th} file after that to obtain the data. Data from a systematic random sample are analysed as though they were derived from a simple random sample.

2.6 STRATIFIED RANDOM SAMPLE

Prior to sampling, the population is divided into mutually exclusive strata based on factors likely to affect the outcome. Then, within each stratum, a simple or systematic random sample is chosen. The simplest form of stratified random sampling is called **proportional** (the number sampled within each stratum is proportional to the total number in the stratum). There are three advantages of **stratified random sampling**.

1. It ensures that all strata are represented in the sample.
2. The precision of overall estimates might be greater than those derived from a simple random sample. The gain in precision results from the fact

that the between-stratum variation is explicitly removed from the overall estimate of variance.

3. It produces estimates of stratum-specific outcomes, although the precision of these estimates will be lower than the precision of the overall estimate.

For example, assume you believe that cats are less likely to be up to date on vaccines than dogs are. You would make up two lists – one of cats and one of dogs – and sample from each list. If 40% of the patients are cats, then 500*0.4=200 cats would be selected, and 300 dogs would be selected.

2.7 CLUSTER SAMPLING

A **cluster** is a natural or convenient collection of elements with one or more characteristics in common. For example:

- a litter is a cluster of piglets,
- a dairy herd is a cluster of cattle,
- a pen in a feedlot is a cluster of cattle, and
- a county is a cluster of farms.

In a cluster sample, the **primary sampling unit** (PSU) is larger than the unit of concern. For example, if you wanted to estimate the average serum selenium level of beef calves in PEI, you could use a cluster sample in which you randomly selected farms, even though the unit of concern is the calf. In a cluster sample, every element within the cluster is included in the sample.

Cluster sampling is done because it might be easier to get a list of clusters (farms) than it would be to get a list of individuals (calves), and it is often less expensive to sample a smaller number of clusters than it is to travel around to collect information from many different clusters.

In this example of cluster sampling, a survey to determine the average serum selenium level of beef calves in PEI was conducted. Fifty herds were selected from a provincial herd list and every calf in each of the 50 herds was bled at weaning. A cluster sample is convenient because it is impossible to get a complete list of beef cattle in PEI, but it is easy to get a list of the beef producers. It is also more practical to sample all cattle on 50 farms than it is to drive around to all ~300 beef farms in PEI and sample a few animals on each farm. Of course, calves within a herd are probably more alike than calves from different farms, so the sampling variation for a given number of individuals is greater than if they had been chosen by simple random sampling.

When a group is not a cluster In cluster sampling, a group is a cluster of individuals. A sample is a cluster sample if the group is the sampling unit and the elements within the group are the unit of concern. When the group is both the sampling unit and the unit of concern, then by definition, the sample is not a cluster sample. For example, the following is **not** a cluster sample: a sample of herds to determine whether or not the herds are infected with a particular disease agent (in this case, the herd is the unit of concern, not the individual animals).

2.8 MULTISTAGE SAMPLING

A cluster might contain too many elements to obtain a measurement on each, or it might contain elements so nearly alike that measurement of only a few elements provides information on the entire cluster. **Multistage sampling** is similar to cluster sampling except that, after the PSUs (*eg* herds) have been chosen, then a sample of secondary units (*eg* animals) is selected. Assume again that you are interested in the serum selenium level of beef calves at weaning, and that within-farm variation is small. That means that you don't need to sample very many cattle on a particular farm to get a good estimate of the serum selenium level of all the calves on that farm. Consequently, you might only sample a small number of individuals on each farm.

If you want to ensure that all animals in the population have the same probability of being selected, two approaches are possible. First, the PSUs chosen might be selected with a probability proportional to their size. In other words, if the herd size is known ahead of time, large herds should have a higher probability of being chosen than small herds. After the number of herds is chosen, you select a fixed number of calves in each herd to get serum samples from. If herd size is not known ahead of time, take a simple random sample of the PSUs and then sample a constant proportion of the calves in each herd. Either approach will ensure each animal has the same probability of selection. If this is not the case, the probability of selection needs to be accounted for in the analysis (see section 2.9.2).

How many herds and how many animals to sample within each herd depend upon the relative variation (in the factor(s) being measured) between herds, compared with within herds, and the relative cost of sampling herds compared with the cost of sampling individuals within herds. In other words, when the between-herd variation is large relative to the within-herd variation, you will have to sample many more herds to get a precise estimate. Multistage sampling is very flexible where cost of sampling is concerned. If you are like most researchers, you are working on a limited budget and, when it is expensive to get to herds, you will want to sample as few as possible. On the other hand, if the cost of processing samples from an individual animal is high relative to the cost of getting to the farm, you will want to sample fewer animals per farm. It is desirable to have the most precise estimate of the outcome for the lowest possible cost. These two desires can be balanced by minimising the product of the variance and the cost. Regardless of the total sample size for the study (n), the variance*cost product can be minimised by selecting n_I individuals per herd according to the following formula:

$$n_I = \sqrt{\frac{\sigma_H^2}{\sigma_I^2} * \frac{c_I}{c_H}} \qquad \textit{Eq 2.1}$$

where n_I is the number of individuals to be sampled per herd, σ_H^2 and σ_I^2 are the between- and within-herd variance estimates and c_H and c_I are the costs of sampling herds and individuals, respectively. The value for n_I needs to be rounded to an integer value and cannot be less than 1. Once the number of individuals per herd has been determined, the number of herds to be sampled is then $n_H=n/n_I$.

Keep in mind that cluster and multistage sampling almost always require more subjects for the same precision than simple random sampling. Example 2.1 describes a stratified multistage sampling approach. Multistage sampling, as the name suggests, can be extended to more than the two levels discussed above.

Example 2.1 Multistage sampling
data=dairy_dis

A study was conducted in the three Maritime provinces of eastern Canada to determine the prevalence of serologic reactions to three infectious diseases of dairy cattle: Johne's disease (Map), enzootic bovine leukemia virus and *Neospora caninum*. The dataset is described in Chapter 27. The study had the following characteristics:

- The **external population** was all dairy herds in the region.
- The **target population** was all dairy herds in the region that participated in an official milk-recording programme (approximately 70%).
- The **sampling frame** was a list of all herds in the target population (provided by the milk-production testing programme).
- Sampling was **stratified** by province with 30 herds being randomly selected within each province.
- Sampling was carried out as **multistage sampling** with the herds being selected first and then 30 cows randomly selected within each of the herds. The sampling frame within each herd was the list of cows on the milk-recording programme.
- The **study population** consisted of the animals selected for participation in the study.
- All random sampling was performed using **computer-generated, random numbers**.

These data will be used in Examples 2.2 through 2.4.

2.9 Analysis of Survey Data

The nature of the sampling plan needs to be taken into account when analysing data from any research project involving a complex sampling plan. (**Note** Although referred to as 'survey' data, the concepts discussed in this chapter apply equally to the analysis of data from analytic studies based on complex sampling plans.). There are three important concepts that have been raised in the above discussion of various sampling plans: stratification, sampling weights and clustering. In addition to these, the possibility of adjusting estimates derived from finite populations must be considered.

2.9.1 Stratification

If the population sampled is divided into strata prior to sampling, then this needs to be accounted for in the analysis. For example, in a study of the prevalence of Johne's disease in cattle herds, the herds might be divided into dairy and beef. The advantage of such stratification is that it provides separate stratum-specific estimates of the

outcome of interest. If the factor upon which the population is stratified is related to the outcome (*eg* prevalence of Johne's in the two strata), then the standard error (SE) of the overall prevalence estimate might also be lower than if a non-stratified sample was taken. Correct accounting for the stratified nature of the sample requires that the total population size in each stratum be known in order to get the sampling weights correct (section 2.9.2).

In Example 2.2, the *Neospora* data have been analysed ignoring the stratification by province, and then by taking it into account.

Example 2.2 Analysis of stratified survey data
data=dairy_dis

Valid test values for *Neospora caninum* were obtained from 2,425 cows. A simple estimate (treating the sample as a simple random sample) of the overall seroprevalence was 0.1905 (19.05%) and the SE of that estimate was 0.0080 (0.80%).

If the data are stratified by province, the seroprevalence estimates are as follows:

		Seroprevalence	
Province	Number of samples	Prevalence	SE (prevalence)
1	810	0.1012	0.0106
2	810	0.2111	0.0143
3	805	0.2596	0.0155
Overall	2425	0.1905	0.0080

There are considerable differences across the provinces in terms of the seroprevalence of *N. caninum*. The SE of the overall estimate from the stratified sample is slightly smaller than when the data were treated as a simple random sample, but the difference is minimal. Stratification alone does not change the overall point estimate of the prevalence. **Note** This analysis is provided for pedagogical purposes only. It would not be correct to ignore the sampling weights (section 2.9.2) given that the non-proportional sampling was carried out across strata.

2.9.2 Sampling weights

Although probability sampling requires that a formal random process be used to select the sample, it does not imply that all units sampled have the same probability of selection. If a sample of herds is selected from a target population and a sample of cows is selected within each of those herds, then the probability of selection for any given cow can be computed as:

$$\text{p(selection)} = \frac{n}{N} * \frac{m}{M} \qquad \textit{Eq 2.2}$$

where n is the number of herds in the sample, N is the number of herds in the target population, m is the number of cows that were selected from the sampled herd, and M is the number of cows in that herd. For example, assume that 10 herds are selected out of 100 in a region and that in each herd, 20 animals are sampled. If herd A is an 80-cow herd, the probability that a cow in that herd will ultimately end up in the sample is:

$$10/100 * 20/80 = 0.025 \ (2.5\%)$$

Similarly, if herd B is a 200-cow herd, the probability that a cow in that herd will be in the sample is:

$$10/100 * 20/200 = 0.01 \ (1\%)$$

These different probabilities of selection need to be taken into account in order to obtain the correct point estimate of the parameter of interest.

The most common way of forming sampling weights is to make them equal to the inverse of the probability of being sampled. This value reflects the number of animals that each of the sampled individuals represent. For example, a cow in herd A would actually represent 1/0.025=40 cows in total. A cow in herd B would have a sampling weight of 1/0.01=100 because she had a much smaller probability of selection.

In Example 2.3, the overall prevalence of *Neospora* has been computed taking sampling weights into consideration.

Example 2.3 Analysis of weighted survey data

data=dairy_dis

Cows within the study population had different probabilities of being selected for the sample. Two factors influenced this:

- the probability that the herd would be selected
- the probability that the cow would be selected within the herd.

Herd selection probability: Within each province the probability of a herd being selected was 30 divided by the total number of herds on the milk-recording programme in the province. For example, herd 2 was in province 3, in which there were 242 herds on milk recording. Consequently, the probability of this herd being selected was 30/242=0.1240 (12.40%).

Cow selection probability: Within each herd, the probability of a cow being selected was the total number of cows sampled within the herd divided by the total number of cows in the herd on the day the herd list was generated. For example, 27 samples were obtained in herd 2, from the 128 cows on the herd list. A cow in this herd (*eg* cow # 86) has a selection probability of 27/128=0.1875 (18.75%).

Overall selection probability: The overall selection probability for cow 86 in herd 2 was the product of the above two probabilities: 0.1240*0.1875=0.0232 (2.32%).

Sampling weights: The sampling weight applied to cow 86 in herd 2 was the inverse of the overall selection probability: 1/0.0232=43.02. Effectively, the results from this cow were considered to represent 43 cows in the population.

Taking the sampling weights into consideration, the overall estimate of the prevalence of *N. caninum* was 0.2020 (20.20%), with an SE of 0.0095 (0.95%). Incorporating weights into the analysis has changed the point estimate of the prevalence and has also increased the SE.

2.9.3 Clustering

Cluster sampling and multistage sampling involve the sampling of animals within groups. Animals within groups are usually more alike (with regard to the outcome being measured) than animals chosen randomly from the population. From a statistical perspective, this means that these observations are no longer independent and this lack of independence must be taken into account in the analysis. Failure to do so will almost always result in estimated SEs that are smaller than they should be.

Clustering occurs at multiple levels. For example, udder quarters are clustered within a cow while the cows are clustered within a herd. In Chapters 20-22, we discuss techniques for evaluating the amount of clustering at each of the possible levels. However, when analysing survey data, one often wants to simply deal with the clustering as a nuisance factor in order to obtain correct estimates of the SEs. The simplest approach to this is to identify the PSU (*eg* herd) and adjust the estimate for all clustering effects at levels below this (*eg* clustering within cows and within herds).

Computation of the appropriate variance estimates in the presence of clustering is not a straightforward matter and requires software specifically designed for the process. One approach to the computation is to use a 'linearisation variance estimate' based on a first-order Taylor series linear approximation (Dargatz, 1996). That is the approach used in Example 2.4, in which the overall prevalence of *Neospora* has been estimated taking the within-herd clustering into account. Herds were the PSU and cows were sampled within herds.

2.9.4 Finite population correction

In most epidemiologic studies, sampling is carried out **without replacement**. That is, once an element has been sampled, it is not put back into the population and potentially sampled again. If the proportion of the population sampled is relatively high (*eg* >10%), then this could substantially increase the precision of the estimate over what would be expected from an 'infinite-sized' population. Consequently, the estimated variance of the parameter being estimated can be adjusted downward by a **finite population correction** (*FPC*) factor of:

$$FPC = \frac{N-n}{N-1} \qquad \textit{Eq 2.3}$$

where N is the size of the population and n is the size of the sample.

Note An *FPC* should not be applied in cases where multistage sampling is carried out, even if the number of PSUs sampled is >10% of the population.

2.9.5 Design effect

The overall effect of the nature of the sampling plan on the precision of the estimates obtained can be expressed as the **design effect** (referred to as deff). The deff is the

Example 2.4 Analysis of multistage survey data
data=dairy_dis

The dairy disease data were sampled in a **multistage** manner with herds being the **primary sampling unit**. If the multistage nature of the sample was taken into account (in addition to the stratification and sampling weights), the overall prevalence estimate remains at 0.2020 (20.20%) but the SE increases to 0.0192 (1.92%).

A summary of the estimates of the overall seroprevalence taking various features of the sampling plan into account is shown below.

	Seroprevalence	
Type of analysis	Estimate	SE
Assuming it was a simple random sample	0.1905	0.0080
Taking stratification into account	0.1905	0.0079
Taking stratification and sampling weights into account	0.2020	0.0095
Taking stratification, sampling weights and clustering into account	0.2020	0.0192

The last row contains the most appropriate estimates for the seroprevalence (and SE) of *Neospora caninum*. The design effect from this analysis was 5.5 which indicates that correctly taking the sampling plan into consideration produces an estimate of the variance of the prevalence which is 5.5 times larger than the estimate would have been if a simple random sample of the same size (n=2,425) had been drawn.

ratio of variance obtained from the sampling plan used to the variance that would have been obtained if a comparable-sized, simple random sample had been drawn from the population. A deff >1 reflects the fact that the sampling plan is producing less precise (larger variance) estimates than a simple random sample would have. (Of course, a simple random sample is often impossible to obtain.) The deff of the sampling plan computed in the *Neospora* study is also presented in Example 2.4.

2.10 Sample-size determination

The choice of sample size involves both statistical and non-statistical considerations. Non-statistical considerations include the availability of resources such as time, money, sampling frames, and some consideration of the objectives of the study. Interestingly, cost can be factored into sample-size calculations, and the greater the cost per sampled element, the smaller the sample size when the budget is fixed.

Statistical considerations include the required precision of the estimate, the variance expected in the data, the desired level of confidence that the estimate obtained from sampling is close to the true population value (1-α) and, in analytic studies, the power (1-β) of the study to detect real effects.

2.10.1 Precision of the estimate

Whether you want to determine the proportion of cull cows at slaughter that test positive for Johne's disease or to estimate the average weight of beef calves at weaning, you must determine how precise an estimate you want. The more precise you wish to be, the larger the sample size you will require. If you want to know how many cull cows are Johne's positive within ±5%, you will have to sample more cows than if you were only interested in obtaining an estimate within ±10%. Likewise, if you wanted your estimate of the average weaning weight to be within 2 kg of the real population value, you would need to weigh more calves than if you only needed to be within 5 kg of the true population mean.

2.10.2 Expected variation in the data

The natural variation inherent in the data must be taken into account when calculating sample size. The variance of a simple proportion is $p*q$, where p is the proportion of interest and q is $(1-p)$. Consequently, to estimate the sample size necessary to determine a proportion, then (paradoxical as it might seem) you must have a general idea of the proportion that you expect to find.

The measure of variation used for the estimation of the required sample size of a continuous variable such as weaning weight is the population variance (σ^2). We often don't know what the standard deviation (σ) is, so we have to estimate it. One way to do this is to estimate the range that would encompass 95% of the values and then assume that range is equal to 4σ. For example, if you think that 95% of calves would have weaning weights between 150 kg and 250 kg, then a rough estimate of the σ would be (250–150)/4=25 kg, and the variance would be 625 kg.

2.10.3 Level of confidence

In descriptive studies, we must decide how sure we want to be that the **confidence interval** (CI) from your estimate will include the true population value. Similarly, in analytic studies, we must decide on the certainty we want that any difference we observe between two sampled groups is real and not due to chance. This is referred to as **confidence** and it is most commonly set to 95% (Type I (α) error rate of 5%).

2.10.4 Power

The **power** of a study is the ability of it to find an effect (*eg* a difference between two groups) when a real difference of a defined magnitude exists. For example, if the real difference in weaning weights between male and female calves is 20 kg, then a study with a power of 80% would detect a difference of this magnitude (and declare it statistically significant) 80% of the time. To increase the power, it is necessary to increase the sample size. The Type II (β) error rate is 1-power.

Precision and power have been presented as two separate issues although they arise from the same conceptual basis. Sample sizes can be computed using either approach,

although they will produce different estimates.

2.10.5 Sample-size formulae

The formulae for sample size required to estimate a single parameter (proportion or mean), or to compare two proportions or means, are shown below the following definitions:

Z_α $Z_{0.05}$=1.96 The value of Z_α required for confidence=95%.
Note This is a 2-tailed test. (Z_α is the (1-α/2) percentile of a standard normal distribution)

Z_β $Z_{0.80}$=-0.84 The value of Z_β required for power=80%
Note This is a 1-tailed test.

L=the precision of the estimate (also called the 'allowable error' or 'margin of error') equal to ½ the confidence interval

p=*a priori* estimate of the proportion (p_1, p_2 – estimates in the two groups in an analytic study)

q=1-p

σ^2=*a priori* estimate of the population variance

μ=*a priori* estimate of the population mean (μ_1, μ_2 – estimates in two groups)

Estimating proportions or means
n=sample size

To estimate a sample proportion with a desired precision:

$$n = \frac{Z_\alpha^2 pq}{L^2} \qquad \textit{Eq 2.4}$$

To estimate a sample mean with a desired precision:

$$n = \frac{Z_\alpha^2 \sigma^2}{L^2} \qquad \textit{Eq 2.5}$$

Comparing proportions or means
n=sample size per group

To compare two proportions:

$$n = \frac{\left[Z_\alpha \sqrt{(2pq)} - Z_\beta \sqrt{p_1 q_1 + p_2 q_2}\right]^2}{(p_1 - p_2)^2} \qquad \textit{Eq 2.6}$$

To compare two means:

$$n = 2\left[\frac{(Z_\alpha - Z_\beta)^2\sigma^2}{(\mu_1 - \mu_2)^2}\right] \qquad \textit{Eq 2.7}$$

Note The formulae shown above are approximations and most software will compute sample sizes using more 'exact' formulae.

Sampling from a finite population

If you are sampling from a relatively small population, then the required sample size (n') can be adjusted downward using the following *FPC* factor:

$$n' = \frac{1}{1/n + 1/N} \qquad \textit{Eq 2.8}$$

where n=the original estimate of the required sample size in an infinite population and N=the size of the population.

It is useful to make this finite population adjustment when computing the sample size for a simple or stratified random sample if the sampling fraction exceeds 10%. It is only applied to descriptive studies, not to analytic studies or controlled trial sample size calculations.

Example 2.5 shows the calculation of a sample size for a study comparing two proportions.

Example 2.5 Sample size for comparing proportions

Assume that you want to determine if a vaccine (administered at the time of arrival) reduces the risk of respiratory disease in feedlot steers. For the vaccine to be worth using, you would want it to reduce the risk from the current level of 15% to 10% of animals affected. You want to be 95% confident in your result and the study should have a power of 80% to detect the 5% reduction in risk.

$p_1 = 0.15$ $p_1 = 0.10$ $p_2 = 0.125$

$q_1 = 0.85$ $q_1 = 0.90$ $q_2 = 0.875$

$Z_{0.05} = 1.96$ $Z_{0.80} = -0.84$

$$n = \frac{\left[1.96\sqrt{2 * 0.125 * 0.875} - (-0.84)\sqrt{0.15 * 0.85 + 0.10 * 0.90}\right]^2}{(0.15 - 0.10)^2}$$

$$= 676$$

Consequently, you would require 1,352 (676*2) animals with 676 being vaccinated and the rest not vaccinated. A sample size derived using exact formulae is 726 animals per group.

2.10.6 Adjustment for clustering

In veterinary epidemiologic research, we often deal with clustered data (*eg* cows clustered within herds) with units within the cluster (*eg* cows) being more similar to each other with respect to the outcome than observations drawn randomly from the population. If our study is taking place exclusively at the lower (cow) level, with the factor of interest distributed at the cow level independent of the herd, and the outcome measured at the cow level, this clustering does not present a problem when computing the necessary sample size. Such a situation arises when conducting a controlled trial of a treatment that is randomly assigned to cows within herds (ensuring that treatment allocation is independent of herd) and the outcome is measured at the cow level (*eg* days from calving to conception in dairy cows).

However, if the factor of interest is something that occurs at the herd level (*eg* barn type: freestall vs tiestall), then the number of herds in the study becomes a more critical concern than the number of cows (even though the outcome is measured at the cow level). The total sample size will need to be increased with the magnitude of the increase depending on:

1. the degree to which observations within a herd are similar (measured by a parameter called the intra-cluster (or intra-class) correlation coefficient) (section 21.2.1) and,
2. the number of cows sampled per herd (having many cows sampled within a herd is of little value if the cows within a herd are very similar). The formula for adjusting the sample size is:

$$n' = n(1 + \rho(m-1)) \qquad \textit{Eq 2.9}$$

where n' is the new sample size, n is the original sample size estimate, ρ is the intra-cluster correlation coefficient and m is the number of cows sampled per herd. See Chapter 20 for further discussion of this issue. In Example 2.6, the sample size estimate from Example 2.5 is adjusted for a group-level study.

If the factor of interest is measured at the cow level (*eg* parity), but also clusters within herds (*ie* some herds have older cows than other herds), then the required sample size will lie somewhere between the simple estimate (ignoring clustering) and the much more conservative estimate required for herd-level variables. Which of these two extremes it lies closest to will depend on how highly 'clustered' the factor of interest is within herds.

2.10.7 Adjustment of sample size in multivariable studies

If you want to consider confounding and interaction (Chapter 13) in your study, you generally need to increase your sample size (Smith and Day, 1984). If the confounder is not a strong confounder (odds ratio (*OR*) with disease and exposure between 0.5 and 2), then about a 15% increase is needed. If it is a stronger confounder, then a greater increase in study size should be used. For continuous-scaled confounders, consider the correlation of the confounder with the exposure variable ρ_{ce}. The increase in sample size is $(1-\rho_{ce}^2)^{-1}$. For k covariates, the approximate increase is:

$$n' = n\left(\frac{1+(k-1)\rho_{ce}^2}{1-\rho_{ce}^2}\right) \qquad \textit{Eq 2.10}$$

where ρ_{ce} is an average correlation between the confounder and the exposure variable of interest. Thus, for five covariates with a ρ_{ce} approximately equal to 0.3, the increase in study size is 50%.

A similar approach was used by Hsieh et al (1998). They started with a simple approach to estimating sample size for one covariate and then modified this for the multivariable situation using the **variance inflation factor** (*VIF*).

$$n' = n*VIF \qquad \textit{Eq 2.11}$$

where $VIF = 1/(1-\rho^2_{1,2,3,\ldots,k})$.

Note that $\rho^2_{1,2,3,\ldots,k}$ is the squared multiple correlation coefficient (between covariate 1 and the remaining k-1 variables) or, the proportion of variance of factor 1 that is explained when it is regressed on the other k-1 variables. In general, as ρ increases, then the multiple correlation increases, as does the *VIF*. The approach to estimating the *VIF* is the same for both continuous and binary covariates.

Example 2.6 Sample size with clustering
data=none

If it is not possible to randomly assign the vaccine or placebo to steers within a pen and then keep track of individuals through their feeding period, then you might want to conduct the study by randomly assigning some pens to be vaccinated and other pens to receive the placebo. Rates of respiratory disease tend to be highly clustered within pens and, from previous work, you know the intra-class correlation (ρ) for respiratory disease in pens in feedlots is about 0.3

Assuming that there are about 50 steers in each pen, the revised sample size that you will need will be:

$$\begin{aligned} n' &= n(1+\rho(m-1)) \\ &= 676(1+0.3(50-1)) \\ &= 10613 \end{aligned}$$

Consequently, you will need 10,613 steers per group or 10,613/50=212 pens allocated to each group. This very large increase in sample size is a function of the fact that the intra-cluster correlation for respiratory disease is quite high and we are using a large number of observations (50) in each pen.

2.10.8 General approaches to sample-size estimation

As indicated in section 2.10.5, computing sample size for analytic studies (*eg* comparing two means) can be done either by specifying the desired power of the study to detect a difference of a defined magnitude, or by specifying the desired width of the CI for the difference being estimated (*ie* a precision-based approach). In simple situations, these calculations are relatively straightforward. Two approaches to generalising these calculations for more complex study designs are described below.

Precision-based sample-size computations
The general formula for the width of a confidence interval of a parameter is:

$$par \pm Z * \mathrm{SE}(par) \qquad \textit{Eq 2.12}$$

where *par* is the parameter being estimated, Z is the desired percentile of the normal distribution and SE*(par)* is the SE of the parameter estimate.

Note Z is being used as a large sample approximation for the t-distribution, and for simplicity's sake will be used throughout these examples.

For linear regression models, the SE of any parameter can take the general form of:

$$\mathrm{SE}(par) = \sigma * c \qquad \textit{Eq 2.13}$$

where σ is the estimated standard deviation from the model and c is a value which will depend on the design of the study. For example, for estimating a mean in a single sample:

$$c = \sqrt{1/n} = 1/\sqrt{n} \qquad \textit{Eq 2.14}$$

where n is the sample size.

For a comparison of means from two samples:

$$c = \sqrt{2/n}$$

where n is the sample size in each of the two groups.

The formulae for the CI can be inverted to solve for n. For example, to estimate the difference between two means with the CI of the estimate being $2L$ units long (*ie* $\pm L$), then:

$$L = Z * \sigma * \sqrt{2/n} \qquad \textit{Eq 2.15}$$

Based on this, the sample size required is:

$$n = \frac{2Z^2\sigma^2}{L^2} \qquad \textit{Eq 2.16}$$

Eq 2.15 is the 2-sample analogue of Eq 2.5.

Note Unlike in Eq 2.7, we have not specified a Z_{β}, nor have we specified hypothesised 'true' values for the two means. The sample size estimated is the one required to provide a confidence interval (for the difference) with a desired width ($2L$), regardless of what the actual difference is.

This approach can be generalised to any sort of sample-size estimation, provided that the structure of c can be determined. This is based on the design of the study. For example, computing the sample size required to evaluate a two-way interaction between two dichotomous variables is equivalent to evaluating mean values in each of four possible groups (formed by the possible combinations of the two variables). Consequently:

$$c = \sqrt{4/n}$$

and the sample size required will be:

$$n = \frac{4Z^2\sigma^2}{L^2}$$

This leads to the useful guideline that a study in which you want to evaluate interactions among dichotomous variables needs to be 4 times as large as is required to estimate main effects.

Power calculation by simulation

An approach to power calculation that is applicable to almost any analytical situation is one that is based on simulation. In general, you simulate a large number of datasets that are representative of the type that you are going to analyse and then compute the proportion of times that the main factor you are interested in has a P-value below the level you have set for significance (*eg* 0.05). This approach can be applied to multivariable regression-type models as well as simpler unconditional analyses.

There are two approaches to generating the simulated datasets. In the first (and simplest) approach, you might want to evaluate the power of a study which you have already conducted. For example, let's assume that you have conducted a controlled trial of pre-milking teat-dipping as a means of reducing the frequency of clinical mastitis cases in dairy cows. You did the study in 600 cows (300 in the treatment group and 300 in a control group), with data from one full lactation for each cow. Your outcome (Y) is the number of mastitis cases in each lactation and you are confident that this followed a Poisson distribution. (See Chapter 18 for details of Poisson regression.) Although you randomly assigned cows to the two treatment groups, you still want to control for parity in your analysis so ultimately you fit a Poisson model of the following form:

$$\ln E(Y) = \beta_0 + \beta_1(\text{parity}) + \beta_2(\text{treatment})$$

When you analysed the data, the coefficient for treatment was -0.23 (suggesting that treatment reduced the frequency of mastitis), but it was not significant and you want to determine what power the study had to detect an effect of the magnitude that you found.

The steps involved in determining the power by simulation are:

1. For each observation in the dataset, compute the predicted value based on the coefficients from the model and the particular X values (parity and treatment) for the observation.
2. Generate a random value for the outcome from a Poisson distribution with a mean at the predicted value. (In this case, you don't need to worry about the variance of the distribution because the mean and variance of a Poisson distribution are equal.)
3. Reanalyse the data and note the P-value for the coefficient for the treatment (β_2) effect.
4. Repeat steps 1-3 many times (*eg* 1,000) and determine the proportion of datasets in which the P-value for the treatment effect is <0.05. This is an estimate of the power of the study to detect a true effect corresponding to β_2=-0.23.

Note This post-hoc power calculation has been presented because it is the simplest example of the use of simulation methods for sample-size calculation. In general, post-hoc power calculations are not useful (Smith and Bates, 1992).

If you want to compute sample sizes prior to conducting a study, the process is similar except that you start by creating a hypothetical dataset based on an expected final model. This means that you will need to specify the distributions of the X variables, the size of the dataset, the hierarchical structure of the data (if it is hierarchical in nature; see Chapters 20-22) and all of the relevant variance estimates. A paper outlining the general procedure is available (Feivesen, 2002). An example of the determination of the power of an already-completed study is shown in Example 2.7.

2.11 SAMPLING TO DETECT DISEASE

Sampling to detect the presence (or confirm the absence) of disease is fundamentally different than sampling to estimate a parameter such as the prevalence of disease. If you want to be absolutely certain that a disease is not present in a population, then the only option is to test the entire population (and even this only works if the test you have is perfect). As this is rarely feasible, we rely on the fact that most diseases, if present in a population, will exist at or above some minimal prevalence. For example, we might think that if a contagious disease was present in a population, it would be very unlikely that less than 1% of the population would be infected. Based on this, you can compute a sample size required to be reasonably confident that you would detect the disease if the prevalence was 1% or higher.

If you are sampling from a finite population (*eg* <1,000 animals), then the formula to determine the required sample size is:

$$n = \left(1 - (\alpha)^{1/D}\right)\left(N - \frac{D-1}{2}\right) \qquad \textit{Eq 2.17}$$

where:

n = required sample size

α = 1–confidence level (usually α=0.05)
D = estimated minimum number of diseased animals in the group (population size*minimum expected prevalence)
N = population size

Example 2.7 Power calculation by simulation
data=pig_adg

You have carried out a study to evaluate the effects of internal parasites (ascarids) and respiratory diseases on growth rates in swine. You carry out a regression analysis to evaluate the effects of the presence of adult worms (observed in the intestinal tract at slaughter) on the pig's average daily gain (adg). In this regression analysis, you also adjust for the effects of the sex of the pig and the farm of origin. The important results from that regression analysis are:

- the coefficient for the presence/absence of worms is -7.7 suggesting that pigs with worms in the intestinal tract gained 7.7 gm/day less than pigs without worms.
- the P-value for the coefficient was 0.25 so you have relatively little confidence that the estimate was really different from 0.
- the standard error of prediction for adg was 46.9 gms/day (this represents the standard deviation of predicted results – see Chapter 14).

Your study was carried out in 341 pigs (114 with worms and 227 without) and you want to know how much power such a study had to detect an effect **IF** the real effect of worms was to reduce growth rates by 7.7 gm/day.

You generate 1,000 datasets with randomly generated adg values. For each pig in each dataset, the adg value is drawn from a normal distribution with the following characteristics:

- it has a mean value that corresponds to the predicted value from the real data that you started with (*ie* based on the pig's worm status, sex and farm of origin)
- it has a standard deviation of 46.9 gmday

You analyse each of these new datasets and determine the proportion that gave a P-value for the worms' coefficient that was $\leq$0.05. It turns out that the power was 0.218 (21.8%).

If the true effect of worms was -7.7 gm/day, a study based on 114 positive pigs and 227 negative pigs only had a 21.8% chance of finding a significant effect of worms. This value compares reasonably closely to a power estimate of 29.9% based on a simple comparison of two groups (computations not shown).

If you are sampling from an infinite population, then the following approximate formula can be used:

$$n = \ln\alpha / \ln q \qquad \textit{Eq 2.18}$$

where n=the required sample size, α is usually set to 0.05 or 0.01, q=(1–minimum expected prevalence).

If you take the required sample and get no positive results (assuming that you set α to 0.05), then you can say that you are 95% confident that the prevalence of the disease in

the population is below the minimal threshold which you specified about the disease in question. Thus, you accept this as sufficient evidence of the absence of the disease. Example 2.8 shows the calculation of the required sample size to determine freedom from *Mycoplasma* in a sow herd.

Example 2.8 Sample size for freedom from disease

Assume that you want to document the absence of *Mycoplasma* from a 200-sow herd and that, based on your experience and the literature, a minimum of 20% of sows would have seroconverted if *Mycoplasma* were present in the herd.

$$n = 200 \quad \alpha = 0.05 \quad D = 40$$

$$n = \left(1 - (\alpha)^{1/D}\right)\left(n - \frac{D-1}{2}\right)$$

$$= \left(1 - (.05)^{1/40}\right)\left(200 - \frac{40-1}{2}\right)$$

$$= (0.072)(180.5)$$

$$= 13.02 \cong 13$$

If you test 13 sows and get all negative test results, you can state that you are 95% confident that the prevalence of *Mycoplasma* in the herd is <20%. As you don't believe that the disease would exist at a prevalence <20%, you are confident that it is not present. **Note** This assumes the test is 100% sensitive and specific. See Chapter 5 for a discussion of test characteristics.

SELECTED REFERENCES/SUGGESTED READING

1. Dargatz DA, Hill GW. Analysis of survey data. Prev Vet Med 1996; 28: 225-237.
2. Feivesen AH. Power by simulation. The Stata Journal 2002; 107-124.
3. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat in Med 1998; 1623-1634.
4. Levy PS, Lemeshow S. Sampling of populations: methods and applications. 3d ed. New York: John Wiley & Sons, Inc., 1999.
5. Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology 1992; 3: 449-452.
6. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984; 13: 356-365.
7. VanLeeuwen J, Keefe G, Tremblay R, Power C, Wichtel JJ. Seroprevalence of infection with *Mycobacterium avium* subspecies *paratuberculosis*, bovine leukemia virus, and bovine viral diarrhea virus in Maritime Canada dairy cattle. Can Vet J 2001; 42: 193-198.

SAMPLE PROBLEMS

1. Sampling strategies

The dataset smpltype contains data about the weight gain of 1,114 piglets raised in a 'batch-production system' on six farms in Denmark. On these farms, sows were 'batch farrowed' and a group of piglets was then moved at about three weeks of age from the farrowing barn to the weaner facility. They stayed in this facility until they were approximately nine weeks old and then were moved to the finishing barn. This dataset has data on their growth rates from birth up to their transfer to the finishing barn. The data are a subset of the more complete dataset ap2 (which in turn are part of a larger dataset collected by Dr Håkan Vigre of Denmark). A description of all of the variables in the dataset is included in Chapter 27.

In addition to the original data, this dataset contains indicator variables that identify pigs that were part of a simple random sample, a systematic random sample, a stratified random sample, a cluster sample and a multistage sample.

a. First, compute the overall population mean for daily weight gain (and its SE). Although these data are a 'census' of the whole study population, they are a sample of all pigs going through these six farms, so it is legitimate to compute an SE of the mean.
b. Simple random sample

   A simple random sample of 100 pigs was selected using computer-generated random numbers. This was only possible as we had the complete population enumerated (*ie* in real life, this would not have been possible).

   i. What is the estimate (and its SE) of the daily weight gain of piglets up to the time of transfer, based on the simple random sample?
c. Systematic random sample

   The farms were visited in the following sequence: 3, 6, 1, 4, 2, 5 on the day that a batch of pigs was being transferred from the weaner barn to the finishing barn. As the pigs were run down the alleyway, the 7^{th} pig was sampled and then every 11^{th} pig was sampled. (The order they ran down the alleyway is given in the variable -barn_ord-.) This gave a sample of 101 pigs over the six farms so the last pig was dropped from the sample to give a final sample size of 100 pigs.

   i. What is the estimate (and its SE) of the daily weight gain of piglets up to the time of transfer?
   ii. Do you expect this estimate to be more or less precise than the one based on the simple random sample?
   iii. Is this a biased estimate? If so, why?
d. Stratified random sample

   The population was divided into four strata based on the parity of the piglets' dam. The strata were parities 1, 2, 3-4, 5+. Within each stratum, a simple random sample of 25 pigs was selected using computer-generated random numbers. Once again, this was only possible because we had the complete population enumerated.

   i. What is the estimate (and its SE) of the daily weight gain of the piglets?
      1. First, compute this without paying attention to what the sampling

probabilities were.
2. Second, incorporate sampling probabilities into your estimate. Which of these two estimates is better? Why?

ii. Is this estimate more, or less precise than the one from the simple random sample? Why?

iii. What is the main advantage of a stratified random sample over a simple random sample in this instance?

e. Cluster sample

Two herds (# 2 and # 6) were randomly chosen. All pigs being transferred from the weaner barn to the finishing barn were selected for the sample giving a total sample size of 460 pigs.

i. What is the estimate (and its SE) of the daily weight gain of the piglets?
1. First, ignore the fact that herds were randomly selected before the piglets were.
2. Second, take the sampling plan into account in the analysis. What effect does this have on the precision of the estimate?

ii. Do you need to take sampling weights into account in this analysis?

f. Multistage sample

The same two herds (#2 and # 6) were selected, but within each herd, 50 pigs were randomly selected, giving a sample size of 100 pigs.

i. What is the estimate (and its SE) of the daily weight gain of the piglets?
1. First, ignore the fact that herds were randomly selected before the piglets were.
2. Second, take the sampling plan into account in the analysis. What effect does this have on the precision of the estimate?

ii. Do you need to take sampling weights into account in this analysis?

2. Sample sizes – population means

You are interested in studying aggressive behaviour in dogs and evaluating whether or not spaying (ovario-hysterectomy) has an influence on that behaviour. You have developed an 'aggression index' which measures the level of aggressive tendencies in a dog. The scale ranges from 0 (absolutely no aggression) to 10 (the proverbial 'junk-yard dog') and can take on non-integer values based on the values from a series of observations. From previous work, you think that scores in intact (non-spayed) bitches are approximately normally distributed with the mean score being about 4.5 and with 95% of bitches scoring between 1 and 8.

a. How large a sample do you need to take if you want to determine the mean aggression index value for a new population that you are about to start working with? **Note** You have not been given an estimate of the standard deviation of the distribution, so you will have to use the available data to estimate one.

b. If you think that spaying increases the mean aggression index by 0.5 units, how large a sample will you need to take to be 80% certain of finding a significant difference (if the true difference is 0.5 units) if you want 95% confidence in your result? How much power would a study with 100 bitches in each group (spayed and non-spayed) have to detect a difference of 0.5 units?

3. Sample sizes – proportions
   You are about to start a research project evaluating risk factors for *Neospora caninum* infection in dairy herds. Previous work has suggested that the presence of a dog on the farm might be a risk factor and that the prevalence of *N. caninum* antibodies in dairy herds is approximately 10% in farms without a dog and 30% in farms with a dog.
   a. Assuming that approximately one-half of all farms have a dog (*ie* your best guess as to the overall population prevalence is 20%), how many cows would you have to test to get an estimate of the overall prevalence if you wanted to be 95% certain that your estimate was within 5% of the true prevalence? (Assume you could take a simple random sample from the study population.)
   b. If you wanted to estimate the prevalence within a single 100-cow dairy herd that had a dog, with an allowable error of 10%, how many cows would you need to sample?
   c. Ignoring the fact that the prevalence of *N. caninum* antibodies almost certainly clusters within herds, how many cows would you need to include in your study if you wanted to detect a difference of 10% versus 30% for cows exposed to a dog compared with those not exposed? (Assume a power of 80%.)
   d. You know that *N. caninum* antibodies cluster within herds and you guess that the intra-cluster correlation coefficient is about 0.3. It is also important to note that 'presence of a dog' is a herd-level variable. What impact does this have on your sample size derived in 'c.' if you assume that the average herd size is 50 cows?
   e. While your main interest is in the effect of dogs as a risk factor for infection, you are going to investigate a total of 10 possible risk factors in your study. Assuming that a regression of dog ownership on the other nine factors produces an $p^2(e^2)$ of 0.2. What effect does this have on your sample size estimate?
4. Sample sizes – detecting disease
   A sheep research station that you work with undertook some procedures to eradicate Maedi-Visna from their flock of about 1,000 ewes. Once they thought that they were free of the condition, they decided to check if they really were. They would be satisfied provided that they could be 95% certain that the prevalence in the flock was <1%.
   a. How many sheep do they have to bleed if you assume that a population of 1,000 is essentially 'infinite'?
   b. How does your estimate change if you treat the population as 'finite'?

3

QUESTIONNAIRE DESIGN

OBJECTIVES

After reading this chapter, you should be able to:

1. Plan a questionnaire with appropriate content.
2. Write well-crafted questions for that questionnaire.
3. Format the questionnaire for ease of administration and coding.
4. Pre-test the questionnaire to identify weak points.
5. Code data from the questionnaire as a precursor to data entry.

3.1 INTRODUCTION

Questionnaires are one of the most commonly used tools for collecting data in veterinary epidemiologic research. The terms **questionnaire** and **survey** are often used interchangeably, but we will use them as follows.

> Questionnaire: A data collection tool that can be used in a wide variety of clinical and epidemiologic research settings.
>
> Survey: An observational study designed to generate descriptive information about an animal population. Surveys often use questionnaires as a data-gathering tool.

This chapter will focus on the design of questionnaires regardless of whether they are to be used in a survey or other type of research study. Further discussion of surveys is presented in Chapter 7.

The development of a questionnaire is a complex process involving consideration of many aspects of its design. These are discussed below.

> Every questionnaire must be handcrafted. It is not only that questionnaire writing must be 'artful'; each questionnaire is also unique and original. A designer must cut and try, see how it looks and sounds, see how many people react to it, and then cut again, and try again.
> (Converse and Presser, 1986)

3.1.1 Study objectives

In order for the questionnaire to be effective, it must be carefully planned with consideration given to a number of design elements. First and foremost, it is essential that the objectives and information requirements of the study be established. This process could involve consultation with subject 'experts', and with the ultimate 'users' of the information (if the data are being collected for use by another group, *eg* policymakers). Members of the population to be surveyed should also be consulted in this phase of the planning process. If previous questionnaires covering the subject matter of interest have been published, copies of these questionnaires should be obtained. Previous questionnaires are particularly valuable if a formal validity assessment of the questionnaire has been carried out, but unfortunately, this is not often the case in animal-health studies.

3.1.2 Focus groups

Focus groups consisting of 6-12 people provide an opportunity for a structured form of consultation with members of the intended study population, the end users (target population) and/or the interviewers. An independent moderator can ensure that the focus group stays on topic and the discussion is not dominated by one or two individuals. Focus groups can offer insight into attitudes, opinions, concerns, experiences of the various stakeholders and help to clarify objectives, data requirements, research issues to

be addressed, salient definitions and concepts. To be sure the information is preserved and to avoid ambiguity, the group discussion should be audio or video recorded.

3.1.3 Types of questionnaire

Questionnaires can be **qualitative** or **quantitative**. The former are sometimes referred to as 'explorative' questionnaires and consist primarily of open questions (see section 3.3) designed to allow the participant to express freely their views and thoughts on the subject matter. Qualitative questionnaires can be used in the hypothesis-generation phase of research when it is necessary to identify all of the issues pertaining to the research subject. These types of questionnaire are often administered through interviews and could be taped (with permission) to allow for a detailed evaluation of the content of the material discussed at a later time. Qualitative questionnaires will not be discussed further in this chapter and the reader is referred to Creswell (1998) for more details and Vaarst et al (2002) for a recent example .

Quantitative, or structured, questionnaires are designed to capture information about animals, their environment, their management *etc*. They are more often used in veterinary epidemiology than qualitative questionnaires. All examples used in this chapter are derived from a structured questionnaire designed to capture information about veterinary use of post-operative analgesics in dogs and cats (Dohoo and Dohoo, 1996a,b).

3.1.4 Methods of administration

Questionnaires can be administered through a face-to-face interview, a phone interview, as a mailed questionnaire, or as an internet-based questionnaire. The advantages of a face-to-face interview are that the purpose of the study can be fully explained, a high participation rate can usually be obtained, and audio-visual aids can be used (*eg* photos of medications when ascertaining what products have been used on a farm). Face-to-face interviews also help to develop a rapport between the investigator and participant which might be important if ongoing participation in the study is required. The disadvantages of this approach are that they are time consuming, expensive, geographically limited to areas close to interviewers and might be subject to interviewer bias. This last problem can be avoided, at least in part, by careful training of interviewers.

Telephone interviews share many of the advantages of face-to-face interviews (*eg* high response rate, opportunity to explain the study) but are less time consuming and less expensive. They might be less susceptible to interviewer bias than face-to-face interviews (*eg* no visual cues can be given) but are limited in terms of time that a participant can be expected to spend on the questionnaire. There are also many issues related to telephone communication which need to be considered (*eg* some potential study participants might not have a phone or might have an unlisted number).

Mailed questionnaires are commonly used because they are inexpensive and, being administered by the respondent, have no potential for interviewer bias. However, they are more likely to suffer from low response rates, there is no ability to control

who completes them and they are completely inappropriate if the respondents have poor literacy. A mean response rate (actually a 'risk') of approximately 60% has been reported from a survey of 236 mailed health-related surveys (Asch et al, 1997) although there are many examples of 70%+ response rates. Selection bias is a serious concern if the response rate is low (see Chapter 12), but being able to collect data relatively easily from a widely dispersed study population makes this an attractive option for many studies.

Internet questionnaires have become feasible recently and might even be less expensive than mailed questionnaires. They have the additional advantage that responses can go directly into an electronic database with no data coding and entry required. However, they suffer from the same drawbacks as mailed questionnaires and, in addition, are applicable only to respondents who have access to the internet. Care must also be taken to prevent individuals from completing multiple copies of the questionnaire. A text dealing with the design of internet surveys has recently been published (Dillman, 1999).

3.2 Designing the question

When drafting questions, you must keep in mind: who is responding, whether or not the data are readily available, the response burden (*ie* the length and complexity of the questionnaire), the complexity, confidentiality and sensitivity of the data being collected, the reliability of the data (*ie* validity of question), whether the interviewer or respondent might find any of the topics embarrassing, and ultimately how the data will be processed (coding and computer entry).

Responding to a question usually involves four distinct processes: understanding the question, retrieval of information (from memory or records), thinking and/or making a judgement if the question is at all subjective, and communicating the answer (written or verbal). All aspects must be considered for each question. Once a draft of a question is prepared, ask yourself:

1. Will the respondent understand this question? (The question must be clearly worded in a non-technical manner.)
2. If the question deals with factual information, will the respondent know the answer to the question or have to seek out additional information to be able to answer it? (If additional information is required, the respondent might skip the question or fabricate an answer.)
3. Does answering the question involve a subjective decision? (If it does, is there any way to make it less subjective?) If the question deals with opinions or beliefs, it is bound to be subjective in nature. Special care will be required in the design of these questions to ensure they elicit the desired information.
4. Are the possible responses clear with an appropriate method of recording the response?

Questions can be classified as open (if there are no restrictions on the type of response expected) or closed (if the response has to be selected from a pre-set list of answers).

Both types are discussed below. Regardless of the format, questions can be regarded as a diagnostic test and can be evaluated using the same methods discussed in Chapter 5.

3.3 Open question

In general, open questions (also referred to as open-ended questions) are more often applicable to qualitative than quantitative research because they generate information that might not be applicable for standard statistical analyses. By their nature, open questions allow the respondent to express their opinion. Sometimes we might attach a 'comments' section on a closed question for this purpose.

One type of open question used in quantitative research, particularly for capturing numerical data, is the 'fill-in-the-blank' question. If possible, it is preferable to capture numerical data as a value (*ie* continuous variable) rather than as part of a range. For example, knowing that a dog weighs 17 kg is preferable to simply knowing which of the following ranges the weight falls in: (<10, 10-20, 20-30, >30 kg). Numerical data can be categorised during analysis if need be.

However, in some circumstances, such as when seeking sensitive information (*eg* total family income), a respondent might be more willing to indicate a category (range) than to give a specific numerical value. When capturing numerical data, it is important to specify the units being used (*eg* lb, kg), and it is often desirable to give the respondent a choice of measurement scale (*eg* inches or cm). Example 3.1 shows an open question with an expected numerical response.

Example 3.1 Open question

3. Year of graduation from veterinary school: ☐

Some categorical data are better captured using fill-in-the-blank questions if the range of possible responses is not known before the questionnaire is administered (*eg* for breed of cow: Angus or Angus cross-breed or Angus-Charolais-cross are all possible valid answers).

3.4 Closed question

In designing closed questions (also called closed-ended questions), the researcher can choose from a range of possible options. They include:

- checklist questions (*ie* check all options that apply)
- two-choice/multiple-choice questions
- rating scale questions (*ie* rate the response on a defined scale)

- ranking questions (*ie* rank the options in order of priority).

The advantages of closed questions are that they are generally easier for the respondent to answer (while maintaining consistent responses) and it is easier to code the responses (prior to data entry).

However, closed questions are difficult to design and there is always a risk that closed questions might either oversimplify an issue or elicit answers where no knowledge or previous opinion exists. Sometimes a closed question might request information in a format that is different from what a respondent usually uses (*eg* you might ask for herd-average milk production based on litres per cow per day while the producer assesses milk production using average 305-day production values).

3.4.1 Checklist question

A checklist question is similar to a multiple-choice question except that the respondent is asked to check all responses that apply (so they need not be mutually exclusive or jointly exhaustive). They are equivalent to having a series of 'yes/no' questions for each category. Consequently, each option on the list requires a separate variable in the database.

3.4.2 Two-choice/multiple-choice question

In two-choice/multiple-choice questions it is important to have categories that are **mutually exclusive** (*ie* no overlap) and **jointly exhaustive** (*ie* cover all possibilities). The addition of a category of 'other - please specify' (semi-open question) as the last choice can ensure that the options are jointly exhaustive. However, if the question has been well designed, there should not be a lot of responders using this option. It is recommended that the list of possible choices not exceed five in face-to-face or telephone-interview questionnaires and 10 in mailed/internet questionnaires. There is some evidence that respondents more frequently choose items at the top of a list. This problem can be avoided by having multiple versions of the questionnaire with varying orders to these questions. However, this adds complexity to the data-coding process. Data derived from a two-choice/multiple-choice question can be stored as a single variable in the database (see Example 3.2).

Example 3.2 Multiple-choice question (questionnaire sent only to veterinarians doing some companion animal practice)

6. Type of practice (check one only):
   1. Mixed ☐
   2. Small animal exclusively ☐
   3. Feline exclusively ☐
   4. Referral (please specify type) ☐ _____
   5. Other (please specify type) ☐ _____

3.4.3 Rating question

Rating questions require the respondent to assign a value based on some pre-defined scale. Responses might be ordinal, such as a Likert scale in which the respondent states their level of agreement with a statement (*eg* strongly agree, agree, neither agree nor disagree, disagree and strongly disagree) or recorded on a more continuous numerical scale (*eg* a scale of values from 1 to 10) as in Example 3.3. Continuous data can also be captured using a visual analog scale in which the respondent puts a mark on a line of a given length and the rating assigned is based on how far along the line the mark is (Houe et al, 2002).

Example 3.3 Rating question

In your opinion, how severe would the pain be in dogs in the first 12 hours after each of the following surgeries if no post-operative analgesics were given? Estimate the pain on a 10-point scale where 1 equals no pain at all and 10 equals the worst pain imaginable (circle one number).

11.	Major orthopedic surgery	1 2 3 4 5 6 7 8 9 10 don't know
12.	Repair of ruptured cruciate	1 2 3 4 5 6 7 8 9 10 don't know
13.	Abdominal surgery (non-OHE)	1 2 3 4 5 6 7 8 9 10 don't know
14.	Ovario-hysterectomy (OHE)	1 2 3 4 5 6 7 8 9 10 don't know
15.	Castration	1 2 3 4 5 6 7 8 9 10 don't know
16.	Dental surgery	1 2 3 4 5 6 7 8 9 10 don't know

There are several issues to be considered when developing rating questions. If there are distinct categories, you must decide how many categories there should be and whether or not there should be a middle 'neutral' category (*eg* neither agree nor disagree). It has been suggested that the scale contain a minimum of 5 to 7 points in order to avoid a serious loss of information resulting from translating an underlying continuous response into a series of categories (Streiner and Norman, 1995). For data on a numerical scale, respondents might be unwilling to select values at either end of the scale, particularly if many values (*eg* 1 through 10) are available. It is also advisable to provide an option for 'don't know/no opinion' or 'not applicable' in order to differentiate these responses from ones in which no answer was recorded (*ie* missing data). For practical purposes, data obtained from a rating question with a minimum of five points are often treated as continuously distributed (interval) data in subsequent analyses.

Some rating scales consist of a series of questions with two or more options for each question. Results from this series of questions could be combined to create one or more rating-scale variables. This combination process could be a simple summation of the scores (provided all questions are answered), an average score (provided all questions had the same scale) or could be based on more complex multivariable techniques such as factor analysis (discussed briefly in Chapter 15).

3.4.4 Ranking question

Ranking format questions ask the respondent to order all of the possible responses (or a subset of responses) in some form of rank order (Example 3.4). They are often difficult for respondents to complete, especially if the list of choices is long because all the categories must be kept in their mind at once. In face-to-face interviews, cards with the various responses on them can be prepared and provided to the respondent. This might simplify the ranking process because the respondent only has to choose between a pair of responses at one time (and repeat the process until the cards are in the appropriate rank order).

Example 3.4 Ranking question

Please rank the following as sources of your knowledge of recognition and control of post-operative pain in dogs and cats (1= most important source, 6 = least important source).

37. Undergraduate veterinary school
38. Post-graduate training
39. Journal articles
40. Continuing-education lectures/seminars
41. Experience gained while in practice
42. Discussion with other veterinary practitioners

Rank intervals are unknown to the respondent and might not be equal (*ie* the difference between 2 and 3 is not the same as between 1 and 2). Respondents could frequently assign 'tied' rankings (*ie* the respondent lists two items as one) if they have difficulty choosing between two options. Decisions about how the data will be analysed (including how tied ranks will be handled) should be made before the questionnaire is administered. Computing average ranks for various options assumes that the ranks were approximately equally spaced and this might not be the case. Averaging ranks is also a problem if some possible categories have been omitted as these would influence how the respondent might rank the options that were listed. Alternatively, the proportion of respondents who rank an option highly (*eg* proportion who assign a rank of 1 or 2 to each option) might be computed.

3.5 WORDING THE QUESTION

The wording used in questions has a major impact on the validity of the results from those questions. Vaillancourt et al (1991) recommend that questions not exceed 20 words. It is important to avoid the use of abbreviations, jargon and complex or technical terminology. At all times, bear in mind who the respondent is and what level of technical knowledge they have. For example, 'How many fatal cases of neonatal diarrhea occurred during the time period?' is a poorly worded question if the respondent

is a dairy producer. 'How many calves died from scours during January?' would be more appropriate.

Make the question as specific as possible. For example, if asking for information about annual milk production, specify the time frame (*eg* January 1, 2002 to December 31, 2002) and clearly define how milk production is to be measured (*eg* total weight of bulk tank shipments).

Avoid double-barrelled questions. For example, asking 'Do you think BVD is an important disease that producers should vaccinate for?' is really asking two questions (one about the importance of BVD and one about the utility of vaccination). These issues should be separated into two questions.

Avoid 'leading' questions. Asking a question such as 'Should dogs be allowed to suffer in pain after castration without the benefit of analgesics?' might very likely produce a biased response compared with a more neutral question such as 'Do you think dogs should be given analgesics following castration?'

3.6 Structure of questionnaire

Questionnaires should begin with an introduction explaining the rationale and the importance of the questionnaire, and how the data will be used. In it, you should also assure the respondent of the confidentiality of their answers. Telling the respondent approximately how long it will take to complete the questionnaire will help to improve response rate (provided the questionnaire has been kept to an acceptable length). In mailed questionnaires, the introduction might be incorporated into the first page, but it is usually desirable to have it as part of a separate cover letter that is sent with the questionnaire. For interview format questionnaires, the information must be provided verbally at the start of the interview.

After the introduction, it is a good idea to start with questions that build confidence in the respondent. If it is necessary to give instructions to the respondent, make sure they are clear and concise. Highlight them in some way (*eg* **bold typeface**) to draw attention to them. Remember that people only read instructions if they think they need help.

Questions should be grouped either according to subject (housing, nutrition) or chronologically (calving, breeding period, pregnancy diagnosis). Within a section, questions might follow a 'funnel' approach in which the subject matter is increasingly specific and focused. Pairs of questions which capture essentially the same information ('date of installing a milking system' and 'age of milking system') might be included at different locations in the questionnaire either for verification of critical information or as a general check on the validity of the questionnaire responses.

It is important that mailed (or internet) questionnaires be visually appealing and easy to complete. Professional-looking questionnaires will enhance the respondents perspective on the importance of the study (Salant and Dillman, 1994).

When designing the form layout, consider ease of data coding and entry in order to minimise mistakes and reduce the required effort. If at all possible, questions should be pre-coded (*ie* the numerical codes assigned to possible responses are printed beside the various options). It is advisable to leave space on the questionnaire (*eg* a column down the right-hand edge of the page) to allow for the recording of all responses that are to be entered into a computerised database. This will allow data-entry personnel to simply read down a column of responses rather than having to jump around the page (see Example 3.5).

Example 3.5 Coding questionnaires

The space at the right allows for direct coding of responses on the questionnaire.

	For office use only
1. Sex 1. Male 2. Female	1. [ ]
2. Age ☐ years	2. [ ]
3. Year of graduation from veterinary school ☐	3. [ ]

3.7 PRE-TESTING QUESTIONNAIRES

All questionnaires need to be pre-tested before applying them to the study population. Pre-testing allows the investigator to identify questions that are confusing, ambiguous or misleading and to determine if there are any problems with the layout of, or the instructions on, the questionnaire. When you pre-test a questionnaire, you can determine if there are questions that respondents will be unable or unwilling to answer or perhaps identify additional categories required for multiple-choice questions. It also serves to estimate the time that would be required to complete it.

The first step in pre-testing the questionnaire is to have colleagues or experts in the field evaluate it to ensure all important issues are identified and covered. A single pre-testing on a small sample from the study population can be used to obtain feedback on the clarity of questions. This might be done by having the respondent complete the questionnaire as it will be done in the study and then discussing any problematic aspects. Alternatively, a 'think-aloud' pre-test can be carried out in which the respondent explains all of their thought processes as they work through the questionnaire. It is desirable to have a second pre-test in which the questionnaire is readministered to the same test group of respondents in order to assess the repeatability of questions. The time interval between the two pre-tests needs to be long enough that the respondent does not recall how they answered questions the first time, but short enough that the information being sought is unlikely to have changed. A test-retest evaluation is only valid if the questionnaire is not

changed much after the first pre-test. It will also require quite a few more respondents if the repeatability of the questions is to be evaluated.

3.8 Data coding and editing

Before administering any questionnaire, procedures for coding of responses and computer data entry should be considered. When coding responses, it is wise to have a single value to represent missing values. Do not simply leave these blank as, subsequently, it will be impossible to differentiate items that were not answered on the questionnaire from those that were missed in coding or data entry. A unique value (*eg* -999) that could not be a legitimate answer to any of the questions should be used for missing values. Consistency of coding is important and, because it is convenient to analyse no/yes (dichotomous) variables coded as 0/1, it is advisable to use this coding from the start.

Coding of responses is best accomplished directly on the paper forms (either mailed questionnaires or data capture forms used in interviews). Do not attempt to combine coding and data entry into a single step. It is a good idea to use a distinctive colour of ink for recording all codes on the forms so it is easy to differentiate writing done by the coder from that done by the respondent or interviewer.

Computer data entry can be done using specialised software or general purpose programs such as spreadsheets and database managers. The advantage of specialised software is that it allows you to set validation criteria easily (such as acceptable ranges for values in a given variable) that preclude entry of illogical values. One useful public domain program for data entry is EpiData (freeware http://www/epidata.dk). Spreadsheets must be used with caution. While they are convenient and easy to set up for data entry, the ability to sort individual columns in the spreadsheet makes it possible to completely destroy the data (*ie* responses from one individual will no longer be on the same row). General-purpose database managers are useful and allow greater manipulation of the data. However, because most data will ultimately be transferred to a statistical package for verification and analysis, it is advisable to perform all data manipulations in that statistical package, where it is easier to document and record all procedures carried out. The process of data verification and processing is discussed further in Chapter 25.

SELECTED REFERENCES/SUGGESTED READING

1. Asch DA, Jedrziewski MK, Christakis NA. Response rates to mail surveys published in medical journals. J Clin Epidemiology 1997; 50: 1129-1136.
2. Converse JM and Presser S. Survey questions: handcrafting the standardized questionnaire, Sage Publications, 1986.
3. Creswell JW. Qualitative inquiry and research design - choosing among five traditions. London: Sage Publications, 1998.
4. Dillman DA. Mail and internet surveys: the tailored design method. 2d ed. London: John Wiley & Sons, 1999.
5. Dohoo SE, Dohoo IR. Post-operative use of analgesics in dogs and cats by Canadian veterinarians. Can Vet J 1996a; 37: 546-551.
6. Dohoo SE, Dohoo IR. Factors influencing the post-operative use of analgesics in dogs and cats by Canadian veterinarians. Can Vet J 1996b; 37: 552-556.
7. Houe H, Ersbøll AE, Toft N, Agger JF. , eds. Veterinary epidemiology from hypothesis to conclusion. Copenhagen: Samfundslitteratur KVL Bogladen, 2002.
8. Salant P, Dillman DA. How to conduct your own survey. London: John Wiley & Sons, 1994.
9. Streiner DL, Norman GR. Health measurement scales; a practical guide to their development and use. 2d ed. Oxford University Press, 1995.
10. Vaarst M, Paarup-Laursen B, Houe H, Fossing C, Andersen HJ. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J Dairy Sci 2002; 85: 992-1001.
11. Vaillancourt JP, Martineau GP et al. Construction of questionnaires and their use in veterinary medicine. Proc of Soc. Vet. Epidem and Prev Med, 1991.

4

MEASURES OF DISEASE FREQUENCY

OBJECTIVES

After reading this chapter, you should be able to:

1. Explain the different ways of measuring disease frequency and differentiate among counts, proportions, odds and rates.

2. Describe the difference between incidence and prevalence and when each should be used.

3. Describe the difference between risk and rate as applied to measures of incidence.

4. Elaborate upon the concepts of 'cause-specific measures', proportional morbidity/ mortality rates and case fatality rates.

5. Apply all of the above concepts and select the appropriate measures of disease frequency to be used in specific circumstances.

6. Compute the appropriate measures when provided with the necessary data and calculate exact and/or approximate confidence intervals.

4.1 Introduction

Measurement of disease (or event) frequency is the basis for many epidemiological activities. These include routine surveillance, observational research and outbreak investigations, among others. In observational studies, measuring the frequency of a disease and an exposure, and subsequently linking (or associating) the exposure and the disease are the first steps to inferring causation. The hypothesis we test is described qualitatively but the process involves quantification and begins with measurement of events and exposures.

Morbidity and mortality are the two main categories of events for which frequency measures are calculated. However, there are other events of interest such as culling (the premature removal of animals from a herd or flock), survival to weaning, and pregnancy (*eg* the probability of an animal becoming pregnant within a specified time period). The format for calculating these is the same as it is for morbidity and mortality.

Because both morbidity and mortality are strongly associated with animal (or herd) attributes, and different diseases have different impacts, we usually calculate these measures for specific host attributes (*eg* age, sex, and breed) and for specific diseases (*ie* outcomes of interest).

4.1.1 Some factors affecting the choice of frequency measure

Study period When selecting a measure of disease frequency for use in a study, it is important to consider both the study period and the risk period. The study period is the period of time over which the study is conducted. It is usually measured in terms of calendar time, but sometimes the study period is a point in time. In either instance, the study period could be specified in calendar time or by the event at which the data are collected (*eg* at slaughter or at birth).

Risk period The risk period is the time during which the individual could develop the disease of interest. Thus, an important question is: how long is the risk period? For example, for diseases such as retained placenta in dairy cows, the risk period is short – a day or two at most; whereas, for diseases such as lameness or foot problems, it could be very long.

Both the risk and study period relate to whether the population is deemed to be closed or open (see section 4.4.1). However, disregarding this, diseases with a short risk period (relative to the study period) are good candidates for risk measures. Diseases with long risk periods are likely candidates for rate-based measures. These two approaches to measuring the incidence of disease are discussed in section 4.3.

4.2 Count, proportion, odds and rate

Before discussing specific measures of disease frequency, it is necessary to review the mathematical forms that these measures can take. These include counts, proportions, odds and rates.

Count This is a simple enumeration of the number of cases of disease or number of animals affected with a condition in a given population. Because the size of the population is not taken into consideration, counts of events are of **very** limited use for epidemiologic research.

Proportion This is a ratio in which the numerator is a subset of the denominator. For example, if 200 cows are tested for enzootic bovine leukosis (EBL) and 40 of them are positive, the proportion positive is 40/200=0.2 (or 20%). Prevalence (section 4.7) and risk (sections 4.3, 4.4) are both proportions. In the former, both the numerator and denominator are measured at a point in time. In the latter, the numerator relates to the number of new cases over a period of time so, although proportions have no units, the time period must be specified for the proportion to make sense.

Odds This is a ratio in which the numerator is not a subset of the denominator. For example, if there are three stillborn animals and 120 live births, the odds of stillbirth is 3:120=0.025:1 or 25 stillbirths to 1,000 live births. The odds of EBL (based on the data given above) is 40/160=0.25 (or 1:4).

Rate A rate is a ratio in which the denominator is the number of animal-time units at risk. For example, if there are 30 cases of kennel cough in a 100-dog kennel over a three-month period, the incidence rate is 30/(100*3)=0.1 cases per dog-month. Note the 300 dog-months in the denominator.

Note The term 'rate' is often used in a general sense to refer to all types of measures of disease frequency. Strictly speaking though, it should only be used to refer to measures based on the concept of animal-time units. Similarly, we often say that animals with a high 'chance' of having or getting the disease have a 'high risk' although the underlying measure of frequency might not be a risk.

4.3 INCIDENCE

Incidence relates to the number of new events (*eg* new cases of a disease) in a defined population within a specific period. Because they deal with new cases of disease, studies based on incident cases of disease are used to identify factors associated with an animal becoming ill. Although incidence deals with 'new cases' of disease, it does not necessarily imply just the 'first case' within an animal. For some diseases (*eg* clinical mastitis in dairy cows), multiple cases are possible within an animal, either by involving different quarters of the udder or recurring in the same quarter after a period of absence from that quarter.

For reasons perhaps related to their unique susceptibility, or due to the effect of the first disease occurrence in the animal, animals that develop one case of a disease are often at a much higher risk of developing a subsequent case. Thus, it might be preferable to count only the first case in terms of a disease frequency measure but to enumerate separately the number of occurrences per animal in the study period.

There are three ways of expressing incidence:

- incidence count
- incidence risk (R)
- incidence rate (I).

Incidence count is the simple count of the number of cases of disease observed in a population. It is often used to describe the frequency of a disease in a population in which the disease did not previously exist (*eg* country X has had 12 cases of bovine spongiform encephalopathy (BSE)). It might also be used for some common diseases (*eg* case counts of *Salmonella* in humans) but without data on the number of samples/animals examined, there are limits to the inferences we can make from count data. Incidence counts are rarely used in epidemiologic research unless they are combined with information about the population at risk (*eg* Poisson regression, Chapter 18).

Incidence risk An incidence risk is the probability that an individual animal will contract or develop a disease in a defined time period. Because risk is a probability, it is dimensionless (that is, it has no units) and ranges from 0 to 1. Although risk is dimensionless, the time period to which the risk applies must be specified. For example, the risk of a cow having a case of clinical mastitis in the next year is very different (*ie* much higher) than the risk of having a case in the next week. In addition, only the first occurrence of a disease in the time period of interest is relevant because, once an animal has had one case, it contributes to the numerator of the proportion and what happens to it after that is irrelevant. Risk is used in studies in which making individual predictions is the objective. For example, a study might determine that the probability that a seven-year-old boxer will develop some form of detectable neoplasia over the next year is 14%. Incidence risk is sometimes referred to as **cumulative incidence**. In the context of survival analysis, survival (S) is defined as: S=1-R.

Incidence rate An incidence rate is the number of new cases of disease in a population per unit of animal-time during a given time period. It has units of 1/animal-time, and is positive without an upper bound. If a cattery housing 50 cats has 72 cases of upper respiratory disease over a period of a year, the incidence rate is 72/50, which is 1.44/cat-year (or 0.12/cat-month). Incidence rates are used in studies designed to determine what factors are related to diseases and what the effects of those diseases are. Incidence rates are sometimes referred to as **incidence density**. A related concept is the hazard rate which expresses the theoretical limit of I as the time period approaches zero. Hazard rates are used in survival analysis.

4.4 Calculating risk

Determining the number of new cases requires a clear case definition (*ie* what criteria need to be met for a 'case' to be considered as such) and a surveillance programme capable of identifying all such cases. Risk is most commonly computed at the animal level (*eg* the probability of an eight-year-old dog developing lymphosarcoma within the next year) but can be computed at other levels of aggregation (*eg* the probability of

a dairy herd becoming infected with *Strep. agalactia* in a one-year period). The latter requires a case definition of what constitutes an infected herd.

Risk (R) of disease is estimated as:

$$R = \frac{\text{number of new cases of disease in a defined time period}}{\text{the population at risk}} \qquad \textit{Eq 4.1}$$

4.4.1 Population at risk

While counting the new cases of disease presents some challenges, estimating the population at risk can be even more difficult. The population at risk might be considered 'closed' or 'open'. Regardless of whether the population is closed or open, only animals free of the disease at the start of the study period are considered to be at risk.

Closed population A closed population is one in which there are no additions to the population for the duration of the study and few to no losses. The duration of the study might be defined in terms of calendar time (*eg* a herd of dairy cows followed for the next year) or in terms of some life event (*eg* all cows in a dairy herd followed for the first two months of lactation – regardless of when the lactation starts – to determine the risk of ketosis). Only disease-free animals in the population at the start of the study period are considered to be at risk and are monitored for the outcome of interest. Animals which are lost to follow-up during the study period are called **withdrawals** and the simplest way of dealing with them is to subtract half of the number of withdrawals from the population at risk when computing R (this assumes that, on average, the withdrawals leave halfway through the study period). This correction for withdrawals is derived from (or related to) actuarial life-table methods. Unless there are no withdrawals, the risk estimate is biased. Nonetheless, provided the number of withdrawals is small relative to the population size being studied, the bias is small.

Open population An open population is one in which animals are leaving and entering the population throughout the study period. For example, if you wanted to determine the frequency of lymphosarcoma over a one-year period in a population of dogs served by a single veterinary clinic (assuming that all cases are diagnosed at the veterinary clinic), the population at risk would be an open population of dogs that were served by that clinic. An open population is considered to be stable if the rate of additions and withdrawals and the distribution of host attributes are relatively constant over time.

It is not possible to compute risk directly from an open population but it can be estimated from I (section 4.6). Risk can also be estimated in open populations using methods for the analysis of 'survival' data (Chapter 19).

Sometimes we can define a follow-up period after a specified exposure/event in a manner that converts an open population to a closed population. For example, dairy and swine herds are inherently open in the sense that new animals enter the at-risk group (this use of open is not the same as saying that a farmer does or does not purchase new

'outside' animals). However, if we observe a set of animals, *eg* post-partum, for a full, defined risk period, then the population becomes closed.

4.5 Calculating incidence rates

Incident rates (I) are calculated as:

$$I = \frac{\text{number of cases of disease in a defined time period}}{\text{number of animal-time units at risk during the time period}} \qquad \textit{Eq 4.2}$$

An **animal-time unit** is one animal for a defined period of time (*eg* a cow-month, a dog-day (not to be confused with the 'dog days' in August)).

Incidence rates can be calculated using only the first occurrence of disease for any given animal (and from then on they are not considered to be at risk), or using all occurrences of disease. For example, a neoplastic disease would likely occur only once in an animal's lifetime but some infectious diseases such as mastitis can occur more than once in a dairy cow. However, even for diseases that might occur multiple times, we might only be interested in an animal's first case of mastitis as risk factors for a first case might be different from risk factors for recurrences.

Note The inverse of I ($1/I$) is an estimate of the average time to the occurrence of the disease if the population is closed, or open and stable, providing the outcome is inevitable (all animals achieve it if they live long enough).

As with calculating the number of animals at risk for R, there are several methods for calculating animal-time units at risk for I. The exact method is always preferred, but often the information is not available for you to use the exact method and an approximation must be substituted.

Exact or approximate methods can be adapted for situations when animals are at risk for multiple disease episodes, as opposed to only one disease episode per animal. The important thing to remember is that if you are only interested in the first case of disease, then, after the animal contracts the disease of interest, it is **no longer** at risk! It no longer contributes to the pool of animal-time units at risk, even if it remains in the herd or study.

Exact calculation An exact calculation requires that the exact amount of animal-time contributed by each member of the study population be known. Example 4.1 presents a simple exact calculation.

Approximate calculation If only one case of disease per animal is considered, then I is calculated as:

$$I = \frac{\text{cases}}{(\text{start} - 1/2\ \text{sick} - 1/2\ \text{wth} + 1/2\ \text{add}) * \text{time}} \qquad \textit{Eq 4.3}$$

Example 4.1 Exact incidence rate calculation

Assume four previously healthy animals were observed for exactly one month (30 days). The history for each individual was as follows:

1 animal not sick at all		1.00	animal-month at risk
1 animal sick on day 10		0.33	animal-months at risk
1 animal sick on day 20		0.67	animal-months at risk
1 animal sold on day 15		0.50	animal-months at risk
Total 'population at risk'	=	2.50	animal-months at risk
Total new cases of disease	=	2	
$I = 2/2.5$	=	0.80	cases/animal-month

where: cases = # of new cases
start = # at risk at start of study period
sick = # developing disease
wth = # withdrawn from the population
add = # added to the population
time = length of study period (same for all animals).

If multiple cases of disease per animal are possible, then I is calculated as:

$$I = \frac{\text{cases}}{(\text{start} - 1/2\ \text{wth} + 1/2\ \text{add}) * \text{time}} \qquad \textit{Eq 4.4}$$

Note For relatively rare diseases, the second formula might be used even if the investigator is only interested in 'first cases' because the adjustment to the average population at risk by removing those cases will be very small.

In general, if the risk period is much shorter than the study period, using risk as a measure of disease is appropriate. If the risk period is longer than the study period, then I is a more appropriate measure of disease incidence and the question of whether only one case, or all cases of disease will be counted must be considered.

4.6 Relationship between risk and rate

Another approach to estimating risk is to use the functional relationship between R and I. If complete data are available for a closed population then:

$$R = A/N \text{ and } I = A/(N\Delta t)$$

so

$$R = I\Delta t$$

where A = number of cases, N = population at risk and Δt = length of study period.

However, if only an average rate I is available for a population, then assuming that I is constant over the time period:

$$R = 1\text{-}e^{-I\Delta t} \qquad \textit{Eq 4.5}$$

For example, Table 4.1 shows data obtained from 100 animals followed for a two-year period and the estimation of the annual R based on the average annual I.

Table 4.1 Estimation of R from average I

Year	Population at risk	Cases	Annual I
1	100	22	0.22
2	78[a]	18	0.23

[a]Only 78 at risk because 22 had already had the disease.
Two-year R=22+18/100=0.4.
Average annual I=0.225 cases/animal-year.
Estimated annual $R=I\text{-}e^{-0.225}$=0.20.

Note If $I\Delta t$ is small (*eg* <0.1) then $R \approx I\Delta t$. For example, if I=0.01 cases/animal-year, then the estimated annual $R \approx 0.01$.

4.7 PREVALENCE

Prevalence relates to cases of disease existing at a specific point in time rather than new cases occurring over a period of time. Hence, the prevalence count is the number of individuals in a population that have an attribute or disease at a particular time.

The prevalence proportion (P) (also referred to simply as prevalence) is calculated as:

$$P = \frac{\text{cases}}{\text{par}} \qquad \textit{Eq 4.6}$$

where cases = # of cases of disease in a population at a point in time
par = # of animals in the population at risk at the same point in time.

For example, if you bleed 75 horses from a large riding stable and test for equine infectious anemia (swamp fever) and three test results are positive, P is:

$$P = \frac{3}{75} = 0.04 = 4\%$$

Relationship between prevalence and incidence In a stable population in which I of a disease remains constant (which it rarely does for contagious diseases), P (at any point in time) and I and disease duration (D) are related as follows:

$$P = \frac{I * D}{I * D + 1} \qquad \textit{Eq 4.7}$$

For example, if the incidence rate of subclinical mastitis in a dairy herd is 0.3/cow-year (*ie* 30 new infections/100 cows per year) and the mean duration of an infection is three months (0.25 year), then we would expect P to be:

$$P = \frac{0.3 * 0.25}{0.3 * 0.25 + 1} = 0.07 = 7\%$$

so on any given day throughout the year, we would expect 7% of cows to have subclinical mastitis.

A series of prevalence studies is often used to determine I of diseases which are not easily detected on the basis of clinical signs. This is particularly relevant for determining the rate at which animals become infected with a certain pathogen. For example, by bleeding a group of cats at regular intervals and testing for feline leukemia virus, the rate at which cats are becoming infected can be estimated.

Note P is less useful than I for research into risk factors for diseases because factors that contribute to either the occurrence of disease or its duration will both affect prevalence.

Example 4.2 shows the calculation of various measures of P, R and I.

4.8 Mortality statistics

These statistics are calculated in exactly the same way as P, R and I. The disease event of interest in these statistics is, by definition, death. The term **mortality rate**, strictly speaking, refers to the incidence rate of mortality. However, it is often misused to describe the risk of mortality. You should be alert to this and interpret the literature accordingly. Overall, the mortality rate describes the number of animals that die from all causes in a defined time period and is analogous to I except that the outcome of interest is death. Mortality rate is calculated in the same way as I.

The cause-specific mortality rate, as one would expect, describes the number of animals that die from (or with) a specific disease during a defined time period. This is also calculated the same as I.

Mortality statistics can describe the number of deaths due to a disease or the number of deaths with a disease, but it is often difficult to determine the specific cause of death. For example, if a recumbent cow regurgitates and contracts aspiration pneumonia and then dies, did it die:

- due to recumbency?
- due to pneumonia?
- with pneumonia?

Usually the 'cause' will be the factor which is deemed to be the proximate cause (*ie* the straw that broke the back). As indicated above, that might be a difficult decision to make.

Example 4.2 Calculation of risk and rate

You are interested in determining the frequency of new intramammary infections (IMI) with *Staph. aureus* in dairy cattle so you identify five cows in a dairy herd, follow them for one full lactation (10 months) and culture milk samples at months 0 (calving), 2, 4, 6, 8 and 10 (dry-off). The results are presented in the table below. A cow is only considered to have a new intra-mammary infection if it was negative on the preceding sample.

	Sampling times						Total months at risk	
Cow	0	2	4	6	8	10	First case only	All cases
A	0	**X**	0	0	**X**	X	2	6
B	0	0	0	–	–	–	4	4
C	X	0	0	**X**	X	X	0	4
D	0	0	0	0	0	0	10	10
E	0	0	**X**	0	**X**	X	4	6

where:

X = positive culture

X = positive culture that represents a new IMI

0 = negative culture

– = cow removed from herd

par = population at risk

a) risk of infection during first 2 months of lactation
par = 4 cows
new IMI = 1 cow
2-month R = 1/4 = 0.25

b) risk of infection during lactation
par = 4 - 1/2 (1 withdrawal) = 3.5 cows
new IMI = 2 cows
lactation R = 2/3.5 = 0.57

c) rate of IMI – considering first cases only
par = 20 cow-months
new IMI = 2 first cases
I = 2/20 = 0.1 cases/cow-month
= 1 case/cow-lactation

d) rate of IMI – considering all new IMI
par = 30 cow-months
(*eg* cow A at risk for months 0 to 2 and 4 to 8)
new IMI = 5 cases
I = 5/30 = 0.17 cases/cow-month
= 1.7 cases/cow-lactation

e) lactation risk estimated from lactation rate (first cases only)
I = 1 case/cow-lactation
$R = 1\text{-}e^{-1} = 0.63$

f) prevalence at dry-off
par = 4 cows
existing IMI = 3
P = 3/4 = 0.75

Note We are using the sampling time as the time of occurrence (or withdrawal). Some might prefer to use the midpoint between samplings; we have not done this to keep the calculations simple.

4.9 OTHER MEASURES OF DISEASE FREQUENCY

Virtually all disease frequency measures can be defined in terms of *P*, *R* and *I* provided the outcome of interest, the population at risk and the study period are adequately defined. However, a few specific terms that appear frequently in the literature warrant some attention. Most of these are referred to as rates but are really measures of risk.

4.9.1 Attack rates

Attack rates are used to describe the frequency of disease in outbreak situations. They are computed as the number of cases divided by the size of the population exposed. Consequently, they are really a measure of risk. Attack rates (risk) are used in situations such as outbreaks where the risk period is limited and all cases arising from the exposure are likely to occur within that risk period.

4.9.2 Secondary attack rates

Secondary attack rates are used to describe the 'infectiousness' (or ease of spread) of living agents. The assumption is that there is spread of an agent in the aggregate (*eg* herd, family) and that not all cases are a result of a common-source exposure. When the latent period is long, it is often difficult to distinguish between animal-to-animal spread and that due to common exposure (*eg* BSE in cattle). Secondary attack rates are the number of cases minus the initial case(s) divided by the population at risk.

4.9.3 Case fatality rates

The case fatality rate describes the proportion of animals with a specific disease that die from it (within a specified time period). It is actually a 'risk' measure (*ie* a proportion) instead of a 'rate' and is often used to describe the impact of epidemic-type diseases or the severity of acute diseases for affected individuals.

4.9.4 Proportional morbidity/mortality rates

These rates are used when the appropriate denominator is unknown and they are calculated by dividing the number of cases (or deaths) due to a specific disease by the number of cases (or deaths) from all diseases diagnosed. Proportional morbidity/mortality rates are often used for diagnostic laboratory data and are subject to variation in the numerator or the denominator. Hence, they are less preferable than measures of risk.

4.10 CONFIDENCE INTERVALS

Either approximate or exact confidence intervals (CIs) can be computed for proportions (risk and prevalence) and rates.

Approximate CIs are computed by determining the mean (μ) and its standard error (SE) of the parameter of interest. The lower and upper bounds of the CI are then:

$$\mu - Z_\alpha * \text{SE} \quad , \quad \mu + Z_\alpha * \text{SE} \qquad \textit{Eq 4.8}$$

where Z_α is the (1-α/2) percentile of the standard normal distribution.

In small samples, or in situations where the frequency of disease is very low (or very high), the approximate CIs might be misleading (and lower bounds might be negative). In these cases, exact CIs based on probabilities derived from the binomial distribution (for proportions) or the Poisson distribution (for rates) will be more appropriate.

Example 4.3 shows the calculation of approximate and exact CIs for a prevalence proportion and exact CIs for some estimated incidence rates.

4.11 Standardisation of risks and rates

4.11.1 Accounting for differences in populations

Often our intent is to describe the occurrence of disease in a manner that allows valid inferences to be made about factors which affect the frequency of specific diseases. Frequently, host factors are confounders and bias the comparison of risks (rates) whether they be from different geographical areas or have a different exposure history. This confounding can be prevented by standardising the risks or rates. See Chapter 13 for a more complete discussion of confounding.

'Technical' aspects

A population might be divided into strata (denoted by the subscript *j*), based on one or more host characteristics (*eg* age, sex, geographical location). The overall frequency of disease in the population is a function of the host factor distribution (denoted here as H_j) and the rates (I_j) or risks of disease (R_j) in each of the strata. The H_j for risks is N_j / N (the proportion of the study group or population in that stratum) and for rates the H_j is T_j / T (the proportion of animal-time in that stratum). Specifically, the crude risk (R) in a population is:

$$R = \sum H_j R_j \qquad \textit{Eq 4.9}$$

where $H_j = N_j / N$

And the crude rate (I) is:

$$I = \sum H_j I_j \qquad \textit{Eq 4.10}$$

where $H_j = T_j / T$.

Note For simplicity, for the rest of this discussion, we will primarily refer to rates, but the methods are equally applicable to risks.

Example 4.3 Confidence intervals for proportion and rate
data=dairy_dis (herd 1)

Prevalence data for several infectious diseases were obtained from a sample of dairy herds. See Example 2.1 or Chapter 27 (dairy_dis) for a more complete description of these data.

Approximate and exact CIs for the prevalence proportion of leukosis and Johne's disease in herd 1 (27 cows) in this dataset were computed.

Disease type		Number of positive samples	P	95% CI	
Leukosis	approximate	22	0.815	0.658	0.971
	exact			0.619	0.937
Johne's	approximate	3	0.111	-0.016	0.238
	exact			0.024	0.292

This shows that approximate CIs might go beyond the theoretically possible boundaries of 0 and 1.

Note The approximate CIs shown were computed using a *t*-distribution, not the *Z*-distribution shown in Eq 4.8, because of the small sample size.

Incidence rates were computed by assuming that:

- the age of each cow (in years) was her current lactation number plus 2.
- all infections arose immediately before the cow was tested (*ie* her period of risk was equal to her age). (This is a very untenable assumption for these two diseases and has been done only for the sake of this example.)

Exact CIs for the incidence of these two disease rates were then determined based on the Poisson distribution.

Disease	Number of positive samples	Cow-years at risk	I	95% CI	
Leukosis	22	158	0.139	0.087	0.211
Johne's	3	158	0.019	0.004	0.056

Differences in disease rates (I) between populations of animals might be due to different distributions of host characteristics (H_j) or to actual differences in the stratum-specific rates (I_j). We can remove the effect of differences in host characteristics by 'standardising' the risks or rates. We can carry out this standardisation by using a set of standard rates (I_j) from a referent population (called **indirect standardisation**) or by using a set of H_j from a standard population (called **direct standardisation**).

4.11.2 Indirect standardisation of rates

One method to control the potential confounding effect of host characteristics when comparing rates from different populations is to compute standardised morbidity/mortality ratios (*SMR*). These are based on a set of stratum-specific rates from a reference, or standard, population (Is_j) together with the observed proportion of animal-time in each of the strata in the study group. The process is called indirect standardisation. It is very useful if the actual stratum-specific rates are not available for the study population or if the estimates of those rates are based on small sample sizes.

The standard rates from the reference population will allow us to calculate the adjusted, or expected rate (I_e) as:

$$I_e = \sum H_j Is_j \qquad \textit{Eq 4.11}$$

The expected number of cases in the study population (denoted as if the reference population rates apply) is:

$$E = T * I_e \qquad \textit{Eq 4.12}$$

where T is the total time at risk.

If A is the observed number of cases in the area, the ratio A/E is the standardised morbidity rate ratio (similarly I/I_e=*SMR*). To obtain the indirect standardised rate (I_{ind}), we use the overall rate in the standard population (Is) multiplied by the *SMR*.

$$I_{ind} = Is * SMR \qquad \textit{Eq 4.13}$$

The standard error (SE) of the log of the standardised rate ratio [ln*SMR*] is:

$$\mathrm{SE}[\ln SMR] = 1/\sqrt{A} \qquad \textit{Eq 4.15}$$

and the confidence limits for the *SMR* can be calculated using:

$$e^{[\ln SMR] \pm Z_\alpha * \mathrm{SE}} \qquad \textit{Eq 4.14}$$

Example 4.4 demonstrates the indirect standardisation of rates.

4.11.3 Indirect standardisation of risks

We can use the same strategy for rates as described above for risks. The only difference is that H_j is based on the proportion of animals in each stratum instead of the proportion of animal-time. The expected number of cases, if the reference population risks apply to the study group's distribution of animals, is $E=N*Rs$ where Rs is the overall risk in the standard population. The ratio of observed to expected cases, A/E, is the standardised morbidity risk ratio. Again, the indirect standardised risk for the area is $Rs*SMR$. The variability of an *SMR* based on risks is somewhat more complex than one based on rates and, because most standardisation is done on rates, the formulae for variance will not be given here.

Example 4.4 Indirect standardisation of rates

Assume that you have data on the herd rate of tuberculosis (*ie* incidence rate of herds found to be positive) from two geographical regions which you would like to compare. However, the proportion of dairy and beef herds differ in the two regions and you know that this factor influences the rate of herd infections. You obtain a set of standard incidence rates based on data from the whole country and they are:

- the rate in beef herds is 0.025 cases/herd-year,
- the rate in dairy herds is 0.085 cases/herd-year, and
- the overall rate is 0.06 cases/herd-year.

In Region A, you have data from 1,000 herds over one year and in Region B, data on 2,000 herds for one year. The data are:

Type	Number of cases	Number of herd-years (T)	Observed rate (I_j)	Herd-years distribution (T_j)	Standard rate (Is_j)
Region A					
Beef	17	550	0.031	0.55	0.025
Dairy	41	450	0.091	0.45	0.085
Total	58	1000			
Overall rate*			0.058		0.052
SMR = 0.058/0.052 = 1.12					
Indirect standardised rate (I_{ind}) = 0.06 * 1.12 = 0.067					
Region B					
Beef	10	500	0.020	0.25	0.025
Dairy	120	1500	0.080	0.75	0.085
Total	130	2000			
Overall rate*			0.065		0.07
SMR = 0.065/0.07 = 0.93					
Indirect standardised rate = 0.06 * 0.93 = 0.056					

* Overall rate is the sum of the stratum-specific rates times the T_j distribution (eg overall observed rate in Region A=(0.031*0.55)+(0.091*0.45)=0.058 (except for slight rounding errors).

Although the stratum-specific rates in Region A are higher than in Region B, the crude overall rate would suggest (incorrectly) a lower rate in Region A (0.058 vs 0.065) whereas the standardised rates show (correctly) a higher rate in Region A (0.067 vs 0.056).

4.11.4 Direct standardisation of rates

A second way of addressing the problem is through direct standardisation. Here we use a standard distribution of the population time-at-risk in each level (stratum) of the confounder (or combination of confounders) for the factor(s) of interest (*ie* the Ts_j). The direct standardised rate (I_{dir}) is:

$$I_{dir} = \sum Ts_j I_j \qquad \textit{Eq 4.16}$$

where Ts_j is the proportion of the total subject time-at-risk allotted to the j^{th} stratum of subjects.

A major drawback to the direct method is that there is no adjustment for the variance of the stratum-specific rates, they all have equal weight even if they are based on a very few animals. Example 4.5 presents the calculation of direct standardised rates.

Example 4.5 Direct standardisation of rates

Using the same data presented in Example 4.4, and a suitable reference population which had a cattle type time-at-risk distribution (Ts_j) of:

- beef 40%
- dairy 60%.

Direct standardised rates can be computed as:

Cattle type	Observed rate (I_j)	Reference population distribution (Ts_j)	Product (I_j* Ts_j)
Region A			
Beef	0.031	0.4	0.012
Dairy	0.091	0.6	0.055
Direct standardised rate (I_{dir})			0.067
Region B			
Beef	0.02	0.4	0.008
Dairy	0.08	0.6	0.048
Direct standardised rate			0.056

Standardisation has once again revealed that the rate of tuberculosis is actually higher in Region A.

To express the variability of the direct standardised rate, the SE is:

$$\mathrm{SE}(I_{\mathrm{dir}}) = \sqrt{\sum \left(Ts_j^2 * I_j * S_j / N_j\right)} \qquad \textit{Eq 4.17}$$

where S_j=1-I_j

The confidence interval can be calculated using:

$$I_{\mathrm{dir}} \pm Z_\alpha * \mathrm{SE}(I_{\mathrm{dir}}) \qquad \textit{Eq 4.18}$$

The direct standardisation of risks proceeds in an analogous manner to that of rates. The actual proportion of animals (Hs_j) in each category in the reference population is used instead of the proportion of animal-time (Ts_j) in each category.

4.12 Application

There are a number of areas where rate standardisation is really useful. It allows us to compare a set of rates without being concerned about whether or not they are confounded – provided we can measure the confounders. Rate standardisation works best when the confounders are categorical in nature.

One example stems from work in Ireland on tuberculosis. There, one measure of progress of the control programme is to monitor the annual risk (actually, prevalence) of lesions in supposedly tuberculosis-free cattle at slaughter. A number of factors affects the lesion risk. Two of the more important factors are slaughter plant (not all plants do an equally good job at finding lesions) and class of animal slaughtered (cows tend to have higher lesion prevalence than heifers, steers or bulls). Season also has an effect. One might think that, on an annual basis, season would cancel out but, if the slaughter distribution shifted seasonally, this would impact the lesion risk. Thus, with approximately 18 major slaughter plants, four classes of animal and four seasons, we would have 288 strata for each year. For each stratum, one needs the number slaughtered and the number of tuberculous lesions found (from which the stratum-specific risks can be computed). Then the number of cattle in that stratum is expressed as a proportion of the total slaughtered (*eg* using national data from a 10-year period as the standard population). We then have the H_j and an R_j for each stratum which are combined to compute a direct standardised annual risk. In this manner, the annual lesion risks could be compared without concern about the effects of season, animal class, or slaughter plant biasing them.

SELECTED REFERENCES/SUGGESTED READING

1. Bendixen PH. Notes about Incidence Calculations in Observational Studies. Prev Vet Med 1987; 5: 151-156.
2. Vandenbroucke JP. On the rediscovery of a distinction. Am J Epidemiol 1985; 121: 627-628.
3. Rothman KJ, Greenland S. Modern Epidemiology, Chapter 3. 2d ed. Philadelphia: Lippincott-Raven, 1998.
4. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research, Chapters 6 and 7. Principles and Quantitative Methods. London: Lifetime Learning Publications, 1982.

SAMPLE PROBLEMS

1. You are interested in determining the frequency of feline leukemia virus (FELV) infection and feline leukemia-related diseases (FLRD) in a cattery. The cattery has the following number of cats on the 15th of each month during a year:

Month	# of cats	Month	# of cats	Month	# of cats
January	227	May	165	September	195
February	203	June	134	October	218
March	198	July	153	November	239
April	183	August	179	December	254

The total number of cat-months for the year would therefore be 2,348 with an average monthly population of 195.7.

The following are relevant pieces of information about the cattery and its disease situation.

- On January 15, you bleed all of the cats and find that 63 are positive (antigen test) for feline leukemia virus.
- During the year, 16 cats develop one of the feline leukemia-related diseases (FLRD) and 12 of these cats die.
- Cases of FLRD last an average of three months before the cat dies or recovers – cats that recover are at risk of developing another case of FLRD.
- an additional 13 cats die of other causes.

Compute the following parameters:

a. *I* of FLRD
b. *P* of FELV infection on January 15
c. Estimated *P* of clinical cases of FLRD (at any time during the year)
d. The overall mortality rate
e. The FLRD specific mortality rate
f. The FLRD case fatality rate
g. The estimated risk of an individual cat developing FLRD.

2. A pig farmer has 125 sows and on March 10 *Actinobacillus pleuropneumonia* is first diagnosed in his barn. Between then and July 12, a total of 68 pigs develop clinical signs with 24 of them being treated twice. The condition responds well to antibiotic therapy and only four pigs die, but the pigs are so unproductive after the outbreak that the owner goes out of business and becomes a real estate salesman. What was *I* and *R* of clinical disease during this outbreak?
3. A recently published survey of sheep diseases in Canada reported on losses determined from a survey of producers as well as on findings reported from diagnostic laboratories across the country. For diarrheal diseases, the laboratories reported the following etiologies.

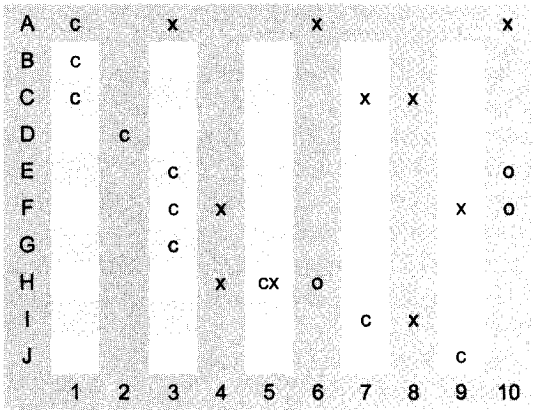
E. Coli	294
Salmonella	33
Cryptosporidia	10
Enterotoxemia	51

What is the proportional mortality due to *Salmonella*? Is this a good indication of the importance of *Salmonella* as a cause of diarrhea in sheep? Why?

4. Assume that you want to measure the frequency of clinical mastitis in a dairy herd. You have the resources to record data for a 10-week period in this herd and because clinical mastitis is more common in early lactation than later, you decide to follow only those cows which calve during that period. The data you collect are shown below. Compute both *R* and *I* of clinical mastitis.

You can assume that:

- all events occur at the beginning of the week in which they are registered.
- cows are not considered at risk for the week in which the case of mastitis occurs and for one week afterward.
- for computing *I*, multiple cases of mastitis are considered.



c = calving

x = case of disease (mastitis)

o = cow culled or died (not mastitis)

5

SCREENING AND DIAGNOSTIC TESTS

OBJECTIVES

After reading this chapter, you should be able to:

1. Define accuracy and precision as they relate to test characteristics.

2. Interpret three measures of precision for quantitative test results; calculate and interpret kappa for categorical test results.

3. Define epidemiologic sensitivity and specificity, calculate their estimates and their standard errors (or confidence intervals) based on all members or subsets of a defined study population.

4. Define predictive values and explain the factors that influence them.

5. Know how to use and interpret multiple tests in series or parallel.

6. Define and explain the impact of correlated test results (*ie* tests that are not conditionally independent).

7. Know how to choose appropriate cutpoints for declaring a test result positive (this includes receiver operating characteristics curves and likelihood ratios).

8. Be able to use logistic regression to control the effects of extraneous variables and produce stratum-specific estimates of sensitivity and specificity.

9. Estimate sensitivity and specificity when no gold standard exists.

10. Describe the main features influencing herd-level sensitivity and specificity based on testing individual animals.

11. Describe the main features influencing herd-level sensitivity and specificity based on using pooled specimens.

5.1 INTRODUCTION

Most of us think of tests as specific laboratory test procedures (*eg* a liver enzyme, serum creatinine, or blood urea nitrogen test). A test, more generally, is any device or process designed to detect, or quantify a sign, substance, tissue change, or body response in an animal. Tests can also be applied at the herd, or other level of aggregation. Thus, for our purposes, in addition to the above examples of tests, we can consider clinical signs (*eg* looking for a jugular pulse), questions posed in the history-taking of a case work-up (*eg* how long since previous calving), findings on the general inspection or routine examination of an animal or premises (*eg* a farm inspection for state of hygiene), or findings at post-mortem examination of carcasses as tests. Indeed, tests are used in virtually all problem-solving activities and therefore, the understanding of the principles of test evaluation and interpretation are basic to many of our activities. Several discussions of the application and interpretation of tests are available (Greiner and Gardner 2000a,b; Linnet, 1988; Martin, 1984; Tyler, 1989; Seiler, 1979).

If tests are being considered for use in a decision-making context (clinic or field disease detection), the selection of an appropriate test should be based on the test result altering your assessment of the probability that a disease does or does not exist and that guides what you will do next (further tests, surgery, treat with a specific antimicrobial, quarantine the herd *etc*) (Connell and Koepsell, 1985). In the research context, understanding the characteristics of tests is essential to knowing how they effect the quality of data gathered for research purposes. The evaluation of tests might be the stated goal of a research project or, this assessment might be an important precursor to a larger research programme.

5.1.1 Screening vs diagnostic tests

A test can be applied at various stages in the disease process. Generally, in clinical medicine, we assume that the earlier the intervention, the better the recovery or prognosis. Tests can be used as **screening tests** in healthy animals (*ie* to detect seroprevalence of diseases, disease agents or subclinical disease that might be impairing production). Usually the animals or herds that test positive will be given a further in-depth diagnostic work-up, but in other cases, such as in national disease-control programmes, the initial test result is taken as the state of nature. For screening to be effective, early detection of disease must offer benefits to the individual, or 'programme', relative to letting the disease run its course and being detected when it becomes clinical. **Diagnostic tests** are used to confirm or classify disease, guide treatment or aid in the prognosis of clinical disease. In this setting, all animals are 'abnormal' and the challenge is to identify the specific disease the animal in question has. Despite their different uses, the principles of evaluation and interpretation are the same for both screening and diagnostic tests.

5.2 LABORATORY-BASED CONCEPTS

Throughout most of this chapter, the focus will be on determining how well tests are able to correctly determine whether individuals (or groups of individuals) are diseased

or not. However, before starting the discussion of the relationship between test results and disease status, we should address some issues related to the ability of a test to accurately reflect the amount of the substance (*eg* liver enzyme or serum antibody level) being measured and how consistent the results of the test are if the test is repeated. These concepts include analytic sensitivity and specificity, accuracy and precision.

5.2.1 Analytic sensitivity and specificity

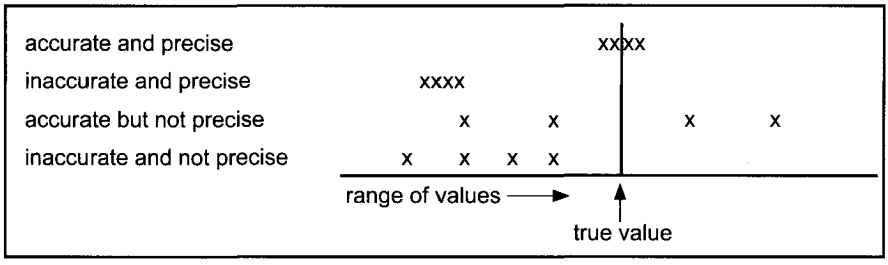
The analytic sensitivity of an assay for detecting a certain chemical compound refers to the lowest concentration the test can detect. In a laboratory setting, specificity refers to the capacity of a test to react to only one chemical compound (*eg* a commonly used test in the dairy industry to identify the presence of antibiotic (β-lactam) inhibitors in milk). The analytic sensitivity of the test is 3 ppb for penicillin, meaning that the test can detect levels of penicillin in milk as low as 3 ppb. The test reacts primarily to β-lactam antibiotics but will also react with other families at higher concentrations, such as tetracyclines. Thus, the test is not specific to just β-lactam antibiotics. Epidemiologic sensitivity and specificity (discussed starting in section 5.3) depend (in part) on laboratory sensitivity and specificity, but are distinctly different concepts (Saah and Hoover, 1997). The epidemiologic sensitivity answers: Of all milk samples that actually have penicillin residues, what proportion tests positive? The epidemiologic specificity answers this question: Of all the milk samples that don't have penicillin residues, what proportion gives a negative result?

5.2.2 Accuracy and precision

The laboratory accuracy of a test relates to its ability to give a true measure of the substance being measured (*eg* blood glucose, serum antibody level). To be accurate, a test need not always be close to the true value, but if repeat tests are run, the average of the results should be close to the true value. On average, an accurate test will not overestimate or underestimate the true value.

The precision of a test relates to how consistent the results from the test are. If a test always gives the same value for a sample (regardless of whether or not it is the correct value), it is said to be precise. Fig. 5.1 shows the various combinations of accuracy and precision.

Fig. 5.1 Laboratory accuracy and precision



Results from tests that are inaccurate can only be 'corrected' if a measure of the inaccuracy is available and used to adjust the test results. Imprecision can be dealt with by performing repeated tests and averaging the results. Correct calibration of equipment and adherence to standard operating procedures are essential to good accuracy and precision; however, the details are beyond the scope of this book.

5.2.3 Measuring accuracy

Assessing accuracy involves running the test on samples with a known quantity of the substance present. These can be field samples for which the quantity of the substance has been determined by a generally accepted reference procedure. For example, the accuracy of an infrared method for determining milk urea nitrogen (MUN) level in milk samples was recently determined by comparing those results with those obtained from a 'wet-chemistry' analysis (Arunvipas et al, 2002). Alternatively, the accuracy of a test can be determined by testing samples to which a known quantity of a substance has been added. The possibility of background levels in the original sample and the representativeness of these 'spiked' samples make this approach less desirable for evaluating tests designed for routine field use.

Variability among test results (*ie* an estimate of precision) might be due to variability among results obtained from running the same sample within the same laboratory (**repeatability**) or variability between laboratories (**reproducibility**). Regardless of which is being measured, evaluating precision involves testing the same sample multiple times within and/or among laboratories. Methods for quantifying the variability in test results are discussed in the following two sections. A much more detailed description of procedures for evaluating laboratory-based tests can be found in Jacobson (1998).

5.2.4 Measuring precision of tests with quantitative outcomes

Some commonly used techniques for quantifying variability, or for expressing results of comparisons between pairs of test results are:

- coefficient of variation
- Pearson correlation coefficient
- concordance correlation coefficient (*CCC*)
- limits of agreement plots.

The **coefficient of variation** (CV) is computed as:

$$\text{CV} = \frac{\sigma}{\mu} \qquad \textit{Eq 5.1}$$

where σ is the standard deviation among test results on the same sample and μ is the average of the test results. It expresses the variability as a percentage of the mean. The CV for a given sample can be computed based on any number of repeat runs of the same test and then these values can be averaged to compute an overall estimate of the CV (see Example 5.1).

Example 5.1 Measuring agreement - quantitative test results
data=elisa_repeat

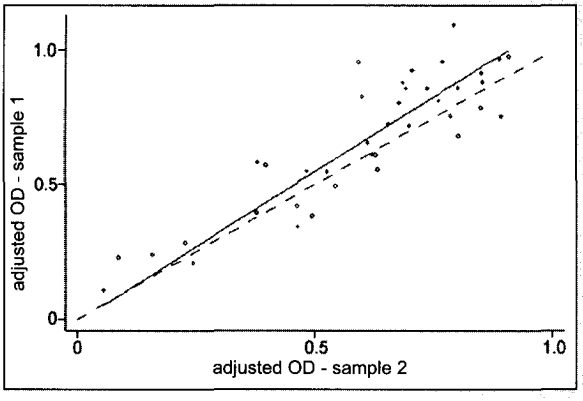
A set of 40 individual cow milk samples was tested for parasite antibodies six times using an indirect microtitre ELISA based on a crude *Ostertagia ostertagi* antigen. Both raw and adjusted optical density (OD) values are recorded in the dataset (see Chapter 27 for description of adjustment method). The results were used to evaluate the precision and repeatability of the test.

The CV for each sample was computed based on the six replicate values and then averaged across the 40 samples. The mean CV was 0.155 for the raw values and 0.126 for the adjusted values suggesting that the adjustment process removed some of the plate-to-plate variability.

Pearson correlation was used to compare values from replicates 1 and 2. The correlation was 0.937 for the raw values and 0.890 for the adjusted values.

Comparing replicates 1 and 2, the *CCC* was 0.762 for the raw values and 0.858 for the adjusted values, suggesting much better agreement between the two sets of adjusted values (than between the two sets of raw values). Fig. 5.2 shows a *CCC* plot.

Fig. 5.2 Concordance correlation plot



Note Data must overlay dashed line for perfect concordance.

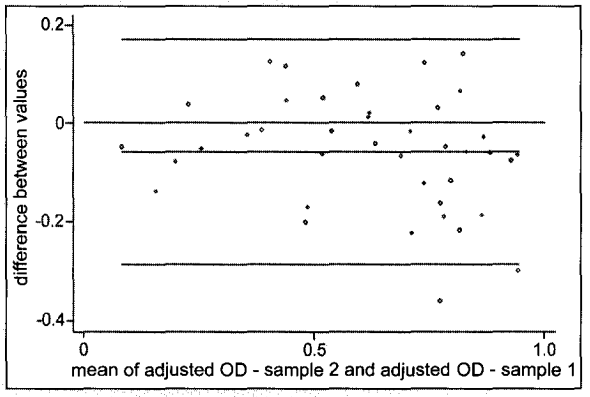
There appears to be a greater level of disagreement between the two sets of values at high OD readings compared with low OD readings.

(*continued on next page*)

Example 5.1 (*continued*)
data=elisa_repeat

The limits of agreement plot for the same data is shown in Fig. 5.3. It indicates that most of the differences between the replicates fell in the range of +0.18 and -0.30 units.

Fig. 5.3 Limits of agreement



A Pearson **correlation coefficient** measures the degree to which one set of test results (measured on a continuous scale) varies (linearly) with a second set. However, it does not directly compare the values obtained (it ignores the scales of the two sets of results) and for this reason, it is much less useful than a concordance correlation coefficient for comparing two sets of test results (see Example 5.1). Both of these statistics are based on the assumption that the data are normally distributed.

As with a Pearson correlation coefficient, a **concordance correlation coefficient** (Lin, 1989) can be used to compare two sets of test results (*eg* results from two laboratories), and it better reflects the level of agreement between the two sets of results than the Pearson correlation coefficient does. If two sets of continuous-scale test results agreed perfectly, a plot of one set against the other would produce a straight line at a 45° angle (the equality line). The *CCC* is computed from three parameters. The location-shift parameter measures how far the data are (above or below) from the equality line. The scale-shift parameter measures the difference between the slope for the sample data and the equality line (slope=1). (The product of the location-shift and scale-shift parameters is referred to as the **accuracy parameter**.) The usual Pearson correlation coefficient measures how tightly clustered the sample data are around the line (slope). The *CCC* is the product of the accuracy parameter and the Pearson correlation coefficient. A value of 1 for the *CCC* indicates perfect agreement. Example 5.1 shows a concordance correlation plot for two sets of ELISA results. The *CCC* has recently been generalised to deal with >2 sets of test results and to work with categorical data (King and Chinchilli, 2001).

Limits of agreement plots (also called Bland-Altman plots) (Bland and Altman, 1986) show the difference between the pairs of test results relative to their mean value. Lines that denote the upper and lower difference values that enclose 95% of the points are added to the plot. They indicate the range of differences between the two sets of test results. This method is also useful to determine if the level of disagreement between the two sets of results varies with the mean value of the substance being measured and can also be used to identify the presence of outlying observations. A limits of agreement plot is presented in Fig. 5.3.

5.2.5 Measuring precision and agreement of tests with a qualitative outcome

All of the above procedures are useful if the quantity of interest is measured on a continuous scale. If the test results are categorical (dichotomous or multiple categories), a kappa (or weighted kappa) statistic can be used to measure the level of agreement between two (or more) sets of test results. Obviously, the assessments must be carried out independently of each other using the same set of outcome categories. The data layout for assessing agreement is shown in Table 5.1 for a 2X2 table (larger 'square' tables are also used).

Table 5.1 Layout for comparing results from two qualitative (dichotomous) tests

	Test 2 positive	Test 2 negative	Total
Test 1 positive	n_{11}	n_{12}	$n_{1.}$
Test 1 negative	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	

In assessing how well the two tests agree, we are not seeking answers relative to a gold standard (section 5.3.1) as this might not exist, but rather whether the results of two tests agree with each other. Obviously, there will always be some agreement due to chance, and this must be considered in the analysis. For example, if one test was positive in 30% of subjects and the other test was positive in 40%, both would be expected to be positive in 0.4*0.3=0.12 or 12% of subjects by chance alone. So, the important question is: what is the level, or extent, of agreement beyond what would have been expected by chance? This question is answered by a statistic called Cohen's kappa. We can calculate the essential elements of kappa very easily. They are:

observed agreement = $(n_{11} + n_{22})/n$

expected agreement (chance) = $[(n_{1.} * n_{.1})/n + (n_{2.} * n_{.2})/n]/n$

actual agreement beyond chance = observed - expected

potential agreement beyond chance = (1 - expected)

kappa = actual agreement beyond chance/potential agreement beyond chance.

A formula for calculating kappa directly is:

$$\text{kappa} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{1.}n_{2.} + n_{.2}n_{.1}} \qquad \textit{Eq 5.2}$$

Formulae for the standard error and test of significance are available elsewhere (Bloch and Kraemer, 1989; Kraemer and Bloch, 1994). Before assessing kappa, we should assess whether there is test bias. This would be indicated by the proportion positive to each test differing (*ie* $p_1 \neq p_2$, where p_1 and p_2 represent the proportion positive to tests 1 and 2, respectively). Because the data are paired this can be assessed by McNemar's test or an exact binomial test for correlated proportions.

$$\text{McNemar's } \chi^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21}) \qquad \textit{Eq 5.3}$$

A non-significant test would indicate that the two proportions positive do not differ. If significant this test suggests a serious disagreement between the tests and thus the detailed assessment of agreement could be of little value.

The magnitude of kappa is influenced by the extent of the agreement as well as by the prevalence (*P*) of the condition being tested for. When the latter is very high or very low (outside of the range 0.2 to 0.8), the kappa statistic becomes unstable (*ie* difficult to rely on and/or interpret). Common interpretations of kappa, when applied to a test that is subjective in nature (*eg* identifying lesions on an X-ray), are as follows:

<0.2	slight agreement
0.2 to 0.4	fair agreement
0.4 to 0.6	moderate agreement
0.6 to 0.8	substantial agreement
>0.8	almost perfect agreement.

Example 5.2 shows the computation of kappa for assessing agreement between indirect fluorescent antibody test (IFAT) results for infectious salmon anemia (ISA) when the test was performed in two different laboratories.

For tests measured on an ordinal scale, computation of the usual kappa assumes that any pair of test results which are not in perfect agreement are considered to be in disagreement. However, if a test result is scored on a five-point scale, a pair of tests with scores of 5 and 4 respectively, should be considered in 'less disagreement' than a pair of scores of 5 and 1. Partial agreement can be taken into account using a weighted kappa in which pairs of test results that are close are considered to be in partial agreement (through a weight matrix which specifies how much agreement should be assigned to them). Example 5.3 shows the data layout and the results of an unweighted and weighted kappa for comparing two sets of IFAT results for the ISA virus in salmon.

Example 5.2 Agreement among dichotomous test results
data=ISA_test

Kidney samples from 291 salmon were split with one-half being sent to each of two laboratories where an IFAT test was run on the sample. IFAT results were expressed as 0 (negative) or 1+, 2+, 3+, or 4+. They were subsequently dichotomised so that all scores of 1+ and higher were considered positive. The data were:

	IFAT 2 positive	IFAT 2 negative	Total
IFAT 1 positive	19	10	29
IFAT 1 negative	6	256	262
Total	25	266	291

The McNemar's χ^2 test had the value 1.00 (P=0.317; the binomial P-value was 0.45) indicating that there is little evidence that the two laboratories found different proportions positive.

observed agreement = 0.945
kappa = 0.674
95% CI of kappa = 0.132 to 0.793
expected agreement = 0.832
SE(kappa)[a] = 0.0584
P < 0.001

Thus the level of agreement appears substantial and is statistically significantly better than that expected due to chance. However, the CI is wide, reflecting considerable uncertainty about the estimate.

[a] There are a number of formulae for the SE; the one used here is attributed to Fleiss (1981).

5.3 The ability of a test to detect disease or health

The two key characteristics we estimate are the ability of a test to detect diseased animals correctly (its sensitivity), and at the same time to give the correct answer if the animal in question is not diseased (its specificity). For pedagogical purposes, we will assume that animals are the units of interest (the principles apply to other levels of aggregation). Further, we will assume that a specific 'disease' is the outcome although other conditions such as pregnancy, premature herd removal (culling), having a specified antibody titre, or infection status could be substituted in a particular instance. To initiate this discussion, it is simplest to assume that the test we are evaluating gives only dichotomous answers, positive or negative. This might be a bacterial culture in which the organism is either present or absent, or a question about whether or not a dairy farmer uses a milking machine with automatic take-offs. In reality, many test results provide a continuum of responses and a certain level of response (colour, test result relative to background signal, level of enzyme activity, endpoint titre *etc*) is selected such that, at or beyond that level, the test result is deemed to be positive.

5.3.1 The gold standard

A gold standard is a test or procedure that is absolutely accurate. It diagnoses all of the

Example 5.3 Agreement among ordinal test results
data=ISA_test

The data described in Example 5.2 were used except the original ordinal data were retained (5-point scale).

	IFAT 2				
IFAT 1	Neg	+	++	+++	++++
Neg	256	5	0	1	0
+	8	2	0	2	0
++	2	1	0	4	0
+++	0	0	2	2	0
++++	0	0	0	3	3

Simple (unweighted) kappa=0.45 (assumes that all test results which were not identical as being in disagreement).

A weighted kappa was computed (Fleiss, 1981) in which test results were:
- identical: weighted as complete agreement
- 1 level apart: weighted as 70% agreement
- 2 levels apart: weighted as 30% agreement
- >2 levels apart: weighted as complete disagreement.

(One should, of course, explain the motivation behind the weights used.)

Weighted kappa=0.693, SE(kappa)=0.046.

The weighted kappa still suggests only moderate agreement but is a better reflection of the agreement between the two sets of tests than the unweighted test is.

specific disease that exists and misdiagnoses none. For example, if we had a definitive test for feline leukemia virus infection that correctly identified all feline leukemia virus-infected cats to be positive and gave negative results in all non-infected cats, the test would be considered a gold standard. In reality, there are very few true gold standards. Partly this is related to imperfections in the test itself, but a good portion of the error is due to biological variability. Animals do not immediately become 'diseased', even sub-clinically, when exposed to an infectious, toxic, physical or metabolic agent. Usually, a period of time will pass before the animal responds in a manner that produces a detectable or meaningful change. The time period for an animal's response to cross the threshold and be considered 'diseased' varies from animal to animal.

Traditionally, in order to assess a new test we need the gold standard. Often, however, because of practical difficulties, we must use the accepted diagnostic method which might be closer to a 'bronze', or in the worst case a 'tin' standard. This can produce considerable difficulties in test evaluation, and thus, in recent years, we have begun to use statistical approaches (*eg* maximum likelihood estimation methods) to help estimate the two key test characteristics in the absence of a gold standard (see section 5.9).

5.3.2 Sensitivity and specificity

The concepts of sensitivity and specificity are often easier to understand through the use of a 2X2 table, displaying disease and test results in a sample of animals.

Table 5.2 Data layout for test evaluation

	Test positive (T+)	Test negative (T-)	Total
Disease positive (D+)	a (true positive)	b (false negative)	m_1
Disease negative (D-)	c (false positive)	d (true negative)	m_0
Total	n_1	n_0	n

The sensitivity of a test (*Se*) is the proportion of diseased animals that test positive. It is described statistically as the conditional probability of testing positive given that the animal is diseased [p(*T*+|*D*+)], and is measured by:

$$Se = \frac{a}{m_1} \qquad \textit{Eq 5.4}$$

The specificity of a test (*Sp*) is the proportion of non-diseased animals that test negative. It is described statistically as the conditional probability of testing negative given that the animal does not have the disease of interest [p(*T*-|*D*-)] and is measured by:

$$Sp = \frac{d}{m_0} \qquad \textit{Eq 5.5}$$

For future purposes, we will denote the false positive fraction (*FPF*) as 1-*Sp* and the false negative fraction (*FNF*) as 1-*Se*. From a practical perspective, if you want to confirm a disease, you would use a test with a high *Sp* because there are few false positives. Conversely, if you want to rule out a disease, you would use a test with a high *Se* because there are few false negatives.

The estimation of *Se* and *Sp* of an indirect ELISA test for detecting bovine fetuses persistently infected (PI) with the bovine virus diarrhea (BVD) virus is shown in Example 5.4. A blood sample is taken from the cow in late lactation and tested for antibodies to the virus. If they are present at a high level, the fetus is deemed to be persistently infected with the BVD virus.

5.3.3 True and apparent prevalence

Two other terms are important descriptors of the tested subgroup. One denotes the actual level of disease that is present. In screening-test jargon, this is called the **true prevalence** (*P*); in clinical epidemiology, this is referred to as **prior prevalence**, or **pre-test prevalence**. *P* is a useful piece of information to include in our discussion of test evaluation because it will affect the interpretation of the test result. In Example 5.4, P=p(D+)=m_1/n=233/1673=0.139 or 13.9%.

In contrast to the 'true' state, unless our test is perfect, the test results will only provide an estimate of the true prevalence and, in screening-test jargon, this is called

the **apparent prevalence** (*AP*). In Example 5.4 $AP=p(T+)=n_1/n=800/1673=0.478$ or 47.8%. In clinical epidemiology, this might be referred to as a **post-test prevalence**.

5.4 Estimating test sensitivity and specificity

5.4.1 Characteristics of the sampled population

Sensitivity and specificity represent average values of the test characteristics and as such, we can expect their levels to vary from one subgroup of the population to another.

Example 5.4 Sensitivity, specificity and predictive values
data=bvd_test

The data used for this example came from a study done to evaluate an ELISA test for the diagnosis of bovine fetuses persistently infected (PI) with BVD virus. See Chapter 27 for a more complete description of this dataset. The test was designed to work on both milk and blood samples, but the data used here relate only to the blood sample results. The mean optical density was 0.92 units. Thus, for this example a fetus was deemed to be test positive if the optical density of the blood test was greater than 0.92 units. (This is not an optimal cutpoint for this test, but is used for illustration.)

		Test +	Test -	
PI+	(D+)	178	55	233
PI-	(D-)	622	818	1440
		800	873	1673

For purposes of description, the 178 animals are called true positives, the 622 are false positives, the 55 are false negatives and the 818 are true negatives. We will assume here that the study subjects were obtained using a simple random sample.

In this example,

- *Se* = 178/233 = 76.4% 95% CI = (70.4% to 81.7%)
- *Sp* = 818/1440 = 56.8% 95% CI = (54.2% to 59.4%)
- *FNF* = 1-0.764 = 23.6%
- *FPF* = 1-0.568 = 43.2%
- *P* = 233/1673 = 13.9%
- *AP* = 800/1673 = 47.8%
- *PV+* = 178/800 = 22.3% 95% CI = (19.4% to 25.3%)
- *PV-* = 818/873 = 93.7% 95% CI = (91.9% to 95.2%)

Note The confidence intervals are exact based on the binomial distribution.

Consequently, when estimating *Se* and *Sp*, it is important that the study population to which the gold standard diagnostic procedure is applied be representative of the target population (*ie* those animals to whom the test will be applied in the future). This representativeness refers to the attributes of the animals being tested including their age, breed, sex *etc* as host and environmental factors might influence the ability of a test to detect disease. In fact, often it is useful to stratify the results based on the more important of these factors in order to obtain more valid stratum-specific estimates. In addition, it is important that the study group contains an appropriate spectrum of disease (*eg* severity, chronicity or stage of development). Certainly, the test characteristics might differ in different stages of the disease process; for example, tests for Johne's disease work much better once the animal is clinically ill as opposed to only being infected with the organism *Mycobacterium avium* subsp *paratuberculosis* (Map).

5.4.2 Study designs for determining sensitivity and specificity

In some settings, the two groups of *D*+ and *D*- animals (or samples from them) are available and the new test can be evaluated on them. In other instances, a sample of animals is obtained and the test and gold standard are applied to all the sampled animals. In still other circumstances, only a sub-sample of test positive and negative animals is submitted to the gold standard test.

If a pre-determined set of *D*+ and *D*- animals is used for *Se* and *Sp* estimation, then these statistics can be treated as binomial parameters for purposes of calculating variances and confidence limits. Common software will usually provide either asymptotic or exact confidence limits. Predictive values (*PV*s) (section 5.5) have no meaning with fixed m_1 and m_0 sample sizes (*ie* when *P* is determined by the investigator).

If a cross-sectional sample of animals is used for the study population, with complete verification of true health status on all study animals, then the same approach can be used. Here, predictive values, true and apparent prevalence are all meaningful and treating each of these as a binomial proportion allows calculation of variance and CIs.

In either case, it is advantageous to have a spectrum of host attributes and clustering units (if any) present (*ie* animals from a number of different farms). The results should be assessed for differences in *Se* or *Sp* by host attributes using logistic regression (see section 5.8). Blind assessment and complete work-ups of all animals are useful aids to prevent bias in the estimates. When *Se* and *Sp* are estimated based on samples obtained from several animals within a number of farms, adjustment of the standard errors (SEs) for the clustering effect should be made.

If a cross-sectional sample of animals is tested, but only a subset of the test positive and negative animals are assessed for their true health status, this feature must be built into the ensuing estimates of *Se* and *Sp*. Predictive values are unbiased. In addition, it is vitally important that selection of animals for verification be independent of their true health status. In this instance, if we denote the fraction (*sf*)of the test positives that are verified as sf_{T+}, and that of the test negatives as sf_{T-}, then the corrected estimate of *Se* is:

$$Se_{corr} = \frac{a/sf_{T+}}{a/sf_{T+} + b/sf_{T-}} \qquad \textit{Eq 5.6}$$

and the corrected estimate of *Sp* is:

$$Sp_{corr} = \frac{d/sf_{T-}}{d/sf_{T-} + c/sf_{T+}} \qquad \textit{Eq 5.7}$$

(See Example 5.5.) If $sf_{T+}=sf_{T-}$ no correction to *Se* or *Sp* is needed.

Example 5.5 Estimating *Se* and *Sp* using a validation subsample
data=none

Suppose that, at slaughter, we examine 10,000 cattle for the presence of lesions consistent with bovine tuberculosis (TB). We find lesions in 242 cattle. A detailed follow-up is done on 100 of the animal specimens with lesions and on similar tissue specimens from 200 of the 'clear' animals. In the animals with lesions, 83 are confirmed as bovine tuberculosis, whereas two of the 200 clear animals are found to have tuberculosis. The data are shown here.

		Lesion+	Lesion-
TB+	(D+)	83	2
TB-	(D-)	17	198
		100	200

and

$$sf_{T+} = 100/242 = 0.413$$
$$sf_{T-} = 200/9758 = 0.0205$$

From these we can calculate Se_{corr} and Sp_{corr}

$$Se_{corr} = \frac{83/0.413}{83/0.413 + 2/0.0205} = \frac{200.9}{298.5} = 0.672$$

with approximate variance of (0.672*0.328)/85=0.003 and

$$Sp_{corr} = \frac{198/0.0205}{0.9716/0.0205 + 17/0.413} = \frac{9658.5}{9941.2} = 0.9716$$

with approximate variance of (0.9716*0.0284)/215=0.00013

The variances of these 'corrected' proportions are calculated using only the number of verified individuals in the variance formulae (*ie* the $a+b$ verified animals for Se_{corr} and the $c+d$ verified animals for Sp_{corr} (Table 5.2).

5.4.3 Precision of sensitivity and specificity estimates

When designing a study to estimate the *Se* and/or *Sp* of a test, we need to consider the number of animals that is required to obtain a specified precision for each estimate. These form the basis for estimating the 95% (or other specified level) CIs as shown in Example 5.4. For *Se*, estimates within ±5% might suffice, whereas for screening low-risk populations, much larger sample sizes are needed as *Sp* estimates need to be within at least ±0.5% of the true value. In a diagnostic setting, *Sp* estimates within 3-5% of the true value should suffice. See Chapter 2 for details on sample size.

5.5 PREDICTIVE VALUES

The *Se* and *Sp* are characteristics of the test. However, these terms do not tell us directly how useful the test might be when applied to animals of unknown disease status. Once we have decided to use a test, we want to know the probability that the animal has or does not have the disease in question, depending on whether it tests positive or negative. These probabilities are called **predictive values** and these change with different populations of animals tested with the same test because they are driven by the true prevalence of disease in the study population as well as by the test characteristics. In this discussion, we assume the group of subjects being tested is homogeneous with respect to the true prevalence of disease. If not, then the covariates that affect disease risk should be identified and separate estimates made for each subpopulation.

5.5.1 Predictive value positive

With data as shown in Table 5.2, the predictive value of a positive test (*PV+*) is the probability that given a positive test, the animal actually has the disease; this might be represented as $p(D+|T+)$ or a/n_1. The predictive value of a positive test can be estimated using the following formula:

$$PV+ = \frac{p(D+)*Se}{p(D+)*Se + p(D-)*(1-Sp)} \qquad \textit{Eq 5.8}$$

This formula explicitly indicates that the true prevalence of disease in the tested group affects the *PV+*.

5.5.2 Predictive value negative

In a similar manner, the *PV* of a negative test (*PV-*) is the probability that given a negative test, the animal does not have the disease (*ie* $p(D-|T-)$). From Table 5.2 this is $PV-=d/n_0$. The predictive value of a negative test result can be estimated using the following formula:

$$PV- = \frac{p(D-)*Sp}{p(D-)*Sp + p(D+)*(1-Se)} \qquad \textit{Eq 5.9}$$

Estimates of *PV+* and *PV-* are shown in Examples 5.4 and 5.6. **Note** These values represent the predictive values given the *P* observed in the study population.

Example 5.6 Effect of prevalence on predictive values
data=bvd_test

In order to examine the impact of a change in *P* on the outcome of a test, we will use the values of *Se* and *Sp* from Example 5.4 and specify three scenarios where the true prevalence varied from 50% to 5% and then to 1%. For pedagogical purposes, we demonstrate the calculations for the 50% prevalence scenario in a 2X2 table. A simple way to proceed to obtain these results is to construct a fictitious population of 1,000 animals with 500 being 'diseased' (*ie* PI+) and 500 being PI- based on the true prevalence of 50%. Then, we calculate 76.4% (*Se*) of 500 and fill in the 382 true positives. Finally, we calculate 56.8% (*Sp*) of 500, fill in the 284 true negatives, and complete the table.

	Test +	Test -	
PI+	382	118	500
PI-	216	284	500
	598	402	1000

From these data:
PV+ = 382/598 = 63.9% The probability that a cow testing positive will truly have a PI+ calf is 63.9%
PV- = 284/402 = 70.6%. The probability that a cow testing negative will truly have a PI- calf is 70.7%

Comparable values if the prevalence is 5% or 1% are:

Prevalence (%)	PV+ (%)	PV- (%)
5	8.5	97.9
1	1.8	99.6

As you can see, the *PV+* drops off rapidly as *P* falls, but the *PV-* rises.

Because we are more often interested in the 'disease' side of the question, there is a measure of the probability that an animal that tests negatively is actually diseased. It is called the positive predictive value of a negative test or $PPV\text{-}=b/n_0$ or 1-(*PV-*).

5.5.3 Increasing the predictive value of a positive test

One way to increase the predictive value of a positive test is to use the test on animals where the *P* in the population being tested is relatively high. Thus, in a screening programme designed to ascertain if a disease is present, we often might slant our testing towards animals that are likely to have the disease in question. Hence, testing culled animals, or animals with a particular history, is a useful way of increasing the pre-test (prior) probability of disease.

A second way to increase *PV+* is to use a more specific test (with the same or higher *Se*), or change the cutpoint of the current test to increase the *Sp* (but this would decrease the *Se* somewhat also). As *Sp* increases, *PV+* increases because *b* approaches zero (fewer false positives). A third, and very common way to increase *PV+* is to use more than one

test. Here the result depends on the method of interpretation as well as the individual test characteristics.

5.6 USING MULTIPLE TESTS

5.6.1 Parallel and series interpretation

Using two tests represents the simplest extension of more than one test although the principles discussed below hold true for multiple tests. Suppose we have two different tests for detecting a disease. In Example 5.7, we use the results from the IFAT test for infectious salmon anemia (*Se*=0.784, *Sp*=0.951) and the polymerase chain reaction (PCR) test for the same disease (*Se*=0.926, *Sp*=0.979). If both tests are carried out, the results can be interpreted in one of two ways. With **series** interpretation, only animals that test positive to both tests are considered test positive. With **parallel** interpretation, animals that test positive to one test, the other test or both tests are considered test positive. Series interpretation increases *Sp* but decreases *Se*; whereas parallel testing increases *Se* and decreases *Sp*.

Example 5.7 Series versus parallel interpretation
data=ISA_test

The data in this example are from the ISA_test dataset. The tests we are comparing are the indirect fluorescent antibody test (IFAT) and the polymerase chain reaction (PCR) test, with clinical disease status (see dataset description Chapter 27) used as the gold standard. The observed joint distributions of test results and virus presence are shown below along with the four possible test interpretation criteria.

	Number of fish by test-result category				Totals
IFAT result	+	+	0	0	
PCR result	+	0	+	0	
Diseased fish	134	4	29	9	176
Non-diseased fish	0	28	12	534	574
Series interpretation	+	0	0	0	
Parallel interpretation	+	+	+	0	

Se of IFAT only = 138/176 = 0.784

Sp of IFAT only = 546/574 = 0.951

Se of PCR only = 163/176 = 0.926

Sp of PCR only = 562/574 = 0.979

Se of series interpretation = 134/176 = 0.761

Se of parallel interpretation = (134+4+29)/176 = 0.949

Sp of series interpretation = (28+12+534)/574 = 1.000

Sp of parallel interpretation = 534/574 = 0.930

Note If tests are going to be interpreted in series, it often makes sense to first test all animals with the test that is less expensive and/or more rapid, and then test all test positives with the second test. This is referred to as **sequential testing** and it provides the same results as simultaneous testing, but at lower cost, because only those subjects/ samples positive to the first test are followed-up with the second test.

5.6.2 Correlated test results

Given the previous discussion on parallel and series interpretation, one might think that virtually 100% *Se* would be obtainable with two-to-three tests used in parallel or 100% *Sp* with three-to-four tests used in series. However, Example 5.7 uses observed values, not ones we might expect assuming conditional independence of tests. The expected distributions of results, if the tests were independent, are shown in Table 5.3.

Table 5.3 Expected Se and Sp levels with combined tests for ISA assuming conditional independence (Example 5.7)

	Sensitivity		Specificity	
Interpretation	Expected	Observed	Expected	Observed
Parallel	0.784+0.926-0.784*0.926 = 0.984	0.949	0.951*0.979=0.931	0.930
Series	0.784*0.926=0.726	0.761	0.951+0.979-0.979*0.951=0.999	1.000

The expected *Se* for parallel interpretation is slightly higher than observed and slightly lower than observed for series interpretation. The expected and observed values for *Sp* are virtually identical. Note that **conditional independence** assumes that, in *D*+ animals, the probability of a positive test result to test 2 is the same in samples that test negative to test 1 as it is in those that test positive to test 1. A similar assumption exists in *D*- individuals. More likely, and as observed with these data, especially if the tests are biologically related (*eg* both antibody tests), if test 1 is negative, the result on test 2 is more likely to be negative than if test 1 was positive. In this instance, we would describe the test results as dependent, or correlated (Gardner et al, 2000).

The extent of the dependence can be calculated as shown below and in Example 5.8.

1. Denote the observed proportion of *D*+ animals with a positive test result to both tests as p_{111} (more generally p_{ijk}; *i* denoting test 1 result, *j* denoting test 2 result, and *k* denoting disease status (1=diseased, 0=non-diseased).
2. In the *D*+ group, and using the sample estimates of *Se* for tests 1 and 2 respectively, (Se_1 and Se_2), the covariance is:
$$\text{covar}(+) = p_{111} - Se_1 * Se_2$$
3. Similarly, in the *D*- group and using the sample estimates of Sp_1 and Sp_2, the covariance is:
$$\text{covar}(-) = p_{000} - Sp_1 * Sp_2$$
The usual circumstance would be that these covariances would be positive, indicating dependence. In a more formal sense, if one calculates an odds ratio (*OR*) on the data from the *D*+ group (*OR*+) and separately on the *D*- group

Example 5.8 Estimating covariance between test results
data=ISA_test

Using the *Se* and *Sp* estimates obtained in Example 5.7, the covariance in the *D*+ and *D*- groups are:

D+ group: $\text{covar}(+) = p_{111} - Se_1 * Se_2 = 0.761 - 0.726 = 0.035$

D- group: $\text{covar}(-) = p_{000} - Sp_1 * Sp_2 = 0.930 - 0.931 = -0.001$

There is a slight positive covariance in the *D*+ group, but it is sufficiently small that the correction will not materially affect the results when the tests are used in combination. There is virtually no covariance in the *D*- group.

(*OR*-), these *OR*s describe the above two covariances respectively, because, if the tests were conditionally independent, the *OR*s would equal 1. Similarly, if the test results are conditionally independent, the kappa statistic in data from *D*+ and *D*- individuals would both equal 0.

4. Given dependence, the *Se* and *Sp* resulting from parallel interpretation of two tests are:

$$Se_p = 1 - p_{001} = 1 - (1 - Se_1) * (1 - Se_2) - \text{covar}(+)$$

$$Sp_p = p_{000} = Sp_1 * Sp_2 + \text{covar}(-)$$

From series interpretation of two tests these are:

$$Se_s = p_{111} = Se_1 * Se_2 + \text{covar}(+)$$

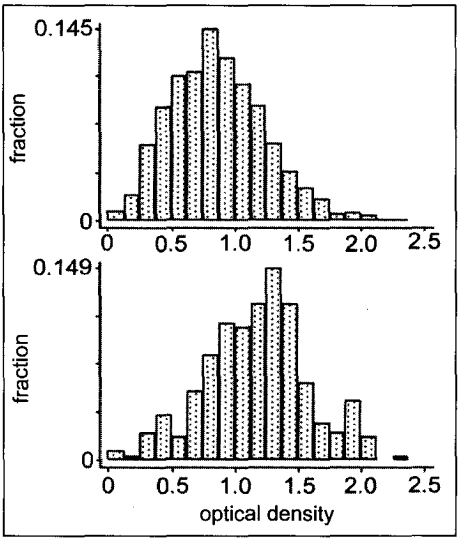
$$Sp_s = 1 - p_{110} = 1 - (1 - Sp_1) * (1 - Sp_2) - \text{covar}(-)$$

Functionally, this means that the gains/losses from using either of these approaches are not as great as predicted under conditional independence. It can also affect the choice of tests to be used. For example, a more optimal outcome might arise from choosing two independent tests with lower sensitivities than two dependent tests with higher sensitivities.

5.6.3 Setting cutpoints for declaring a test result positive

For many tests, the substance being evaluated (*eg* urea in milk, serum calcium, liver enzymes) is measured on a continuous scale or with semi-quantitative (ordinal) results. These items need cutpoints (also called cut-offs or thresholds) to determine what level of result indicates a positive test result. This is also true for many serologic titres. In reality, there is often an overlap in the distribution of the substance being measured between healthy and diseased animals and we usually select a cutpoint that optimises the *Se* and *Sp* of the test. The dilemma is depicted in Fig. 5.4. As will be demonstrated (section 5.6.5), it is often useful to use the actual result when assessing the health status of the tested subject(s).

Fig. 5.4 Overlap between healthy and diseased animals



The distribution of OD in PI+ and PI- calves overlaps considerably. Thus, whatever cutpoint we choose to denote a calf as test positive, there will be both false positive and false negative calves as shown in Example 5.9. PI- calves with test results at or above a given cutpoint are false positives and PI+ calves with test results below the cutpoint are false negatives. If we raise the cutpoint, the *Sp* will increase (false positives decrease) and the *Se* will decrease (more false negatives). Lowering the cutpoint has the opposite effect. Thus, the choice of cutpoint to use will depend on the relative seriousness of either a false negative or a false positive test result.

If one has to choose among multiple cutpoints, graphical procedures (see section 5.6.4) might be used to help choose an optimal cutpoint. Alternatively, it is possible to use the actual test result value by computing likelihood ratios (see section 5.6.5) and avoid having to select a specific cutpoint.

5.6.4 Receiver operating characteristic curves

A receiver operating characteristic (ROC) curve is a plot of the *Se* of a test versus the false positive rate (1-*Sp*) computed at a number of different cutpoints to select the optimum cutpoint for distinguishing between diseased and non-diseased animals. The 45° line in Fig. 5.5 represents a test with discriminating ability that is no better than chance alone. The closer the ROC curve gets to the top-left corner of the graph, the better the ability of the test to discriminate between diseased and non-diseased animals. (The very top-left corner represents a test with a *Se* of 100% and a *Sp* of 100%).

Use of an ROC curve has the advantage over a 'one cutpoint value' for determining *Se* and *Sp* in that it describes the overall ability of the test to discriminate diseased from non-diseased animals over a range of cutpoints. The area under the ROC curve

Example 5.9 Impact of changing the cutpoint on *Se* and *Sp*
data=bvd_test

By varying the optical density (OD) cutpoints for the BVD test results, the following estimates of *Se*, *Sp* and likelihood ratios were obtained (based on all samples, n=2162).

Optical density cutpoint	PI+ category percentage	Sensitivity cumulative percentage[a]	PI- category percentage	Specificity cumulative percentage[b]	Sensitivity + Specificity
≥0.0	6.76	100.00	16.91	0.00	100.00
≥0.5	4.63	93.24	17.44	16.91	110.15
≥0.7	13.17	88.61	21.64	34.34	122.95
≥0.9	16.01	75.44	18.29	55.98	131.42
≥1.1	23.13	59.43	13.29	74.27	133.70
≥1.3	18.51	36.30	6.75	87.56	123.86
≥1.5	7.83	17.79	3.62	94.31	112.10
≥1.7	3.56	9.96	1.01	97.93	107.89
≥1.9	6.05	6.41	0.85	98.94	105.35
≥2.1	3.60	0.36	0.21	99.79	100.15
>2.1	0.00	0.00	0.00	100.00	100.00

[a] from highest to lowest OD category.

[b] from lowest to highest OD category.

Clearly, as the cutpoint for a test to be declared positive is increased, *Se* decreases and *Sp* increases. If the 'costs' of errors (*ie* false negative versus false positive) are equal, then the maximum separation of PI+ and PI- individuals is at a setting of >1.1 where the sum of *Se* and *Sp* is maximum.

(AUC) can be interpreted as the probability that a randomly selected *D*+ individual has a greater test value (*eg* optical density) than a randomly selected *D*- individual (again assuming the distribution of the test statistic in the *D*+ group is higher than that in the *D*- group). If an estimate of the SE of the AUC is available, it is useful for sample-size considerations when designing studies to evaluate tests (see Greiner et al, 2000).

Of course, depending on the seriousness of false negative versus false positive results, one might want to emphasise test results in one particular region of the ROC curve (*eg* an area that constrains *Se* (or *Sp*) within defined limits). Given equal costs to test result errors, the optimal cutpoint is that with *Se*+*Sp* at a maximum, and this occurs where the curve gets closest to the top left corner of the graph (or alternatively, the farthest away from the 45° line).

Both parametric and non-parametric ROC curves can be generated. A non-parametric curve simply plots the *Se* and (1-*Sp*) using each of the observed values of the test result as a cutpoint. A parametric ROC curve provides a smoothed estimate by assuming

that latent variables representing the *Se* and (1-*Sp*) at various cutpoints follow a specified distribution (usually binormal). Example 5.10 shows both parametric and non-parametric ROC curves for the bvd_test data. An alternative to ROC curves for simultaneously evaluating how *Se* and *Sp* vary as the cutpoint is changed is to plot *Se* and *Sp* against various cutpoints (see Fig. 5.6).

5.6.5 Likelihood ratios

A likelihood ratio (*LR*) for a positive test result (*LR*+) is the ratio of the post-test odds of disease divided by the pre-test odds. Recall that, in general, an odds is *P*/(1-*P*) so an *LR* of a positive test result is the odds of disease given the test result divided by the pre-test odds:

$$LR+ = \frac{PV+/(1-PV+)}{P/(1-P)} = \frac{Se}{1-Sp} \qquad \textit{Eq 5.10}$$

where *P*=prevalence or p(*D*+) in the group being tested. Consequently, *LR*s reflect how our view changes of how likely disease is when we get the test result.

The value of the *LR* approach (not to be confused with likelihood ratio tests as used in Chapter 16) is that it can be calculated for each cutpoint when the test result is a continuous, or ordinal, variable. Thus, the LR_{cp} at a selected cutpoint (*ie* **cutpoint-specific *LR*** approach) generalises to:

$$LR_{cp}+ = \frac{Se_{cp}}{1-Sp_{cp}} \qquad \textit{Eq 5.11}$$

where cp denotes the cutpoint at or above which the test is considered positive. In this context, the *LR*+ can be viewed as the probability of a diseased individual having a test result as high as observed compared with the probability of the same result in a non-diseased subject. The *LR* for a negative test result (*LR*-) at a given cutpoint is the ratio (1-*Se*)/*Sp*. It denotes the probability of the negative result from a diseased relative to that of a non-diseased subject. Examples of *LR*s at various cutpoints are shown in Example 5.11.

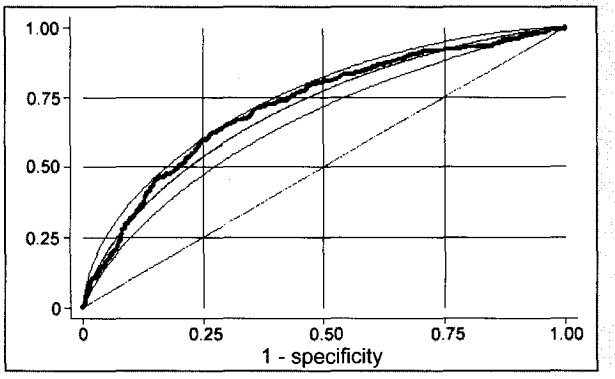
The *LR* makes use of the actual test result (as opposed to just being positive) and gives a quantitative estimate of the increased probability of disease given the observed result. For example, at the cutpoint >1.1, the *LR*+ is 2.31, meaning that a cow that tests positive at this cutpoint is 2.3 times more likely to have a PI+ calf than you thought it was prior to testing. **Note** Technically, we should state that the odds, rather than the probability, of the disease has gone up 2.6 times but if the disease is rare, then odds≈probability. This approach makes use of the fact that in general the *LR* increases as the strength of the response (test result) increases.

Often, researchers in a diagnostic setting prefer to calculate *LR*s based on the **category-specific** result (LR_{cat}) as opposed to the cumulative distributions (Giard and Hermans, 1996).

Example 5.10 ROC curves
data=bvd_test

Fig. 5.5 shows both non-parametric (thick line) and parametric (thin line) ROC curves along with 95% CI curves for the parametric ROC.

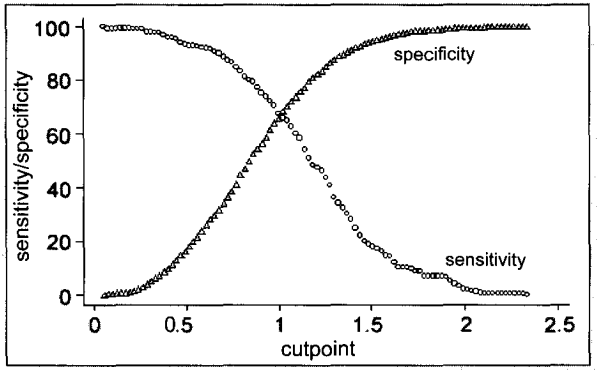
Fig. 5.5 Parametric and non-parametric ROC curves



Area under curve = 0.7038; SE (area) = 0.0166

Alternatively, a graph of the *Se* and *Sp* of a test can be plotted against various possible cutpoints as is shown in Fig. 5.6.

Fig. 5.6 Se and Sp plotted against cutpoints



As can be seen, obtaining an *Se* much greater than 70% entails accepting quite a low *Sp* (and vice versa).

Example 5.11 Likelihood ratios
data=bvd_test

Cutpoint-specific and category-specific likelihood ratios for the bvd_test data

Optical density cutpoint	PI+ category (%)	Cumulative sensitivity (%)	Positive likelihood ratio[a]	PI- category (%)	Cumulative specificity (%)	Negative likelihood ratio[a]	Category-specific likelihood ratio
≥0.0	6.76	100.00	1.00	16.91	0.00		0.40
≥0.5	4.63	93.24	1.12	17.44	16.91	0.40	0.27
≥0.7	13.17	88.61	1.35	21.64	34.34	0.33	0.61
≥0.9	16.01	75.44	1.71	18.29	55.98	0.44	0.88
≥1.1	23.13	59.43	2.31	13.29	74.27	0.55	1.74
≥1.3	18.51	36.30	2.92	6.75	87.56	0.73	2.74
≥1.5	7.83	17.79	3.13	3.62	94.31	0.87	2.16
≥1.7	3.56	9.96	4.81	1.01	97.93	0.92	3.52
≥1.9	6.05	6.41	6.02	0.85	98.94	0.95	7.12
≥2.1	3.60	0.36	1.67	0.21	99.79	1.00	17.14
>2.1	0.00	0.00		0.00	100.00	1.00	

[a] Based on cumulative distributions (*ie* sensitivity and specificity)

Here the *LR* is:

$$LR_{cat} = \frac{P(result/D+)}{P(result/D-)} \qquad \textit{Eq 5.12}$$

In either format the *LR* is useful because it combines information on both sensitivity and specificity and it allows the determination of post-test from pre-test odds of disease as shown:

$$\text{post-test odds} = LR * \text{pre-test odds} \qquad \textit{Eq 5.13}$$

When interpreting the post-test odds, we need to be aware of whether the LR_{cp} or LR_{cat} is being used. The former gives the post-test odds for an animal testing positive at that level or higher, whereas the latter gives the post-test odds for animals testing positive in that specific category (or level) of test result. The process of computing the category-specific post-test probability is as follows, assuming that, prior to testing, you thought there was a 2% probability of the cow having a PI+ fetus and that the test OD was 1.97 (LR_{cat}=7.12):

1. convert the pre-test probability to pre-test odds
   pre-test odds = 0.02 / 0.98 = 0.0204
2. multiply the pre-test odds by the likelihood ratio to get the post-test odds
   post-test odds = 0.0204 * 7.12 = 0.145
3. convert the post-test odds to a post-test probability
   post-test probability = 0.145 / (1 + 0.145) = 0.127

After obtaining a test result of 1.97, your estimate of the probability that the cow is carrying a PI+ fetus is 12.7%.

The variance of the $\ln LR_{cp}$ is:

$$\text{var}(\ln LR_{cp}) = \left(\frac{1 - PV+_{cp}}{(a+c)_{cp}} + \frac{1-P}{n} \right) \qquad \textit{Eq 5.14}$$

and a (1-α)% CI is:

$$LR_{cp} * e^{\pm Z_\alpha \sqrt{\text{var}(\ln LR_{cp})}} \qquad \textit{Eq 5.15}$$

See Greiner and Gardner (2000a) for related discussion, and Greiner et al (2000) for the relationship between *LR*s and the ROC.

5.7 Estimating the true prevalence of disease

If the *Se* and *Sp* of a test are known, the true prevalence of disease in a population is estimated by:

$$p(D+) = \frac{AP - (1-Sp)}{1 - [(1-Sp) + (1-Se)]} = \frac{AP + Sp - 1}{Se + Sp - 1} \qquad \textit{Eq 5.16}$$

where *AP* is the apparent prevalence of disease.

For example, if *AP*=0.150 and *Se*=0.363, *Sp*=0.876, then our estimate of true prevalence is 0.108 or 10.8%.

5.8 Sensitivity and specificity estimations using logistic regression

While the *Se* and *Sp* are often considered characteristics of a test, there is increasing evidence that for many tests, the *Se* and *Sp* vary with the characteristics of the population to which they are applied. For example, the specificity of serologic tests for *Brucella abortus* is higher when the test is used in populations in which no calfhood vaccination is used compared with vaccinated populations. Often it is important to know what characteristics of a population affect the *Se* and *Sp* of a test (some might prefer to think of factors relating to the occurrence of false negative or false positive results). If there are few such factors to be considered, you can stratify on these and estimate the *Se* and *Sp* in each stratum. However, when there are several factors to investigate, stratification rapidly runs into problems of inadequate sample size and it is more convenient to use a logistic regression approach (Coughlin et al, 1992; Lindberg et al, 1999; Lindberg et al, 2001). For details on logistic regression see Chapter 16.

We begin by creating a dichotomous variable representing the test outcome (positive or

negative) at each selected cutpoint of the test result. Logistic regression (see Chapter 16) can then be used to model the test outcome at each cutpoint as a function of the true health status variable (X_{ts}) as well as the factors that might affect the *Se* and *Sp*. This can either be done by carrying out separate logistic regressions using the *D*+ and *D*- animals (as shown in Example 5.12) or by including the true health status variable (X_{ts}) in the model. In the latter approach it might be necessary to include interaction terms between X_{ts} and the other factors to allow for the fact that those factors might have different effects in *D*+ and *D*- animals. Non-significant factors might be eliminated, but the variable representing the true health status of the animal must remain in the model. For a given set of factor values, the *Se* of the test at the selected cutpoint will be:

$$Se = \frac{e^{\mu}}{1+e^{\mu}} \qquad \textit{Eq 5.17}$$

where $\mu=\beta_0+\beta_1 X_{ts}+\Sigma\beta_j X_j$ when $X_{ts}=1$ and the X_j are the other factors in the model (or alternatively $\mu=\beta_0+\Sigma\beta_j X_j$ from a model based only on *D*+ animals).

Example 5.12 Estimating *Se* and *Sp* with logistic regression models
data=bvd_test

Using the bvd_test data, the effects of calving season, specimen type, breed, stage of gestation and parity on the *Se* of the ELISA were evaluated. The outcome shown here was the logistic model based on the *D*+ animals (n=281) and the ELISA result dichotomised at the test result 1.0. Specimen type, breed and parity were removed from the model because they were not statistically significant. The coefficients of interest are:

	Coef	SE	Z	P	95% CI	
Month of gestation	0.697	0.097	7.20	0.000	0.507	0.887
season=spring	0.722	0.347	2.08	0.037	0.043	1.401
season=summer	0.673	0.538	1.25	0.212	-0.383	1.728
season=fall	0.468	0.508	0.92	0.357	-0.527	1.463
constant	-4.013	0.636	-6.31	0.000	-5.260	-2.767

The sensitivity at cutpoint 1.0 for a calf at seven months' gestation in the fall is

$$\mu = -4.013 + 7 * 0.697 + 0.468$$
$$= 1.334$$

Thus,

$$Se = \frac{e^{1.334}}{1+e^{1.334}} = \frac{3.796}{4.796} = 0.79$$

Similarly, the *Sp* was estimated using a model based only on *D*- animals and found to be 0.68.

The positive coefficient for month of gestation indicates that the sensitivity of the procedure was higher later in the gestation period. Comparable assessments could be made for other values of the factors of interest (*eg* comparing seasons). Similarly, other cutpoints could be selected as the outcome to adjust the *Se* and *Sp* as deemed necessary.

The specificity of the test is:

$$Sp = 1 - \frac{e^{\mu}}{1+e^{\mu}} \qquad \textit{Eq 5.18}$$

where $\mu=\beta_0+\Sigma\beta_j X_j$ because $X_{ts}=0$ (or alternatively $\mu=\beta_0+\Sigma\beta_j X_j$ from a model based only on *D*- animals).

One can use the same approach to estimate predictive values but in that case, the outcome is the true disease status and the test result is one of the explanatory variables. Examples of this are discussed elsewhere (Greiner and Gardner, 2000, pp 19-20).

5.9 Estimating *Se* and *Sp* without a gold standard

So far, in this chapter, we have assumed that a gold standard procedure is available that detects disease and the non-diseased state perfectly. Often such a procedure is unavailable and we are unsure of the disease state values m_1 and m_0 (see Table 5.2).

5.9.1 Assuming disease-free status

A commonly used method to estimate *Sp* when disease is known to be infrequent (say, less than 2%) is to assume that all of the test positive animals are false positives (*ie* $Sp=1-AP$). If a portion of the test positives are found (or known) to be true positives, then the *AP* can be adjusted accordingly. For example, in Ireland, about four animals per 1,000 test positive to the skin test for bovine tuberculosis; hence, the *Sp* of this test cannot be less than 1-0.004=0.996 (99.6%).

5.9.2 Standard test sensitivity and specificity available

If the *Se* and *Sp* of a reference test (Se_{ref} and Sp_{ref}, respectively) are known, then from the data in a 2X2 table based on the new test results (but with disease status determined by the reference test), we could estimate the Se_{new} and Sp_{new} of the new test using the syntax of Table 5.2 as follows (Staquet et al, 1981; Enoe et al, 2000):

$$Se_{new} = \frac{n_1 Sp_{ref} - c}{nSp_{ref} - m_0} \qquad \textit{Eq 5.19}$$

$$Sp_{new} = \frac{n_0 Se_{ref} - b}{nSe_{ref} - m_1} \qquad \textit{Eq 5.20}$$

We could also estimate *P* using

$$P = \frac{n(Sp_{ref} - 1) + m_1}{n(Se_{ref} + Sp_{ref} - 1)} \qquad \textit{Eq 5.21}$$

Variance formulae are available (Gart and Buck, 2003). This procedure assumes that, conditional on the true disease state, the new test and the reference test are independent. In reality, this is not likely true, thus reducing the value of this approach.

5.9.3 Maximum likelihood estimation

If no gold standard test is available, then it might be possible to estimate the *Se* and *Sp* of two or more tests provided that sets of samples from at least two populations, with different prevalences of disease, have been tested using both tests (Enoe et al, 2000). The minimum requirement is to have two sets of test results from each of two populations. With these data, there are six unknown parameters to be estimated: the two test sensitivities, the two test specificities, and the two population prevalences. The data generate two (one for each population) 2X2 tables of combined test results so they represent 6 degrees of freedom (df), once the sample size in each population is fixed.

An approach originally developed by Hui and Walter (1980), uses maximum likelihood estimation procedures to determine the set of parameter estimates (for *Se*, *Sp* and *P*) that make the observed data most likely. As six parameters are being estimated from 6 df, it is not possible to carry out any assessment of how well the derived estimates fit the observed data. However, the procedure can be extended to more than two tests and more than two populations, in which case some evaluation of the procedure is possible (Enoe et al, 2000). Recently, an online computer program for carrying out the maximum likelihood estimations (using either a Newton-Raphson algorithm or an EM algorithm) has been made available (Pouillot et al, 2002).

The procedure is based on three critical assumptions.

1. The tests must be independent (*ie* no conditional dependency as described in section 5.6.2).
2. The *Se* and *Sp* must be constant across all of the populations evaluated.
3. The prevalence of disease in the two populations must be different. (Provided there is some difference in prevalence between the two populations, convergence usually occurs. However, as the difference in prevalence gets smaller, the CI for the estimates increases dramatically).

Violation of these assumptions invalidates the parameter estimates. However, if data from more than two tests, or more than two populations are available, it is possible to evaluate the validity of some of those assumptions.

Based on the data presented in Example 5.7, the maximum likelihood estimates (based on the Newton-Raphson algorithm) of the *Se* and *Sp* of the IFAT and PCR and the two population prevalence estimates are shown in Table 5.4. Because these diseased and non-diseased populations were selected based on clinical signs, it is likely that the test will perform better in these two populations than in other populations. Consequently, the *Se* and *Sp* estimates in Table 5.4 are probably overestimates. Because *P* in the non-diseased population has been estimated to be 0, the *Sp* estimates are exactly the same as those shown in Example 5.7.

Table 5.4 Maximum likelihood estimates of Se and Sp and population prevalences from ISA test results database

	Prevalence		Sp		Se	
	Diseased	Non-diseased	IFAT	PCR	IFAT	PCR
Estimate	0.950	0	0.951	0.979	0.823	0.974
Lower 95% CI	0.899	NA[a]	0.930	0.964	0.756	0.924
Upper 95% CI	0.976	NA	0.966	0.988	0.874	0.991

[a] Not applicable

Note Data from Example 5.7.

5.9.4 Bayesian methods

Bayesian methods offer a more flexible approach to estimating *Se* and *Sp* in the absence of a gold standard test. Bayesian estimates can incorporate prior (independent) estimates of the *Se* and *Sp* of the tests into the process. They can also be used to relax the requirement of having data from multiple populations, or to build in factors which account for the conditional dependence among test results. However, discussion of these procedures is beyond the scope of this book.

5.10 Herd-level testing

If a herd, or other aggregate of individuals, is the unit of concern, and a single test of the group (*eg* a culture of a bulk-tank milk sample for *Strep. agalactia* in a dairy herd) is taken to classify the group as test positive or test negative, the previously described approach to test evaluation and interpretation applies directly. The group becomes the unit of concern rather than the individual.

However, frequently, we are asked to certify the health status of a herd, or group of animals based on test results compiled from a number of individuals. In this instance, in addition to the *Se* and *Sp* of the test at the individual level, three factors interplay in determining the *Se* and *Sp* at the group level – namely, the frequency of disease within infected groups, the number of animals tested in the group, and the number of reactor animals per group that will designate a positive or negative herd. Once the *Se* and *Sp* of the procedure at the group level are known, the evaluation of the predictive values of positive and negative herd results follows the same pattern as already described (Martin et al, 1992; Christensen and Gardner, 2000).

As mentioned, herd sensitivity (*HSe*) and herd specificity (*HSp*) are influenced by the individual level *Se* and *Sp*, within herd *P*, and the threshold number, or percentage, of positive tests that denote the herd, or group, as test positive. For simplicity, we assume only one test is used; however, multiple tests and repeat testing results can make up the herd test and one need only establish their combined *Se* and *Sp*. The probability of obtaining a positive test is:

$$AP = p(T+) = P * Se + (1-P)(1-Sp) \qquad \textit{Eq 5.22}$$

If a herd is infected, then a positive test could arise correctly based on $P*Se$, or it could arise correctly, but for incorrect reasons, because of the $(1-P)(1-Sp)$ component.

The AP, if disease is present, is: $AP_{pos} = P * Se + (1-P)(1-Sp)$.

Note If the herd is not infected (diseased) then the AP is: $AP_{neg} = (1-Sp)$.

Now, if the critical number of animals testing positive to denote the herd as test positive is k, we can use a suitable probability distribution for AP and solve for the probability of $\geq k$ animals testing positive when n animals are tested. If n/N is less than 0.2, then a binomial distribution is acceptable for sampling of n animals from a total of N animals in a herd; otherwise, the hypergeometric distribution, which provides more accurate estimates, should be used. In the simplest setting, if $k=1$, the easiest approach is to solve the binomial for $k=0$ and take 1 minus this probability to obtain the probability of one or more test positive animals. Thus for $k=1$ and assuming the herd is infected:

$$HSe = 1 - (1-AP_{pos})^n \qquad \textit{Eq 5.23}$$

On the other hand, if the group is disease free, then

$$HSp = Sp^n \qquad \textit{Eq 5.24}$$

In the more general case, if more than k positives are required before a herd is declared positive, the HSe can be estimated as:

$$HSe = 1 - \sum_{0}^{k} C_k^n \left(AP_{pos}\right)^k \left(1 - AP_{pos}\right)^{n-k} \qquad \textit{Eq 5.25}$$

where C_k^n is the number of combinations of k positives out of n animals tested.

The HSp will be:

$$HSp = \sum_{0}^{k} C_k^n \left(Sp\right)^{n-k} \left(1 - Sp\right)^k \qquad \textit{Eq 5.26}$$

Both HSe and HSp are estimates of population parameters that apply to herds with the underlying conditions and characteristics used to determine the estimates.

The general findings from studying herd test characteristics are:

1. If n is fixed, HSe increases with P and/or AP, providing $Se>(1-Sp)$.
2. As n increases, HSe increases. Gains in HSe from increasing n are especially large if $AP<0.3$.
3. With fixed n, HSe increases as Sp decreases (noted earlier).
4. HSp decreases as Sp decreases.

A program called Herdacc (©D Jordan, 1995) is available at http://epiweb.massey.ac.nz to perform 'what-if' calculations to see how changing the sample size, the number

required to consider a herd positive or the statistical distribution (binomial or hypergeometric) affects the results. An example of estimating *HSe* and *HSp* is shown in Example 5.13.

5.10.1 Herd test characteristics based on pooled specimens

Often, to reduce cost, or when individual results are not needed or individual samples are not available, specimens from a number of animals might be pooled and tested as one sample. Such an approach is most efficient when P is low. If the laboratory is limited in terms of mass or volume of sample, one needs to be aware of the effects of sampling from the primary specimen (*eg* issues of homogeneity of mixing), as well as the effects of dilution of the substance being tested for (perhaps to below the laboratory *Se*), and the increased possibility of having extraneous cross-reacting substances added to the pool because of the inclusion of material from more animals (the latter might or might not be a 'likely' event).

If the number of animals in the pool (m) is moderately large, the *Se* of the test based on the pooled sample (*PlSe*) is likely less than *Se;* pooled *Sp* is denoted *PlSp*.

Christensen and Gardner (2000) showed that *HSe* based on r pooled samples, each containing material from m animals is:

$$HSe = 1 - [(1 - (1 - P))(1 - Se) + (1 - P)^m PlSp]^r \qquad \textit{Eq 5.27}$$

If the herd is *D-*, then the herd *Sp* based on the pooled sample (*HSp*) is $(PlSp)^r$, and if no clustering occurs within pools, $PlSp=Sp^m$. Thus, if pooled testing is performed on a number of assumed *D-* herds, then $HAP=1-HSp=1-(PlSp)^r$ which allows one to solve for the unknown *PlSp*. Similarly because $Sp=PlSp^{1/m}$, increasing r or m increases the *HSe* and decreases *HSp* in the same manner as increasing n when testing individuals within a group. At present, the optimal choice of r and m should be investigated on a case-by-case basis. An example of estimating *HSe* and *HSp* based on pooled specimens is shown in Example 5.14.

Example 5.13 Estimating herd *Se* and *Sp*

We will assume that we are testing herds with an average of 60 adult cattle for the presence of *Mycobacterium avium* subsp *paratuberculosis* (*Map*) using the ELISA. This test has an estimated *Se* of 0.391 and *Sp* of 0.964. We will assume that if *Map* is present, the true prevalence at the time of testing is 12%. Thus the *AP* in the herds with disease will be:

$$AP_{pos} = p(T+) = P * Se + (1-P)(1-Sp)$$
$$= 0.12 * 0.391 + (0.88)(1-0.964) = 0.0786$$

and the *AP* in the disease-free herds will be:

$$AP_{neg} = 0.036$$

Now, assume that the critical number of positive-testing animals to denote a herd as test positive is k=2. For the purposes of this example, we will use the binomial probability distribution to solve for the probability of ≥2 positive-testing animals when n=60 animals are tested (assuming an infinite population). The probability of $k \geq 2$ is found by first solving for the probability that $k<2$.

$$p(k<2) = \sum_{0}^{k} C_k^n AP^k (1-AP)^{n-k}$$

The probability that k=0 is:

$$p(k=0) = C_0^{60} * (0.079)^0 * (1-0.079)^{60}$$
$$= 1 * 1 * 0.921^{60} = 0.0072$$

The probability that k=1 is:

$$p(k=1) = C_1^{60} * (0.079)^1 * (1-0.079)^{59}$$
$$= 60 * 0.079 * 0.921^{59} = 0.037$$

The sum of these two probabilities is 0.044. Hence, the probability of two or more animals testing positive in a herd with P=0.12 is 1-0.044=0.956, which gives us the *HSe* estimate.

For *HSp*, we would assume the herds are disease free, thus the probability of 0 or 1 reactors is the sum of these two probabilities.

Given a herd is disease free, the probability that k=0 is:

$$p(k=0) = C_0^{60} * (0.964)^{60} (1-0.964)^0$$
$$= 1 * 0.111 * 1 = 0.111$$

and the probability that k=1 is:

$$p(k=1) = C_1^{60} * (0.964)^{59} (1-0.964)^1$$
$$= 60 * 0.115 * 0.036 = 0.248$$

Hence the *HSp* is 0.359.

With an *HSe* of 95%, we can be confident that we will declare the herd as infected if it is infected. However, with the *HSp* of only 36%, we will declare 64% of *Map*-free herds as infected, so the test needs to be used with great care.

Example 5.14 Estimating *HSe* and *HSp* from pooled specimens

We can suppose that we are going to test herds for *Map* using pooled fecal culture. Fecal culture has an estimated *Se* of 0.647 and *Sp* of 0.981. Suppose we wish to pool fecal samples from five cows together and we will use six pooled samples per herd. Hence m=5 and r=6.

If the herd is *D*-, then the herd *Sp* based on the pooled sample (assuming homogenous mixing) is:

$$(HSp) = (PlSp)^r = (Sp^m)^r = (0.981^5)^6 = 0.5624$$

If the herd is infected with a true prevalence of 12%, and assuming no dilution effect, then *HSe* is:

$$\begin{aligned} HSe &= 1 - [(1 - (0.88))(0.353) + (0.88)^5 * 0.909]^6 \\ &= 1 - [0.311 + 0.480]^6 \\ &= 1 - 0.245 = 0.755 \end{aligned}$$

As with individual testing, the *Se* at the herd level is increased by testing more animals through the use of pooled samples but the *Sp* at the herd level is decreased. One could compare the two approaches ignoring costs and then add the cost information to the final decision-making process.

SELECTED REFERENCES/SUGGESTED READING

1. Arunvipas P, VanLeeuwen J, Dohoo IR, Keefe GP. Evaluation of the reliability and repeatability of automated milk urea nitrogen testing. Can J Vet Res 2002; 67: 60-63.
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1: 307-310.
3. Bloch DA, Kraemer HC. 2x2 kappa coefficients: measures of agreement or association. Biometrics 1989; 45: 269-287.
4. Christensen J, Gardner IA. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Prev Vet Med 2000; 45: 83-106.
5. Connell FA, Koepsell TD. Measures of gain in certainty from a diagnostic test. Am J Epidemiol 1985; 121: 744-753.
6. Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. J Clin Epidemiol 1992; 45: 1-7.
7. Enoe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev Vet Med 2000; 45: 61-81.
8. Fleiss JL. Statistical methods for rates and proportions. 2d ed. John Wiley and Sons New York, 1981
9. Gardner IA, Stryhn H, Lind P, Collins MT. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev Vet Med 2000; 45: 107-122.

10. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol 2003; 83: 593-602.
11. Giard RWM, Hermans J. The diagnostic information of tests for detection of cancer: the usefulness of the likelihood ratio concept. Eur J Cancer 1996; 32: 2042-2048.
12. Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev Vet Med 2000a; 45: 3-22.
13. Greiner M, Gardner IA. Application of diagnostic tests in veterinary epidemiologic studies. Prev Vet Med 2000b; 45: 43-59.
14. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver operating characteristic analysis for diagnostic tests. Prev Vet Med 2000; 45: 23-41.
15. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-171.
16. Jacobson RH. Validation of serological assays for diagnosis of infectious diseases. Rev sci tech 1998; 17: 469-486.
17. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Stat Med 2001; 20: 2131-2147.
18. Kraemer HC, Bloch DA. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med 1994; 13: 876-880.
19. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255-268.
20. Lindberg A, Groenendaal H, Alenius S, Emanuelson U. Validation of a test for dams carrying foetuses persistently infected with bovine viral-diarrhoea virus based on determination of antibody levels in late pregnancy. Prev Vet Med 2001; 51: 199-214.
21. Linnet K. A review on the methodology for assessing diagnostic tests. Clin Chem 1988; 34: 1379-1386.
22. Martin SW. Estimating disease prevalence and the interpretation of screening test results. Prev Vet Med 1984; 2: 463-472.
23. Martin SW, Shoukri MM, Thoburn MA. Evaluating the health status of herds based on tests applied to individuals. Prev Vet Med 1992; 14: 33-44.
24. Pouillot R, Gerbier G, Gardner IA. "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. Prev Vet Med 2002; 53: 67-81.
25. Saah AJ, Hoover DR. Sensitivity and specificity reconsidered: the meaning of these terms in analytical or diagnostic settings. Am Intern Med 1997; 126: 91-94.
26. Seiler RJ. The non-diseased reactor: considerations on the interpretation of screening test results. Vet Rec 1979; 105: 226-228.
27. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. J Chron Dis 1981; 34: 599-610.
28. Tyler JW, Cullor JS. Titers, tests, and truisms: Rational interpretation of diagnostic serologic testing. J Am Vet Med Assoc 1989; 194: 1550-1558.

SAMPLE PROBLEMS

Exercise 1

1. Using the dataset isa_test, let's examine the agreement between tests.
   First, repeat the kappa test on -ifat1- and -ifat2—this is testing agreement between two laboratories. At this point, we are using the original results of the two tests.
   a. First, repeat the kappa.
   b. This kappa assumes any disagreement is total disagreement, so let's weight the agreement as follows:
      1 level apart 80% agreement
      2 levels apart 50% agreement
      3 levels apart 10% agreement
      4 levels apart 0% agreement
   What is the extent of agreement using this approach?
2. Okay, now examine the agreement of the PCR result with the dichotomised results of -ifat1-.
   a. First display the data in a 2X2 table, and test for equality of the number positive using McNemar's test.
   b. Then, if this test is not significant examine kappa. Does this value make sense given the data?
3. Repeat 2. comparing PCR with the dichotomised histologic results. Comment on the results.

Exercise 2

1. Using the dataset bvd_test; examine the sensitivity and specificity of the ELISA test at the OD cutpoints of ≥ 0.8 and ≥ 1.7. The idea of this test is to test the dam and see if the result was of value for predicting the infection status of the fetus. Comment on the results.
   Note We have already created the dichotomous variables from OD; they are labelled co_5....co_1.7 so, you don't need to generate new variables.
2. Is the OD associated with the specimen tested (*ie* milk versus blood)? Explain your answer.
3. a. If you used the test on blood at the cutpoint of ≥ 0.8 to test 1,000 pregnant cattle that had a true prevalence of 3% PIs, what would the positive and negative predictive values be? (You need to do this manually.)
   b. What if you used the test on blood at the cutpoint of ≥ 1.3 instead of ≥ 0.8?
4. Use the ROC approach to evaluate the sensitivity and specificity of the ELISA test at various optical densities as well as the overall ability of the test to differentiate diseased from non-diseased animals. Here we can leave OD as a continuous variable.
   a. First compute an overall ROC curve for the ELISA. What do you think of its predictive ability?
   b. Use the AUCs to compare the ELISA on milk versus the results of ELISA on blood samples.
5. Divide the OD into categories by using cutpoints from 0.5 to 2.1 in units of 0.2. Now compute the likelihood ratio for positive and negative tests at each of these cutpoints. Interpret these results.

6. a. Identify the impact of parity, breed and season of testing in six-to-eight-month gestation females on the sensitivity and specificity of the ELISA test for identifying PI+ calves using blood as the specimen. If, as an example, season affects the test characteristics, should one adjust for season such that the characteristics are maintained at a constant level across seasons, or just accept that the test characteristics will fluctuate by season in a predictable manner?
   b. Do these factors operate the same in cows with a PI+ calf as in cows with a PI- calf?

Exercise 3

1. Use the program Herdacc to estimate the *HSe* and *HSp* under the following situations:
   - Sensitivity=0.8
   - Specificity=0.9
   - Sample size=10
   - Population size=200
   - Sample without replacement
   - Within-herd prevalence estimates of 1%, 5%, 10%, 20% and 50%
   - Cutpoints (k) of 1, 2, or 3.

   a. What cutpoint (k) would you choose if you were testing for a disease of very low prevalence (<6%)?
   b. What cutpoints would you use if you were testing for a disease with prevalence above 19% when you wanted to limit the number of false positive herd results?
2. We are going to test herds of cattle for the presence of *E. coli* 0157. We will pool the feces from 3 (k=3) animals and test 10 (r=10) pools per herd.
   a. If the individual specimen-level sensitivity is 30% and the specificity is 95%, what would you expect the herd sensitivity and specificity to be?
   b. Do you have any suggestions about modifying the number of pools or the number of samples per pool to increase the overall value of the test?

6

MEASURES OF ASSOCIATION

OBJECTIVES

After reading this chapter, you should be able to:

1. Calculate and interpret the following measures of association:
   - risk ratio
   - odds ratio
   - incidence rate ratio
   - risk difference (attributable risk)
   - attributable fraction (exposed)
   - population attributable risk
   - attributable fraction (population).

2. Understand when to use each of the above measures of association.

3. Correctly use the concepts of strength of association and statistical significance when presenting research results.

4. Understand the basis for the common methods of computing significance tests and confidence intervals.

6.1 INTRODUCTION

Measures of association are used to assess the magnitude of the relationship between an exposure to a disease (*eg* a potential 'cause') and a disease. In contrast, measures of statistical significance cannot be used to indicate the magnitude of the effect (*ie* the strength of association) because they are heavily dependent on sample size.

In general, the material in this chapter will focus on comparing the frequency of disease in exposed subjects with the frequency of disease in subjects not exposed. Depending on study design, disease frequency can be expressed as:

- incidence risk (cohort study design)
- incidence rate (cohort study design)
- prevalence (cross-sectional study design)
- odds (cohort or cross-sectional study design).

Conversely, in case-control study designs, the objective is to compare the odds of exposure in two groups, those with the disease under investigation (the cases) and those without the disease under investigation (the controls).

If disease frequency has been measured as risk, the data for measuring the strength of association between exposure and disease are summarised in Table 6.1.

Table 6.1 Presentation of incidence risk data

	Exposure		
	Exposed	Non-exposed	
Diseased	a_1	a_0	m_1
Non-diseased	b_1	b_0	m_0
	n_1	n_0	n

where:
a_1 = the number of exposed animals that have the disease
a_0 = the number of exposed animals that do not have the disease
b_1 = the number of non-exposed animals that have the disease
b_0 = the number of non-exposed animals that do not have the disease.

If disease frequency has been measured as rates, the data for measuring the strength of association between exposure and disease are summarised in Table 6.2.

Table 6.2 Presentation of incidence rate data

	Exposure		
	Exposed	Non-exposed	
Number of cases	a_1	a_0	m_1
Animal-time at risk	t_1	t_0	t

Note For simplicity, we will refer to the frequency of disease in animals, but these could also be measured in groups of animals (*eg* number of herds affected). We will also refer to associations as though we believe them to be causal. Criteria for inferring causation are reviewed in Chapter 1.

6.2 Measures of association

The strength of an association between an exposure and a disease is usually expressed using a 'relative' effect measure which is computed as a ratio of two estimates of disease frequency. There are three common ratio measures of association: the risk ratio (*RR*), the incidence rate ratio (*IR*) and the odds ratio (*OR*). The appropriate measure of association depends on the study design and its corresponding measure of disease frequency.

6.2.1 Risk ratio

RR is the ratio of the risk of disease in the exposed group to the risk (*R*) of disease in the non-exposed group.

$$RR = p(D+|E+)/p(D+|E-)$$
$$= (a_1/n_1)/(a_0/n_0) \qquad \textit{Eq 6.1}$$

Risk ratio (also known as relative risk) can be computed in cohort studies and, in some cases, cross-sectional studies. It cannot be used in case-control studies because the p(*D*+) is an arbitrary value determined by the number of cases and controls included in the study.

RR ranges from 0 to infinity. A value of 1 means there is no association between exposure and disease:

RR < 1 exposure is protective (*eg* vaccines)
RR = 1 exposure has no effect (*ie* null value)
RR > 1 exposure is positively associated with disease.

Risk ratio says **nothing** about how much disease is occurring in the population. The actual frequency of the disease can be quite low, but the *RR* can be high. For example, in Table 6.3, which summarises the records from a large (hypothetical) herd of Hereford cattle over five years, the risk of 'cancer eye' in the herd is low: 40/6000=0.0067, but the risk of cancer eye in cattle with white eyelids is 3.8 times that of cattle with pigmented lids.

Table 6.3 Data on ocular carcinoma and eyelid pigmentation from a hypothetical longitudinal study of a large herd of Hereford cattle

		Eyelids		
		Non-pigmented	Pigmented	
Ocular carcinoma	Present	38	2	40
	Absent	4962	998	5960
		5000	1000	6000

$RR = (38/5000)/(2/1000) = 3.8$

As noted, *RR* can be computed from cross-sectional studies. Cross-sectional studies normally measure the prevalence of disease, but in certain situations (*eg* a short period of risk of disease that has been completed for all animals) the prevalence might be a valid estimate of the incidence risk. In this situation, *RR* can be used. In other situations, the term prevalence ratio (*PR*) would be preferred. It is computed in the same way as *RR* (and the term *RR* is sometimes used instead of *PR*).

6.2.2 Incidence rate ratio

The incidence rate ratio (*IR*) is the ratio of the disease frequency (measured as incidence rate) in an exposed group to the incidence rate in a non-exposed group.

$$IR = (a_1/t_1)/(a_0/t_0) \qquad \textit{Eq 6.2}$$

IR can only be computed from studies in which an incidence rate can be calculated (*ie* cohort studies). It is sometimes referred to as the incidence density ratio. *IR* ranges from 0 to infinity. A value of 1 means there is no association between the exposure and disease, with values <1 indicating protection and values >1 indicating an increased rate of disease in the exposed group.

Table 6.4 presents some hypothetical data on teat pre-dipping and cases of clinical mastitis in dairy herds.

Table 6.4 Data on cases of mastitis and pre-dipping in a hypothetical dairy herd

	Not pre-dipped	Pre-dipped	
# of cases of mastitis	18	8	26
# of cow-months	250	236	486

$IR = (18/250)/(8/236) = 2.12$

In this example, the rate of mastitis is 2.1 times higher in cows whose teats are not pre-dipped than in cows whose teats are pre-dipped prior to milking.

6.2.3 Odds ratio

The *OR* is the odds of the disease in the exposed group divided by the disease odds in the non-exposed group.

$$\begin{aligned} OR &= \text{odds}(D+|E+)/\text{odds}(D+|E-) \\ &= (a_1/b_1)/(a_0/b_0) \\ &= (a_1 b_0)/(a_0 b_1) \end{aligned} \qquad \textit{Eq 6.3}$$

Alternatively, it can be calculated as the odds of exposure in the diseased group divided by the odds of exposure in the non-diseased group.

$$\begin{aligned} OR &= \text{odds}(E+|D+)/\text{odds}(E+|D-) \\ &= (a_1/a_0)/(b_1/b_0) \\ &= (a_1 b_0)/(a_0 b_1) \end{aligned} \qquad \textit{Eq 6.4}$$

Based on the data in Table 6.3, the *OR*=(38/2)/(4962/998)=3.82.

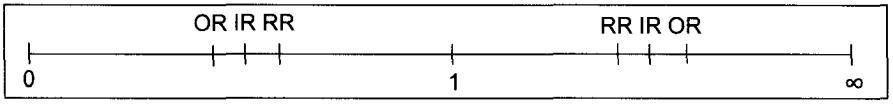
Note The odds ratio is the only measure of association that exhibits this 'symmetry' which enables you to switch the exposure and the disease (outcome). Consequently, *OR* is the only measure of strength of association applicable to case-control studies. (Because disease frequency in the sample is artificially established in case-control studies, the relative risk is not an appropriate measure of strength of association.)

The interpretation of *OR* is the same as *RR* and *IR*. An *OR*=1 indicates no effect while values <1 and >1 are indicative of reduced risk (protection) and increased risk, respectively.

6.2.4 Relationships among *RR*, *IR* and *OR*

In general, the relationships among *RR*, *IR* and *OR* is such that *IR*s are further from the null value (1) than *RR*s, and the *OR*s are even further away as can be seen in Fig. 6.1.

Fig. 6.1 General relationship among RR, IR and OR



RR* and *OR

If the disease occurs infrequently in the underlying population (prevalence or incidence risk <5%), *OR* is approximately equal to *RR*. In this situation,

$$RR = \frac{\dfrac{a_1}{a_1+b_1}}{\dfrac{a_0}{a_0+b_0}} \approx \frac{\dfrac{a_1}{b_1}}{\dfrac{a_0}{b_0}} = OR$$

because if the disease is rare, a_1 is very small and $a_1 + b_1$ approaches b_1 and a_0 is very small so $a_0 + b_0$ approaches b_0.

Similarly, if *RR* in a population is close to the null (*ie* $RR \approx 0$) then *RR* and *OR* will be very close. (If $RR=1$, then $RR=OR$). *OR*s are commonly used because they can be derived easily from logistic regression analyses (Chapter 16). It is difficult to derive *RR* from multivariable analyses, but two approaches to do this have been described (Zhang, 1998; Wacholder, 1986).

RR* and *IR

RR and *IR* will be close if the exposure has a negligible impact on the total time at risk in the study population. This occurs if the disease is rare or if *IR* is close to the null value ($IR=1$). (See Chapter 4 for details on role of time at risk in computation of incidence rates.)

OR* and *IR

OR is a good estimator of *IR* under two conditions. If controls are selected in a case-control study using 'cumulative' or risk-based sampling (*ie* controls selected from all non-cases once all cases have occurred – see Chapter 9), then *OR* will be a good estimate of *IR* only if the disease is rare. However, if controls are selected using 'density' sampling (*ie* a control selected from the non-cases each time a case occurs), then *OR* is a direct estimate of *IR*, regardless of whether or not the disease is rare.

6.3 Measures of effect

The effect (or impact) of a risk factor on a disease is usually expressed using an 'absolute' effect measure which is computed as the difference between two measures of disease frequency. The effect can be computed just for the exposed group or for the whole population. Although we use the term 'effect', it is well to remember that we are measuring associations. Thus, the 'effect' will only be the result of exposure if the association is causal.

6.3.1 Measures of effect in the exposed group

Even when an exposure is very strongly associated with disease occurrence (*eg* smoking and lung cancer in humans), typically some disease cases occur in the non-exposed population (lung cancer does occur rarely in non-smokers). The incidence in the non-exposed population can be viewed as the 'baseline' level of risk for individuals if the exposure were completely absent from the population. To evaluate the effect of an exposure on disease frequency in exposed subjects, we can consider both the absolute difference in risk between the exposed and non-exposed groups (risk difference (*RD*)) and the proportion of disease in the exposed group that is attributable to the exposure (attributable fraction (AF_e)). Both these measures incorporate the baseline risk in the non-exposed population, and assume that all other risk factors are common to both the exposed and non-exposed groups (*ie* absence of confounding, see Chapter 13).

Risk difference, incidence rate difference

RD is the risk of disease in the exposed group minus the risk of disease in the non-exposed group. It is also referred to as the **attributable risk**.

$$\begin{aligned} RD &= \mathrm{p}(D+|E+) - \mathrm{p}(D+|E-) \\ &= (a_1/n_1) - (a_0/n_0) \end{aligned} \qquad \textit{Eq 6.5}$$

RD indicates the increase in the probability of disease in an exposed group, beyond the baseline risk, that results from the exposure.

The incidence rate difference (*ID*) can similarly be calculated as the difference between two incidence rates:

$$ID = (a_1/t_1) - (a_0/t_0) \qquad \textit{Eq 6.6}$$

Difference measures are interpreted as follows:

RD or *ID* < 0 exposure is protective
RD or *ID* = 0 exposure has no effect
RD or *ID* > 1 exposure is positively associated with disease.

Attributable fraction (exposed)

The AF_e expresses the proportion of disease in exposed individuals that is due to the exposure, assuming that the relationship is causal. Alternatively, it can be viewed as the proportion of disease in the exposed group that would be avoided if the exposure were removed. AF_e can be calculated from either incidence data in both exposed and non-exposed groups, or directly from the *RR*.

$$\begin{aligned} AF_e &= RD/\mathrm{p}(D+|E+) \\ &= [(a_1/n_1) - (a_0/n_0)]/(a_1/n_1) \\ &= (RR-1)/RR \\ &\cong (OR-1)/OR \qquad \text{(approximate } AF_e\text{)} \end{aligned} \qquad \textit{Eq 6.7}$$

These calculations assume that exposure is positively associated with disease, and values for attributable fraction range theoretically from 0 (where risk is equal regardless of exposure; *RR*=1) to 1 (where there is no disease in the non-exposed group and all disease is due to the exposure; *RR*=∞). If exposures are negatively associated with disease, attributable fraction can be calculated in the same manner by regarding 'lack of exposure' to the protective factor as the factor that enhances risk. One example of this approach is estimation of vaccine efficacy. In case-control studies when actual disease frequencies in the exposed and non-exposed groups are unknown, attributable fraction can be approximated by substituting the *OR* for *RR* (as shown in Eq 6.7).

Vaccine efficacy is one form of AF_e with 'not vaccinated' equivalent to being 'factor positive' (*E*+). For example, if 20% of non-vaccinated animals develop disease [p(*D*+|*E*+)=0.20] and 5% of vaccinated animals develop disease [p(*D*+|*E*-)=0.05], the following can be calculated:

$$RD = 0.20 - 0.05 = 0.15$$
$$AF_e = 0.15/0.20 = 0.75 = 75\%$$

The vaccine has prevented 75% of the cases of disease that would have occurred in the vaccinated group if the vaccine had not been used. This is known as vaccine efficacy.

6.3.2 Measures of effect in the population

Measures of effect are useful for deciding which exposures are important contributors to the total disease experienced in a population, and which are trivial. For example, there might be a strong association between neonatal beef-calf loss and the use of prophylactic neomycin boluses at calving, but if the practice of giving neonatal calves a neomycin bolus is infrequent, it does not contribute much to neonatal mortality in beef calves. On the other hand, a relatively weak risk factor that is common might be a more important determinant of neonatal mortality in the population as a whole. In terms of national or regional disease-control programmes, information about the effect of a factor in the total population is useful in allocating resources for health-promotion and disease-control programmes.

Population attributable risk

PAR is analogous to *RD*, in that it indicates a simple difference in risk between two groups. However, the focus of *PAR* is the increase in risk of disease in the entire population that is attributable to the exposure. Therefore it is calculated as the overall observed risk (combining exposed and non-exposed groups) in the population minus the baseline risk (risk in the non-exposed). Clearly, *PAR* is determined by both the strength of the association and the frequency of exposure to the risk factor.

$$\begin{aligned} PAR &= \mathrm{p}(D+) - \mathrm{p}(D+|E-) \\ &= (m_1/n) - (a_0/n_0) \\ &= RD * \mathrm{p}(E+) \end{aligned} \qquad \textit{Eq 6.8}$$

Note *PAR* might also be called the risk difference (population), but generally isn't.

Population attributable fraction

Population attributable fraction (AF_p) is analogous to AF_e, but is focused on the disease in the entire population rather than the exposed group. Assuming a causal relationship, AF_p indicates proportion of disease in the whole population that is attributable to the exposure, and would be avoided if the exposure were removed from the population. It is calculated as the ratio of *PAR* to overall risk p(*D*+) in the population, and again is a function of the strength of the association and the prevalence of exposure.

$$\begin{aligned} AF_p &= PAR/\mathrm{p}(D+) \\ &= \frac{\mathrm{p}(E+)(RR-1)}{\mathrm{p}(E+)(RR-1)+1} \end{aligned} \qquad \textit{Eq 6.9}$$

The AF_p can be estimated from unmatched data in a case-control study using:

$$AF_p = AF_e\left(\frac{a_1}{m_1}\right) \qquad \textit{Eq 6.10}$$

Note When based on rates, the measures of effect in the exposed group or in the population relate to proportional or absolute changes in the rates, but not necessarily to the proportion or number of cases. This technical difference arises because the exposure might affect the timing (*ie* when) of disease occurrence but not the actual number of cases. Thus, the actual number of cases could be constant but the time at risk, and hence the rate, would differ.

Example 6.1 shows sample calculations of all these parameters. Table 6.5 presents a summary of the measures of association that can be computed from various study designs.

Example 6.1 Measures of association

Assume that you want to determine if being over-conditioned (*ie* fat) at the time of calving affects a cow's risk of developing ketosis. A body condition score (BCS) of 4.0 or above would be considered over-conditioned. You carry out a cohort study in a single large dairy herd (your population of interest) and all cows are observed from the time of calving through the first four months of lactation (the period at which they are at risk of developing ketosis). In addition to recording the number of cows in each BCS group that developed and did not develop ketosis, you record the number of cow-months at risk. Once a cow had a case of ketosis, she stopped contributing to the number of cow-months at risk. This occurred, on average, at two months' post-calving.

	BCS		
	≥ 4	< 4	
Ketosis +	60	157	217
Ketosis -	41	359	400
cows	101	516	617
cow-months	284	1750	2034

101 'fat' cows contributed 284 cow-months at risk and had 60 cases of ketosis.

516 'normal' cows contributed 1,750 cow-months at risk and had 157 cases of ketosis.

(continued on next page)

Example 6.1 *(continued)*

Measures of disease frequency	**Practical interpretation**
$R = p(D+) = 217/617 = 0.352$	35% of all cows had ketosis
$R_{E-} = p(D+\|E-) = 157/516 = 0.304$	30% of normal cows had ketosis
$R_{E+} = p(D+\|E+) = 60/101 = 0.594$	59% of fat cows had ketosis
$I = 217/2034 = 0.11$	0.11 cases of ketosis per cow-month in whole population
$I_{E-} = 157/1750 = 0.09$	0.09 cases of ketosis per cow-month in normal cows
$I_{E+} = 60/284 = 0.21$	0.21 cases of ketosis per cow-month in fat cows
Measures of association	
$RR = 0.594/0.304 = 1.95$	Fat cows were 1.95 times as likely to develop ketosis as normal cows
$IR = (60/284)/(157/1750) = 2.34$	The rate of ketosis in fat cows was 2.34 times higher than the rate in normal cows
$OR = (359*60)/(157*41) = 3.35$	The odds of ketosis in fat cows was 3.35 times higher than the odds in normal cows
Measures of effect	
$RD = 0.594-0.304 = 0.290$	For every 100 fat cows, 29 had ketosis due to them being fat (assuming a causal relationship)
$AF_e = 0.290/0.594 = 0.488$	49% of the ketosis occurring in fat cows was attributable to them being fat
$PAR = 0.352-0.304 = 0.048$	For any 100 cows in this population, five had ketosis that was attributable to them being fat
$AF_p = 0.048/0.352 = 0.136$	14% of the ketosis in the population was attributable to some cows being fat

Table 6.5 Summary of calculation of various measures of association by study type

	Cross-sectional	Cohort study	Case-control
RR	X	X	
IR		X	
OR	X	X	X
RD	X	X	
AF_e	X	X	X[b]
PAR	X	X[a]	
AF_p	X	X[a]	X[c]

[a] The *PAR* and AF_p can be estimated from a cohort study provided that an independent estimate of the p(*D*+) or the p(*E*+) in the source population is available.

[b] Estimated using *OR* as an approximation of *RR*.

[c] Estimated using *OR* as an approximation of *RR* and an independent estimate of p(*E*+|*D*+).

6.4 HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

The material presented in previous sections has focused on the computation of point estimates of parameters. Investigators usually want to evaluate the statistical significance of parameters as well and there are three general approaches to doing this.

- A standard error (SE) of the parameter can be computed to provide a measure of the precision of the point estimate (*ie* how much uncertainty there is in the estimate).
- A significance (hypothesis) test can be carried out to determine if the point estimate is significantly different from some value specified by the null hypothesis test.
- A confidence interval (CI) for the estimate can be computed.

What follows is a non-technical introduction to hypothesis-testing and confidence intervals in the context of unconditional (*ie* one exposure and one outcome) associations. These procedures are based on a classical (sometimes denoted 'frequentist') approach to statistics. An alternative approach, one based on Bayesian statistics, is less commonly used (see Chapter 23).

Note Throughout this section, all references to parameters in the text and in the formulae will refer to estimates derived from the data unless otherwise stated. 'Population parameters' (*ie* true, unknown values) will be referred to as such in the text.

6.4.1 Standard error

For some of the parameters described in previous sections, estimates of the variance of the parameter can be computed directly and the square root of this variance is the estimated SE of the parameter. For example, based on the incidence rate data presented in Table 6.2, SE of *ID* is:

$$\mathrm{SE}(ID) = \sqrt{\frac{a_1}{t_1^2} + \frac{a_0}{t_0^2}} \qquad \textit{Eq 6.11}$$

For other population parameters, it is not possible to directly compute their variance although methods for estimating the variance are discussed in section 6.4.3.

6.4.2 Significance (hypothesis) testing

Significance (hypothesis) testing is based on the specification of a null hypothesis about the population parameter(s). The null hypothesis is usually that there is no association between the factor and the outcome which means that measures of difference (*eg ID*) will be 0 or that ratio measures (*eg IR*) will be 1.

An alternative hypothesis is stated and it can be of a one-tailed or two-tailed nature. For example, if we have disease incidence rates in two groups (exposed and non-exposed), the usual two-tailed hypothesis is that I in the exposed group is different than in the

non-exposed group (*ie* it could be higher or lower). We are interested in finding out if there is statistical evidence to support a difference in rates that could be in either direction. A one-tailed hypothesis would be that I is higher in the exposed group than in the non-exposed group. We either do not believe that it is possible that I could be lower in the exposed group or we have no interest in this possible outcome. (An alternative one-tailed hypothesis would be that the rate is lower, and we are not at all interested in the possibility of the rate being higher.) In general, the use of one-tailed hypotheses is much harder to justify than the use of two-tailed hypotheses, so they should be used with caution.

The next step in the hypothesis-testing process is to compute a test-statistic (*eg* a t-statistic, a Z-statistic or a χ^2-statistic). From the expected distribution of this test statistic, a P-value is determined. The P-value is the probability that the test statistic would be as large or larger (in absolute value) than the computed test statistic, if the null hypothesis were true. A small P-value indicates that, if the null hypothesis were true, it is unlikely (*ie* low probability) that you would obtain a test statistic as large or larger than the one you have obtained. In this case, it is usual to reject the null hypothesis.

P-values, while conveying useful information, are limited in their ability to convey the full picture about the relationship being evaluated. They are often dichotomised into 'significant' or 'non-significant' based on some arbitrary threshold (usually set at 0.05) but this entails a huge loss of information about the parameter of interest. Knowing that an effect was 'significant' provides neither indication of the actual probability of observing the test statistic computed, nor information about the magnitude of the effect observed. Reporting the actual P-value solves the first problem but not the second. The second issue will be discussed under confidence intervals (see section 6.4.3).

Test statistics

There are four commonly used types of test statistic for evaluating associations between exposure and disease: Pearson χ^2, exact test statistics, Wald tests and likelihood ratio tests.

Pearson χ^2 is the most commonly used test statistic for the comparison of proportions. For data laid out as shown in Table 6.1, the equation for Pearson χ^2 is:

$$\chi^2 = \sum_{\substack{\text{all}\\ \text{cells}}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \qquad \textit{Eq 6.12}$$

where: obs = observed value in each cell of the table, and
exp = expected value for the cell=row total * column total/grand total.
(For example, the expected value for the cell with obs=a_1 is $n_1 * m_1/n$).

The Pearson χ^2 has an approximate χ^2 distribution provided all expected cell values are >1 and 80% (or 3 of 4 in a 2X2 table) are >5.

Note A closely related χ^2 statistic, the Mantel-Haenszel χ^2 differs from Pearson χ^2 only

by a multiplier of $n/(n\text{-}1)$ which is negligible for moderate to large values of n. The Mantel-Haenszel χ^2 is used more commonly in the analysis of stratified data (Chapter 13).

In some cases, exact probabilities for test statistics can be computed based on the distribution of the data. In these cases, the P-values are derived directly from the permutations of the data rather than by relying on an assumed distribution (*eg* normal or χ^2) for the test statistic. For example, an exact test statistic for a 2X2 table (*eg* testing the significance of an *RD* or an *RR*) can be obtained from the hypergeometric distribution. First, the hypergeometric probability of every possible table with the same row and column totals as the observed data is computed. Fisher's exact P-value is the sum of the probabilities of all tables with smaller hypergeometric probabilities than the observed table. In general, computation of exact statistics is computationally demanding so, historically, they have been used most commonly for relatively small datasets where approximations based on large numbers of observations are unsatisfactory.

Wald statistics are appropriate provided the sample size is moderate to large (see guideline for Pearson χ^2 above). The general formula for a Wald statistic is computed as:

$$Z_{Wald} = \frac{\theta - \theta_0}{\mathrm{SE}(\theta)} \qquad \textit{Eq 6.13}$$

where $\mathrm{SE}(\theta)$ is the estimated standard error of θ, and θ_0 is the value of θ specified in the null hypothesis (this is often zero). Under the null hypothesis, a Wald statistic is assumed to have a normal distribution (or a χ^2 distribution for the square of the statistic).

Likelihood ratio tests (*LRT*) are based on the likelihood of a parameter (θ). The likelihood of a parameter [$\mathrm{L}(\theta)$] is the probability (density) of obtaining the observed data, if θ is the true value of the population parameter. A likelihood ratio (*LR*) compares the likelihood of the estimated θ with the likelihood of θ_0 (the value of θ specified in the null hypothesis). An *LRT* is computed as follows and, provided the sample size is reasonably large, it has an approximate χ^2 distribution.

$$LRT = -2(\ln LR) = -2\left(\frac{\ln \mathrm{L}(\theta)}{\ln \mathrm{L}(\theta_0)}\right) \qquad \textit{Eq 6.14}$$

Note In some cases it is possible to derive an exact probability for an *LRT* rather than rely on the χ^2 approximation. In general, *LRT*s are superior to Wald tests. *LRT*s are discussed further in Chapter 16.

6.4.3 Confidence intervals

Confidence intervals (CIs) reflect the level of uncertainty in point estimates and indicate the expected range of values that a parameter might have. Although a CI covers a range of possible values for an estimated parameter, values close to the centre of the range

are much more likely than those at the ends of the range. While we use an estimated SE and a specific percentile of a test statistic distribution to compute a CI, a CI generally conveys more information than simply presenting a point estimate of a parameter and its P-value because it clearly shows a range of likely values for the population parameter. Specifically, a 95% CI means that if we were to repeat the study an infinite number of times under the same conditions and create a CI for each study, 95% of these CIs would contain the true parameter value.

If the 95% CI includes the null value (*eg* 1 for *RR*, *IR* or *OR*, 0 for *RD*, *ID*), it suggests that the parameter is not statistically significant from the null at a P-value of 0.05. However, this surrogate significance test is an 'under-use' of CI because it doesn't fully use all the information contained in the CI.

Computing confidence intervals

As with hypothesis tests, CIs can be computed using either exact probability distributions or large sample approximations. Exact CIs are based on the exact probabilities of the distributions underlying the parameter (binomial for proportions, Poisson for rates and hypergeometric for odds ratios). They are generally employed when dealing with relatively small sample sizes although increasing computer power has made the computation of exact CIs for most measures of association feasible for moderate to large sample sizes. An approximation of an exact CI (although it seems illogical that such an entity can exist) for *OR* is Cornfield's approximation (Cornfield, 1956). Computation of this CI is an iterative process and it is used less now that it is possible to directly compute exact confidence intervals.

Large sample approximations require an estimate of the variance of the parameter. As indicated above, this can be computed directly for some parameters but needs to be estimated for others. This approximation is most commonly done using a Taylor series approximation. Alternatively, a test-based method (sometimes referred to as the delta method) can be used (Kleinbaum et al, 1982) but it generally results in confidence intervals that are too narrow and will not be discussed further.

The variance of *RD* can be computed directly as:

$$\text{var}(RD) = \frac{\frac{a_1}{n_1}\left(1-\frac{a_1}{n_1}\right)}{n_1} + \frac{\frac{a_0}{n_0}\left(1-\frac{a_0}{n_0}\right)}{n_0} \qquad \textit{Eq 6.15}$$

This variance estimate can then be used to derive a CI for the risk difference (Eq 6.18).

For a ratio measure (*eg IR*), the parameter estimate and CI are computed on the log scale (*ie* CI for $\ln\theta$) and then exponentiated to obtain the CI on the original scale. However, there is no simple expression for the var($\ln\theta$), so it must be estimated. One approach to estimating the variance of a parameter is to use a first-order Taylor series approximation in the estimation procedure. The formulae for Taylor series approximation estimates of the var($\ln RR$) and var($\ln OR$) are:

$$\text{var}(\ln RR) = \frac{1}{a_1} - \frac{1}{n_1} + \frac{1}{a_0} - \frac{1}{n_0} \qquad \textit{Eq 6.16}$$

$$\text{var}(\ln OR) = \frac{1}{a_1} + \frac{1}{a_0} + \frac{1}{b_1} + \frac{1}{b_0} \qquad \textit{Eq 6.17}$$

Once an estimate of the variance has been obtained, the general formula for the confidence interval of a difference measure (θ) is:

$$\theta \pm Z_\alpha \sqrt{\text{var}(\theta)} \qquad \textit{Eq 6.18}$$

For a ratio measure, the general formula is:

$$\theta * e^{\pm Z_\alpha \sqrt{\text{var}(\ln \theta)}} \qquad \textit{Eq 6.19}$$

Note A CI for *OR* that is based on the Taylor series approximation of the variance is sometimes referred to as Woolf's approximation.

Example 6.2 presents a variety of point estimates and CIs for parameters computed in Example 6.1.

Example 6.2 Confidence intervals for measures of association

The following table presents a variety of CIs computed for some of the measures of association computed in Example 6.1

Measure of effect	Point estimate	Type of CI	CI Lower bound	CI Upper bound
ID	0.122	direct	0.066	0.177
IR	2.354	exact	1.719	3.190
RD	0.290	exact	0.186	0.393
RR	1.952	exact	1.587	2.402
OR	3.346	exact	2.108	5.329
		Woolf 's (Taylor series)	2.157	5.192
		Cornfield's	2.161	5.181
		Test based	2.188	5.117

Direct or exact CIs were computed for *ID*, *IR*, *RD* and *RR*. A variety of CIs were computed for *OR* for comparison purposes. The exact CIs are the widest, followed by Woolf's and Cornfield's approximations (which were similar). The test-based CI was the narrowest and these are not recommended for general use.

SELECTED REFERENCES/SUGGESTED READING

1. Cornfield J. A statistical problem arising from retrospective studies. Berkeley CA: Third Berkeley Symp, 1956.
2. Hammell KL, Dohoo IR. Mortality patterns in infectious salmon anemia virus outbreaks in New Brunswick, Canada. Journal of Fish Diseases 2003; accepted.
3. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research, Chapters 8 and 9. Principles and quantitative methods. London: Lifetime Learning Publications, 1982.
4. Rothman KJ, Greenland S. Modern epidemiology, Chapter 4. 2d ed. Philadelphia: Lippincott-Raven, 1998.
5. Wacholder, S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J of Epidemiol 1986; 123: 174-184.
6. Zhang, J. A method of correcting the odds ratio in cohort studies of common outcomes. J Am Med Assoc, November 18, 1998, 280: 1690-1691.

SAMPLE PROBLEMS

The file fish_morts contains data about mortalities in salmon sea cages in the Bay of Fundy. It is a small subset of data from an investigation of factors related to outbreaks of Infectious Salmon Anemia (Hammell and Dohoo, 2003). In order to know how many fish are dying in large sea cages containing anywhere from 5,000 to 20,000 fish, the producer has a diver go down to the bottom of the cage periodically to collect all of the dead fish. The data in this file are from one dive in each of 236 cages. The variables in the dataset are as follows:

Variable	Description
cage_id	cage identifier numbered 1-236
days	# of days since previous dive (ie # of days over which mortalities collected)
morts	# of mortalities found on the dive
fish	estimated # of fish in the pen
feed	type of feed (1=dry feed, 0=wet feed)

1. Compute two new variables:
   a. **fishdays** the number of fish-days since the previous dive in each cage.
   b. **mr** the daily mortality rate for each cage (expressed in morts/100,000 fish-days).
2. Compute the mean, standard deviation and median mortality rates.
3. Generate a histogram with ten 'bars' to evaluate the distribution of mortality rates.
4. Create a 0/1 variable called **hilow** that classifies cages according to whether or not they have a mortality rate above or below the median value. Add value labels to the two categories.
5. What is the relative risk of being classified as a 'high' mortality cage if the cage was fed dry feed compared with wet feed?
6. What proportion of the high mortality cages that were fed dry feed could have been prevented from being high mortality cages if they had been fed wet feed?
7. What proportion of the high mortality cages in the whole population could have been prevented from being high mortality cages if the whole population was fed wet feed?
8. How do the CIs for the above three estimates (questions 5, 6 and 7) change if you compute test-based CIs?
9. What is the *IR* for mortalities in dry-feed cages compared with wet-feed cages.
10. Overall, what proportion of mortalities could have been prevented by feeding only wet feed? Why is this value different from the value computed in question 7?

7

INTRODUCTION TO OBSERVATIONAL STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Differentiate between descriptive and explanatory studies.
2. Describe the general strength and weaknesses of experimental versus observational study designs for the identification and evaluation of causal factors.
3. Design a cross-sectional study which takes into account the strengths and weaknesses of this study type.
4. Identify circumstances in which a cross-sectional study is the most feasible observational study design.

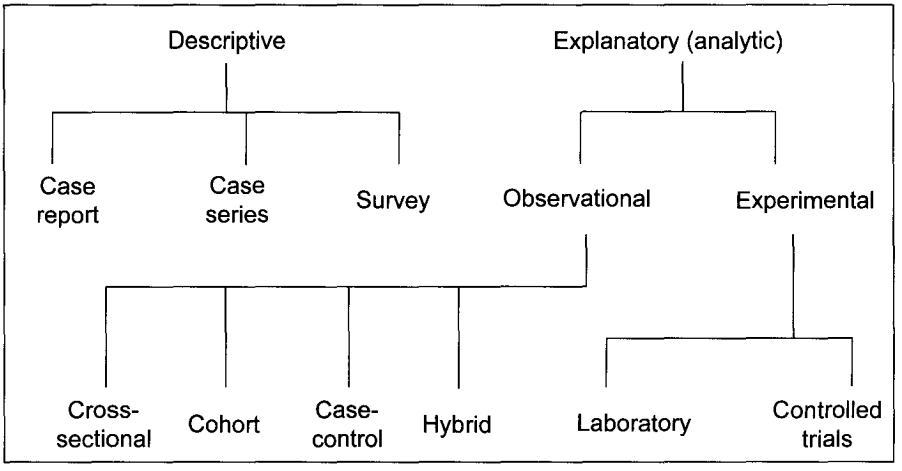
7.1 INTRODUCTION

A general theme throughout this text is that a major preoccupation of epidemiologists is to identify causal factors that can be manipulated to prevent disease, or minimise its harmful effects. We continue that theme here and ask the question 'how best to go about the task?' The overall objectives of the research and the context in which the study will be conducted will have a major impact on the choice of study type. Hence, in this section, we provide an overview of the range of study types available for use by animal-health researchers.

7.1.1 Descriptive versus analytic studies

Research studies can be classified into two large categories: descriptive and analytic (see also Fig. 7.1). Descriptive studies are designed solely to describe animal-health-related phenomena. In descriptive studies, no comparisons are made between study groups (*eg* exposed versus non-exposed or treated versus not-treated) and consequently, no conclusions about associations between exposures and outcomes can be made. Descriptive studies include case-reports, case-series reports and surveys. These are described in more detail in section 7.2.

Fig. 7.1 Schematic representation of study types



Analytic (or explanatory) study designs are ones in which the investigators set out to make comparisons between groups of study subjects (animals, herds *etc*). These comparisons allow the investigator to make inferences about relationships between exposures of interest (*eg* risk factors, treatments *etc*) and outcomes of interest (*eg* disease occurrence, productivity effects *etc*). Analytic studies can be subdivided into experimental and observational studies.

7.1.2 Experimental versus observational studies

Experimental studies are those in which the investigator controls the allocation of the study subjects to the study groups (*eg* treated versus not treated, exposed to a risk factor versus non-exposed). In contrast, in observational studies, the investigators try not to influence the natural course of events for the study subjects, but confine their activities to making careful observations (which might include collection of a variety of samples) about the study subjects with particular attention paid to the exposure and outcomes of interest.

The choice between experimental and observational approaches might be evident early on in the thought process; however, it is often valuable to consider the range of study designs available rather than fixing on the study design too early and trying to fit the investigation of the problem within the chosen design. Experiments often are the preferred choice if the treatment is straightforward and easily controllable, such as a vaccine trial or an evaluation of the efficacy of a specific therapeutic agent such as a hormone or antibiotic. A major advantage of the experimental approach is the ability to control potential confounders, both measured and unmeasured, through the process of randomisation. Observational studies usually are the preferred study design if the exposure(s) is more complex, and not easily controllable by the researcher either for practical, ethical, or economic reasons. They have the advantages that a much wider array of hypotheses can be tested, and in most instances the subjects will be exposed to the risk factor whether the study is done or not (Table 7.1). Maclure (1991) suggested some taxonomic axes or hierarchy for study design. He concluded that if a controlled trial (experimental) of a specified intervention is 'do-able' then, this is the preferred approach.

Experimental studies can be broadly classified as laboratory based or controlled trials. The former are carried out under strictly controlled conditions (*eg* laboratory studies). These have the advantage that the investigator has almost complete control over the experimental conditions (*eg* type of animal used, environmental conditions, timing, level and route of exposure, method of outcome assessment *etc*). Evidence of an association between an exposure and a factor obtained from this type of study provides the best evidence of causation, but given the very artificial environment in which they are conducted, the relevance of the results to 'real-world' conditions is often much more in doubt. Laboratory-based studies do not fall within the realm of epidemiologic studies and will not be discussed further in this text. Controlled trials are ones in which the investigator 'controls' the allocation of the subjects to the study groups, but which are carried out under natural 'real-world' conditions. The design of these types of study is discussed in Chapter 11.

Observational studies include cross-sectional (section 7.4), cohort (Chapter 8), case-control (Chapter 9) and hybrid (Chapter 10) studies. Observational studies can often take advantage of the fact that exposed subjects already exist and therefore with an appropriate design the impact of these exposures can be investigated without having to expose needlessly specifically selected study subjects to the exposure. It would be a stretch to imply that these are 'natural' experiments but the fact that subjects are

Table 7.1 Characteristics of various study types

Type of study	Level of difficulty	Level of investigator control	Strength of 'proof' of causal association	Relevance to 'real-world' situations
Descriptive				
Case report	very easy	very low	na	low to high
Case series	easy	very low	na	low to high
Survey	moderate	moderate	na	high
Explanatory - experimental				
Laboratory	moderate	very high	very high	low
Controlled trial	moderate	high	very high	high
Explanatory - observational				
Cross-sectional	moderate	low	low	moderate
Cohort	difficult	high	high	high
Case control	moderate	moderate	moderate	high

na = not applicable (associations cannot be evaluated in descriptive studies)

being exposed and the outcomes are happening, begs the question of why not seize the opportunity to capture data that can help assess any association between the exposure and the outcome. Kalsbeek and Heiss (2000) have noted that most empirical knowledge has been based on observations of incomplete samples (*ie* selected subgroups) of human (subject) experience. When it is impractical to study the entire population sampling issues must be considered and indeed, these form the basis of the different observational approaches introduced here and discussed in detail in Chapters 8-10. Observational studies make up a substantial proportion of the research carried out by veterinary epidemiologists.

7.2 Descriptive studies

As noted above, descriptive studies are not designed to evaluate any associations between exposures and outcomes of interest. However, unusual observations noted in a descriptive study often form the basis of a hypothesis which can be further investigated in an analytic study. Three forms of descriptive studies are: case reports, case series reports and surveys.

Case reports generally describe a rare condition or an unusual manifestation of a more common disease. They might be based on only one or a very few cases. The very fact that they are based on unusual cases might limit their relevance to typical 'real-world' conditions. However, these unusual observations might generate useful hypotheses to

be investigated in analytic studies. In some case reports, the authors draw conclusions about the cause or the outcome or the relative merit of the therapy used. These hypotheses are purely the author's conjecture as no data to support such a conclusion are available directly from a case report.

A case series generally present a description of the usual clinical course of the condition of interest. As such it might provide valuable information about the prognosis of the condition, provided the cases described are representative of all cases in the population. As noted, the features of the series might help the researcher posit hypotheses about causal or prognostic factors for the outcome in question, but the case series usually has no direct data on these factors.

Surveys are conducted to estimate, with some specified precision, the frequency and distribution of selected outcomes in defined populations. In many cases, their principal objective is to provide data about the frequency of occurrence of a disease of interest in a specific population. The two main design issues which need to be considered in designing a survey are the sampling protocol to be used (see Chapter 2) and the design of the data-collection instrument (see Chapter 3). If a survey collects information about both an outcome of interest and potential exposures of interest, it then becomes a cross-sectional analytic study (section 7.4), not a descriptive study because it can be used to evaluate associations between exposures and outcomes.

7.3 Observational analytic (explanatory) studies

Analytic, (also called explanatory) observational studies have an explicit formal contrast as part of their design. All analytic studies differ from descriptive studies in that the comparison of two (or more) groups is the foundation of their design. As noted above, observational studies differ from experiments in that the researcher has no control over the allocation of the study animals to the two (or more) groups being compared.

7.3.1 Prospective versus retrospective

Analytic studies can also be classified as prospective or retrospective. In prospective studies, only the exposure might have happened at the time the study starts. The design of prospective studies will include information-gathering techniques so that all the necessary data are recorded as part of the study itself, or the study could build on available data sources, supplementing these data as necessary. In retrospective studies, both the exposure and the outcome will have occurred when the study begins and typically these studies rely on pre-recorded data from one or more secondary sources. The availability of these data is an advantage, but often the quality and scope of the data are also limitations of the retrospective approach. Here again, selecting a suitable study design can maximise the information gained from the data available.

The choices of observational analytic study design have traditionally been among one of three approaches. In a cross-sectional study (section 7.4) a sample of study subjects

is obtained and then the prevalence of both disease and exposure are determined. Such studies are described as non-directional in contrast to prospective and retrospective.

In a cohort study (Chapter 8), a single sample of study subjects with heterogeneous exposure, or two or more groups defined by known exposure status, is obtained, and the incidence of the outcome in the follow-up study period determined. While these are usually prospective in nature, in select cases, with sufficient information recorded in routine databanks, they might be carried out retrospectively.

In a case-control study (Chapter 9), subjects with the outcome of interest are identified and the exposure history of these cases contrasted with the exposure history of a sample (often randomly selected from a defined source) of non-cases (also called the control subjects). These studies could be carried out retrospectively using a databank of cases that have already occurred or prospectively, in which cases are enrolled in the study as they occur. Because subjects are selected based on their outcome status, they differ from cohort studies, in which subjects are selected based on exposure status. Variations on these themes are described under the heading of hybrid study designs in Chapter 10.

Cross-sectional studies are of lower rank than other observational studies because of the inability to refute reverse-causation (*ie* determine which came first, the exposure or the outcome – see section 7.4.2); hence, when possible, other study designs should be investigated. Case-control and cohort studies are better for valid causal inferences than cross-sectional studies because of the longitudinal nature of their designs and their use of incidence data, both of which should allow refutation of reverse-causation, cohort designs being superior to case-control studies in this regard. Non-randomised intervention studies (sometimes called quasi-experiments) are ranked below case-control and cohort designs but above cross-sectional studies for causal inference purposes. The issue of random allocation of subjects to interventions is discussed in section 4 of Chapter 11.

7.4 Cross-sectional studies

Due to their ease of implementation, cross-sectional studies are one of the most frequently chosen study designs in veterinary epidemiology. Perhaps because the basic design is straightforward, there is very little written concerning details of design, at least relative to what is written regarding cohort and case-control studies. The basis of the design is that a sample of subjects is obtained and their exposure and outcome status at that point in time ascertained. Thus, the outcome frequency measure is inherently prevalence. As pointed out in Example 7.4, researchers might design questions to obtain 'incidence-like' data, but often problems remain in terms of making valid causal inferences.

If the researcher wants to make inferences about the frequency of the outcome or the prevalence of exposure in a target population, then the study subjects should be obtained by a formal random sampling procedure. The exact random process selected

can vary but could include stratified, cluster or multistage sampling as discussed in Chapter 2. However, if the primary objective is to evaluate potential associations between the exposure(s) and outcome(s) of interest, a non-random sample of study subjects is often obtained purposively. Some authors decry this non-random approach because the design is open to considerable selection bias. However, this selection bias generally limits the external validity of the study (ability to extrapolate results to other populations) rather than the internal validity. Biases that affect external validity are of less concern than those that affect internal validity.

Although the study design can support the investigation of a variety of potential causal factors and a number of outcomes, in practice one or two selected outcomes are chosen and a set of potential causal factors are selected for investigation. A potential drawback to this study design is that often the search for potential causes is not very focused and thus, a lot of data-mining for significant factors is used in the analysis stage. One also needs to decide if the association between exposure and outcome in the full study population or in defined subgroups is the principal goal. In the latter instance, researchers need to ensure that adequate numbers of study subjects in the defined groups are available to provide reasonable power for assessing the hypotheses.

The two major limitations of a cross-sectional study design are related to the fact that the outcome measure is prevalence (section 7.4.1) and that it is often difficult or impossible to determine if exposure occurred before the outcome (problem of reverse-causation – section 7.4.2).

7.4.1 Prevalence as an outcome

By its nature, a cross-sectional study measures prevalence of exposure and outcome. Consequently, it is often difficult to disentangle factors associated with persistence of the outcome (or persistence of study subjects with the outcome) and factors associated with developing the outcome in the first instance (*ie* becoming a new case). Animals with a factor which contributes to their survival once they have the disease of interest will be included in a cross-sectional study more frequently than animals without the factor (by virtue of the fact that the factor keeps them alive longer). Consequently, it will appear that the factor is associated with the disease and the investigators might incorrectly conclude that it is a 'risk factor' or cause of the disease.

7.4.2 The reverse-causation problem

Because both the exposure and outcome of interest are measured at the same time, cross-sectional studies are best suited for time-invariant exposures such as breed, sex, or permanent management factors. In these cases, the investigator can be certain that the exposure preceded the outcome (one of the fundamental criteria for establishing causation). When the exposure factors are not time-invariant, it is often very difficult to differentiate cause and effect (or the so-called reverse-causation problem). For example, if one is studying the relationship between a management factor (*eg* hoof trimming) and the frequency of hoof disorders, if the association is positive, one cannot differentiate between herds that initiated hoof-trimming in response to a problem

with hoof disorders and those that developed the disease because of the management factor. The more changeable the exposure, the worse this issue becomes. If the factor truly is preventive and often implemented when the disease has occurred, or reached a threshold frequency, the positive and negative associations could cancel each other leaving the factor appearing to be independent of the outcome.

7.4.3 Repeated cross-sectional studies versus a cohort study design

Sometimes it is necessary to follow a population over time and here one must consider performing repeated cross-sectional samplings of the population versus a cohort approach. Briefly, if the objective is to follow specific individuals over time then the cohort approach is preferable. However, depending on the length of the study period, the remaining individuals in the study will become increasingly different from the existing population at that time (for example they will be much older and in many instances represent a highly selected subgroup of those originally studied). If the objective relates more to the events and associations within the population at different periods of time, then a series of repeated cross-sectional studies might be the preferred approach. In this design, depending on the sampling fraction, most of the study subjects selected at different points in time will not have been included in prior samples. However, with larger sampling fractions, sufficient subjects might be selected in two or more samplings to allow within-study subject comparisons over time (see Diehr et al, 1995, for methods to help choose between these two approaches).

7.5 Examples of cross-sectional studies

In this section, we discuss four published cross-sectional studies that highlight some of the strengths and weaknesses of this study design. Example 7.1 demonstrates the value of random sampling in allowing for the analysis of data at multiple levels, and the evaluation of both time variant and invariant exposures, and the use of information about the potential duration of exposure to attempt to clarify the directionality of possible causal associations.

In Example 7.2, the authors used a combination of non-random and random sampling to achieve their objectives.

The study described in Example 7.3 used repeat visits to study farms over the period of the year. As the population was dynamic, at the animal level, the study could be viewed as a repeated cross-sectional census of the cattle on the study farms.

In the study described in Example 7.4, the authors attempt to obtain incidence data in a one-time cross-sectional study. However, the outcome data might have been a mixture of incident and prevalent cases and the reverse-causation issue between management factors and the outcome was still a problem in the study.

Example 7.1 Time variant and invariant factors in cross-sectional studies.

Atwill et al (1996), used a random sample of 3,000 of the 39,000 equine operations (based on a 1988 census), in New York State in a cross-sectional study of risk factors for *Ehrlichia risticii* (ER) . The study was conducted in 1991-93 and was designed to identify counties with a high prevalence of ER as well as to identify host, management and environmental factors consistent with an oral (helminth mediated) route of transmission. Data were obtained from personal interviews with owners, blood samples from horses and resource maps for geographic information. The use of a random, state-wide sample allowed for analyses to be carried out at the county, farm and horse levels. If a purposive study had been done, it might very well not have included sufficient counties or farms for analyses at those levels. A wide range of both time invariant (*eg* breed of horse) and time variant (*eg* frequency of deworming) were evaluated. Of particular interest was the evaluation of environmental characteristics (elevation and proximity to large bodies of water) as risk factors that might relate to helminth mediated transmission. While these factors are time invariant for the farm, they might be time variant for the individual horse because they often moved between farms. The authors attempted to clarify the directionality of these associations by carrying out three sets of analyses based on the length of time that horses had been on the farm. Among the many results reported was an association between county and risk of seropositivity. Given the geographic attributes of the counties (low elevation level and proximity to large bodies of water), the authors concluded that this was consistent with helminth vector spread of the disease (others had found similar geographic associations). However, at the farm level, the farms with the highest risk of seropositivity had a low elevation but no proximity to standing or running water. This tended to cast doubt on the helminth hypothesis.

Example 7.2 Random and non-random sampling in cross-sectional studies

This study (McDermott et al, 1987a,b) was carried out to estimate the frequency of two selected diseases and identify 'associated factors' as well as the impact of the two diseases on production in a rural development project area in Sudan. Three of five (out of seven) accessible areas were included in the study. Two cattle camps within each of the three areas were selected on the basis of accessibility. Within each camp, individual family herds were selected systematically; if that herd owner was absent, the adjacent herd was selected. Finally, within each herd, 25 animals were selected in proportion to the sex and age structure of the herd. Consequently, areas and camps were sampled purposively based on the investigators' knowledge of the area and logisitic concerns, but herds and animals were sampled using a form of random sampling (systematic sampling). The authors discussed random versus non-random sampling strategies in this context and defended their non-random process given the practical and logistical limitations (while stressing the need for a knowledge of the populations being sampled). The systematic sampling of cattle by age and sex was designed to obviate the problem of owners presenting very healthy animals (or more likely, older diseased animals) for the study.

Two research teams visited each location, one to sample the animals and one to conduct the interview with the owners. A central laboratory tested the samples for brucellosis and contagious bovine pleuropneumonia. Data analyses were performed at the individual animal level. A specific hypothesis about a breed association was not confirmed, but the results was confounded by differential missing information by breed. Although written records for past events were not available, the knowledge of the owners was deemed satisfactory in this context. In a subsequent paper based on the same study the authors discuss the extent to which a cross-sectional study could provide useful data for livestock development plans. Based on the fact that the associations detected in the study, and the impact of the diseases were consistent with known effects in other areas, they concluded that the study design was useful (and perhaps the only feasible approach).

Example 7.3 Repeated cross-sectional sampling

This example demonstrates the additional power and enhanced inference-making possible by using repeated cross-sectional surveys (O'Callaghan et al, 1999). The study population was small holder dairies in the Kenyan highlands. Six of 15 dairy societies were selected purposely and then 15 farms within each society were randomly selected giving a sample of 90 farms. Each farm was visited monthly for 12 months. A comprehensive initial farm survey was conducted on risk factors for Theileriosis. At each visit, a survey was conducted of all animals present on the day of visit and blood samples (for titres) obtained. At the farm level, this study could be described as a single-cohort study. However, at the animal level, the population was dynamic. Some animals were present at most farm visits while others entered and left the study. Further, although formal farm surveys were not conducted at each visit, the researchers were able to ascertain the management practices actually used, as distinct from the replies to the initial survey that tended to describe the management practices recommended for that area. The monthly samplings also allowed the investigators to better demarcate when new (or repeated) exposures to *T. parva* occurred, and hence obtain incidence data at the animal level.

Example 7.4 Attempts to obtain incidence data

These authors (Wells et al, 1999) used a cross-sectional sample of dairy herds with more than 30 cows to assess the incidence of papillomatous digital dermatitis (PDD) and investigate herd-level risk factors. Incidence data were derived by asking the herd managers for the number of cows that had 'shown clinical signs' of PDD in the previous 12 months (it is not clear if these were new or continuing cases of PDD). Herds were later categorised into those with >5% versus <5% of cows affected. Nonetheless, when making inferences about potential causal associations, the possibility of factors being an effect of PDD, rather than a cause (a reverse-causation between hoof-trimming and PDD level), was acknowledged by the authors.

SELECTED REFERENCES/SUGGESTED READING

1. Atwill ER, Mohammed HO, Lopez JW, McCulloch CE, Dubovi EJ. Cross-sectional evaluation of environmental, host, and management factors associated with risk of seropositivity to *Ehrlichia risticii* in horses of New York State. Am J Vet Res 1996; 57: 278-85.
2. Diehr P, Martin DC, Koepsell T, Cheadle A, Psaty BM, Wagner EH. Optimal survey design for community intervention evaluations: cohort or cross-sectional? J Clin Epidemiol 1995; 48: 1461-1472.
3. Kalsbeek W, Heiss G. Building bridges between populations and samples in epidemiological studies. Annu Rev Public Health 2000; 21: 147-169.
4. Maclure M. Taxonomic axes of epidemiologic study designs: a refutationist perspective. J Clin Epidemiol 1991;44: 1045-1053.
5. McDermott JJ, Deng KA, Jayatileka TN, El Jack MA. A cross-sectional cattle disease study in Kongor Rural Council, Southern Sudan. I. Prevalence estimates and age, sex and breed associations for brucellosis and contagious bovine pleuropneumonia. Prev Vet Med 1987a; 5: 111-123.
6. McDermott JJ, Deng KA, Jayatileka TN, El Jack MA. A cross-sectional cattle disease study in Kongor Rural Council, Southern Sudan. II. Brucellosis in cows: associated factors, impact on production and disease control considerations. Prev Vet Med 1987b; 5: 125-132.
7. O'Callaghan CJ, Medley GF, Peter TF, Mahan SM and Perry BD. Predicting the effect of vaccination on the transmission dynamics of heartwater (cowdria ruminatium infection). Prev Vet Med 1999; 42: 17-38.
8. Wells SJ, Garber LP, Wagner BA. Papillomatous digital dermatitis and associated risk factors in US dairy herds. Prev Vet Med 1999; 38: 11-24.

8

COHORT STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Describe the major design features of risk-based and rate-based cohort studies.
2. Differentiate between open and closed study populations.
3. Identify hypotheses and population types that are consistent with risk-based cohort studies.
4. Identify hypotheses and population types that are consistent with rate-based cohort studies.
5. Elaborate the principles used to select and measure the exposure.
6. Design and implement a valid cohort study for studying a specific hypothesis.

8.1 INTRODUCTION

The word **cohort**, from its Latin root, refers to a group of subjects that has a defined characteristic in common. In epidemiologic study design, the characteristic of interest is the exposure status. Usually, the selection of the study groups is based directly on the exposure status (*eg* when we select a group of exposed and a group of non-exposed individuals). However, we might select a single group of subjects that we believe will be heterogeneous with respect to the exposure(s) of interest and then determine their exposure status. We denote this design as a **single cohort** or **longitudinal** study. Once selected, we ensure the study subjects do not have the outcome(s) of interest at the start of the follow-up period, and then compare the incidence of the outcome in the groups defined by exposure status during the specified follow-up time period. Note that the study subjects could be individuals or aggregates of individuals, such as litters, pens or herds. Comprehensive reviews of cohort study design and analysis are available (Prentice, 1995; Samet and Munoz, 1998; Rothman and Greenland, 1998).

8.2 BASIS

Each specific study presents its own unique challenges, but the starting point for all studies is to clearly and concisely state the hypothesis to be tested. This includes defining the **exposure**(s), **outcome**(s) and **follow-up period** in the study subjects (*ie* animals, herds or other aggregates) and the setting (*ie* context) of interest. If sufficient biological facts are known, such hypotheses should also indicate the amount of exposure that is likely needed to trigger the effect, and how long after an **exposure threshold** is reached before one might reasonably expect to see disease from that exposure arise (*ie* the **induction period**). Clarifying the study objectives often helps us decide whether current or past exposure is relevant, whether lifetime exposure or exposure in a narrower window of time is important, whether repeated measures of exposures are required and if so, how to handle changes in exposure status.

Depending on the availability of suitable records, cohort studies could be performed **prospectively** or **retrospectively**. Prospective studies imply that the outcome has not occurred at the time the study starts. They often provide the opportunity for more detailed information-gathering and attention to recording the details of interest than retrospective studies. Retrospective cohort studies imply that the follow-up period has ended when the study subjects are selected based on their exposure status. Retrospective studies require suitable existing databases and are often of more limited scope than prospective studies.

When choosing two or more exposure groups, it is desirable that they be obtained from the same identifiable population. Often, these exposure groups are chosen purposively, not randomly. The study subjects in these exposure groups might not be equal with respect to risk factors other than exposure, and this needs to be taken into account in the study design in order to prevent confounding (see section 8.6).

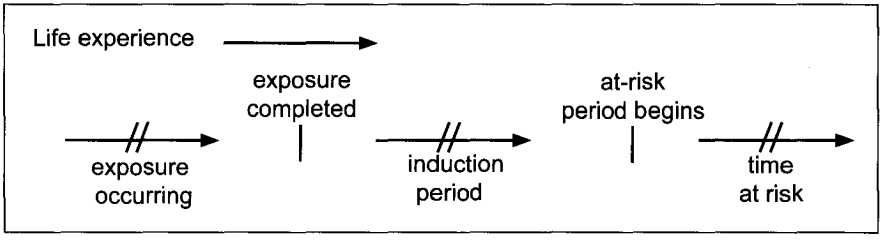
8.3 THE EXPOSURE

In cohort studies, our objective is to identify the consequences of a specific exposure factor. The exposure refers to any potential cause of disease and might range from infectious or noxious agents to housing, management or feed-related factors. Exposure status can be measured on a dichotomous scale (exposed or non-exposed), an ordinal scale (low, medium, or high dose), or a continuous scale (organisms per gm of feces, ppm of a toxin in air or water, gm of colostrum ingested *etc*). Exposure can be expressed separately in terms of dosage and duration or as a combination of the two (*ie* perhaps their product). The exposure might be a permanent factor or a factor that can change over time.

8.3.1 Permanent exposures

These exposures include factors such as sex, breed or whether or not a calf received sufficient colostrum within 12 hours of birth. Permanent and 'one-time' exposures are relatively easy to measure, but even here a moment's thought would suggest that defining 'sufficient' or 'inadequate' with respect to colostrum intake in a calf might be more complex than it first appears to be. In any event, for factors where the exposure is based on a threshold or dosage, the amount of exposure necessary to deem an individual as being 'exposed' needs to be clearly stated as exposed time at risk does not begin until the criteria for completing the exposed state have been met. If the outcome event occurs during the time period before exposure is completed, it should not be included in the analysis; as exposure has not been completed, it could not have caused the outcome. These issues are shown graphically in Fig. 8.1.

Fig. 8.1 Life experience with exposure, induction period and time at risk



An example of a cohort study with a permanent exposure factor is presented in Example 8.1.

8.3.2 Non-permanent exposures

When discussing exposures that can change during the study period, it is useful to recall the criteria for completing 'exposure' as there might be time and/or dose components necessary before the study individuals are deemed to be exposed. If one type of exposure ends and another type of exposure begins, there might be a lag effect from the first exposure. Diseases occurring within this period should be attributed to the former, not the latter, exposure.

Example 8.1 A cohort-study design with a permanent exposure factor

Suppose we want to investigate the association between congenital infection with *Neospora caninum* and subsequent fertility and abortion in dairy cattle (see Thurmond and Hietala (1997) for an example). We will assume that we can obtain a sufficient number of cattle in local herds and that we have sufficient time and money to complete a multiyear study. We will further assume that we have a perfect test for *Neospora* infection and clear criteria for 'fertility' and 'abortion'. We might begin by testing at birth to identify infected calves, then follow these through their breeding and the subsequent lactation. If we have no losses to our study groups, we could compute and contrast the age at pregnancy, and the risk of abortion in the congenitally infected and non-infected (at birth) groups. One might need to decide if calves infected after birth need to be identified and excluded from the non-exposed group (see Example 8.2).

For an exposure that can change over time (for example, the type of housing experienced by a cow over two lactations), both the timing and the order of the exposures might be important to measure and analyse. This adds further complexity to the exposure factor. Sometimes a simple summary measure of exposure will suffice (*eg* days spent on concrete versus dirt flooring), whereas in other studies more complex measures of exposure are needed (*eg* the number of days spent housed in different stall designs where the stall size and the flooring material also might need to be considered). Neutering is often a factor of interest and here the age at neutering as well as the fact of neutering could be important. Examples 8.2 and 8.3 are descriptions of studies with exposures that change over time.

Example 8.2 A cohort-study design with a non-permanent exposure factor

In a follow-up example on *Neospora caninum* infection, we might develop a new hypothesis concerning post-natal infections. Thus, we would monitor the study calves for the acquisition of infection after birth by testing them at birth to ensure they were not congenitally infected and then test at regular (*eg* three-month) intervals thereafter. Abortion following first conception might be the outcome of interest in this study. In addition, we might wish to stratify the data based on the timing of infection (*eg* the actual age at infection or whether the infection occurred before or during pregnancy). In either instance, this would be a closed population and thus, a risk-based analysis would be appropriate (see section 8.5).

8.3.3 Determining exposure time

If the timing and nature of exposure are obvious, then exposure time continues to accumulate until the event of interest occurs, or the study period ends or, if there are losses to follow-up, until the last date exposure status is known (in this instance use the midpoint of the last period if the precise time is unknown). If the measure of exposure is a composite (*eg* 'total hours confined' determined from the hours per day confined

Example 8.3 A more complex cohort-study design with a non-permanent exposure factor

As a second example, we might investigate the association between stall confinement and stereotypical behaviour in horses. Here we have a much more complex exposure to assess. Some of the axes of exposure could relate to the size and design of the stall (*eg* stall size, construction, lighting and whether or not the horse can see other horses), the duration of recent stall confinement, and the history of previous confinement. It is also possible for exposure to stall confinement to be intermittent (*eg* a horse is confined during the colder months but on pasture during the summer).

We would need to define the criteria for being exposed to stall confinement (*eg* how many days in a stall is considered to be confined or exposed?). If a stereotypy occurs before the 'exposure' is completed, it would be excluded from the analysis. Once the criteria of being exposed are met, the number of days the horse is confined is then accumulated for the purposes of computing incidence rates. Subsequently, if a horse develops a stereotypy, its exposure category at the time of that occurrence is used for the purpose of calculating the incidence rate. If the animal's housing is changed, then after considering any lag effects, the number of days spent in each exposure category is accumulated in the overall denominator of the rate for that category.

multiplied by the number of days confined), then it might be advantageous to study the two components separately in the same model because their effects might differ (it might be the number of days confined and not the hours of confinement per day that increases the risk of a stereotypy; see Example 8.3).

If the exposure status can change during the study period, an individual can accumulate animal-time in both exposed and non-exposed groups. In addition, if an induction period is known, then technically, until that period is over, the experience of otherwise-exposed individuals should be added to the non-exposed group. Some researchers prefer to discard the experience during the induction period for exposed individuals because of uncertainties about the duration of the induction period. In the face of uncertainty about these effects, this is likely the best choice to make providing there is sufficient time at risk in the non-exposed group to maintain precision. Similarly, if previously exposed individuals become non-exposed, one would only add the non-exposure time (of otherwise-exposed individuals) to the non-exposed cohort if there was strong evidence that the period of risk for the outcome of interest was of limited duration. In Example 8.3, if the horse ceases to be confined, it would only start to accumulate time in the non-confined group if it was assumed that the effect of confinement on stereotypy development ended as soon as confinement ended (*ie* there was no lag effect).

8.3.4 Measuring exposure on a continuous scale

Typically, individuals are classified as exposed or non-exposed (*ie* a dichotomous exposure) or perhaps into an ordinal level of exposure category. The outcome frequency will then be determined within each exposure category. Exposure might also be measured

and classified on a continuous scale. As in other instances, maintaining the continuous scale has advantages because the categorisation of a continuous exposure variable usually results in loss of information. Thus, one might relate the outcome frequency (*ie* risk or rate) to the continuous exposure scale using an appropriate regression model. If categorisation of exposure is deemed to be sufficient, there should be some evidence available to help decide the appropriate cutpoints for exposure categories. As before, more than one axis of exposure measurement is useful. For example, we might assess the risk of the outcome according to the maximum daily exposure, or the median daily exposure (these would be the cutpoints for exposure categorisation). If lag effects are minimal, when different exposure categories exist for the same individual, the exposure category assigned to an individual is that level of exposure the individual was in at the time the outcome event occurred. The prior exposure time at that level is accumulated for that individual as well as any time at risk in the other levels of exposure for that individual. The more information that can be collected on exposure, such as its level(s), when it started, and when (if) it stopped adds credibility for causal relationships is more useful for preventive action/management intervention, and enhances our biological understanding of the problem.

8.4 Sample-size aspects

Usually, sample-size determination assumes that we want an equal number of exposed and non-exposed individuals. There is nothing magical about this assumption and, if cost or other practicalities dictate different sample sizes in different exposure categories, then this can be accounted for. The risk-based approach for sample size estimation, as shown in Chapter 2, is often sufficient for planning purposes even if the population is open and a rate-based study must be used.

8.5 The nature of the exposure groups

When selecting two or more exposure groups, it is best if the groups come from one identifiable population that has numerous characteristics in common other than exposure. The 'population' might be real, or it could be a virtual population as in the 'group of dogs at a clinic' or the 'group of farms served by one veterinary practice.' If exposure groups are defined at the start of the study and this does not change, it is called a **fixed cohort**. If there are no additions and few or no losses, then the fixed cohort is deemed to be **closed** (section 4.4.1). This allows the calculation of risks and average survival times (times to endpoint).

In many cohort studies, the population is **open** in that some or all of the individuals in the cohorts will change over time and hence, they will be observed for only a portion of their at-risk period. Individuals might be lost from, or added to, the study and/or the exposure status of each individual can change over time. In this situation, one needs to accumulate the amount of exposure time and non-exposure time contributed by each individual. Open populations require a rate-based approach to study design.

8.5.1 Risk-based (cumulative incidence) designs

In a 2X2 table, the summary format for a closed-population cohort study is:

	Exposed	Non-exposed	Total
Diseased	a_1	a_0	m_1
Non-diseased	b_1	b_0	m_0
Total	n_1	n_0	n

In this design, we select n_1 exposed and n_0 non-exposed individuals from the N_1 exposed and N_0 non-exposed individuals in the target population. Having ensured that none of the study group has the outcome (a) at the start of follow-up, we follow or observe all study subjects for the full period of risk. During the study, we observe a_1 exposed subjects developing disease out of the n_1 exposed subjects and a_0 non-exposed diseased subjects out of the n_0 non-exposed subjects. Overall, we observe a total of m_1 diseased and m_0 non-diseased subjects. The study population data are used to estimate the two risks (R) of concern, namely:

$$R_1 = a_1/n_1 \qquad \text{and} \qquad R_0 = a_0/n_0$$

8.5.2 Rate-based (incidence density) designs

In this design, the initially selected exposed and non-exposed subjects each contribute an amount of 'at-risk' time to the denominator of the rates until they develop the outcome, or are lost to the study or their observation ends (*eg* the study is terminated). If new individuals are added to the study group, or if the exposure status of individuals changes during the follow-up period, then the appropriate amount of time at risk is added to either the exposed or non-exposed categories. As noted earlier, individuals do not contribute exposed time at risk until they have qualified as 'exposed' and until the induction or lag period are completed.

The summary format for an open-population cohort study is:

	Exposed	Non-exposed	Total
Diseased	a_1	a_0	m_1
Animal-time at risk	t_1	t_0	t

All subjects in the study group are followed for the duration of their risk within the study period, and we observe a_1 exposed cases of disease out of t_1 animal-time units of exposure and a_0 non-exposed cases out of t_0 non-exposed animal-time units. Here t_1 is the sum of all of the exposed time at risk for each of the individuals that were 'exposed' for some time prior to, or during, follow-up. Similarly t_0 is the summed time at risk in the non-exposed category. The two rates (I) of interest we wish to estimate would be:

$$I_1 = a_1/t_1 \qquad \text{and} \qquad I_0 = a_0/t_0$$

If the follow-up time is relatively short, the rates will be used to measure disease frequency. If the follow-up time is so long that assumptions about a constant rate over the entire study period are highly suspect, survival analysis methods should be used to analyse the data (see Chapter 19).

8.6 Ensuring exposed and non-exposed groups are comparable

If the study subjects in the exposure groups are not comparable with respect to factors related to the outcome and exposure, a biased assessment of the exposure–outcome association can result. In general, one or more of the following three approaches can be used to help ensure that the exposed and non-exposed groups are comparable in all relevant aspects other than their exposure status.

8.6.1 Exclusion/restricted sampling

Here we identify variables likely to be confounders (see Chapter 13 for a discussion of confounding) and then select both exposed and non-exposed study subjects so that they have only one level of these variables (*eg* use only one age, one breed, or one sex of animal). In other circumstances, the criteria for study entry are restricted (*eg* only steers in defined feedlots) and applied to both exposed and non-exposed groups. This serves to reduce the background variability, or noise, and might help reduce confounding from unknown factors.

8.6.2 Matching

Here we identify major confounding variables and then select the non-exposed subjects so that they are the same as the exposed subjects with respect to these variables. One method of selection is **one-on-one matching** (*eg* select the next listed, non-exposed animal (if using existing records) provided they are of the same age, breed *etc*). Another method is **group matching** which ensures an overall balance. These two approaches lead to different forms of analysis (see section 13.6). Matching can help achieve greater study efficacy as well as confounding control in cohort studies.

8.6.3 Analytic control

Here we identify and measure the important confounders and then use analytic control (*eg* ranging from Mantel-Haenszel-type stratification to multivariable regression approaches) to adjust for these confounders (see Chapters 13 and 16), and hence obtain unbiased measures of association. Information on other exposures/confounding factors also should be as accurate as possible because misclassification of these confounders seriously reduces our ability to control confounding.

8.7 FOLLOW-UP PERIOD AND PROCESSES

This is a very important aspect of study validity and the follow-up process must be unbiased with respect to exposure status. This often requires some form of blinding process as to exposure status. This can be implemented in both prospective and retrospective studies (although the latter has more limited options). Unless the study period is short, it is helpful to enumerate and characterise the population at risk at specified times during the study. If passive surveillance for cases is used, then cases are identified when reported. With active surveillance and regular evaluation, it is feasible to get more accurate data on time of outcome occurrence. The date of event occurrence should be as accurate as possible, because inaccurate recording increases the possibility of serious measurement error. Collecting ancillary information is useful to manage issues such as loss to follow-up because of culls/sales, and to assess if censorship is unrelated to exposure. In a closed cohort, it is important to trace as many 'losses' as possible in order to ascertain their last known health and exposure status. If the percentage of study subjects lost becomes large (some use >10% as a cutpoint), it will begin to cast doubt on the validity of study findings.

8.8 MEASURING THE OUTCOME

Each study will need explicit protocols for determining the occurrence and timing of outcome events. Clear definition(s) of **diagnostic criteria** are useful to ensure as few diagnostic errors as possible (*eg* what constitutes stereotypical behaviour). Ensuring **blindness** as part of the diagnostic process is helpful to equalise diagnostic errors but this does not reduce them.

The outcome is measured as incidence in a cohort study. This requires at least two tests: the first at the start of the period to ensure that the animals did not have the disease, and the second to investigate whether or not the disease developed during the observation period. If the study group is screened regularly for the outcome event, then the time of occurrence of the outcome should be placed at the midpoint between examinations. If clinical diagnostic data are used to indicate the outcome event, you must remember that these are based on time of diagnosis not on time of occurrence of disease. For diseases that might remain subclinical for months or years, ignoring this difference could lead to inferential errors.

One of the advantages of a cohort study is that we can assess multiple outcomes. However, if multiple outcomes are assessed, some might be significantly associated with the exposure by chance alone. In this instance, it might be best to consider the study as hypothesis-generating not hypothesis-testing, unless the outcomes were specified *a priori*.

8.9 ANALYSIS/INTERPRETATION

8.9.1 Risk-based cohort analysis

If the study population is closed, we can measure the average risk of disease(s) and survival times during the follow-up period. Bivariable analyses are shown in Chapter 6, stratified analyses in Chapter 13 and multivariable logistic models in Chapter 16. Retrospective single-cohort studies that were analysed using a risk-based approach are presented in Examples 8.4 and 8.5.

8.9.2 Rate-based cohort analyses

If the study population is open, rates are used to measure disease frequency and a Poisson regression model (Chapter 18) is appropriate for the analysis. The incidence of disease is expressed relative to the amount of time at each level of exposure, not to the number of exposed (or non-exposed) individuals. Example 8.6 contains an example of a rate-based cohort study of colic in horses.

Example 8.4 Retrospective single-cohort study – closed population

Risk factors for metritis in Danish dairy cows (Bruun et al, 2002)
A retrospective single-cohort (longitudinal) study of factors affecting metritis occurrence in the first 30 days of lactation was conducted in Denmark using data collected during 1993-1994 (Bruun et al, 2002). Data on herd size, breed, parity and treatment of disease were obtained from the Danish Cattle Database. Management and production-facility data were collected using a questionnaire, conducted as a telephone interview in 1994. The study included 2,144 herds from three regions in Denmark (102,060 cows). Herd-level exposure variables included: herd size, housing, flooring, grazing, calving measures, and calving supervision. Cow-level exposure variables were: parity, breed, calving season and whether the cow had been treated by a veterinarian for dystocia or retained placenta, reproductive disease, ketosis, milk fever or dry-cow mastitis.

This study population can be considered fixed in that the exposure status was considered permanent within a lactation and all cows were observed for the full 30-day risk period – few cows are culled during this stage of lactation, and no 'new' cows were added, so it was a closed population.

Example 8.5 Retrospective cohort study – closed population

Musculoskeletal injuries in Thoroughbred horses during races (Cohen et al, 1999)
This example allows the reader to compare different approaches (cohort and case-crossover studies) to answer the same general question – in this instance, factors affecting leg injuries in racehorses. See Example 10.2 for the comparison study.

This was a retrospective cohort design. The study population was selected from a larger cohort of horses that raced on four tracks in Kentucky between January 1, 1996 and October 25, 1997. Prior to each race, each horse was examined by a Kentucky Racing Commission veterinarian and a summary score indicative of increased injury risk was recorded. This score was dichotomised by the researchers, and records for horses with an elevated risk and one randomly selected horse deemed to be at no increased risk in that race were selected for study. A major analytical feature was that, over the study period, horses could be included many times and their injury risk status could change.

Any horse that raced on one of the four tracks was eligible for the study and, although the horse population itself likely changed during the study period, all horses were observed for the full risk period (*ie* the race) and hence, the study population was closed allowing a risk-based analysis.

Example 8.6 Prospective cohort study – open population

Prospective study of equine colic incidence and mortality (Tinker et al, 1997)
Data from 31 farms with more than 20 horses each were maintained for one year. Descriptive information on 1,427 horses were collected at the outset and updated every three months allowing horse-time at risk to be determined for each horse. The crude *I* for colic was 10.6/100 horse-years but this varied from 0 to 30/100 horse-years across farms. Fourteen horses had more than one colic episode and the colic-specific mortality rate was 0.7/100 horse-years. The rates of colic differed by breed, use and age but not by gender.

SELECTED REFERENCES/SUGGESTED READING

1. Bruun J, Ersbøll AK, Alban L. Risk factors for metritis in Danish dairy cows. Prev Vet Med 2002; 54: 179-190.
2. Cohen ND, Mundy GD, Peloso JG, Carey VJ, Amend NK. Results of physical inspection before races and race-related characteristics and their association with musculoskeletal injuries in Thoroughbreds during races. J Amer Vet Med Assoc 1999; 215: 654-661.
3. Prentice RL. Design issues in cohort studies. Stat Methods Med Res 1995; 4: 273-292.
4. Rothman KG, Greenland S. Modern Epidemiology. 2d ed. Philadelphia: Lippincott-Raven, 1998; pp 79-91.
5. Samet JM, Munoz A. eds. Cohort studies. Epidemiol Rev 1998; 20: 1-136.

6. Tinker MK, White NA, Lessard P, Thatcher CD, Pelzer KD, Davis B, Carmel DK. Prospective study of equine colic incidence and mortality. Equine Vet J 1997; 29: 448-53.
7. Thurmond MC, Hietala SK. Effect of congenitally acquired *Neospora caninum* infection on risk of abortion and subsequent abortions in dairy cattle. Amer Jour Vet Res 1997; 58:1381-1385.

9

CASE-CONTROL STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Describe the major design features of risk-based and rate-based case-control studies.
2. Identify hypotheses and population types that are consistent with risk-based case-control studies.
3. Identify hypotheses and population types that are consistent with rate-based case-control studies.
4. Differentiate between open and closed primary-base and secondary-base case-control studies.
5. Elaborate the principles to select and define the case series.
6. Implement the principle features to select controls in open and closed primary-base case-control studies.
7. Implement the principle features to select controls in open secondary-base case-control studies.
8. Design and implement a valid case-control study for studying a specific hypothesis.

9.1 INTRODUCTION

The essence of the case-control study design is to select a group of cases and a group of non-cases (*ie* controls), and compare the frequency of the exposure factor in the cases with that in the controls. The cases are the study subjects that have developed the disease or outcome of interest, whereas the controls have not developed the disease or outcome of interest at the time they are selected. It is important to stress that a case-control study is not a comparison between a set of cases and a set of 'healthy' individuals, but between a set of cases and a set of individuals whose exposure to the factor of interest reflects the exposure in the population from which the cases were drawn. An overview of key design issues is available elsewhere (Breslow and Day (1980); Rothman and Greenland (1998)). Although the study designs are described as though an individual animal is the unit of interest, the designs also apply to aggregates of individuals such as litters, pens, barns or herds.

9.2 THE STUDY BASE

The study base is the population from which the cases and controls are obtained. If the cases and controls come from a well-defined **target population** for which there is, or could be, an explicit listing of sampling units (*ie* potential study subjects), this population is denoted as the **primary study base**. If the cases and controls come from a referral clinic, laboratory or central registry, these sources, which are one step removed from the actual source population, are referred to as a **secondary study base**. Explicitly defining the target or source population can be difficult when using a secondary base but, in so far as is possible, the controls should be derived from the population that gave rise to the cases in a manner such that they reflect the distribution of exposure in that base. Often it is useful to identify the factors which would lead (or exclude) cases to (from) the secondary-base registry. For example, there could be a large number of animals in the source population that develop the disease of interest but which will not be entered in the secondary base because of the animals' lack of economic value, or the owners' attitudes towards secondary medical care (if the secondary base is a specialised or referral hospital). In such an instance, we would attempt to select controls from non-cases with other disease(s) that will likely have similar referral patterns to the cases.

9.2.1 Open versus closed study populations

Variations in the study design are necessary depending on whether one is conducting the study in an open or closed population. As noted in section 4.4.1, a **closed** population has no additions during the study period and few or no losses. Populations are more likely to be closed if the risk period for the outcome is of limited duration (*eg* as in bovine respiratory disease in feedlot calves). **Open** populations could have both additions and losses during the risk period and are more likely to arise when the risk period for the outcome of interest is long (*eg* a study of risk factors for lymphosarcoma in cattle, or a study of risk factors for stereotypy in horses). Sometimes it is possible to convert an open population to a closed population. For example, a study of risk factors for mastitis

in dairy cows over one calendar year would likely have to contend with new cows being added and original cows being lost part way through their lactation. However, if the hypothesis is to identify risk factors for first occurrence of mastitis in the initial 60 days of lactation, by following a defined group of cows for the full 60 days after they calve, we have created a closed population. Only cows that calve in the herd(s) and are followed for the full 60-day period are included in the study. Closed populations can support risk-based case-control designs; open populations require a rate-based design.

9.2.2 Nested case-control study designs

In describing the source population in a case-control study, the term **nested** usually implies that the entire source population from which the cases are drawn has been enumerated and followed such that the case series represents all of the cases, or a known fraction thereof, from this population. When this is true, by knowing the sampling fractions of cases and controls, we can estimate the frequency of disease by exposure status, a feature that is absent in almost all other types of case-control studies. However, whether or not the study is truly nested in an explicitly definable population, it is useful to think of all case-control designs in this context even if the source population is not explicitly listed (*eg* as in a secondary-base study, section 9.6.3).

9.2.3 Keeping the cases and controls comparable

Reducing the number of extraneous factors that can adversely affect the study, many of which are unknown, is always a good strategy. Both **exclusion** and **inclusion** criteria can be used for this purpose, and should apply to both cases and potential controls. In addition, as with cohort studies, there are three general approaches to preventing confounding by 'known factors'. The first is exclusion or **restricted sampling**. For example, if breed is a likely confounder, you might include only one breed in the study, the dominant one in the source population. Hence, there could not be any confounding by breed. What we would lose in this approach is the ability to generalise the results to other breeds or to assess interactions with the exposure across breeds.

Matching on known confounders is a second strategy frequently used to prevent confounding and, to a lesser extent, to increase efficiency (*ie* power of the study). Unfortunately, matching often does not work well for either of these objectives in case-control studies (section 13.6). However, if matching is used, then a conditional analysis of the data is required (section 16.14). Third, we can use **analytic control** as a strategy for the control of confounding. Here we measure the confounders and use multivariable techniques to prevent confounding. This is our preferred choice, often working in concert with restricted sampling (see Chapters 12 and 13 for more detail).

9.3 THE CASE SERIES

The key elements in selecting the case series include the definition of the disease (the required diagnostic criteria for the outcome), the source of the cases, and whether to include only incident or both incident and prevalent cases.

9.3.1 Case definition

The actual diagnostic criteria will vary depending on the outcome, but they should include specific, well-defined manifestational (*ie* clinical) signs and, when possible, documented diagnostic criteria so that they can be applied in a uniform manner. In some instances, it might be desirable to subdivide the case series into one or more subgroups based on 'obvious' differences in the disease manifestation, especially if the causes of the different forms of the disease might differ. We need to be careful in imposing detailed diagnostic criteria for the cases in the sense that the case series in the study could become increasingly different from the majority of cases of that disease if high levels of time commitment and money are required to complete the diagnostic work-up. Thus, a case series of autoimmune disease in dogs obtained from a referral hospital might differ from the majority of autoimmune cases seen in private practice. Nonetheless, there is merit in a set of very specific diagnostic criteria for the cases as preventing false positives will reduce any bias in the measure of association caused by lack of sensitivity in the detection of cases (section 12.6.5).

9.3.2 Source of cases

A major decision is whether the cases will all (or most) be from a defined population (a primary-base study), or if they will be obtained from a secondary source such as clinic or registry records (a secondary-base study). Sampling directly from the source population has the advantage that it avoids a number of potential selection biases, but it is more costly than using secondary data. The challenge is to obtain as complete coverage as possible with respect to case ascertainment. This design is moderately common in veterinary medicine because farms with good records allow virtually complete enumeration of animals and events (although one might have to choose between 'owner-diagnosed' and 'veterinary-diagnosed' cases). As noted, depending on the outcome, the study design might allow these populations to be considered as closed thus allowing risk-based analyses. In secondary-base studies, the challenge is to conceptualise the actual source population and design the study to obtain a valid sample of non-cases to serve as controls.

9.3.3 Incident versus prevalent cases

The issue of selecting incident versus prevalent cases seems fairly clear; there is virtually unanimous agreement that only incident cases be used for the study. There could be specific circumstances in which the inclusion of prevalent cases can be justified, but this would be the exception, not the rule. The problem with prevalent cases is that it is difficult to separate the factors that relate to 'getting' the disease, from the factors that relate to 'having' the disease (*ie* duration). Thus, a 'beneficial' factor that increases survival in affected animals could appear to be a risk factor for the disease if prevalent cases are included. Also, because we are uncertain about when a prevalent case began, it is more difficult to focus the search temporarily for causal factors than it is for incident cases.

9.3.4 Exposure and covariate assessment

When ascertaining exposure status and information on confounders, it is preferable to obtain the greatest accuracy possible, even if that leads to different data-collection processes between the cases and controls. Failing that, the process should have comparable accuracy in both groups. Usually this approach is implemented by using the same process for obtaining exposure and confounder data in cases and controls and, where possible, having the data collectors blinded to case status.

Many times the exposures that are studied are not permanent and can change over time. If a subject's exposure history changes during the study period, the case's exposure status should be that which existed at the time of event occurrence. For controls, their exposure status at the time of their selection as controls is required.

9.4 Principles of control selection

The selection of appropriate controls is often one of the most difficult aspects of a case-control design. The key guideline for valid control selection is that they should be representative of the exposure experience in the source population. Controls are subjects that would have been cases if the outcome had occurred. Hence, the more explicitly the source population can be defined, the easier it is to design a valid method for selection of controls (Wacholder et al, 1992a,b).

The major principles are:

- Controls should come from the same study base (population) as the cases.
- Controls should be representative of the source population with respect to exposure.
- In open populations, controls should mirror the exposure time of the non-case subgroup in the population.
- The time during which a non-case is eligible for selection as a control is the time period in which it is also eligible to become a case if the disease should occur.

The implementation of these principles depends on the study design, so we shall begin our discussion with the traditional risk-based design.

9.5 Selecting controls in risk-based designs

The traditional approach to case-control studies in veterinary medicine has been a risk-based (cumulative incidence) design. In this approach, the controls are selected from among those animals that did not become cases up to the end of the risk period. An individual can be selected as a control only once. This design is appropriate if the population is closed and the risk period for the outcome in an individual has ended before subject selection begins. It fits situations such as outbreaks from infectious or

toxic agents where the risk period for the disease is short and essentially all cases that will arise from that exposure will have occurred within the defined study period (*eg* a point-source food-borne outbreak, or bovine respiratory disease occurrence post-arrival in a feedlot – see Example 9.1). Because the risk period has (for practical purposes) ended, the study cases represent virtually all of the cases that would arise from the defined exposure even if the study period were extended. It should, however, be noted that if the population is actually open, the use of this risk-based sampling strategy can lead to significant bias.

Example 9.1 Prospective risk-based case-control study

This study investigated associations of viral and mycoplasmal antibody titres with respiratory disease in calves in Ontario feedlots (Martin et al, 1999). Blood samples were taken from calves in 32 different groups on arrival at the feedlot and again 28-35 days after arrival. The high-risk period for bovine respiratory disease (BRD) in feedlot calves usually lasts less than four weeks and on average about 30% of calves develop BRD. Because all calves were observed for the full risk period, this study population was closed and a risk-based design with controls being those calves not developing BRD within 28 days of arrival was appropriate. A feature of the design is that, although all calves were bled at both times (arrival and 28-35 days later), the researchers only determined serological titres on the cases and an equal number of controls, thus reducing the number of serological analyses by 20-40% and converting the study from a single-cohort study to a prospective case-control study in a defined population.

9.5.1 Sampling issues in risk-based studies

The closed-source population can be categorised with respect to exposure and outcome as shown below (upper-case letters denote the population, lower case the sample):

	Exposed	Non-exposed	Total
Cases	A_1	A_0	M_1
Non-cases	B_1	B_0	M_0
Total	N_1	N_0	N

The cases are those that arose during the study period, whereas the controls are those that remained free of the outcome during the study period. The controls should be selected such that there is an equal sampling fraction of exposed and non-exposed controls (*ie* sampling is independent of exposure status).

Usually, all or most of the cases (M_1) are included in the study. There are B_1 exposed non-cases and B_0 non-exposed non-cases in the source population from which we select our study control subjects b_1 and b_0. We want to select the controls such that the sampling fractions (*sf*) in these two groups of non-cases are equal, *ie*:

number of exposed controls in the sample is $b_1 = sf(B_1)$, and
number of non-exposed controls in the sample is $b_0 = sf(B_0)$.

In a primary-base study, an equal sampling fraction would be obtained by random selection from the non-case population (*ie* from the group that remains free of the disease at the end of the study period). In a secondary-base study, it could be achieved by selecting controls randomly from the other non-cases in the registry. There is one caveat in selecting controls in a secondary-base study, however – in order to keep the sampling fractions equal, one should sample from sets of diagnostic categories of non-cases that are not associated with the exposure(s) of interest. As we point out subsequently, most secondary study bases are open and hence a rate-based design should be used to select controls.

If censoring of study subjects is not independent of exposure, a rate-based sampling approach (see section 9.6), coupled with the usual unmatched risk (odds ratio) calculations, will provide a more consistent estimator of the risk ratio than sampling from the non-case group at the end of the risk period. Non-independent censoring might, for example, be common in studies of risk factors for diseases in many food-animal species where 'removal' of study subjects is under the owner's control.

In risk-based studies, the measure of association is the odds ratio (*OR*) which is a valid measure of association in its own right, and also estimates the ratio of risks (*RR*) if the outcome is relatively infrequent (*eg* <5%) in the source population (see Chapter 6).

9.6 SELECTING CONTROLS IN RATE-BASED DESIGNS

Because the populations we study are often open, the case-control designs for these populations should use a rate-based approach (**incidence rate** (*I*)) which seeks to ensure that the time at risk is taken into account when the controls are selected.

9.6.1 Sampling issues in rate-based studies

We can visualise the classification of the open-source population with respect to the number of cases and the time at risk in each of the exposure levels in the population as shown below (upper-case letters denote the population, lower case the sample):

	Exposed	Non-exposed	Total
Cases	A_1	A_0	M_1
Animal-time at risk	T_1	T_0	T

To help understand incidence rate designs, it is useful to think about how the two key frequencies would be measured, and what animals would be included in a cohort study of the population. Recall that, in a cohort study, if we wanted to study the relationship of exposure to the rate of outcome, the two rates of interest would be:

$$I_1 = A_1/T_1 \quad \text{and} \quad I_0 = A_0/T_0$$

where A represents the number of incident cases and T represents the animal-time at risk in each exposure group. The drawback to the cohort approach is that all animals

in the study population must be followed. In a case-control study, the control series is used to reflect the animal-time exposure experience without the full enumeration of the population or the time at risk. Thus, the cases are those subjects that experience the outcome in the hypothetical cohort study. The controls are selected from non-cases such that the denominators T_1 and T_0 can be estimated validly. Recall that we estimated B_1 and B_0 in the risk-based design.

The main goal of a rate-based study is to select controls with an exposure distribution that matches that in the source population. This information allows us to estimate the ratio of the rates in exposed and non-exposed animals without knowing T_1 and T_0. Hence, we select controls such that the sampling rate (*sr*) is equal in the exposed and non-exposed populations. That is, the ratio of the number of exposed controls (b_1) in our sample divided by animal-exposure-time equals the number of non-exposed controls (b_0) in our sample divided by the non-exposed animal-time. Thus, in our sample,

$$\frac{b_1}{T_1} \approx \frac{b_0}{T_0} \quad \text{or} \quad \frac{b_1}{b_0} \approx \frac{T_1}{T_0} \qquad \textit{Eq 9.1}$$

If we achieve the equality in Eq 9.1, then:

$$\frac{a_1/b_1}{a_0/b_0} \approx \frac{A_1/T_1}{A_0/T_0} \qquad \textit{Eq 9.2}$$

That is, the ratio of the exposed cases to exposed controls divided by the ratio of the non-exposed cases to non-exposed controls in the study population estimates the ratio of the incidence rates (*IR*) in exposed and non-exposed individuals in the source population. This ratio can also be viewed as the odds of exposure in the cases compared with the odds of exposure in the controls and is called the cross-product ratio or odds ratio (*OR*). In this design, the *OR* estimates the *IR* and no assumption about rarity of outcome is necessary for a valid estimate. The penalty for the efficiency of this process is that the statistical precision is lower than if the hypothetical cohort study had been done.

9.6.2 Sampling primary-base open-population controls

The easiest way to ensure valid selection of controls in this instance is to randomly select controls from the source population. In the **unmatched** selection procedure, the probability of selecting each control should be proportional to the time at risk, as it is the amount of time at risk in the exposed and non-exposed groups that we should mirror in our controls. If time-at-risk data are available, controls can be selected at the end of the study period using the time at risk as the basis (probability) for their selection. The time at risk would be known in well-defined populations such as herds or flocks with complete records for all animals. For example, in a case-control study of risk factors for bovine leukosis, if herds on milk-recording systems were used for the study, it would be possible to obtain time-at-risk data for each cow and hence, sample accordingly.

If the time at risk is not known, controls can be selected at defined points throughout the study period from the **risk set** (those non-cases eligible to become cases at that point in

time). The number of controls per period can vary and need not have a constant ratio to cases. If the exposure and covariate characteristics of the population don't change over the study period (*ie* they are stable), the sample *OR* estimates the *IR*.

Another method of obtaining controls is by selecting a specified number of non-cases from the risk set **matched**, time-wise, to the occurrence of each case. This is called **incidence density sampling**. The controls are randomly selected at the time the case arises from those non-cases eligible to become cases at that point in time. The number of controls per case can vary and need not have a constant ratio. If the level of exposure might be related to calendar time, then the matched design needs to be analysed as such, otherwise the data can be analysed by unmatched procedures. In this design, the *OR* estimates the *IR* whether or not the population is stable.

Regardless of the selection process, animals initially identified as controls can subsequently become cases. Their data are treated as independent in the analysis (*ie* if this happens, their control data reflect their exposure and covariate status at the time they were selected as a control, and when they are a case, their exposure and covariate status reflect the situation at that time). Because we are sampling directly from the source population, there should be no exclusions of potential controls because of exposure status.

9.6.3 Sampling secondary-base population controls

When a clinic, laboratory or other registry is the source of the cases, we have a secondary-base study. In such studies, selecting **non-cases** from the same registry is preferable to obtaining them from different sources. The basic tenet is that the controls should reflect the exposure patterns in the population of potential cases that would have entered that registry had they developed the disease or outcome of interest. The problem is knowing whether having the exposure of interest would lead non-cases to the registry and hence, the controls would have an excess of exposure (*eg* exposure increases the chance of being in the registry by increasing the risk of being admitted for non-case subjects). To avoid bias with respect to the distribution of exposure in controls, exposure should not be related to admission of non-cases to the registry; hence, we should select controls from diagnostic categories that are not associated with exposure. For the same reason, subjects with disease events that are intermediate between the exposure and the outcome are not eligible as controls.

Diagnostic category exclusions for controls should only relate to admissions during the study period time frame, and not to previous admissions (if the individual was admitted for a condition related to exposure before the study period, that individual should still be eligible as a control in the study period provided its reason for entry into the registry at this time is independent of exposure). Some recommend that controls should only be selected from those diagnostic categories for which data exist to show that they are not related to the exposure of interest. Most researchers have tended to use less stringent exclusion criteria for independence and select controls from diagnostic categories that are not known to be associated with exposure. Regardless, it is usually preferable to select controls from a variety of non-case diagnostic categories.

Often we obtain the controls randomly from all the non-case admissions listed up to the end of the study period, having excluded those non-case categories that are associated with the specified exposure(s). This might seem like a risk-based sampling strategy but the non-case subjects can be listed in the registry numerous times because of admission for the same, or different, non-case diseases. This essentially reflects their exposure time at risk. It is also possible to select controls randomly from the non-cases in the registry at regular intervals throughout the study period. Thus, if a three-year study period was used and 300 controls were to be selected, 8 or 9 individuals would be selected each month, from all the non-cases in the registry up to that point. The period of time in which an animal is eligible to be a control should be the same as that in which it is eligible to be a case should the event occur; hence, controls can become cases in both selection processes. If the outcome rate or exposure level in the source population(s) varies with calendar time, then one should stratify on time in the analysis to prevent bias. If the population is stable, the sample *OR* estimates the *IR*.

Alternatively, as in primary-base studies, one might **match** for 'time at risk' by selecting a specified number of controls listed in the registry after a case occurrence. If the exposure level is constant over the study period, an unmatched analysis can be performed (if there is no matching for other reasons), otherwise a matched analysis should be pursued. Bear in mind that the process of selecting controls in open populations means that the same animal can be selected more than once – in this instance depending on how many times it was admitted to the registry. If only first incident cases are included in the study, these animals cannot be selected as controls after they have developed the disease of interest.

Example 9.2 shows a secondary-base case-control study.

9.7 Other sources of controls

The following two procedures can be used in either primary- or secondary-base studies.

9.7.1 Neighbourhood controls

When random sampling is not possible in a primary-base study, choosing neighbours of cases might suffice. This means that a matched analysis should be conducted if neighbourhood is related to exposure. In a secondary-base study, neighbours can also be selected as controls but their suitability needs to be established according to the study context. Selecting neighbours could introduce a bias and might cause overmatching in some studies. For example, in a primary-base study of factors related to *Salmonella* spp in bulk milk tanks on dairy farms, the closest farm was used as a control. However, often these farms were owned by relatives of the case farm owner and many times farm implements and food items were shared between case and non-case farms. Thus, overmatching was likely present (West et al, 1988). Non-cases housed next to cases within a barn might be suitable, spatially matched, controls in some studies.

Example 9.2 Secondary-base case-control study

This study concerns environmental tobacco smoke (ETS) as a risk factor for malignant lymphoma in pet cats (Bertone et al, 2002). Feline leukemia (FL) could serve as a model for non-Hodgkin's lymphoma in humans. The study population was obtained from a large veterinary teaching hospital in Massachusetts and consisted of 80 FL positive cats and 114 controls drawn from cats with renal disease. Cats with renal disease were selected as likely being more representative of the source population for the cases and because there was no known association of renal disease with ETS. ETS exposure history and covariate information were obtained by a mailed questionnaire and related to the two-year period before the diagnosis of the FL or renal disease. Approximately equal response rates (65%) were achieved in the case and control groups. ETS exposure classification included: ever versus never exposed, years of exposure, number of smokers in the household, number of cigarettes smoked per day and cumulative variables such as those that resulted from years of exposure*average number of cigarettes smoked per day.

In this secondary-base study, the implied source population was deemed to be open. Non-cases were obtained from only one diagnostic category (not a generally recommended practice) although the authors of this study defended this due to a known lack of association with ETS, plus the fact that cats with renal failure would have undergone extensive laboratory tests – similar to the extent of testing in the cases. All cats with renal disease were selected for the study – likely none of them had been admitted for FL but, in theory, they could have been.

9.7.2 Random-digit dialing

This approach can be used to obtain controls if human subjects are being studied. For example, the telephone number of controls might be matched to cases by area code and the first three digits. There are numerous hidden problems with this approach including time of calling, business versus home phone *etc.* If used, then the 'matching' should be accounted for in the analysis if there is any chance that it is related to the exposure.

9.8 The issue of 'representativeness'

Is it important that the cases be representative of all the cases and that the controls be representative of all the non-cases? No! The cases and controls can be restricted in any logical manner the investigator chooses. This might restrict extrapolation of results but will not affect validity. However, the restriction defines the source population and it is this source population that the controls should be representative of. This might be more understandable if we recall that cohort studies can be conducted in subgroups of the population that have a non-representative attribute or exposure status.

9.9 More than one control group

Some have attempted to balance a perceived bias with a specific control group by using more than one control group. However, if this is done, it needs to be very clearly defined as to what biases are likely to be present in each control group and how one will interpret the results – especially if they differ dramatically from one control group to another. The use of more than one control group also adds complexity to the analyses. If we choose more than one control group, the different control groups should be compared with respect to exposure. If they do not differ significantly, it ensures that, if a bias is present, the control groups have the same net bias. If they differ however, we often are not sure which is the correct group to use. The general experience is that the value of more than one control group is very limited.

9.10 More than one control per case

There is nothing magical about having just one control per case. In fact, if the information on the covariates and exposure is already recorded (*ie* in a sense, free), one might use all of the qualifying non-cases in the registry as controls to avoid issues of sampling. In addition, when the number of cases is small, the precision of estimates can be improved by selecting more than one control per case. There are formal approaches for deciding on the optimal number, but usually the benefit of increasing the number of controls per case is small; often 3-4 controls per case is the practical maximum (Breslow and Day, 1987).

9.11 Analysis of case-control data

The analysis of risk-based and rate-based case-control sampling designs proceeds in a similar manner. For displaying the data, we will assume that we observe a_1 exposed cases and b_1 exposed controls, and a_0 non-exposed cases and b_0 non-exposed controls. There is a total of m_1 cases and m_0 controls. Remember that we cannot estimate disease frequency by exposure level directly because the m_1/m_0 ratio was fixed by sampling design. In a 2X2 table the format is:

	Exposed	Non-exposed	Total
Cases	a_1	a_0	m_1
Controls	b_1	b_0	m_0

Chapter 6 outlines the analysis of these data including hypothesis-testing, estimating the odds ratio, and developing confidence intervals for the odds ratio.

If matching was used to select controls, then a conditional analysis should be performed – see section 16.14 for a discussion and examples.

SELECTED REFERENCES/SUGGESTED READING

1. Bertone ER, Snyder LA, Moore AS. Environmental tobacco smoke and risk of malignant lymphoma in pet cats. Am J Epidemiol 2002; 156: 268-273.
2. Breslow NE, Day NE. Statistical methods in cancer research Volume I The analysis of case-control studies. IARC Lyon France, 1980.
3. Breslow NE, Day NE. Statistical methods in cancer research Vol II The design and analysis of cohort studies Chapter 7. IARC Lyon France, 1987.
4. Martin SW, Nagy E, Armstrong D, Rosendal S. The associations of viral and mycoplasmal antibody titres with respiratory disease and weight gain in feedlot calves. Can J Vet Res 1999; 40: 560-570.
5. Rothman K, Greenland S. Modern epidemiology 2d ed Philadelphia: Lippincott-Raven, 1998; pp 93-114.
6. Wacholder S. Design issues in case-control studies. Stat Methods Med Res 1995; 4: 293-309.
7. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. Am J Epidemiol 1992a; 135: 1029-1041.
8. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. Am J Epidemiol 1992b; 135: 1042-1050.
9. West AM, Martin SW, McEwen SA, Clarke RC, Tamblyn SE. Factors associated with the presence of *Salmonella* spp in dairy farm families in Southwestern Ontario. Can Jour Pub Hlth 1988; 79: 119-123.

10

HYBRID STUDY DESIGNS

OBJECTIVES

After reading this chapter, you should be able to:

1. Describe the key features of each of three hybrid study designs (case-cohort, case-crossover, case only).

2. Be able to identify exposures, outcomes and contexts for which these designs are appropriate.

3. Identify contexts in which a two-stage study is appropriate and design the basic sampling strategy.

10.1 Introduction

In this chapter we describe three variants of observational study design and a two-stage design that can be used in conjunction with traditional cross-sectional, cohort and case-control studies. Each design has its own unique advantages and disadvantages and, although still relatively infrequently used, researchers should be aware of the potential of these study designs. The case-cohort design incorporates the strengths of the cohort approach with the efficiency of a case-control design. The case-crossover is an elaboration on the crossover experimental design that allows the researcher to use only cases in the study. Similarly, the case-only study design allows for inferences to be made from studies in which only data from cases are available. Two-stage designs are useful as validation studies and also to enhance the cost-effectiveness of a traditional study design.

10.2 Case-cohort studies

10.2.1 Basis

This study design requires the listing of all subjects in a defined cohort. Then a subsample of the fully enumerated cohort is taken to serve as the comparison or control group. Initial exposure and covariate information is obtained on this subsample as per a single-cohort study. Cases (*ie* those developing the diseases of interest) are derived from this subsample and from the full cohort.

A major advantage of the case-cohort approach is that the one control group can provide the basis of comparison for a series of outcomes, thus allowing the investigation of associations among more than one disease and a defined exposure (as in a regular cohort study), but without having to follow the entire population at risk.

10.2.2 Design issues

If the original cohort is closed (section 8.5), then a risk-based design, which is particularly suited to studying permanent risk factors, can be used. In this design, the subsample is selected from the full cohort at the start of the study using random sampling and the subjects in the subsample that did not become cases during the study period serve as the control series. Information about covariates and exposure status is obtained from cases arising outside of the subsample at the time they become cases. In planning this study design, because a proportion of the subjects in the subsample will become cases, the number initially sampled should be adjusted upward to compensate for this. For valid inferences, if losses to follow-up are present, one must assume that the reasons for loss, or the occurrence of competing risks, are not related to the risk of developing the outcome(s) of interest.

If the original cohort is open, or if exposure status can change during the study period, the exposure status of the case is the exposure category that the individual is in when the outcome occurred. In this design, all or a portion of the original subsample that had not developed the disease by the time each case occurred serve as the 'control' for that

case, and their exposure status is obtained at this time. Unfortunately, when the timing of control selection is matched to the time the case occurred, only one outcome can be studied.

10.2.3 Analysis

At the end of the study period, there will be records of the number of cases arising from the subcohort, the number of cases arising outside the subsample and the remaining number of non-cases in the subsample – as in a risk-based cohort study. If a risk-based design is appropriate, one can combine (*ie* add) the two types of case together, and the data can be analysed in a 2X2 table using a case-control format with an odds ratio (*OR*) as the measure of association. If direct estimates of risk are required, then the two types of case need to be differentiated (Rothman and Greenland, 1998). The analysis is more complex if matching of the exposure of cases and controls to the time the case occurred is used (Thomas, 1998). The data can also be analysed using survival methods (Barlow et al, 1999) and programs for these analyses are available in a number of common commercial packages. In Example 10.1, we describe a case-cohort study.

Example 10.1 Case-cohort study

This is an example of a case-cohort study that was used to investigate the role of agricultural pesticide applications on fetal death in humans (Bell et al, 2001). The cohort was the total recorded number of births in 10 counties in California and all cases of fetal death came from this source. A subcohort consisting of a stratified random sample of non-cases was obtained from this cohort. Exposure was measured by noting applications of pesticide within a 1-sq-mile (narrow definition) or a 9-sq-mile area (broad definition) of the mothers' homes. The time of pregnancy and the time of fetal death were also noted.

Now, you might say that this looks like a risk-based case-control study – and so far it is. Although they did not sample in the prescribed case-cohort approach, the authors of the study used the known sampling fraction of controls (sf=0.01) to estimate and identify the likely number of cases that would have arisen in the subcohort had it been sampled in that manner. This allowed the authors to adjust for gestational length and introduce pesticide exposure as a time-varying exposure. The risk-based design would not support these features.

A multivariable proportional-hazard model was used to analyse the time-dependent pesticide exposures and control for the covariates. The analysis contrasted the exposure experience of the case with the exposure experience of the non-case infants, yet unborn, in the population at the time.

No examples of case-cohort studies were found in the veterinary literature.

10.2.4 Suggested applications

If you are interested in studying a number of infrequent outcomes from a specific exposure, think about using the case-cohort design. It has the advantage of the case-control design for studying rare events and the advantage of the cohort design for studying multiple outcomes. One current example from human medicine is the study of adverse post-vaccinal reactions.

10.3 Case-crossover studies

10.3.1 Basis

This is the observational study analogue to the crossover experiment. It is suitable for the situation where the exposure is abrupt (short-lived and well-defined) and the effect is almost immediate (*ie* the outcome will happen temporally close to the exposure, if the exposure was the cause of the outcome). For validity, the design needs to meet the same assumptions about lag effects (none or time limited) and duration of disease (short duration) as in crossover experiments or crossover clinical trials.

The case-crossover design alleviates many of the problems associated with choosing controls in a case-control study. The exposure status of the case at the time of the event occurrence is compared with the exposure status of the same individual at other times. Only subjects that develop the outcome need to be followed; hence, all time-invariant host-related confounders are controlled by this design. This design is only applicable to situations where the exposure status of individuals can change over time, and where the exposure will produce its effects in a short time period. Effect estimates are based on comparing exposure levels just prior to case development with exposure levels at other (control) times (Navidi and Weinhandl, 2002).

10.3.2 Design issues

For validity, we must ensure that the average exposure level has not changed over time and that the distribution of non-host confounders is also stable over time. This has been the major focus of controversy over the use of this design. In the initial protocols, the control times were always earlier than the case times for purposes of obtaining exposure levels. This is an acceptable approach if the first outcome might affect subsequent exposures (*eg* if one is studying the impact of a training schedule on an outcome such as a leg injury). However, it is subject to bias from temporal changes in the level of exposure. Hence, bidirectional designs were used later especially when environmental exposures were studied. In this revised protocol, pre-specified control times are selected both before and after the case-event time. Later, the design for control-period selection was kept symmetrical around the failure time by selecting a control both before and after the case occurrence time (usually equally spaced) in the hope that, if exposure levels were changing over time, the higher and lower exposure values at these times would cancel each other out. These periods could also be matched to the same day of the week

as the case occurrence to avoid any 'day-of-the-week' bias. The current recommended design which gives the best protection against bias from a trend in exposure follows.

We can suppose that a case might occur at any time (t_k) in the period from the first day of follow-up (k=1) to the last study day (k=N). To identify the control period for each case we:

1. Choose a short lag time, (L); can be a multiple of 7 (*eg* 21 days) to adjust for day-of-the-week effects.
2. Let t_k be the failure time for the j^{th} case.
3. If $t_k \leq L$, choose t_{k+L} as the control time.
4. If $t_k > (N\text{-}L)$, choose $t_{k\text{-}L}$ as the control time.
5. For all other values of t_k, randomly choose half of the control periods from before the case time ($t_{k\text{-}L}$) and half from after (t_{k+L}).
6. Repeat steps 2 to 5 for each case.

10.3.3 Analysis

Data should be analysed as a matched case-control study. Equivalently, one could view the sampling times as producing a set of control exposures for each failure time (case occurrence) or as an independent count of cases out of a total of the number of case-exposure and control-exposure periods. Thus, the case count on each sampled day could be modelled as a Poisson random variable whose mean is a function of the exposure level on that day (Navidi and Weinhandl, 2002). Example 10.2 describes two case-crossover studies.

10.3.4 Suggested applications

The case-crossover design should prove useful in situations where the exposure is expected to produce its effect very shortly after exposure occurs. Examples of the use of this design include studying death risk after vigorous exercise, hospital admission risk following exposure to high pollution levels, the role of cellular-phone usage in car accidents, and racing as a cause of injury in horse-racing. The exposure during the period preceding the case is compared with the exposure at the designated 'control' time(s). Decisions about appropriate lag time *etc* are decided on a context, and disease-specific basis (see Example 10.2).

10.4 Case-only studies

In some instances, such as genetic studies, the exposure status of the controls can be predicted without having an explicit control group: the distribution of exposure in the controls is derived from theoretical grounds (*eg* blood-type distribution). Underlying the design, which is highly efficient relative to case-control designs, lies a strong assumption about independence between the gene frequency and other environmental factors. A recent article discusses the limitations of the study design if independence is not, in fact, present (Albert et al, 2001). The interesting feature is that the assessment

Example 10.2 Case-crossover study

This is a study of intensive racing or training as risk factors for musculoskeletal injury in horses (Estberg et al, 1998). The exposure factor was intensive racing or training, and the outcome was catastrophic musculoskeletal injury (CMI). The cases came from diagnostic laboratory records on all horses dying on 14 racetracks in California during a period in 1991 and 1992.

The exercise and racing histories of these horses were obtained from a computerised commercial information service. Exposure (intensive training) was determined by assessing distance, speed and frequency of training over sliding 60-day periods. If the level of training exceeded a defined threshold, the horses were considered to be in an 'at-risk' period for the 30 days after the end of the period evaluated. If a CMI occurred during one of these 'at-risk' periods, it was considered to have occurred in an exposed (*E+*) horse. All other CMI were considered to have occurred in non-exposed (*E-*) horses. Data were analysed using the Mantel-Haenszel pooled estimator for a common incidence rate ratio. This procedure allows control of categorical covariates.

Similarly, Carrier et al (1998) investigated associations between long periods without high-speed workouts and the risk of humeral or pelvic fracture in Thoroughbred racehorses. Their exposure was a lay-up period of two or more months, and the at-risk (hazard) period was either 10 or 21 days after returning to racing. A fracture in these periods was deemed to be exposed to the risk factor, lay-up, whereas a fracture at other times was deemed non-exposed. See Example 8.5 for a cohort study approach to racehorse injuries.

of the interaction term in the usual logistic regression, assuming data from controls are present, can be performed by regressing the probability of the genetic abnormality on the environmental exposure using the data from cases only. It is this aspect that allows for the gain in efficiency because no controls are necessary. Despite its theoretical advantages, no examples of the use of this study design were found in the veterinary literature, and no immediate applications are evident.

10.5 Two-stage sampling designs

A two-stage sampling design can be applied to the traditional cohort, case-control or cross-sectional study designs. Information on the exposure and outcome of concern is gathered on an appropriate number of subjects (*ie* based on sample-size estimates). Then, a sample of the study subjects is selected for a second-stage study in which information on covariates is collected. This approach is very efficient if the cost for obtaining the data on covariates is expensive. The design also fits the situation where a valid measure of the exposure of interest is very expensive, but an inexpensive surrogate measure is available. The surrogate measure is applied to all study subjects, then a more detailed work-up is performed on a subsample of the study subjects to more accurately determine the true exposure status. The approach can also be used to obtain data on variables for which there are numerous missing values. Instead of

assuming that the data are missing at random, the study subjects with missing data can be the subject of a second-stage data collection effort. As discussed in section 12.9, the two-stage approach is the basis of validation substudies.

A key question is: what sample size should be used for the second stage? There are a number of approaches. In cohort studies, we can take a fixed number of exposed and non-exposed subjects. In a case-control study, we could take a fixed number of cases and controls. However, for efficiency, it is better to stratify on the four exposure and disease categories (present in a 2X2 table) and take an equal number of subjects from each of the four categories. This might involve taking all of the subjects in certain exposure-disease categories and a sample of subjects in others. More elaborate sampling regimes are discussed by Schaubel et al (1997) and Reilly (1996). Cain and Breslow (1988) developed the methodology to analyse two-stage data using logistic regression. One obtains the crude measure of association from the first-stage data, and then adjusts this based on the sampling fractions used in the second stage in a similar manner to the approach used to correct for selection bias using sampling fractions (see section 12.4.1). Here the sampling fractions relate to the ratio between the number of subjects in the second- and first-stage samples. The approach to obtaining correct variance estimates is somewhat more complex, but relatively simple to implement if the data are all dichotomous. Techniques for use if the predictors are continuous are available (Schaubel et al, 1997).

Thus, two-stage sampling designs are an efficient way to study exposure-disease associations and correct for confounders, or adjust for information bias, without measuring all variables on all study subjects.

SELECTED REFERENCES/SUGGESTED READING

1. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for indentifying gene-environment interactions. Am J Epidemiol 2001; 15: 687-693.
2. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. J Clin Epidemiol 1999; 51: 1165-1172.
3. Bell EM, Hertz-Picciotto I, Beaumont JJ. Case-cohort analysis of agricultural pesticide applications near maternal residence and selected causes of fetal death. Am J Epidemiol 2001; 154: 702-710.
4. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage samples. Am J Epidemiol 1988; 128: 1198-1206.
5. Carrier TK, Estberg L, Stover SM, Gardner IA, Johnson BJ, Read DH, Ardans AA. Association between long periods without high-speed workouts and risk of complete humeral or pelvic fracture in thoroughbred racing horses: 54 cases. J Am Vet Med Assoc 1998; 212: 1582-1587.
6. Estberg L, Gardner I, Stover SM, Johnson BJ. A case-crossover study of intensive racing and training schedules and risk of catastrophic musculoskeletal injury and lay-up in California thoroughbred racehorses. Prev Vet Med 1998; 33: 159-170.

7. Navidi W, Weinhandl E. Risk set sampling for case-crossover designs. Epidemiology 2002; 13: 100-105.
8. Reilly M. Optimal sampling strategies for two-stage studies. Am J Epidemiol 1996; 143: 92-100.
9. Rothman KJ, Greenland S. Modern epidemiology 2d ed. Philadelphia: Lippincott-Raven. 1998; pp 246-247, 277-278.
10. Schaubel D, Hanley J, Collet JP, Bolvin JF, Sharpe C, Morrison HI, Mao Y. Two-stage sampling for etiologic studies. Am J Epidemiol 1997; 146: 450-458.
11. Thomas D. New techniques for the analysis of cohort studies. Epidemiol Rev 1998; 20: 122-134.

11

CONTROLLED TRIALS

OBJECTIVES

After reading this chapter, you should be able to:

1. Design a controlled trial which will produce a valid evaluation of an intervention being investigated, paying special attention to:
   a. the statement of objectives of the trial
   b. the definition of the study subjects
   c. the allocation of subjects to the interventions
   d. the identification and definition of appropriate outcome variables
   e. ethical considerations in the design and implementation of the trial.

2. Conduct a controlled trial efficiently, while paying special attention to:
   a. masking as a procedure to reduce bias
   b. following all intervention groups adequately and equally
   c. development of appropriate data-collection methods and instruments
   d. proper assessment of the outcomes being measured
   e. correct analysis and interpretation of the results.

11.1 Introduction

A controlled trial is a planned experiment carried out on subjects maintained in their normal (*ie* usual) environment. Particular care must be taken in the design and execution of these studies for two reasons. First, they often involve client-owned animals. Second, the size and scope of many of these studies make it very difficult to replicate them for the purpose of validating the findings.

Controlled trials can be carried out to evaluate therapeutic or prophylactic products, diagnostic procedures and animal-health programmes. Most trials are conducted to assess one specific intervention and, indeed, this is their forte. The outcomes might include specific health parameters (*eg* clinical disease) or measures of productivity, performance or longevity. The groups can be based on assigning individual animals, herds or other groups to the interventions being compared.

Note The term **clinical trial** is often used synonymously for **controlled trial**. However, some authors restrict its use to trials of therapeutic products and/or trials carried out in a clinical setting. We will use the term controlled trials to refer to trials that might evaluate a wide range of products or procedures, and which might be carried out in a wide range of settings, including herds and community-based trials. Because controlled trials can be used to investigate a wide range of products/programmes, we will refer to the factor being investigated (*eg* treatment) as the **intervention**, and to the effect of interest as the **outcome**. Animals, or groups of animals participating in the trial will be referred to as **subjects** (regardless of whether they are individual animals, herds or other populations of animals), and their owners as **participants**.

Controlled trials are, by far, the best way for evaluating animal-health interventions because they allow much better control of potential confounders than observational studies, as well as reducing bias due to selection and misinformation.

> ... the randomised controlled trial is at present the unchallenged source of the highest standard of evidence used to guide clinical decision-making.
>
> (Lavori and Kelsey, 2002)

> There is simply no serious scientific alternative to the generation of large-scale randomised evidence.
>
> (Peto et al quoted in Green (2002)).

In the absence of evidence as to the efficacy and safety of animal-health products and procedures derived from controlled trials, practitioners are left in the unenviable position of making decisions about their use based on extrapolation of data from studies carried out under artificial (laboratory) conditions or on their own limited and uncontrolled experience.

11.1.1 Phases of clinical research

While controlled trials are valuable for assessing a wide range of factors affecting

animal health and productivity (*eg* management practices, nutrition, environmental changes), one of their most common uses is to evaluate pharmacological products (therapeutic and preventive). Consequently, a brief review of the phases of research into the development and evaluation of these products is warranted.

Clinical pharmaceutical research can be divided into four phases. Phase I trials (sometimes referred to as formulation trials) are studies carried out in healthy animals primarily to evaluate safety of the drug (*eg* to determine safe dosage ranges, to identify adverse reactions *etc*).

Phase II trials are the first evaluation of the drug in a small number of animals from the target population (*eg* sick animals). They are used to document the activity of the drug. These studies might involve before/after comparisons because there is often no control group.

Phase III trials are large-scale experimental studies to determine the efficacy of a drug in a normal clinical population, to monitor side effects and to compare the drug with other available treatments. These studies should be based on randomised controlled trials. While generally required to be carried out before the registration of products for human use, they are not necessarily required for registration of animal-health products in all countries.

Studies carried out for the purpose of registration of animal products (Phase III trials) need to be carried out according to **good clinical practice** (GCP) standards. GCP is a standard for the design, conduct, monitoring, recording, auditing, analysis and reporting of clinical studies. A set of these standards, developed under the International Cooperation on Harmonization of Technical Requirements for Registration of Veterinary Medicinal Products can be found at
http://www.aphis.usda.gov/vs/cvb/vich/goodclinicalpractice6-2000.pdf

Phase IV trials are post-registration trials designed to evaluate the most effective way of using a product. They should also be carried out as randomised controlled trials, although they require less documentation than studies used in the product registration process. In the absence of randomised controlled trials carried out prior to registration, they provide the most reliable information about the efficacy of a product in the content of everyday real-world activities.

11.1.2 Key design elements

An important feature in the design of a controlled trial is the development of a detailed study protocol which covers all elements of the study design and execution. Important elements to be considered include:

- stating the objectives
- defining the study population
- allocation of subjects
- specifying the intervention
- masking (blinding)

- follow-up and compliance
- specifying and measuring the outcome
- analysis of trial results
- ethical considerations.

Each of these will be considered below. General references for controlled-trial design include Meinert (1986), Piantadosi (1997) and an issue of Epidemiologic Reviews (Lavori and Kelsey, eds. 2002), from which a number of specific articles are referenced below. Three published studies which will be referred to throughout this chapter are introduced in Example 11.1.

Example 11.1 Controlled trial examples

Throughout the examples in is chapter, reference will be made to three recently published controlled trials to document how various aspects of trial design were implemented.

Eprinomectin trial

A randomised controlled trial was conducted to evaluate the effect of treatment at calving with the anthelminthic eprinomectin (Ivomec® Eprinex® - Merial Inc.) in dairy herds which had some exposure to pasture. Eprinomectin is a broad-spectrum endectocide that is registered for use in dairy cattle, with no milk-withdrawal period required. Cows from 28 herds in two sites (Prince Edward Island and Quebec, Canada) were randomly allocated to be treated at calving with eprinomectin or a visually identical placebo. The primary outcomes from the trial were milk production and reproductive performance during the first six months of lactation. Secondary outcomes included a variety of health parameters (Nødtvedt et al, 2002; Sanchez et al, 2002).

Teflubenzuron trial

A cluster randomised controlled trial was conducted to evaluate the effect of adding teflubenzuron to the feed of Atlantic salmon on sea-lice numbers on salmon in sea-cage sites. Teflubenzuron is a chitin-synthesis inhibitor that stops the sea lice from forming proper exoskeletons during moults. This was a Phase III trial conducted to good clinical practice standards with the results to be used as part of the registration application. Forty sea cages from three sites were pair-matched based on site, cage size and pre-treatment lice burden and then randomly allocated to be fed treated feed or non-medicated feed. The primary outcome was sea-lice burdens at one and two weeks post-treatment. A secondary outcome was weight gain over the same period (Campbell, 2002a).

Hoof-trimming trial

A randomised controlled trial of autumn hoof-trimming was carried out in 3,444 cows in 77 Swedish dairy herds over two winter periods. Interventions were trimming or no-trimming. Thus, no placebo intervention was possible. Hoof measurements and the presence/absence of lesions at the time of treatment (autumn) were recorded. The primary outcomes of the study were the incidence of lameness in the two groups over the winter period, and the presence/absence of lesions at a spring evaluation of all hooves. While all hooves on all four legs were evaluated and trimmed (or not) in both the fall and spring, the analysis focused on lesions in the hind-leg hooves.

11.2 STATING THE OBJECTIVES

The objectives of the trial must be stated clearly and succinctly. This statement should include reference to the intervention being investigated and the primary outcome to be measured. The importance of the latter component can be seen by considering a controlled trial of a vaccine for use in a food-producing species. The design of the trial will vary substantially depending on whether the goal of the vaccination programme is to:

- prevent infections in individuals
- prevent clinical disease or death in individuals
- prevent introduction of the infectious agent into a population (*eg* herd)
- reduce the level of clinical disease/death in a population
- prevent catastrophic outbreaks of disease in the population (while perhaps permitting an endemic disease situation).

A study should have a limited number of objectives: one or two primary objectives and a small number of secondary ones. Increasing the number of objectives will unnecessarily complicate the protocol and might jeopardise compliance. A trial with a very simple design might be able to afford a much larger sample size within a given budget, thus enhancing the power of the study.

This chapter will focus on controlled trials which compare two groups (sometimes referred to as **two-arm** studies), although the principles also apply to studies with more than two 'arms'. The two groups might be a comparison of an intervention with a placebo, or a new intervention versus a standard treatment, or one of many other possible evaluations (comparison of doses, combinations of interventions, timing of interventions *etc*). In general however, when evaluating a new therapy, it should be compared with an existing standard therapy, if one with a documented level of efficacy exists. It is unethical to include a no-treatment group if it will result in undue suffering in animals assigned to that group that they wouldn't experience under appropriate management (*ie* an existing product or procedure exists to reduce or prevent that suffering).

The decision as to whether to use a **positive control** (existing therapy) or **negative control** (placebo) might have profound effects on the animals available for inclusion in a trial and the results of that trial. For example, recent trials of tilmicosin for the treatment of intra-mammary infections during the dry period in dairy cows obtained very different results depending on whether the trial contained a positive or negative control group. In the trial with positive controls (cloxacillin-treated cows) (Dingwell et al, 2003), the cure rates for both groups were reasonable (>60%) but the tilmicosin-treated group had a statistically significantly higher cure rate. In the negative control trial, cure rates were much lower (<30%) (Reeve-Johnson, 2001). One possible explanation of the difference is that farmers, knowing that there was a 50% chance that any cows they put on the latter trial would receive no antibiotic treatment, were only willing to submit cows with chronic or serious *Staph. aureus* infections and which they were already planning on culling. Certainly the nature of the *Staph. aureus* infections in the latter trial was substantially different than in the former.

Studies with more than two arms require justification for each of the interventions being investigated and a larger sample size, although some efficiency in this area can be obtained through the use of factorial designs. A 2X2 factorial design would be one in which a subject is randomly assigned to one of the four intervention groups defined by the combination of two dichotomous interventions. It is a statistically efficient way of evaluating both interventions and the interaction between them.

11.3 The study population

The study population is the collection of subjects in which the trial will be carried out. It should be representative of the reference (or target) population – the population to which you want the results of the trial to apply (see Chapter 2 for more discussion of study and reference populations). The choice of reference population might be important in phase III clinical trials in which the geographic location could play a role in the acceptability of the trial for use in the registration process. Usually, the study population is obtained by seeking volunteer participants either by contacting them directly (*eg* via letter or the media) or by asking veterinarians to nominate some of their clients whose subjects meet the eligibility criteria. While the use of volunteer participants is unavoidable, how well the study population (participants) represent the target and external populations (see section 2.1.3) must be taken under consideration when extrapolating the study results.

11.3.1 Unit of concern

When defining the study population, the first issue is to specify the level at which the intervention is applied (Example 11.2). If an intervention can only be applied at a group level (*eg* to a litter, pen or herd) then the study population consists of all eligible groups (*eg* all sea cages on the study farms in the teflubenzuron trial). The outcome in such a study might be measured at the group level (a group-level study) or at the individual level (a cluster randomised study – discussed in 11.4.2). If the intervention is applied

Example 11.2 Levels of intervention and outcome measurement

Eprinomectin trial
The treatment was randomly assigned at the individual cow level and the outcome consisted of repeated measures (monthly milk production values) at the same level.

Teflubenzuron trial
The treatment could only be administered at the sea-cage level because it was added to the feed. The outcome (lice counts) was measured at the individual fish level. Consequently, this was a cluster randomised trial.

Hoof-trimming trial
The treatment (trimming) was randomly assigned to cows, but the outcome was assessed at the hoof level. These hoof-level measurements were then aggregated to cow-level outcomes.

at an individual level (*eg* all cows in the study herds in the eprinomectin trial), then the outcome must also be measured at that level. In this case, the study population consists of all eligible individuals.

It is important to remember, that controlled trials are based on volunteers (or at least a participant has volunteered them). Participants must volunteer to have their subjects receive either of the interventions as determined by the allocation process. Once a subject has been enrolled, the allocation should be carried out close to the time at which their participation in the study is scheduled to start. In some cases (*eg* controlled trials in livestock), the owner volunteers to participate but then his/her animals can be randomly assigned to the various treatment groups.

11.3.2 Eligibility criteria

Once it has been determined whether individuals or groups will be recruited for the study, eligibility criteria need to be considered with some, or all, of the following factors being considered.

- Animal-handling facilities and personnel must be in place to allow for the necessary sampling during the trial.
- Adequate records must be available to document the subject's past history and to provide outcome measures (if relevant).
- For trials of therapeutic agents, clear case definitions for the disease being treated must be developed to determine which cases are eligible for inclusion.
- For trials of prophylactic agents, healthy subjects are required and procedures for documenting their health status at the start of the trial might be required.
- Subjects in a trial need to be capable of benefiting from the intervention. As much as possible, avoid the 'ceiling effect' (the maximum possible improvement). For example, the start of the teflubenzuron trial was delayed by a month because the general level of sea lice in the Bay of Fundy was slow to build during the summer the trial was carried out. There was no point evaluating the intervention when there were few lice for it to work on. Similarly, the positive-control tilmicosin trial referred to in section 11.2 obtained better response rates by using a design which encouraged the inclusion of subjects capable of responding. On the other hand, restriction of a trial to subjects that are most likely to respond to (benefit from) the intervention will increase the power of the trial but might limit the generalisability of the results.
- Avoid subjects with high risks for adverse effects.

In some cases, if the participants do not meet these criteria at the time of recruitment (*eg* not having adequate records), it might be acceptable to have them agree to meet the standards during the period of the trial.

A narrow set of eligibility criteria will result in a more homogenous response to the intervention and this might increase the statistical power of the study, but reduce the generalisability of the results. A broad set of eligibility criteria will result in a much larger pool of potential applicants. Balancing these two considerations must be done on a case-by-case basis, while adhering to the objectives of the study.

11.3.3 Sample size

The size of the study needs to be determined through appropriate sample size calculations, with attention being paid to both Type I and Type II errors. When computing the power of the study, the magnitude of the effect to be detected should be one which is clinically (and in some cases, economically) meaningful. It is not unusual to increase the required power from 80% to 90% in controlled trials. The basic formulae for sample-size calculation were presented in Chapter 2 and a more complete description has recently been published (Wittes, 2002). Here we discuss a few important issues that impact on sample-size considerations.

Time-to-event data

As was noted previously, the sample size required for qualitative (*eg* dichotomous) outcomes is often much larger than that required for outcomes measured on a continuous scale. Obviously, the choice of outcome(s) and its measurement should reflect the study objectives. A particular consideration in many controlled trials is the need to compute a sample size for a study based on time-to-event data. A discussion of these methods is beyond the scope of this text and the reader is referred to Peduzzi et al (2002) for a discussion of some of the important issues. A description of one software program for computing sample sizes for time-to-event outcomes has recently been published (Royston and Babiker, 2002).

Time for recruitment

In controlled trials, one issue to be faced is the length of time it will take to recruit the required number of study subjects. This is a particularly serious problem for studies on therapies for relatively rare conditions. If an adequate number of subjects is not available at a single site, a multisite trial might have to be planned. Although multicentre trials complicate the protocol and the implementation of the trial, it can enhance the generalisability of the results and also increase the opportunity to identify interaction effects. Two other specific issues related to time for recruitment deserve consideration. First, if recruitment on a study farm lasts longer than one production cycle (*eg* intercalving interval in dairy herds), then an intervention that is related to the production cycle might be reapplied to cows that have already been treated. This might or might not be acceptable (depending on the nature of the intervention), but at the very least will require special consideration in the analyses. Second, if season of treatment is likely to influence the results, then the recruitment period should span at least one full calendar year.

Loss from the study

Loss of subjects from the study might happen for a variety of reasons. Some subjects might be lost to follow up (*eg* moved away or identification tag lost) while others might be non-compliers (participants who do not comply with the protocol). Finally, some subjects might be lost due to competing risks (*eg* die from other diseases while still on the trial). Of these, non-compliance is of particular concern. The effect of non-compliance is greater than simple reduction of the available sample size because non-compliance is not usually a random event. Consequently, it likely affects the estimate of the intervention effect (generally biased towards the null) which further reduces the

power of the study. Once a sample size has been estimated, it is wise to compute the expected power of the study based on different estimates of the potential losses to the study.

Other issues

Three other issues that impact on sample size are discussed further in section 11.9 but should be mentioned here. First, the sample size needs to be increased if you want to carry out meaningful investigations of the effect of the intervention in subgroups of the study population. Second, if interim analyses of the data are planned (sequential designs), the sample-size calculation will have to take this into account. Finally, if a trial has several primary objectives, sample sizes for each objective should be computed and the largest estimate used.

11.4 Allocation of participants

It is clear that a formal randomisation process is the best method for allocating subjects to study groups. Using clinical judgement in the selection of interventions will build bias into any non-randomised trial (clinical judgement applied in selection of therapies will ensure that confounders are unevenly distributed across study groups and hence, bias the trial). It would be virtually impossible to control this bias analytically. However, before discussing formal randomisation procedures we will discuss some alternatives.

11.4.1 Alternatives to randomisation

Historical control trials are ones in which the outcome after an intervention is compared with the level of the outcome before the trial (before/after comparison). For example, a vaccine for neonatal diarrhea might be introduced into a dairy herd and the incidence of diarrhea in the year after vaccination compared with the incidence in the year before. Historical control trials are generally unacceptable. For a historical control trial to have any validity, four criteria must be met.

- The outcome being measured must be predictable (*eg* constant incidence of neonatal diarrhea from year to year).
- There must be complete and accurate databases on the disease of interest.
- There must be constant and specific diagnostic criteria for the outcome.
- There must be no changes in the environment or management of the subjects in the study.

Rarely are any, let alone all, of these criteria met for animal-health problems. An additional limitation of historical control trials is that it is impossible to use blinding techniques. However, a historical control trial of teflubenzuron was carried out as a comparison to the randomised control trial described in the examples in this chapter (Campbell, 2002b). Given the very short duration of the trial and the investigator control over all pre- and post-treatment data, it was considered to be an acceptable trial for evaluating changes in lice numbers following treatment of an entire site (see Example 11.3).

Example 11.3 Assignment of interventions

Eprinomectin trial
Cows were randomly assigned to the two treatment groups within herds using a stratified and blocked randomisation approach. First, cows were stratified by herd. Subsequently, within each 10 successive calvings in a herd (the block), five cows were assigned to each group. The main purpose of this was to remove any effect of seasonal variation in parasite burdens as a potential confounder. Computer-generated random numbers were used for the assignment prior to the cows calving.

Teflubenzuron trial
This was a cluster randomised trial with sea cages assigned to the treatment groups (because the treatment could only be administered at the cage level). Matched pairs of cages were identified (matched on site, and average lice burden) and one cage from each pair was randomly assigned to the treatment group. The outcome (sea-lice counts) was monitored using repeated cross-sectional samplings of each cage.

Even though cages were the unit of allocation, there was still evidence of contamination between subjects. Reduction of the number of lice in half of the cages within a site (*ie* the treated cages) appeared to reduce the level of recruitment of young (free-swimming) sea lice in control cages. Consequently, the percentage reduction from the randomised trial was less than that observed in a historical trial where all cages at a site were treated (and lice counts before and after treatment compared). **Note** Given the very short time frame of the historical control study (<3 weeks), the fact that collection of both pre- and post-treatment data was under the control of the investigator, and the objective nature of the outcome (lice counts) most, or all, of the criteria for a historical control to be valid were met.

Hoof-trimming trial
Cows were blocked by breed (three groups), parity (three groups), and calving date by generating a list of cows sorted by these three criteria. Following a coin-toss to decide the allocation of the first animal, cows were systematically assigned to the two interventions (every second cow into a given treatment group). Cows that were assigned to the no-trimming intervention but which required therapeutic trimming at the autumn visit were excluded from the trial.

Systematic assignment of individuals to treatment groups (*eg* alternating assignment) is a reasonable alternative to formal randomisation. Systematic assignment might be based on the use of pre-existing animal identification numbers with odd and even numbers forming the basis of the group assignment. Systematic assignment might make it harder to keep participants and study personnel blind as to the intervention identity, but aside from this, is often just as effective as random allocation (provided outcome assessment is done blindly). If half the subjects are to be allocated to receive the treatment, the initial subject allocation should be random and thereafter every second subject receives the initial allocated intervention. Do not apply the intervention to the first (or last) half of the subjects and the comparison treatment to the remainder.

Outcome adaptive allocation procedures are ones which are designed to ensure that the majority of subjects get the benefit of the best therapy available. The allocation of

subjects is influenced by the experience of previous subjects in the trial. One example is 'play the winner' allocation in which subjects continue to be allocated to an intervention level as long as that treatment is producing beneficial results. As soon as it fails, the allocation switches to the other treatment. These procedures are only suitable if the result of the intervention is clearly identifiable in a very short period after the treatment is administered. They have not been used commonly in animal-health studies.

11.4.2 Random allocation

As indicated, formal randomisation is the preferred method of allocation. It must be noted that random allocation does not mean 'haphazard' allocation and a formal process for generating random intervention assignments (*eg* computer-based random number generator) must be employed. Random allocation should be carried out as close as possible to the start of the study to reduce the possibility of withdrawals after allocation.

Simple randomisation involves each subject being assigned to an intervention level through a simple random process without any further considerations. **Stratified randomisation** (*eg* randomisation within age categories) helps ensure that a potential confounder (age) is equally distributed across study groups. One specific form of stratified randomisation is random allocation of animals within herds (Example 11.3). This ensures that all herd factors that might influence the outcome are balanced across study groups. **Blocked randomisation** requires the random allocation of subjects within blocks of subjects as they enter the trial (Example 11.3). Blocked randomisation can enhance statistical efficiency (ensuring equal numbers of subjects in each study group) and also remove any temporal effects from the study.

Cross-over studies
In a cross-over study, each subject gets both of the interventions (in sequence). However, the first intervention administered is still assigned randomly. This process is only suitable for the evaluation of therapies for chronic conditions where the duration of the intervention effect is relatively short-lived. A 'wash-out' period might be required between interventions. It has the advantage that it increases the power of the study because it permits a more powerful 'paired' analysis of the data. A cross-over trial to study the effect of ionophore treatment (evaluating the effect of monensin) on fecal shedding of Map has been designed and is under way at the time of writing this text (Leslie, KE, personal communication, 2003). This study is based on the finding that shedding of organisms were significantly reduced in chronic cases with ionophore treatment (Brumbaugh et al, 2000). Chronically infected cows are being treated with either ionophore or a placebo for three months followed by a wash-out period (one month), after which the cows are switched to the other treatment.

Factorial designs
This design is particularly well-suited to trials investigating two or more interventions, especially if the interventions might produce synergism or antagonism. Here the various combinations of the treatments (*eg* neither, treatment 1 only, treatment 2 only, both) are assigned to the study subjects. Because the design is usually balanced, the

treatment effects are not confounded (*ie* they are unrelated or orthogonal) and the analyses straightforward. Normally, one should not attempt to assess more than 2-3 interventions as the possible interactions become difficult to interpret.

Cluster randomisation

There are a number of reasons why a cluster of animals (*eg* a herd) should be allocated to an intervention group rather than individual animals. In some instances, it might be the only feasible method. For example, if the intervention is one which is always given at the group level (*eg* medication in the drinking water), then there is no choice. Even if the intervention could be administered at the individual level, it might be impossible to keep track of individuals within the group so assignment of the whole group to one intervention would be appropriate. Cluster randomisation is also appropriate if there is potential for physical spread of a treatment to the control group (*eg* pour-on endectocides when applied to half the cattle in a herd (Barber, 2003)).

In some cases, cluster randomisation might be desirable to prevent contamination between intervention groups. For example, a live virus vaccine administered to some animals in a herd might spread to other animals. For this reason, the role of herd immunity has long been a concern in individual randomised controlled trials of vaccines. If one-half of a herd is vaccinated, it might sufficiently reduce the number of susceptibles in the herd to effectively protect the non-vaccinated animals through herd immunity. On the other hand, it has been argued that leaving half of a herd non-vaccinated might allow a sufficient build-up of infectious organisms such that the vaccinated animals are overwhelmed. Both scenarios would result in estimates of vaccine efficacy that were biased towards the null. Whereas this effect was known to be present in early trials of the polio vaccine in humans, one veterinary study which investigated it was unable to document an effect in that trial (Waltner-Toews et al, 1985). The role of herd immunity would be less important for diseases with widespread, simultaneous exposure to the agent versus within-herd transmission.

Cluster randomised trials are much less statistically efficient than trials with random allocation of individuals and the clustering of individual subjects within the groups needs to be taken into account in analysis (see Chapters 20-23). In a cluster randomised trial, the best scenario for follow-up is if all individuals can be monitored for the duration of the study. If this is not possible, following a randomly selected cohort will be the most statistically powerful approach. If it is not possible to follow individuals, the investigator will have to carry out repeated cross-sectional samplings throughout the follow-up period.

Split-plot designs

A final elaboration of allocation discussed here is a split-plot. This design is useful if there are two or more interventions, one of which needs to be applied at the group level and the other(s) can be assigned to individuals. The analysis must take account of the different degrees of freedom to assess intervention effects at the different levels (*ie* group versus individual).

11.5 SPECIFYING THE INTERVENTION

The nature of the intervention must be clearly defined, but that does not mean that it cannot have a degree of flexibility built into it. A **fixed intervention** (one with no flexibility) must be rigorously defined in the protocol, and is appropriate for assessing new products (particularly in phase III trials). A more **flexible protocol** might be appropriate for products that have been in use for some time and for which a body of clinical application information exists. For example, feedlot cattle might be assigned to one of two antibiotics for the treatment of respiratory disease but the timing of a decision to change antibiotics (*ie* a treatment failure) or stop treatment (*ie* a treatment success) might be left up to the person responsible for the animals provided it fell within a range defined in the protocol (*eg* between three and five days). The initial treatment assignment should still be masked so that clinical decisions are not influenced by knowledge of group allocation.

Clear instructions about how the intervention needs to be carried out are essential, particularly if participants are going to be responsible for some or all of the treatments. In addition, the system of ensuring that the correct treatment goes to the right animal must be kept as simple as possible. Finally, some method of monitoring the intervention administration process should be put in place. Example 11.4 describes a few features of the interventions in the three example studies.

Example 11.4 The intervention

Eprinomectin trial
In this trial, the producer was responsible for administering all treatments to cows at calving. To verify the timing of the administration, the producer was asked to record both the treatment date and the calving date of the cow. Only treatments administered between five days before and 59 days after calving were included in the analysis. The validity of the data recording was evaluated by comparing the recorded calving date with the data obtained from the production-recording programme that each farm was enrolled on.

Teflubenzuron trial
To ensure that the correct feed went to each sea cage, every cage was labelled with a large sign with a letter identifying the cage. All feed bags going to the site were individually labelled with the cage letter. Even with these efforts, there was a very small number of bags of feed that ended up being given to the wrong cages.

Hoof-trimming trial
Autumn trimmings were carried out between September/October and January of each of the two winter periods. 92% of the trimmings were performed by a single trained research technician. Claws were trimmed to meet a specified set of conformation criteria (see original publication for description).

11.6 Masking (blinding)

A key component in the effort to prevent bias in controlled trials is the use of masking (or blinding). In a **single-blind study**, the participant is unaware of the identity of the intervention going to the study subjects. This feature should help ensure equal follow-up and management of subjects in the various intervention levels. In a **double-blind study**, both the participant and the study team (*ie* people administering the interventions and assessing the outcomes) are unaware of intervention assignment. This feature helps ensure equal assessment of the subjects in different intervention levels. In a **triple-blind study**, the investigators analysing the data are also unaware as to which group received which treatment. This feature is designed to ensure that the analysis is conducted in an unbiased manner.

In many cases it is necessary to use a **placebo** to ensure that the relevant individuals remain blind. A placebo is a product that is indistinguishable from the product being evaluated and which is administered to animals in the groups designated to receive the comparison treatment. In many drug trials, the placebo is simply the vehicle used for the drug, but without any active ingredient. This was the nature of the placebo in the eprinomectin trial and both treatment and placebo were dispensed in identical bottles labeled only with a bottle number.

In some cases, even using a placebo might not be adequate to ensure blinding. For example, in trials of recombinant somatotropin in dairy cattle, it has been argued that a placebo is irrelevant because the drug produces such a noticeable change in milk production, anyone working with the cows on a regular basis would know which cows received the treatment.

For controlled trials comparing two drugs or treatment regimes, it might not be possible to make them physically indistinguishable. In this case, double placebos might be used, one to match each product, and each study subject appears to be receiving both treatments although only one would contain the active ingredient. One concern with the use of a placebo is that, even though it might not contain the active ingredient being investigated, it could still have either a positive or negative effect on the study subjects. For example, a placebo vaccine that does not contain the antigen of interest might still induce some immunity as a result of adjuvant in the placebo. On the other hand, vaccination of the control group with a placebo could result in stress in that group that would not be present if no vaccine was given.

11.7 Follow-up/compliance

It is essential that all groups in a controlled trial be followed rigorously and equally (Example 11.5). This is a simpler process if the observation period following the intervention is short, but this time period must be long enough to ensure that all outcomes of interest have been observed and recorded. Regardless of the effort expended on follow-up, it is inevitable that some individuals will be lost to the study

Example 11.5 Follow-up/compliance

Eprinomectin trial
In this trial, follow-up was relatively straightforward for the primary outcome (milk production) because it was routinely recorded for all cows in the participating herds by a milk-production recording agency. However, despite monthly visits, the recording of all reproductive data for all cows was only deemed to be sufficiently complete for reliable analysis, in 20 of the 28 herds.

Teflubenzuron trial
Follow-up was also relatively straightforward in this trial because the observation period was short and the protocol specified exact sampling dates. Lice numbers in two cages at one study site necessitated early withdrawal from the study (and treatment with another product). The protocol stipulated 24 hours' notice to be given before withdrawal and this provided time for the study team to perform one final sampling. Once removed, the assignment of the cages was revealed (they were control cages) and their pair-matched treatment cages were also removed.

Hoof-trimming trial
Recording of lameness treatments between the autumn and spring visits was encouraged by offering reduced cost treatments to the participating producers. At the spring examination, the proportion of cows initially assigned to the treatment groups that was available for follow-up examination was 79% and 87% in the first and second year of the study, respectively. Most losses were due to culling of cows from the study herds.

through drop-out or lack of compliance, and the sample size needs to be adequate to allow for this (see section 11.3.3).

Most important in minimising losses from the study is regular communication with all participants. Incentives to remain in the study might also be provided. These might include financial incentives, provision of information which they might not otherwise have (*eg* detailed udder-health evaluation of a dairy herd provided to participants in a controlled trial of a new dry-cow antibiotic product), or public recognition of their efforts (provided confidentiality concerns have been addressed). For those participants that do drop out, information about study subjects might still be available through routine databases (*eg* milk-production recording programmes) if the participant is willing to provide access. This can be used to either provide some follow-up information or to compare general characteristics of the study subjects withdrawn from the study with those that remained in the study. Nonetheless, because participants in a trial should always have the opportunity to withdraw their animal(s) from a trial, procedures for evaluating those withdrawals should be put in place. This should include methods of documenting the reason for the withdrawal and, potentially, procedures to collect samples from all subjects being withdrawn before their departure.

In addition to maximising retention in a study, effort needs to be expended to determine if study subjects are complying with the protocol. This might be evaluated through

interviews at periodic visits or through collection of samples to test for levels of a drug being investigated. Indirect assessment might be carried out by methods such as collecting all empty containers from products used in a trial. The amount of product (or placebo) used should be appropriate for the number of subjects in the study.

11.8 Measuring the outcome

As indicated above, a controlled trial should be limited to one or two primary outcomes (*eg* disease occurrence in a trial of a prophylactic agent) and a small number (1-3) of secondary outcomes (*eg* productivity, longevity). Having too many outcomes leads to a serious problem of 'multiple comparisons' in the analysis (see section 11.9.1). When selecting outcomes to be measured, those that can be assessed objectively are preferred to subjective outcomes, but the latter cannot always be avoided (*eg* occurrence of clinical disease).

In general, outcomes should be clinically relevant. Intermediate outcomes, (*eg* antibody titres in a vaccine trial) might be useful in determining why an intervention might not produce the desired outcome, but should not be a replacement for a primary, clinically relevant, outcome related to the objectives of the study (*eg* occurrence of clinical disease). Clinically relevant outcomes include the following:

- diagnosis of a particular disease – requires a clear case definition
- mortality – objective but still requires criteria to determine cause of death (and not always relevant)
- clinical signs scores for assessing the severity of disease – difficult to develop reliable scales
- objective measures of clinical disease – (*eg* rectal temperature for assessing severity of respiratory disease in feedlot cattle)
- measures of subclinical disease – (*eg* somatic cell counts as indicators of subclinical mastitis)
- objective measures of productivity/performance – (*eg* milk production, measures of reproductive performance)
- global measures of health – combine scores or occurrences of several diseases.

Outcomes might be measured on a continuous scale, or as categorical data (often dichotomous), or time-to-event measurements (*eg* time to the occurrence of a disease). Studies based on time-to-event data might have greater power than a study based on simple occurrence, or not, of an event in a defined time period. Outcomes might also be measured at a single point in time, or assessed multiple times for each subject (longitudinal data).

11.9 Analysis

Analyses can be carried out either on an **intent-to-treat** basis or a **per-protocol** basis. In an intent-to-treat analysis, data from all subjects assigned to a specific intervention

are included in that intervention regardless of whether or not they completed the study, or whether or not they complied with the protocol. Such an analysis will provide a conservative estimate of the effect of the intervention but might reflect the expected response when the intervention is used in another population with characteristics similar to the study population. In a per-protocol analysis, only subjects which completed the study as outlined in the protocol are included in the analysis. This approach might provide a good measure of response given that the intervention is used as intended but will likely produce a biased estimate of the intervention effect in future use because non-compliance is not likely a random event. Consequently, non-compliers are probably not representative of all participants assigned to that intervention.

An analysis usually starts with a baseline comparison of the groups as a check on the adequacy of the randomisation procedures. This should not be based on an assessment of the statistical significance of the difference among groups, but rather an assessment of their comparability. Differences among the groups, even if not statistically significant, should be noted and taken into consideration in the analyses.

The specific procedures for analysing data from controlled trials will not be covered in this chapter because they are discussed in more detail elsewhere in the book. However, a few specific issues will be touched on.

While randomisation is designed to equally distribute potentially confounding factors across the intervention groups, it might not remove all potential confounding. Whether or not to present adjusted (*ie* results adjusted for potential confounders) or unadjusted results is a subject of active debate. Adjusted results might be less biased estimates if the adjustment procedure has removed any residual confounding (particularly a concern in small trials), but could be more confusing to present to users of the trial results. In some cases, control of other factors might substantially improve the precision of the estimate of the intervention effect by substantially reducing the unexplained variance. For example, in the eprinomectin trial, control for factors such as parity and stage of lactation, which have a considerable effect on level of milk production, substantially reduced the unexplained variation in the regression model, hence, increasing the power of the study.

Because many controlled trials involve repeated assessments of subjects throughout the study period, the problem of some missing observations is common. A detailed discussion of how to manage this issue can be found in Peduzzi et al (2002). Analysis of longitudinal data presents some unique challenges. For a starting point the investigator needs to determine if they are most interested in an average effect following intervention, a change in the effect over time or a total effect. Methods of dealing with repeated measures data are covered in Chapters 21 to 23.

If study subjects are maintained in groups, it is important to account for the effects of those groups. This is particularly important in cluster randomised trials, but might also be important in trials in which randomisation occurred within the group. Procedures for analysing data from groups are presented in Chapters 20-24. Analytical issues from the three example studies are presented in Example 11.6.

Example 11.6 Analysis

Eprinomectin trial
In this trial, a linear mixed model with random effects for herds and cows (observations were repeated measures within cows) was used to account for the 'clustering' of observations within cows and herds when analysing milk production. A Cox proportional hazards model was used to evaluate the effect of the intervention on reproductive performance.

Teflubenzuron trial
In this trial, a mixed linear model with random effects for sea cage was used to account for the clustering of lice counts within cage. No formal adjustment was made for the fact that multiple comparisons were made because separate analyses were carried out for each stage in the life cycle of the sea lice and two post-treatment measurements were made. However, most P-values were <0.005.

Hoof-trimming trial
A generalised linear mixed model (random effects logistic regression) with herd-year groupings as the random effect was used to evaluate the effects of treatment on the outcomes. Other predictors included in the model were factors such as breed, parity, housing type, season, stage of lactation. First-order interactions between treatment and other predictors were evaluated and a backward elimination model-building procedure was used to identify statistically significant predictors.

11.9.1 Multiple comparisons

Controlled trials often give rise to analyses in which 'multiple comparisons' are often made. These can arise from:

- multiple outcome measures being evaluated
- multiple intervention groups within the trial
- the analysis of data from multiple subgroups within the trial
- periodic interim analyses being performed during the trial.

The problem with multiple comparisons is that the **experiment-wise error rate** is often much larger than the error rate applied to each single analysis (usually 5%). This can result in the declaration of spurious effects as significant. There are many procedures for adjusting the analyses to account for these multiple analyses. One of the simplest, a Bonferroni adjustment, requires that each analysis be carried out using an α/k Type I error rate, where α is the normal error rate (often 0.05) and k is the number of comparisons made. However, this results in a very conservative estimate of the statistical significance of each evaluation. Other, less conservative, procedures can be found in standard statistical texts.

The problem of subgroup analyses deserves special attention. While it is tempting to evaluate a wide range of subgroups within a trial to determine where an intervention has its greatest effect and where that effect is statistically significant, only analyses planned *a priori*, should be carried out. Otherwise, there is serious danger of identifying

spurious associations. The sample size of the study also needs to take into consideration the need to carry out these subgroup analyses or the sample size might have insufficient power to detect meaningful effects.

Sequential design studies (also called 'monitored' studies) are those in which periodic analyses of the data are carried out throughout the trial. These analyses are carried out so the trial can be stopped if there is:

- clear (and statistically significant) evidence of the superiority of one intervention over another
- convincing evidence of harm arising from an intervention (regardless of the statistical significance of that finding)
- little likelihood that the trial will produce evidence of an effect, even if carried to completion. (This concern is not relevant if the goal of a trial is to demonstrate that a new product/procedure has the same efficacy as an existing standard therapy.)

Methods for these interim analyses and for adjusting the sample size to accommodate the procedures are beyond the scope of this text but are reviewed in Friedman et al (1998). One such example in human medicine was the recently halted trial of hormone-placement therapy for post-menopausal women (Women's Health Initiative, 2002). The trial was stopped after an average follow-up period of 5.2 years instead of being allowed to run the planned length of 8.5 years because there was statistically significant evidence of increased risk of breast cancer in individuals receiving the therapy.

11.10 Ethical considerations

There are two components to the ethical considerations for controlled trials of animal-health products and procedures. The first is an ethics review by a board whose focus is the ethical treatment of the participants, and the second is a review by an animal-welfare committee whose focus is the well-being of the animal subjects. Specific regulations and guidelines will vary from country to country, but in general the following issues must be considered.

- Is the investigation justifiable? That is, is it likely to produce meaningful results which will ultimately benefit animal health? Has the design of the study been adequately planned to ensure that valid results will be obtained?
- Is the sample size appropriate? In this case, the needs of an adequate sample size to ensure sufficient power for the study will have to be balanced by a desire to minimise the sample size in order to reduce the number of subjects who might receive the less desirable intervention.
- Are procedures in place to minimise the risk and maximise the benefits for participants and subjects in the study? This consideration, and the preceding one might necessitate interim analyses of results, if feasible.
- Are all participants in the trial participating on the basis of informed consent? The provision of informed consent implies that not only have they had the details of the trial provided to them, but this has been done in a manner

that ensures that they understand both the risks and benefits of participating.
- Participants must also have the option to withdraw from the study if they so choose.
- Has adequate provision been made to protect all data to ensure their confidentiality and protect the privacy of the participants?

SELECTED REFERENCES/SUGGESTED READING

1. Barber S. Comment on 'A comparison of persistent anthelmintic efficacy of topical formulations of doramectin, eprinomectin, ivermectin and moxidectin against naturally acquired nematode infections of beef calves' and problems associated with the mechanical transfer (licking) of endectocides in cattle. (Letter to the editor.) Vet Parasitol 112: 255-257, 2003.
2. Brumbaugh GW, Edwards JF, Roussel AJ Jr., Thomson, D. Effect of monensin sodium on histologic lesions of naturally occurring bovine paratuberculosis. J Comp Path 123: 22-28, 2000.
3. Campbell PJ, Hammell KL, Dohoo IR. Historical control clinical trial to assess the effectiveness of teflubenzuron to treat sea lice on Atlantic salmon. Dis Aquat Org 2002a; submitted.
4. Campbell PJ, Hammell KL, Dohoo IR. Randomized control clinical trial to investigate the effectiveness of teflubenzuron to treat sea lice on Atlantic salmon. Dis Aquat Org 2002b; submitted.
5. Dingwell RT, Leslie KE, Duffield TF, Keefe GP, Kelton DF. 2003. Management strategies influencing drying-off efficiency and development of new intramammary infections in the dry period. J Dairy Sci 86: 159-168.
6. Friedman LM, Furberg CD, Demets DL. Monitoring response variables. Fundamentals of clinical trials. New York Springer-Verlag, 1998.
7. Green SB. Design of randomised trials. Epidemiol Rev 2002; 24: 4-11.
8. Lavori PW, Kelsey J eds. Clinical trials. Epidemiol Rev 2002; 24: 1-90.
9. Lavori PW, Kelsey J. Introduction and Overview. Epidemiol Rev 2002; 24: 1-3.
10. Manske T, Hultgren J, Bergsten C. The effect of claw trimming on the hoof health of Swedish dairy cattle. Prev Vet Med 2002; 54: 113-129.
11. Meinert CL. Clinical Trials: design, conduct and analysis. Oxford Oxford University Press, 1986.
12. Nødtvedt A, Dohoo IR, Sanchez J, Conboy G, DesCôteaux L, Keefe G. Increase in milk yield following eprinomectin treatment at calving in pastured dairy cattle. Vet Parasitol 2002; 105: 191-206.
13. Peduzzi P, Henderson W, Hartigan P, Lavori PW. Analysis of randomized controlled trials. Epidemiologic Reviews 2002; 24 :26-38.
14. Piantadosi S. Clinical Trials: A methodological perspective. New York John Wiley and Sons, 1997.
15. Reeve-Johnson, L.G. Assessment of the efficacy of a novel intramammary antibiotic for the treatment of mastitis caused by *Staphylococcus aureus* during the non-lactating period in United States dairy herds. Thesis for the Royal College of Veterinary Surgeons for the Diploma of Fellowship. Royal College of Veterinary

Surgeons, London, England. 2001.
16. Risks and benefits of estrogen plus progestin in healthy post-menopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002; 288: 321-333.
17. Royston P, Babiker A. A menu driven facility for complex sample size calculation in randomized controlled trials with a survival or binary outcome. The Stata Journal 2002; 2: 151-163.
18. Sanchez J, Nødtvedt A, Dohoo IR, DesCôteaux L. The effect of eprinomectin at calving on reproduction parameters in adult dairy cows in Canada. Prev Vet Med 2002; 56: 165-177.
19. Waltner-Toews D, Martin SW, Meek AH, McMillan I, Crouch CF. A field trial to evaluate the efficacy of a combined rotavirus-coronavirus/*Escherichia coli* vaccine in dairy cattle. Can J Comp Med 1985; 49: 1-9.
20. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002; 24: 39-53.

12

VALIDITY IN OBSERVATIONAL STUDIES

Objectives

After reading this chapter, you should be able to:

1. Identify the different types of selection bias and assess whether or not a particular study is likely to suffer from excess selection bias.

2. Determine the likely direction and magnitude of a selection bias through the use of estimates of sampling fractions or sampling odds.

3. Apply the principles of bias prevention in the design of a study; for example, how to avoid detection bias in secondary-base studies.

4. Explain the differences between non-differential and differential misclassification bias in terms of sensitivity and specificity.

5. Correct 2X2 table data for misclassification of exposure, disease or both.

6. Explain why one cannot use the population sensitivity and specificity estimates to correct for disease status misclassification in case-control studies.

7. Understand the basis of correcting for measurement bias.

12.1 Introduction

The key features of study design implementation and analysis reflect our efforts to help ensure that we obtain valid results from our research efforts. The term **validity** relates to the absence of a systematic bias in results; that is, a valid measure of association in the study population will have the same value as the true measure in the target population (except for variation due to sampling error). To the extent that the study population and the target population measures differ, the result is said to be biased. There are three major types of bias:

- selection bias: due to factors affecting the selection of study subjects, or to other factors that relate to the willingness to participate in a research project
- information bias: due to factors relating to obtaining accurate information on the exposure, outcome and covariates of interest, and
- confounding bias: due to the effects of factors other than the exposure of interest on the observed measure of association.

In this chapter, we discuss the nature, impact and prevention of selection and information bias; confounding is discussed in Chapter 13.

The actual population in which the study is conducted is called the **study population** (see section 2.1.3). Because most analytic studies are conducted on non-randomly sampled study populations, there is always some uncertainty about how well the attributes and the associations in the study population reflect the attributes and associations in the larger target population. Once the study group(s) is selected, we must be able to accurately measure the exposure, extraneous factors and outcome of interest if we want to make valid conclusions. In this context, an **internally valid** study will allow us, based on the study data, to make unbiased inferences about the association(s) of interest in the **target population**. **External validity** relates to the ability to make correct inferences to populations beyond the target population. In this regard, while it is certainly desirable that the study and target populations be 'representative' of a larger population beyond the target, one should not sacrifice internal validity in order to gain external validity. In the extreme, there is no value in being able to extrapolate incorrect results. Nonetheless, the best studies have findings which lead to scientific theories that can be generalised to broadly defined populations.

12.2 Selection bias

As described in Chapter 1, the ideal comparison group for causal inferences is the counterfactual group. For example, in a cohort study, the ideal counterfactual group for the exposed group would be the exact same subjects if they had not been exposed. However, as this ideal group is non-existent, we strive to ensure that the study groups, based on exposure status, are totally comparable with respect to all factors that might bias the measure of association.

If it occurs, selection bias happens before the study begins and it results from

the procedures used to obtain study subjects or from factors that influence study participation. The responsible factors influence being in the study in such a way that the composition of the study group(s) differs from that in the target population in a manner that biases the association observed between the exposure(s) and the outcome(s) of interest.

Assume the target and study populations have the structure shown in Table 12.1 (upper-case letters represent the target population, lower case the study population). How can we select the study population to avoid selection bias?

Table 12.1 A representation of the structure of the target and study populations

Target population structure				Study population structure			
	E+	E-			E+	E-	
D+	A_1	A_0	M_1	D+	a_1	a_0	m_1
D-	B_1	B_0	M_0	D-	b_1	b_0	m_0
	N_1	N_0	N		n_1	n_0	n

12.2.1 Sampling fractions and sampling odds

The study population is a sample of the target population. Regardless of whether the study population is a random sample from the target population or not, we can visualise the sampling fractions (*sf*) in each of the four categories of exposure and disease. These are:

$$sf_{11}=a_1/A_1$$
$$sf_{12}=a_0/A_0$$
$$sf_{21}=b_1/B_1$$
$$sf_{22}=b_0/B_0 \qquad \textit{Eq 12.1}$$

where the subscripts refer to the row-cell combination in the 2X2 table structure (row 1, column 1 is the upper left cell: exposed and diseased *etc*). If all four sampling fractions are equal, there is no selection bias. Moreover, if the odds ratio (*OR*) of these sampling fractions (OR_{sf}) equals 1, there is no bias to the odds ratio as a measure of association, even if the four sampling fractions are not equal. Under this latter condition, there is also no bias to the risk ratio (*RR*) if disease is infrequent. See Example 12.1 for an application of this sampling fraction odds ratio. In reality, we rarely know the values of the *sf* so this limits the utility of this approach. However, this approach provides a theoretical basis for understanding the conditions under which bias will or will not occur.

In a risk-based cohort, or longitudinal study, one could also express the sampling odds of disease (so_D) among exposed subjects versus the sampling odds of disease in the non-exposed subjects as:

$$so_{D|E+}=sf_{11}/sf_{21}$$
$$so_{D|E-}=sf_{12}/sf_{22} \qquad \textit{Eq 12.2}$$

Example 12.1 Response bias

In order to demonstrate that non-response can bias an association measure, we first give an example where the non-response is related only to exposure and not to the outcome. In this situation, one would not expect the non-response to bias the measure of association. For this example, we will initially assume the following scenario:

- That in the exposed subjects in the target population, 30% are non-responders (nr) and that the risk of the outcome in the non-responders is the same as that in the responders (r) at 25%.
- That in the non-exposed subjects in the target population, 10% are non-responders and these subjects have the same risk of the outcome as the responders at 12%.

Consistent with these assumptions, suppose the target population structure is:

	$Exposed_r$	$Exposed_{nr}$	$Non\text{-}exposed_r$	$Non\text{-}exposed_{nr}$
D+	175	75	972	108
D-	525	225	7128	792
	700	300	8100	900
Risk	0.25	0.25	0.12	0.12

Given our assumtpions, if we initially contact 100 exposed and 100 non-exposed individuals, the study group will have the following structure:

	Exposed	Non-exposed
D+	18	11
D-	52	79
	70	90

Apart from rounding error, the ratio of risks (*RR*) in the study population (*RR*=2.04) matches the risk ratio in the target population (*RR*=2.00). There is no bias.

Now, given exactly the same response risks, we will assume the risk of the outcome is twice as high in non-responders as in responders in both the exposed and non-exposed groups (non-response is now related to both exposure and outcome).

(*continued on next page*)

Example 12.1 *(continued)*

Under this scenario, the population structure would be:

	$Exposed_r$	$Exposed_{nr}$	$Non\text{-}exposed_r$	$Non\text{-}exposed_{nr}$
D+	133	114	891	198
D-	567	286	7209	702
	700	300	8100	900
Risk	0.19	0.38	0.11	0.22

The ratio of the risks in this population is 0.247/0.121=2.04

As before, if we initially contact 100 exposed and 100 non-exposed individuals, the study group will have the following structure (apart from sampling error):

	Exposed	Non-exposed
D+	13	10
D-	57	80
	70	90

Now the study group risk ratio is 0.19/0.11=1.73, which is a biased estimate of the true association. To link this bias to the sampling fractions, the sampling fractions are:

$$sf_{11}=13/247 = 0.053$$
$$sf_{12}=57/853 = 0.067$$
$$sf_{21}=10/1089 = 0.009$$
$$sf_{22}=80/7911 = 0.010$$

and the odds ratio of the sampling fractions is:

$$OR_{sf} = \frac{0.053 * 0.01}{0.67 * 0.009} = 0.88$$

Thus based on the odds ratio of the sampling fractions, the bias would be expected to be towards the null. We should note that $2*0.88 \approx 1.76$. Relatively speaking, because of the non-response, we have over-sampled the non-exposed cases, or conversely we have under-sampled the exposed cases. For example, the sampling odds for disease among the exposed is 5.89 (0.053/0.009), and among the non-exposed, it is 6.7 giving a ratio of 0.88.

If these selection odds are equal, there is no selection bias. This is the goal of selection strategies in cohort studies. Similarly, in a case-control study, the sampling odds of exposure (so_E) in cases and controls are:

$$so_{E|D+}=sf_{11}/sf_{12}$$
$$so_{E|D-}=sf_{21}/sf_{22} \qquad \textit{Eq 12.3}$$

If these selection odds are equal, there is no selection bias. If the ratio of the sampling odds is greater than 1, then the bias is away from the null; if the sampling odds ratio is less than 1 the bias is towards the null. In practice, these sampling odds might be easier to visualise than the individual sampling fractions. Thus, for example, in a case-control study, we need to ensure that we are no more likely to select for exposure among cases than among non-cases to prevent selection bias. Methods to help achieve this are discussed in section 12.3. The conditions for no bias are somewhat more complex in density cohort studies where animal-time affects the sampling probabilities, but the principles are the same.

12.3 Examples of selection bias

12.3.1 Choice of comparison groups

In cohort studies it is important that the non-exposed group be comparable with the exposed group with respect to other risk factors for the outcome that are related to the exposure. This is more of an issue with the usual two-group (*ie* exposed and non-exposed) cohort design, than with a single-cohort study design. For example, a recent study has documented how the design of a surveillance system can bias the risk of disease by breed type (Ducrot et al, 2003). Similarly, in a case-control study, it is important that the control group reflects either prevalence of exposure in the 'non-case' members of the target population (risk-based study) or the proportion of exposed animal-time at risk for the non-case group in the target population (rate-based study). Comparability is best achieved by random sampling from the entire non-exposed (non-case) population; however, this latter population might be difficult to enumerate in order to construct a sampling frame. In the absence of random sampling, the decisions about how to select the comparison group must include knowledge about the study design and the biology of the problem being investigated as well as the structure and dynamics of the target population. However, a general principle is that the study groups should be selected from the same source within the target population.

12.3.2 Non-response

Non-response bias can be a major problem in both descriptive and analytic studies and its level is often understated (Sandler, 2002). Non-response leads to bias if the association between exposure and the outcome in the responders differs from that in non-responders (hence, the association in the study group differs from that in the target population). Although non-response behaves as a confounding variable, it cannot be directly controlled in the same manner. The stronger the association between exposure and disease and the greater the proportion of non-responders, the greater the potential

bias. In veterinary research, non-response on behalf of the owner could be a surrogate indicator for management, housing, or feeding differences of the owner's animals that could relate to both the outcome and the exposure factor. In studies where humans are the units of concern, willingness to enrol in a study might be related to both the exposure and the outcome, hence the study group produces a biased response.

One way to assess the possible effects of response bias, or if non-response exceeds 20-30%, is to ascertain if the extent of non-response within each group (*ie* the exposure cohorts, or the case and control groups) is approximately equal. If it is not, it creates some doubt about the lack of bias. In addition, it is informative to compare responders and non-responders using whatever information you have, recognising that because the owner won't respond (or collaborate), the data might be limited. This might give additional insight into the comparability of the groups. Minimising non-response is an important step in reducing any possible bias. Example 12.1 shows non-response bias.

Missing data can create a bias similar to non-response, because the researcher must either impute the missing value, drop the variable(s) with missing values (and possibly leave a confounding bias), or drop the observation (and hence, effectively produce a non-response). Thus, minimising missing data and assessing whether the extent of missing data is equivalent in the groups being compared are recommended features of study design.

12.3.3 Loss to follow-up or follow-up bias

Similar to non-response bias, if there is a differential loss to follow-up that is related to the exposure status and the outcome, then bias will result. Thus, the design and implementation of the study protocol should try to minimise losses from the study, and failing that we should try to ensure that both groups are followed as completely as possible and with equal rigour (the latter equalises, but does not reduce, the losses). Unfortunately, the larger the losses, the more difficult it becomes to ensure equality of losses across the study groups. Analytic approaches for assessing the impact of losses from the study are available (Cheung, 2001).

A type of bias that can result from activities during the study period relates to differential management of exposed and non-exposed subjects that develops during the study. More generally, behavioural changes in study subjects as a result of being studied are referred to as the Hawthorne effect (Mangione-Smith et al, 2002). In experimental studies we would use single- or double-blind techniques to help ensure equal follow-up of all study subjects. In an observational study, the role of the researcher is to observe, not alter, the normal (*ie* usual) events experienced by the study subject. However, it is often difficult to 'hide' the reason for the study and the act of enquiring into specific management/housing factors could lead the animal owner to modify his/her protocols in ways that are not obvious to the researcher. This could lead to differential management by exposure status, or at the very least, exposure status changes during the study period. Being aware of this possibility and implementing the study in a manner designed to minimise this bias might be the best prevention. An example of this potential bias is described by Ducrot et al (1998).

12.3.4 Selective entry or survival bias

Sometimes the groups we study are highly selected in that only subjects that possess certain desirable attributes are selected for membership. The analogous problem in studies of humans is called the 'healthy worker' effect, and is a major issue especially in occupational-health studies. In veterinary research, adult food-producing animals (sows, cows *etc*) are highly selected for herd membership on entry (*eg* they might need to meet specific growth rate and fertility criteria) and once admitted to the herd, these animals must maintain certain production standards (*eg* number of piglets produced per year) to remain in the herd. Similarly, horses that are currently racing are a biased subset of all horses that tried to enter the race circuit, and they are very likely to be healthier than all horses that have raced one or more times. As one example of this bias, if we wanted to assess the impact of disease or a new treatment programme on fertility in dairy cows, and we only chose the calving-to-conception interval as an outcome measure, we could get a biased view of the disease or treatment effects. Animals that did not get pregnant would be excluded from the outcome measure. Hence, because these cows did not pass the entry criterion of becoming pregnant, they are excluded from the study yet this failing is a crucial component of assessing the fertility status of the herd.

Entry bias can create problems in study design. For example, in Examples 8.1 and 8.2, we posited studies on the effect of *Neospora caninum* on future abortion and fertility. We suggested identifying a set of congenitally infected calves and a set of non-congenitally infected calves and following them through their first lactation – which would not end until approximately three years of age. If *Neospora* had other negative effects, for example on growth rate, many of the congenitally infected calves might not be selected for breeding as heifers and subsequent herd entry. Thus, if one followed and recorded events only for heifers that achieved pregnancy and herd-entry, the observed impacts of *Neospora* could be seriously biased. Ensuring that our study design selects subjects with outcomes that encompass the full set of important outcome events is important to prevent bias.

With respect to selective survival, if the exposure and disease being studied affect whether or not a food animal remains in the herd (or whether or not a horse is still racing) then a study group drawn from only 'existing' (*eg* racing) animals might give a biased measure of association between the exposure and disease. The premature removal of animals from the original group might be highly correlated with the exposure factor and the outcome, thus leaving the study group as a biased subgroup from the target population. Whenever selective survival is likely to be an issue, the study group(s) should be drawn from animals 'ever' in the herd (or ever raced) during a specified time period, not just from animals that are in the herd (or are racing) at the start of the current study period.

As another example, survival bias can be very common if prevalent cases are used in a case-control or cross-sectional study. If the duration of survival after the disease occurs differs by exposure status, then there will be bias in a cross-sectional study design. Partly for this reason, case-control studies should include only incident not prevalent cases.

Unintentional selection bias from factors affecting entry or survival (*ie* isolation of bacteria) might be at play in many temporal studies of antimicrobial resistance patterns. In this instance, the selection bias arises from using data based on isolates obtained from clinically ill subjects. Because many of the isolates would have been exposed to antimicrobials prior to culturing of tissue specimens, the number and type of bacterial isolates, and their level of antimicrobial resistance (or minimal inhibitory concentrations) might be more a function of what antimicrobials are used and how effective they are at reaching and killing susceptible organisms in the tissue samples that get cultured than of the range of pathogenic organisms or their level of antimicrobial resistance. As the objective of the study is largely descriptive, only samples obtained prior to treatment from randomly selected subjects can provide valid insight into the extent of antimicrobial resistance in the target population. The impact of antimicrobial use on the level of resistance in treated and non-treated subjects could be studied in purposively chosen subjects.

12.3.5 Detection bias

In cohort studies, detection bias is best viewed as a misclassification or confounding bias. It can arise if those assessing the outcome know the exposure status of the study subject and if they alter their assessment of the outcome because of that knowledge. In case-control studies, the central issue in detection bias is one of selection in that animals that have the disease of interest might be misclassified as not having that disease because they were less likely (or never) to be examined for the disease (see section 12.6).

This potential bias is of concern when a large percentage of the cases in a case-control study would be found (and therefore be identified as potential study subjects) only after examination in a screening or diagnostic process where participation is influenced by exposure status (*ie* the act of being assessed is influenced by the exposure status). Given this scenario, the issue is how best to select controls. A frequently suggested guideline is that the controls should be non-cases that have undergone the same screening, but the nature of the exposure, disease and the context of diagnostic testing need to be considered. We give three examples related to detection bias.

Detection bias was at the root of protracted discussions about the appropriate control group for a series of uterine cancer cases in a cancer registry (a secondary-study base). Uterine cancer often leads to bleeding and this bleeding would lead women to request a gynecological examination. Women on estrogen also tended to evidence bleeding and therefore would be examined more frequently than women not on estrogen. Hence, the possibility of detection bias was raised. One set of researchers argued that the controls should be restricted to those women who had been examined because of bleeding and found negative for cancer. The other set of investigators argued that the cancer registry used for the study eventually would contain all of the uterine cancer cases whether they were found subsequent to the examination or not and that the controls should be derived from all women in the registry that had other gynecological cancers. The latter turned out to be the correct approach. The major reason was that many cases of uterine cancer listed in the registry were diagnosed outside of the screening programme and almost all cases did end up being recorded in the registry. The lesson we can learn from this

example is that we should not enforce the general principle that controls should have undergone the same testing regime as the cases unless only a small percentage of cases with the outcome of interest would be diagnosed outside of the testing programme.

As a related example, hip dysplasia in dogs is a condition that is rarely diagnosed without radiographic evidence. Thus, having a radiograph is usually a prerequisite for the diagnosis. In early studies of the association between breed and hip dysplasia, all dogs without recorded hip dysplasia in secondary-base registries were included as controls whether or not they had been radiographed. If the act of being radiographed was related to breed and the presence of dysplasia, then some breeds would be radiographed more frequently leading to more diagnoses of hip dysplasia and over-representation of those breeds in the case-series of the study. This stimulated one group of researchers to perform a case-control study in which the potential control dogs had to have been radiographed in a manner that would lead to a diagnosis of hip dysplasia being recorded if it were present. This included dogs that had been radiographed subsequent to accidental injury. Although this might not have been a totally unbiased control group, the effect of increased probability of detection as a result of being radiographed would be equal in the case and control series. Given that few cases of hip dysplasia would be diagnosed without radiographs, restricting the study population to this group would be the appropriate study design (Martin et al, 1980).

A third example of concern over detection bias that was related to misclassification of the outcome was investigated by Singer et al (2001). These workers were selecting birds with avian cellulitis in the slaughter plant using the presence of certain gross lesions as indicators that the birds had the disease, and then culturing these birds for specific strains of *Escherichia coli*. Their concern was that if certain strains of *E. coli* only produced lesion(s) that were not being detected visually, then these birds would not be selected. Hence, only a biased subset of the *E. coli*-caused lesions would be detected. These workers developed a method to assess possible selection bias based on comparing the findings in the birds that they detected with findings in birds detected independently by the USDA inspectors. In general, it is desirable to have a sensitive and specific set of inclusion criteria when selecting study subjects.

12.3.6 Admission risk bias

Admission risk bias has haunted the validity of secondary-base case-control studies, and is the basis of **Berkson's fallacy**. In this instance, the probability of admission to the registry (secondary-study base) is related to both the disease and the exposure. Thus, the controls might not reflect the actual exposure status of the population from which the cases arose, and there might be an excess (or deficit) of exposure in the controls selected from the registry relative to the target population. While we are aware that this bias has perplexed and continues to perplex the designers of case-control studies, it is important to try and obtain quantitative estimates of the likely degree of bias that different comparison groups might produce. Breslow and Day (1980) note that it is nearly impossible to assess selection bias in any given secondary-base study. This necessarily constrains the inferences that should be drawn based on any given secondary-base case-control study but it does not invalidate the approach. Because we

are rarely fully aware of potential relationships between the exposure and other diseases (*ie* non-cases) that might serve as controls, individual researchers must do their best to counteract the potential bias. In addition, consistency of findings in different secondary-base case-control studies is crucial to good causal inferences.

The major reason for developing Example 12.2 is to demonstrate the impact on associations when exposure impacts on admission to a clinic or registry. Subsequently, we develop guidelines for selecting cases and controls in a manner to minimise the magnitude of bias. In designing case-control studies, we try to counteract this potential bias when specifying the groups that are eligible to serve as sources of controls by excluding all groups of non-cases thought to have an association with exposure. In terms of the direction of bias, as the example will demonstrate, if the risk of hospitalisation (*ie* being in the registry) is greater for the disease of interest than the average risk for the potential controls, the sample odds ratio will be less than the population odds ratio. Thus, if the study data leads to a statistically significant odds ratio, the true association in the source population would be even stronger.

12.4 Reducing selection bias

Most of the specific recommendations for preventing selection bias are contained in the study design chapters and will not be repeated here. However, being aware of the potential pitfalls in selecting study subjects, and conceptualising how these pitfalls might apply to selection of study subjects from the proposed target population is the first step in prevention. In any event, from a selection bias point-of-view, the comparison group in observational studies need not be similar to the exposed (or case) group in all respects except for the exposure (disease) of interest, but rather just with respect to the factors related to the outcome (exposure) that might lead to being included in the study. In a case-control study, if the cases are more likely to be included if exposed, then the controls should have the same extent of bias in the selection procedure (albeit this is hard to implement precisely). In cohort studies where explicit exposed and non-exposed groups are selected, care needs to be taken when selecting the comparison group, and due consideration should be given to minimising loss-to-follow-up, or non-response bias. In case-control studies the principle for control selection is that they should represent the proportion exposed, or the exposure time, in the source population. This is chiefly a problem in secondary-base studies and to circumvent it, we implement the guideline of not selecting controls from non-case diagnostic categories that might be associated with the exposure. In addition, case-control studies should rely on only incident cases, and controls should come from the same source population as the cases. Even with all these precautions, care must be taken in making broad inferences from a single case-control study using secondary databases.

12.4.1 Correcting selection bias

For valid and effective control of selection bias one of two conditions needs to be met: the factors associated with selection must be antecedents of both exposure and disease, or the distributions of exposure and disease must be known in the source population.

Example 12.2 Selection bias in a secondary-base study

The population structure

Suppose you are investigating whether vaccination in dogs within two months of admission to the hospital is associated with autoimmune disease (AID). We will denote vaccination within two months of admission as $E+$. We denote autoimmune disease as D_1 and the group of control dogs (in this case, we will use dogs with acute injuries) as D_0. The latter group would normally be composed of dogs having one or more of a combination of many diseases not associated with vaccination, but not dogs with AID.

We will assume that the structure of the target population with respect to E, D_1 and D_0 is known and is based on the following frequencies:

- 10% of dogs are vaccinated within the last two months; $p(E+)=0.1$
- 1% of dogs develop AID per year independent of E status; $p(D_1)=0.01$
- 3% of dogs get acute injury per year, independent of E status; $p(D_0)=0.03$
- there are 100,000 dogs in the target population.

Under this scenario, the population structure is shown below:

AID, acute injury and exposure status of the target canine population

	E+	E-
D_1	100	900
D_0	300	2700

The *OR* in this population=(100*2700)/(300*900)=1, as expected because the D_1 and D_0 risks were independent of E status.

Now, assume that the probabilities of an ill dog from the target population being admitted to the hospital in a risk-based case-control study are as follows:

$$p(H|D_1+E+)=0.80$$
$$p(H|D_0+E+)=0.40$$
$$p(H|D_1+E-)=0.60$$
$$p(H|D_0+E-)=0.20$$

These admission risks lead to the registry population structure shown below.

(*continued on next page*)

Example 12.2 *(continued)*

Observed data on AID, acute injury and vaccination history in the hospital records

	E+	E-	
D_1	80	540	620
D_0	120	540	660

The *OR* in the study (or registry) population is 80*540/120*540=0.67. This bias is reflected in the OR_{sf} and the sampling odds. Had we not known the admission risks, we might have concluded that vaccination was preventive for AID. To cast these data in terms of our sampling fractions from the original target population – in this instance, these are admission risks:

$$sf_{11} = 80/100 = 0.8$$
$$sf_{12} = 540/900 = 0.6$$
$$sf_{21} = 120/300 = 0.4$$
$$sf_{22} = 540/2700 = 0.2$$

and the *OR* of the sampling fractions is 0.67. Note that the sampling odds for disease within the exposed and non-exposed are:

$$so_{D_1|E+}=0.8/0.4=2$$

and

$$so_{D_0|E+}=0.6/0.2=3$$

leading to the ratio of so=2/3=0.67 reflecting the amount of bias in the measure of association.

Under the first condition, the bias can be controlled in a manner similar to confounding; for example if owner income might cause selection bias in a secondary base case-control study it can be measured and controlled in the analysis. Under the second condition, the bias can be removed by using the reciprocal of the sampling fraction odds ratio. However, if the covariate is an intermediate variable, or an effect of disease, and associated with selection, it should not be controlled as if it were a confounder as this actually would increase the bias. Despite our best efforts to use these approaches, as we often have only vague ideas about the magnitude of the sampling fractions, or sampling odds, quantitative adjustments to the data to correct selection bias usually are not possible.

12.5 Information Bias

The previous discussion was concerned with whether the study subjects have the same exposure-disease association as that which exists in the target population. It assumed that disease and exposure were correctly classified. We now move on to discuss the effects of incorrectly classifying, or measuring, the study subjects' exposure, extraneous factors and/or outcome status. If we are concerned with incorrect classification of categorical

variables, the resultant bias is referred to as **misclassification** bias. If the variables of interest are continuous, then we term the erroneous result as **measurement error** or bias. Information bias is a major problem in epidemiological studies as it can produce errors in our estimates of association, the magnitude and direction of which might not be intuitive. Also, the errors in classification, or measurement, can affect different measures of association differently (*ie* risk ratio versus risk difference). Hence, for our purposes, we will focus only on the effects of information bias on relative measures of association (risk ratios and odds ratios). In the discussion that follows, the study subject could be an individual or a group of individuals, such as a herd.

Several methods to assess and/or correct for misclassification and measurement error have been described, but (apparently) few researchers have applied them. We will discuss them subsequently, but first we review the basics of misclassification – the most studied of information biases.

12.6 Bias from misclassification

Misclassification bias results from a rearrangement of the study individuals into the incorrect categories because of errors in classifying exposure or outcome or both. Non-compliance with an assigned treatment in a clinical trial can also produce misclassification bias, because the subject was not actually receiving the treatment specified. With categorical measures of exposure, outcome, or other covariates, especially dichotomous measures (*ie* exposed or not, diseased or not), the errors of classification can be described in terms of sensitivity and specificity as shown in Chapter 5. Here, sensitivity (*Se*) for a given condition is the probability that an individual with the condition will be classified as having the condition. The complement of *Se* is the false negative fraction (*FNF*). Specificity (*Sp*) is the probability that an individual without the condition will be classified as being without the condition. The complement of *Sp* is the false positive fraction (*FPF*).

12.6.1 Non-differential misclassification of exposure

The tabular data layout is the same as shown in Table 12.1. The true cell values are represented by a_1, b_1, a_0, and b_0, with m_1 diseased and m_0 non-diseased, n_1 and n_0 exposed and non-exposed subjects, respectively. The observed cell values will be denoted with $'$ (*ie* the prime symbol): a_1', b_1', a_0', and b_0'.

If misclassification of the exposure and outcome are independent (*ie* errors in classifying exposure are the same in diseased and non-diseased animals and vice-versa when classifying disease in exposed and non-exposed subjects) then the misclassification is called **non-differential**. With non-differential (non-systematic) misclassification, for disease classification, we have

$$Se_{D|E+}=Se_{D|E-}=Se_D \qquad \text{and/or} \qquad Sp_{D|E+}=Sp_{D|E-}=Sp_D$$

where Se_D is the sensitivity of disease classification and Sp_D is the specificity of disease

classification. For exposure classification we have

$$Se_{E|D+}=Se_{E|D-}=Se_E \qquad \text{and/or} \qquad Sp_{E|D+}=Sp_{E|D-}=Sp_E$$

where Se_E is the sensitivity of exposure classification and Sp_E is the specificity of exposure classification.

How do these errors relate to our observed data? Well, using error frequencies for exposure misclassification denoted as Se_E and Sp_E and assuming $Se_{D+}=Sp_{D-}=100\%$, we would have the observed data shown in Table 12.2 and Example 12.3.

Table 12.2 Relationship between the number of correctly and incorrectly classified subjects by exposure status

True number	Observed number
a_1	$a_1'=Se_E*a_1+(1-Sp_E)*a_0$
a_0	$a_0'=(1-Se_E)*a_1+Sp_E*a_0$
b_1	$b_1'=Se_E*b_1+(1-Sp_E)*b_0$
b_0	$b_0'=(1-Se_E)*b_1+Sp_E*b_0$

Example 12.3 Impact of non-differential misclassification of exposure

We first assume that there is no misclassification, hence the true study population structure in this example is:

	Exposed	Non-exposed	Total
Diseased	90	70	160
Non-diseased	210	630	840
Total	300	700	1000

The true *OR* is 3.86. If we now assumed an exposure sensitivity of 80% and an assumed exposure specificity of 90%, we would expect to have the following observed cell numbers (calculations shown):

	Exposed	Non-exposed	Total
Diseased	90*0.8+0.1*70=79	70*0.9+90*0.2=81	160
Non-diseased	210*0.8+630*0.1=231	630*0.9+210*0.2=609	840
Total	300* 0.8+700*0.1=310	700*0.9+300*0.2=690	1000

Note Exposure misclassification does not affect the disease status totals, only the exposure category totals. As predicted, with non-differential errors the odds ratio has been reduced from 3.86 to 2.57.

For dichotomous exposures and outcomes, non-differential errors always bias the measures of association towards the null (given that the minimum of $Se+Sp > 1$) as shown in Example 12.3. In most studies, researchers assume that any errors of classification are non-differential. However, in case-control studies, the assumption of non-differential errors is often open to question. **Recall bias** in case-control studies is one illustration of (likely) differential errors in that 'affected' subjects (*ie* cases) might have an increased sensitivity, and perhaps a lower specificity than non-affected subjects in recalling previous exposures. The effect of differential errors on the observed measure of association is difficult to predict.

The frequency of both types of error can be reduced by using clear and explicit guidelines and 'double-checking' the exposure status whenever possible (*eg* seeking laboratory confirmation, or other confirmatory records, for exposure status). Because the results of non-differential misclassification are predictable, we often recommend 'blind' techniques to help ensure that the errors are equalised. However, our first preference should be to reduce the frequency of errors rather than depend on blindness to balance the frequency of errors.

12.6.2 Differential misclassification of exposure or outcome

If the errors in exposure classification are related to the status of the outcome under study, the errors are called **differential**. Here, the Se_D and Sp_D of classifying disease status differs over exposure levels, and/or the Se_E and Sp_E differ by disease status:

$$Se_{D|E+} \neq Se_{D|E-} \quad \text{and/or} \quad Sp_{D|E+} \neq Sp_{D|E-}$$

and for exposure classification

$$Se_{E|D+} \neq Se_{E|D-} \quad \text{and/or} \quad Sp_{E|D+} \neq Sp_{E|D-}$$

The resulting bias in the measure of association might be in any direction (*eg* an effect might either be exaggerated or underestimated). A few minutes with a spreadsheet playing 'what-if' will help convince you of this. As noted, recall bias is often cited as an example of differential misclassification.

12.6.3 Correcting for non-differential exposure misclassification

If the error frequencies in exposure classification are known, we can correct the observed classifications. In most circumstances however, we would require a validation study with a gold standard procedure to measure exposure in order to estimate sensitivity and specificity. In the absence of such a validation process, one can use a sensitivity analysis – a 'what-if-the-errors-were-known' process – to investigate the likely range of bias.

Assuming non-differential errors, we can use the following approach to reclassify the study population. As $b_1'+b_0'=b_1+b_0=m_0$, we can solve for b_1 as:

$$b_1 = \frac{b_1' - FPF_E * m_0}{(Se_E + Sp_E - 1)} \qquad \textit{Eq 12.4}$$

where FPF_E=(1-Sp_E). Similarly, we can solve for a_1 as:

$$a_1 = \frac{a_1' - FPF_E * m_1}{(Se_E + Sp_E - 1)} \qquad \textit{Eq 12.5}$$

with b_0 and a_0 determined by b_0=m_0-b_1 and a_0=m_1-a_1 (Rothman and Greenland, 1998). Applying these corrections to our data in Example 12.3, we have:

$$a_1 = \frac{79 - 0.1*160}{(0.8+0.9-1)} = 90 \quad \text{and} \quad b_1 = \frac{231 - 0.1*840}{(0.8+0.9-1)} = 210$$

which can then be used to complete the 2X2 table and compute the true value of the *OR* (*ie* 3.86). This process can be used to correct for differential errors in exposure status by repeating the process separately in each of the case and control groups using the appropriate estimates of Se_E and Sp_E.

This approach to making corrections for exposure misclassification can be applied to data from all study types with the reminder that in a rate-based cohort study, we replace the *b*s with *t*s (animal-time at risk). However, the sobering lesson about classification errors is that small changes in the estimated sensitivity and specificity can produce large changes in the observed data; hence the variability in the data arising from these small changes can be much more dramatic than changes that would be expected from sampling variation. Often when attempting to correct for these errors the sensitivity and specificity estimates could produce 'impossible' results. This means that the values used are not consistent with the data, so the actual error risks must differ from the estimated values being used for the corrections.

Another observation about exposure classification errors is that when exposure prevalence is low, lack of specificity produces more errors than lack of sensitivity. The good news in this instance is that even if exposure is selectively recalled among cases (*ie* recall bias giving a higher Se_E and/or a lower Sp_E in cases in comparison to the controls), the observed measures of association usually will be biased towards the null. Thus, in the presence of recall bias, if an association is found in the study population, it is likely to be even stronger in the target population.

12.6.4 Non-differential misclassification of disease-cohort studies

Here the same concepts of classification errors arise as with exposure misclassification except that we now focus on errors in classifying health status in cohort studies. Example 12.4 shows the impact of non-differential disease misclassification. As with exposure misclassification, disease misclassification in a cohort study can be corrected. Given that Se_D, Sp_D, FNF_D=1-Se_D and FPF_D=1-Sp_D are the health status classification probabilities we can solve for a_1 as:

$$a_1 = \frac{a_1' - FPF_D * n_1}{(Se_D + Sp_D + 1)} \quad \text{and} \quad a_0 = \frac{a_0' - FPF_D * n_0}{(Se_D + Sp_D + 1)} \qquad \textit{Eq 12.6}$$

Example 12.4 Impact of non-differential misclassification of disease

We assume the same target population structure as in Example 12.3. With a disease-classification system having an assumed sensitivity of 80% and an assumed specificity of 90% we would, in expectation, observe the following cell numbers:

	Exposed	Non-exposed	Total
Diseased	90*0.8+0.1*210=93	70*0.8+630*0.1=119	212
Non-diseased	210*0.9+0.2*90=207	630*0.9+70*0.2=581	788
Total	300	700	1000

Note that disease misclassification does not affect the exposure status totals. Also note that the effect of the error has been to reduce the odds ratio from its true value of 3.86 in the source population to 2.19 in the study group. Even though the error risks used here are the same as those used for exposure misclassification (see Example 12.3), the impact on the odds ratio differs because the prevalence of disease differed from the prevalence of exposure. Thus, the impact of errors depends both on the error frequencies and the prevalence of the event (*ie* exposure or outcome) being misclassified.

As we assume a risk-based cohort study in this example, we can correct for disease misclassification. This procedure should not be applied to correcting case-control data for diagnostic errors (see section 12.6.5).

For a_1 and a_0, we have

$$a_1 = \frac{93 - 0.1*300}{(0.8+0.9-1)} = 63/0.7 = 90 \quad \text{and} \quad a_0 = \frac{119 - 0.1*700}{(0.8+0.9-1)} = 49/0.7 = 70$$

We can now complete the 2X2 table and estimate the true population odds ratio. This process can be used to correct for differential errors by separately performing the calculations in the exposed and non-exposed groups with appropriate estimates of Se_D and Sp_D.

Now, b_1 and b_0 can be obtained by subtraction from the marginal exposure totals (n_1 and n_0, respectively). As they stand, these formulae can be applied to risk-based cohort study subjects. They also apply to rate-based cohorts if we substitute t_1 and t_0 for the n_1 and n_0 values and use the concept of false positive rate instead of *FPF* (Rothman and Greenland, 1998).

12.6.5 Non-differential misclassification of disease case-control studies

Because of the often unknown sampling fractions in case-control studies, the previous formulae do not apply to this design unless Sp_D=1.00. In that instance, imperfect disease sensitivity does not bias the *RR* or *IR*, and only biases the *OR* if disease frequency is common. The key here is that it pays to verify the diagnoses of the cases so that there are no false positive cases, as the association measures will not be biased even if the diagnostic Se_D is less than 100%.

Example 12.5 Correcting classification errors in case-control studies

Suppose we conduct a case-control study on the misclassified population used in Example 12.4. We will assume the same Se_D=0.8 and Sp_D=0.9 as before, hence the observed population structure will be:

	Exposed	Non-exposed	Total
Diseased	93	119	212
Non-diseased	207	581	788
Total	300	700	1000

Now, for our study, we assume that we take all the apparent cases (sf_{D+}=1.0) and 20% (sf_{D-}=0.2) of the apparent non-cases as controls. Disregarding sampling error, our misclassified study population will have the structure shown below (the numbers are given to 1 decimal place to avoid rounding errors – you might think of these as 'expected' observed values).

	Exposed	Non-exposed	Total
Diseased	72 TP+21 FP=93.0	56 TP+63 FP=119.0	212.0
Non-diseased	37.8 TN+3.6 FN=41.4	113.4 TN+2.8 FN=116.2	157.6
Total	134.4	235.2	369.6

TP and TN relate to true disease positives and true disease negatives, respectively. Similarly, FN and FP relate to false negative and false positive cases, respectively. The odds ratio is, as before, 2.19. However, in order to 'back-correct' the observed values we would need to use the actual sensitivity and specificity based on the case-control data, not the original levels of Se_D=0.8 and Sp_D=0.9.

The actual sensitivity for these data is

$$(72+56)/(72+56+3.6+2.8)=128/134.4=0.95$$

and the actual specificity for these data is

$$(37.8+113.4)/(37.8+113.4+21+63)=0.64$$

These classification probabilities could be found directly using the formulae in section 12.6.5. For the case-control sensitivity we have

$$Se_{cc}=0.8/(0.8+0.2*0.2)=0.8/0.84=0.95$$

and the case-control specificity will be

$$Sp_{cc}=0.2*0.9/(0.1+0.2*0.9)=0.18/0.28=0.64$$

As indicated earlier, these values are not even close to the population *Se* and *Sp* values. Typically as seen here, sensitivity is increased and specificity is decreased relative to the levels in the source population.

When $Sp_D < 1$ then, in a case-control study, the source population is misclassified. If we take all the apparent cases for our study, we will be including $Se_D * M_1$ of the true cases and $FPF_D * M_0$ false positives as cases. Usually we take only some (sf) of the non-cases as controls, hence ultimately, we will include only a very small number of false negative cases ($sf * FNF_D * M_1$) and a much larger number of true non-cases ($sf * Sp_D * M_0$). Thus, in the study group the case-control sensitivity will be

$$Se_{cc} = Se_D / (Se_D + sf * FNF_D) \qquad \textit{Eq 12.7}$$

and the case-control specificity will be

$$Sp_{cc} = sf * Sp_D / (FPF_D + sf * Sp_D) \qquad \textit{Eq 12.8}$$

As shown in Example 12.5, both of these could be very far from the true population values of sensitivity and specificity (see Rothman and Greenland, 1998, p. 352). Thus external estimates of *Se* and *Sp* cannot be used to correct misclassification in case-control studies. Also, estimates of diagnostic *Se* and *Sp* obtained from case-control study subjects cannot be used to estimate the population *Se* and *Sp* values.

12.6.6 Misclassification of both exposure and disease

So far we have examined misclassification of either exposure or disease but not both simultaneously. For purposes of demonstration of joint misclassification, we will assume independence of errors, by which we mean:

probability of joint misclassification = product of individual misclassification probabilities

A 'pedestrian' way of understanding joint misclassification is to consider each of the four possible cross-classifications of exposure and disease in turn (*ie* individuals in each correctly classified cell of the 2X2 table can be misclassified into up to three incorrect cells). For example, the actually exposed and diseased individuals in the population (A_1) will be classified into the four cells of a 2X2 table with the frequencies shown in the first column of Table 12.3. Similarly, the classification of the A_0, B_1 and B_0 subjects, in terms of B_1', A_0', and B_0', follows the pattern shown in the next three columns.

Table 12.3 The probability of being classified into each exposure-disease category (eg A_1') according to the true exposure-disease state[a] (eg A_1)

Number classified by exposure-health status	Number of subjects by true exposure-health status			
	A_1 (E+ and D+)	A_0 (E- and D+)	B_1 (E+ and D-)	B_0 (E- and D-)
A_1'	$Se_{E\|D+} * Se_{D\|E+}$	$FPF_{E\|D+} * Se_{D\|E-}$	$Se_{E\|D-} * FPF_{D\|E+}$	$FPF_{E\|D-} * FPF_{D\|E-}$
A_0'	$FNF_{E\|D+} * Se_{D\|E+}$	$Sp_{E\|D+} * Se_{D\|E-}$	$FNF_{E\|D-} * FPF_{D\|E+}$	$Sp_{E\|D-} * FPF_{D\|E-}$
B_1'	$Se_{E\|D+} * FNF_{D\|E+}$	$FPF_{E\|D+} * FNF_{D\|E-}$	$Se_{E\|D-} * Sp_{D\|E+}$	$FPF_{E\|D-} * Sp_{D\|E-}$
B_0'	$FNF_{E\|D+} * FNF_{D\|E+}$	$Sp_{E\|D+} * FNF_{D\|E-}$	$FNF_{E\|D-} * Sp_{D\|E+}$	$Sp_{E\|D-} * Sp_{D\|E-}$

[a] If errors are non-differential then, for example $Se_{E|D+} = Se_{E|D-} = Se_E$ as in section 12.6.1.

Similarly, we can examine the observed cell quantities (C_i') in terms of the classification errors that could contribute to them. For example, if we examine the observed value

A_1', as expected, it is the sum of the error frequencies (shown in row 1 in Table 12.3) after each is multiplied by the respective true number of subjects in that category. We can repeat this for all four cells (exposure-outcome combinations) and summarise the observed classification for each cell as:

$$C_i' = \sum_{j=1}^{4} P_{ij} C_j \qquad \textit{Eq 12.9}$$

where P_{ij} is the probability of being classified into cell i when the true classification is cell j. The C_j multipliers are the counts of individuals in each correctly categorised cell of the 2X2 table. Summing these products across the columns (*ie j*) gives the observed cell values for the row (*ie i*). In matrix form, we could write this as $C'=PC$ and hence, the corrected classification is $C=P^{-1}*C'$. For more details, see Table 12.3 in Kleinbaum et al (1982) p 231, and Rothman and Greenland (1998) p 353.

As noted earlier, if one works through many examples using realistic error rates, then it becomes clear that misclassification bias can create much more uncertainty in our measures of association than sampling variation. Thus, we need to pay a great deal of attention to reducing these errors whenever possible. Again, the latter approach to correcting classification errors cannot be used for case-control studies; it is valid for risk-based cohort, but not rate-based cohort studies.

12.6.7 Misclassification of extraneous variables

If a confounder is measured with error, it is impossible to fully control for its confounding effect (Marshall and Hastrup, 1996). Thus, measurement error in the confounder can produce bias in the exposure effect estimates. The bias can be large if the true effect of the exposure is weak and the confounder is strongly related to exposure and the outcome. For example, with very strong confounding, a 10% misclassification of the confounder can reduce the true association measure by almost 50%. Surprisingly, measurement error in the exposure variable mitigates some of the effects of measurement error in the confounder (largely because it often reduces the correlation between the confounder and the exposure). However, in the face of misclassification of the confounder it becomes difficult to know whether or not one should control for the confounder (see Chapter 13). A general recommendation is that the impact of controlling an extraneous variable should only be investigated when no misclassification is present or after adjustments for the errors have been made. In reality, because of the practical difficulties, this recommendation has been followed only infrequently.

Misclassification of a non-confounder can make it appear to be a confounder. For example, assuming the exposure has an effect on the outcome, that the extraneous variable is associated only with the exposure, and that misclassification error of the extraneous variable sums to less than 1 (*ie* $FNF+FPF<1$ – all reasonable assumptions), then there is more bias in the 'corrected' odds ratio than in the crude odds ratio. The same is true if the extraneous variable is related only to disease. The general rule for treating a variable as a confounder is, if misclassification is non-differential within each stratum of the extraneous variable but varies across strata, then treating the variable as a confounder will usually reduce bias. In some instances the bias might increase, but it

will be towards the null. Clearly, one must focus on reducing measurement error in all variables, not just confounders, if valid analyses and inferences are to be made.

12.7 MISCLASSIFICATION OF MULTINOMIAL EXPOSURE OR DISEASE CATEGORIES

With several levels of exposure, the effects of classification errors are less predictable than with dichotomous variables (Veierod and Laake, 2001). Non-differential misclassification might bias measures of association in intermediate exposure levels away from the null, and might even reverse the direction of the odds ratios for these levels. This becomes an important issue when we use regression models because these models allow for error in the measurement of the outcome but assume no error of measurement of the predictor variables. Non-differential underestimation of exposure at high levels might cause a threshold effect of exposure to appear as a dose-response relationship. Likewise, non-differential misclassification of both E and D status when the errors are **not independent** might lead to bias away from the null, particularly when the prevalence of both exposure and disease are low.

12.8 VALIDATION STUDIES TO CORRECT MISCLASSIFICATION

Sometimes it is feasible, and often advisable, to select a subsample of study subjects and verify their exposure and/or disease status. Recall that, for direct estimates of sensitivity and specificity, we are determining the probability of the observed state (D'), given that we know the true state of the individual (D). That is:

$$p(D'=1|D=1)$$

whereas when correcting for misclassification, we are attempting to determine the probability of the true state, given knowledge of the observed state:

$$p(D=1|D'=1)$$

Thus, when validating the subsample, we attempt to determine the true status of the individuals given their observed classification. In the content of screening tests, the latter is called a predictive value (see Chapter 5). In general, we can write:

q_{ij}=probability of being in the correct i^{th} cell when classified into the j^{th} cell.

Thus, given that the subject is classified into cell b_1', we can determine the probability that the subject is really in cell a_1 (for example), using our validation study. To implement this, we can think of a 4X4 matrix of predictive values with the rows being the observed status and the columns the true status, as we did previously with sensitivity and specificity (Table 12.3). We can then write:

$$C_i = \sum_{j=1}^{4} q_{ij} C_j' \qquad \textit{Eq 12.10}$$

as the formula for establishing the true cell values. Here the q_{ij} are predictive values obtained from a validation study (not the *Se* and *Sp* values shown in Table 12.3). Once these are obtained, no matrix inversion is necessary to obtain the corrected values so

this approach is easier to use than the formal approach based on knowing *Se* and *Sp*. However, because this is a predictive value approach and predictive values might vary with true prevalence (of disease, exposure, or both) one should stratify the validation results across major confounders or risk factors.

As noted previously, the major problem with post-hoc adjustments of misclassification is that they are very sensitive to changes in the estimates of the error rates to be used in the correction process. Thus, unless there is an extremely thorough validation procedure, the estimates of error might vary sufficiently such that very different 'corrected' results could arise from applying a range of sensible choices of the correction factor. A few minutes trying 'what-if' adjustments should convince you of this.

Validating the test or survey instrument prior to its widespread use is certainly preferable to trying to correct for misclassification error or measurement error after the fact. An example of this is available in veterinary research (Nespeca et al, 1997).

12.9 Measurement error

Errors in measuring quantitative factors can lead to biased measures of association also. This bias can arise either because the variable is not measured **accurately** (*ie* a systematic bias), or due to a lack of **precision** (see section 5.2.2). In turn, lack of precision might arise from either variability in the test *per se*, or because the substance being measured varies within an individual (for physiological reasons) and consequently, repeated measures are needed to provide a valid overall indicator of the status of the individual (*eg* a mean of two or more samples).

Recent work has shed considerable 'new light' on the issue of measurement error, and the general approach to correcting measurement bias is as follows. Suppose we have two quantitative exposure factors and we wish to estimate their association with a binary or continuous outcome. Allowing that the Y variable could represent the logistic transform of a binary outcome, or a continuous outcome variable, we could express the uncorrected 'naive' model as:

$$Y=B_{0u}+B_{1u}X'_1+B_{2u}X'_2 \qquad \textit{Eq 12.11}$$

where the subscript 'u' indicates that the coefficients are biased because the predictor variables, here denoted as X', are measured with error. One approach for 'correcting' the model results is to perform a validation study on a random subset of the study subjects, and use only the data from this subset in the analysis; we will call this a **validation subset estimate** (VSE) approach. This seems (and is) very wasteful of a lot of data (data on the subjects not in the validation subsample) and might lead to small sample bias (the overestimate of the magnitude of effect), because only large associations will be deemed significant in small sample size studies. As another approach, Rosner et al (1990) developed a procedure called the **regression calibration estimate** (RCE). In this method, we obtain a random subset of the study subjects and perform a validation study so that the true values for X_1 and X_2 are obtained. Now, assuming non-differential measurement errors, we regress each true X variable on the set of observed predictor

variables. That is:

$$X_1=B_0+\lambda_{11}X'_1+\lambda_{12}X'_2 \qquad \textit{Eq 12.12}$$

and

$$X_2=B_0+\lambda_{21}X'_1+\lambda_{22}X'_2 \qquad \textit{Eq 12.13}$$

One then 'corrects' the biased coefficients (B_{1u}) using the matrix of correction factors (λs) multiplied by the biased coefficients as shown in Eq 12.14.

$$B_1, B_2 = B_{1u}, B_{2u}, \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} \qquad \textit{Eq 12.14}$$

The regression models chosen for the X variables depend on the assumed distribution of the X variables (*ie* continuous or binary), and the validity of the approach to correcting measurement errors, in part, depends on the fit of the above models. Spiegelman et al, 1997 have discussed using this approach with an 'alloyed' gold standard.

More recently, Robins et al (1995) developed a **semi-parametric estimate** (SPE) approach for correcting measurement error. It is similar to the RCE approach except that no assumptions are made about the distribution of the Xs. In general, if one is aware of the distributions for the true X variables, the RCE approach is more efficient, but the SPE does not depend on knowing these and hence, fits a wider set of applications.

Sturmer et al (2002) have assessed the performance of these three procedures for correcting measurement error in case-control studies. We are not aware of similar studies for cohort designs but will assume that similar findings might apply. The additional factors studied included: the effects of an imperfect gold standard for validation, correlation of errors, magnitude of true effect, and amount of 'error' in measurement. We summarise the findings below.

Naive model

- very sensitive to magnitude of error – moderate errors produce strong attenuation of effect
- the bias increases with the magnitude of the true effect
- the bias in one coefficient depends on the amount of error in the other variable
- correlated errors could cause bias away from the null for one of the variables
- in the presence of differential errors, the direction of the bias is unpredictable.

VSE models

- standard errors of estimate are large, but most biases are reduced
- not sensitive to correlated or differential errors, nor the true effect size
- hampered by the presence of an imperfect gold standard, and
- suffers from a relative lack of power, and the standard error of estimates are large due to the reduced number of study subjects.

RCE models

- often the bias was similar to, but less than, in naive models
- very sensitive to differential errors and to large error variances and should not be used in these situations.

SPE models

- similar to RCE in correcting the biases
- not sensitive to the magnitude of the error variance
- insensitive to differential errors
- easily accommodates correlated errors, but
- user-friendly software is needed because of the complexity of implementing this approach.

The two most influential factors are the error magnitude and the presence of differential errors. RCE is the easiest approach and uses all of the data, but is limited in its applicability. With the exception of when the gold standard is flawed, the VSE approach performed surprisingly well. Based on this, Chatterjee and Wacholder (2002) have suggested that the VSE approach with equal numbers of cases and controls in the validation sample is a viable method for maximising the precision of estimates with a fixed cost. Very large validation sample sizes are needed to obtain precise estimates of effect in this approach.

As a more generalised approach, Chatterjee and Wacholder suggest the use of two-stage designs where data on inexpensive variables are collected on all study participants and data on more expensive variables collected on a random subset. This approach is outlined in Chapter 10. The best strategy for most studies is to ensure that all measurements are made as accurately and precisely as possible in the first instance. If the measurement error is small relative to the range of the X-variable values in the model, then concern over measurement error is decreased considerably. When this is not feasible, researchers should investigate the possibility of an expanded VSE approach, ensure that there are accurate data on a subset, and/or use one of the two-phase study designs.

12.10 Measurement error in surrogate measures of exposure

Often, epidemiologists focus on the effects of complex exposure factor(s). For example, in studies of the impact of air pollution from oil and gas processing emissions on cattle or wildlife health, what is the appropriate measure of air pollution? Another example is what is the appropriate measure of 'housing' as a risk factor for stereotypy in horses? In these and other examples, the exposure might be a complex mixture of agents (or factors), doses and duration, and it will take considerable thought as to what components of exposure to measure and which to ignore. For example, which of the hundreds of compounds in air pollution does one measure? The most abundant, the least expensive to monitor, or the most toxic? If a number of agents are measured, how will they be modelled? The answers to these questions (yes, there undoubtedly will be more than one correct answer) will largely involve knowing context-specific biological background information.

The decisions about surrogate measures must then be translated into what will be measured, and how the various axes of exposure will be analysed in order to achieve the study objectives. For example, will the exposure be measured and analysed on a continuous scale (the preferred option) or will it be categorised into a dichotomous or ordinal exposure variable? If levels of specific agents are highly correlated, which one should be analysed, or should a composite variable be created? Although categorising continuous data is not the preferred choice, it might reflect the reality of the exposure measurements better than the more refined measures. For example, if most levels of exposure are at or near the laboratory sensitivity of the test procedure, it might be best to dichotomise into non-exposed (for most of the data) and exposed for the limited number of measurements that are clearly above accepted levels of exposure. Of course the measured factors, being surrogates, might still fail to reflect the actual exposure. Thus, even if the variables measured are, in fact, measured without error, we need to be aware that because the variables are surrogates, we could still be left with measurement error in respect of the true exposure.

One solution might be to change the questions asked. Instead of asking about the effects of 'air pollution', ask about the effects of only one measurable component (*eg* sulphur dioxide, then factors such as H_2S or particulates would be extraneous variables), and instead of asking about 'housing' (a general portemanteau variable), ask about hours per week spent indoors in a stall. These more focused questions still require the measurement and control of other factors that might confound or interact with the exposure but the more focused answers might allow better progress towards solving the issue(s).

12.11 Misclassification and measurement errors – impact on sample size

It is apparent that classification and measurement errors can have a serious impact on the measures of association. With non-differential misclassification of categorical variables, the measures are biased towards the null. And, under classical measurement error models, the same is true for continuous variables. This has led some to conclude that in planning a study, the projected loss of power due to these errors should be considered and the sample size increased accordingly (Devine and Smith, 1998). However, if we are using the observed outcome levels from previous studies, these might be biased (towards the null) and hence, sample-size estimates based on these would be too large if the factors can be measured without error. In addition, given that the effects of differential measurement error are difficult to predict, and often the coefficients for measures of association are biased away from the null, the observed P-value will be too small. Hence, there is no consensus on adjusting sample sizes at this time. In practice, the optimal strategy is to ensure that we have minimised all sources of error in our studies.

SELECTED REFERENCES/SUGGESTED READING

1. Breslow NE Day NE. Statistical methods in cancer research vol I. The analysis of case-control studies. IARC Lyon France, 1980.
2. Chatterjee N, Wacholder S. Validation studies: Bias, efficiency, and exposure assessment. Epidemiology 2002; 13: 503-506.
3. Cheung YB. Adjustment for selection bias in cohort studies: An application of a probit model with selectivity to life course epidemiology. Epidemiology 2001; 12: 1238-1243.
4. Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: The impact of ignoring exposure measurement uncertainty. Stat Med 1998; 17: 1375-1389.
5. Ducrot C, Roy P, Morignat E, Baron T, Calavas D. How the surveillance system may bias the results of analytical epidemiological studies on BSE: prevalence among dairy versus beef suckler cattle breeds in France. Vet Res 2003; 34: 185-192.
6. Ducrot C, Calavas D, Sabatier P, Faye B. Qualitative interaction between the observer and the observed in veterinary epidemiology. Prev Vet Med 1998; 34: 107-113.
7. Greenland S, Robins JM. Confounding and misclassification. Am J of Epidemiol 1985; 122: 495-505.
8. Kleinbaum DG, Morgenstern H, Kupper LL. Selection bias in epidemiologic studies. Am J Epidemiol 1981: 113: 452-463.
9. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. Principles and Quantitative Methods. London: Lifetime Learning Publications, 1982.
10. Mangione-Smith R, Elliott MN, McDonald L, McGlynn EA. An observational study of antibiotic prescribing behavior and the Hawthorne effect. Health Serv Res 2002; 37: 1603-1623.
11. Marshall JR, Hastrup JL. Mismeasurement and the resonance of strong confounders: Uncorrelated errors. Am J Epidemiol 1996; 143: 1069-1078.
12. Martin SW, Kirby K, Pennock PW. Canine hip dysplasia: Breed effects. Can Vet J 1980; 21: 293-296.
13. Nespeca R, Vaillancourt JP, Morrow WE. Validation of a poultry biosecurity survey. Prev Vet Med 1997; 31: 73-86.
14. Robins JM, Hsieh F, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured data. J R Stat Soc B 1995; 57: 409-424.
15. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992; 136: 1400-1413.
16. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990; 132: 734-745.
17. Rothman KJ, Greenland S. Modern epidemiology. 2d ed. Philadelphia: Lippincott-Raven, 1998.
18. Sandler DP. On revealing what we would rather hide: The problem of describing study participation. Epidemiology 2002; 13: 117.
19. Singer RS, Atwill ER, Carpenter TE, Jeffrey JS, Johnson WO, Hirsh DC. Selection bias in epidemiological studies of infectious disease using *Escherichia coli* and

avian cellulitis as an example. Epidemiol Infection 2001; 126: 139-145.
20. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard". Am J Epidemiol 1997; 145: 184-196.
21. Stürmer T, Thürigen D, Spiegelman D, Blettner M, Brenner H. The performance of methods for correcting measurement error in case-control studies. Epidemiology 2002; 13: 507-516.
22. Veierod MB, Laake P. Exposure misclassification: bias in category specific Poisson regression coefficients. Stat Med 2001; 20: 771-784.
23. Wacholder S, Carroll RJ, Pee D, Gail MH. The partial questionnaire design for case-control studies. Stat Med 1994; 13: 623-634.

13

CONFOUNDER BIAS: ANALYTIC CONTROL AND MATCHING

OBJECTIVES

After reading this chapter, you should be able to:

1. Apply a working set of criteria to identify potential confounders in an observational study.
2. Use restricted sampling to prevent confounding.
3. Determine appropriate variables for control of confounding using matching and implement the matching process in a cohort study.
4. Determine appropriate variables for control of confounding using matching and implement the matching process in a case-control study.
5. Implement a valid plan for the control of confounding using analytic procedures.
6. Use a causal diagram to identify factors (confounders) needing control.
7. Apply a stratified analysis to a set of categorical variables to evaluate the presence of interaction and assess the extent of confounding.
8. Interpret the likely effect of 'controlling' extraneous factors having specified causal associations with the outcome and exposure.
9. Evaluate the potential of a non-measured confounder to bias the outcome measure.

13.1 Introduction

Confounding has been described as the mixing together of the effects of two or more factors. Thus, when confounding is present we might think we are measuring the association of an exposure factor with an outcome, but the association measure also includes the effects of one or more extraneous factors. Hence, our measure of association is **biased**, or **confounded**. For the purposes of explaining confounding, we will assume that we have identified one factor as the exposure of interest. One or more other factors that are of interest chiefly because they might distort the association measure of interest will be called **extraneous factors**. Sometimes these extraneous factors are measured; sometimes they are not. Our concern arises from the knowledge that, if these factors have an association with both the exposure and outcome of interest, failing to 'control' or 'adjust for' these relationships (see section 13.5) might lead us to inappropriate conclusions about the association between the exposure factor and the outcome. The extraneous factors that produce the bias are called **confounders** or confounding factors.

13.2 Confounding and causation

It has been argued that confounding relates to disease causation and can be explained in terms of the counterfactual observations necessary to demonstrate causation (see section 1.5.1; Greenland and Morgenstern, 2001). Recall that if we are interested in 'exposure' to a specific agent as a potential cause of a disease, we might observe the risk of the disease in individuals who are exposed (R_1). But to demonstrate causation, and measure its strength, we would need to know what the risk would have been if these same individuals had not been exposed (R_0). This is the **counterfactual state** – it does not exist, but if it did it would allow us to estimate the **true causal effect** using a measure of association such as a risk ratio (RR) or odds ratio (OR).

Although the counterfactual state is not observable, the 'ideal experiment' is judged to be the closest practical approximation to the counterfactual state chiefly because the use of randomisation provides a probabilistic approach to the balancing of factors, known and unknown, between the treated and non-treated groups. This experiment would allow us to contrast the frequency of outcome in the exposed (R_{E1}) and non-exposed (R_{E0}) subjects and closely approximate the true causal effect in an unbiased manner. However, in observational studies, the best we can hope to achieve is to obtain a non-exposed group of individuals that we assess to be as similar as possible to the exposed individuals with respect to factors that would affect the outcome, and observe their risk of the disease (R_0'). Recall this was the basis of John Snow's famous comparison of water supply to individual houses on the same streets in London during cholera outbreaks (http://www.ph.ucla.edu/epi/snow.html). In this so-called **natural experiment**, most residents did not know where their water supply came from. Nonetheless, because the non-exposed individuals might not be exactly the same as the exposed subjects, except for exposure, they might differ from the ideal counterfactual group in such a way that $R_0 \neq R_0'$. Hence, our observational measure of association will

be biased (see Example 13.1). This is the essence of confounding in terms of the lack of a true counterfactual observation. Put another way, the study groups being compared differ in the frequency of the outcome for reasons other than the exposure of interest. Our challenge is to identify the factors that 'cause' this difference and prevent them from producing a biased result by using one or more mechanism(s) to control their effects (see section 13.5).

13.3 What extraneous factors are confounders?

Confounders might be defined based on their having distributional differences between study groups. This is a necessary but insufficient criterion of confounding. In addition, it is difficult to implement because we rarely know the true state, and the data from our study populations that we use to establish the distributions could themselves be confounded. Nonetheless, based on a working set of criteria, we could conclude that a factor is a confounder if:

1. it is a cause of the disease, or a surrogate for a cause, and
2. it is associated with the exposure in the source population. In a cohort study, this means that the confounding factor must be associated with the exposure at the start of the study. In a case-control study, it means that the confounding

Example 13.1 A demonstration of confounding

We will begin by using a fictitious example with *Mannheimia hemolytica* (Mh) as the exposure of interest and bovine respiratory syncytial virus (BRSV) as the extraneous factor that we wish to control. The outcome is bovine respiratory disease (BRD), and the context is respiratory disease in feedlots. We will assume the factor (BRSV) whose distribution we plan to 'control' is a confounder in the population. BRSV fulfills the criteria of being a confounder variable as it is related to the exposure and the outcome, it is not intermediate between Mh and BRD on a causal pathway, and it is not an effect of BRD. Our summary (fictitious) of the population structure, ignoring BRSV status, is shown below:

	Mh+	Mh-	Totals	Odds ratio
BRD +	240	40	280	3.3
BRD -	6260	3460	9720	
Total	6500	3500	10000	
Risk (%)	3.6	1.1		

Based on observing the risk of BRD by Mh status and ignoring sampling variation, it appears that individuals with an active Mh infection have 3.3 times higher odds (think of this as 'risk') of contracting BRD than Mh- individuals (this assumes that 1.1% of the Mh+ individuals would have developed BRD in the absence of Mh – an assumed counterfactual argument). But what about the effect(s) of BRSV? If BRSV is a confounder, then some of the crude association attributed to Mh might be due to BRSV.

(*continued on next page*)

Example 13.1 (*continued*)

One way that we stress in this chapter for 'controlling' confounding is to stratify the data according to the levels of the confounding variable(s), or their combinations. Assuming that there are no other confounders, when the data are stratified on BRSV status, the 'true' association between Mh and BRD becomes apparent within strata, as shown below (note that we are denoting the presence of the agent or the outcome with 1 and their absence with 0 as this is a very common practice in computerised datasets):

Population structure		Mh			Stratum-specific odds ratios	Crude odds ratio
BRSV	BRD	1	0			
1	1	220	10	230	2	
1	0	5280	490	5770		
		5500	500	6000		
	Risks	0.04	0.02			
						3.3
0	1	20	30	50	2	
0	0	980	2970	3950		
		1000	3000	4000		
	Risks	0.02	0.01			

Note Ignoring the non-collapsibility of odds ratios (see section 13.7.2), the crude odds ratio differs from the stratum-specific odds ratios, indicating confounding is present so we need to use the stratum-specific odds ratios to estimate the causal association of Mh with BRD.

factor must be associated with exposure in the population from whence the cases came (*ie* it must be associated with the exposure status in the control group), and

3. the factor's distribution across exposure levels cannot be totally determined by the exposure (*ie* it is not an intervening factor) or the disease (*ie* it is not a result of the disease). This criterion is met if the confounding factor precedes, temporally, the exposure. An intervening or intermediate factor should not be treated as a confounding factor, whether it is totally determined by the exposure or not, because this would modify (bias) the association between the exposure and the disease such that the true causal effect is not obtained. Similarly, if the disease produces an outcome (*eg* another disease or change in production), that outcome should not be deemed to be a confounding factor.

It is useful to differentiate between a **population confounder** and a **sample** (*ie* study group) **confounder.** For example, if the factor is known to be a confounder in the population, it should be treated as such in the sample (*ie* controlled) regardless of whether it appears to be a confounder in the sample or not. Conversely, if it is known

not to be a population confounding factor, then it should not be controlled in the sample, even though it appears to be a confounder in the study subjects. Unfortunately, because we often do not know the true state of nature, we must use the data from the study population to make inferences about whether or not a factor is a confounder.

13.4 Criteria for confounding

As mentioned, the statistical approach to defining confounding variables is based on the difference(s) in the distribution of the factor(s) between the groups being studied. More precisely, if we have an exposure factor E, an outcome Y, and an extraneous factor Z (that is not an intervening variable or an effect of the outcome), factor Z is a confounder in a cohort study if:

- Z and E are associated unconditionally, and
- Z and Y are associated in exposure negative animals.

In a case-control study, factor Z is a confounder if:

- Z and E are associated in the controls (not just unconditionally), and
- Z and Y are associated in exposure negative animals.

Although these statistical criteria help us understand the necessary basis for confounding, these statistical criteria are insufficient to determine confounding without some additional assumptions about the lack of other confounders. Hence, we do not use statistical criteria to determine if a factor is a confounder or not. Confounding is said to be present when our measure of association differs from the true value. As the true value is usually unknown, the measure of association obtained after control of all identifiable confounders is deemed to be the true causal association. Because the identification and control of confounders is rarely perfect, some confounding is invariably present, and the important issue is how large the confounding effect is, not whether or not it is present. This becomes a matter of judgement (see section 13.7).

13.5 Control of confounding

As noted here and in the chapters on observational study design, we can prevent or control confounding by using one or more of three procedures: exclusion (restricted sampling), matching, or analytic control.

13.5.1 Exclusion (restricted sampling)

Because confounding is the result of a differential distribution of an extraneous factor between the two (or more) groups being compared, we can prevent confounding by selecting only one level of the extraneous factors for our study. This is called exclusion or restricted sampling, and because every study subject has the same level of the potential confounder, no bias is present. Some restricted sampling is natural; for example, we would only select females for a study of mastitis. In other instances we might deliberately want to restrict our study population to a single breed of study

subjects, or farms that use a specific production-recording scheme. The former would prevent confounding by breed whereas the latter could prevent confounding (from differences in herd characteristics across recording schemes) as well as ensure that specific data required for the study would be easily available. Similarly, we could restrict our study population to those possessing a limited range of production or being between specified ages *etc*. For example, Manske et al (2002), prior to a field trial of the effects of hoof-trimming on claw health in dairy cattle, restricted their study population of herds to selected herd sizes, breed compositions, and membership in an official milk-recording scheme.

When considering restricted sampling based on dichotomous extraneous variables we would usually prefer to admit the low-risk group to the study. Admitting subjects in the high-risk group into the study could make data interpretation more difficult if interaction between the exposure and potential confounder were present.

13.6 Matching

Matching is the process whereby we make the distribution of the 'matched' factor the same in the groups being compared. By making the distributions of these factors the same in both groups, we would prevent confounding and we could increase the power of the study. In randomised trials, matching on selected variables is used to reduce the residual variance and thus give the study more power per study subject. It is not used for prevention of bias, although in small experiments, it might help achieve this because randomisation is not likely to balance all the extraneous variables when the sample size is limited. As an example, in a field trial of hoof-trimming and claw health in dairy cows, Manske et al (2002) 'blocked' (*ie* matched) on breed, parity and stage of lactation before allocating, randomly, the treatment (hoof-trimming) to each cow.

In cohort studies, matching on one or more confounding variables can prevent confounding bias and also result in increased power/precision of the study. Matching on host characteristics such as age, breed and sex is used frequently. For example, Walker et al (1996) in a study of lymphocyte subsets in cats with and without feline immunodeficiency virus, matched on stage of disease (*ie* from asymptomatic to severely ill clinically). Matching was preferred to analytic control because the size of the study groups was limited. An example of the effects of matching in a cohort study is shown in Example 13.2.

Although some gains in precision can result from matching, in observational studies, any gains in statistical efficiency come at a substantial cost. Most importantly,

- it is not possible to estimate the effect of the matched factor(s) on the outcome because its distribution has been forced to be identical in the exposure (cohort study) or outcome (case-control) groups for cases and controls. We can, however, still investigate whether the matching factor acts as an effect modifier (*ie* if it produces interaction with the exposure of interest – see section 13.9).
- matching by some surrogate factors, such as farm, might 'match out' other potentially important exposures in hypothesis-generating studies.

Example 13.2 Matching in a cohort study

In our 'pretend' cohort study, we will sample 500 exposed (Mh+) and 500 non-exposed (Mh-) individuals with frequency matching of the Mh- group for the distribution of the confounder (BRSV) in the exposed study group. Based on the population structure in Example 13.1, among the 500 Mh+ subjects, 85% (*ie* 5500/6500) of the Mh+ group will be BRSV+, and their risk of disease will be 0.04. So, ignoring sampling variation, 17 of the 425 Mh+ and BRSV+ individuals in our study will develop BRD. Of the 75 Mh+ individuals without BRSV, 2% or 2 will develop BRD (expected numbers have been rounded to the nearest whole number).

Now, we need to select the Mh- subjects to match their distribution of BRSV to that in the Mh+ group. Normally, 14% (500/3500) of the 500 Mh- subjects would be BRSV+, but we need to have 85% (425) of them BRSV+. Of these 425 BRSV+ Mh- subjects, 2% develop BRD. Of the 75 Mh- subjects who are BRSV-, 1% or 1 develops the disease.

Note The observed stratum-specific odds ratios are equal to 2 (except for rounding errors), the same as in the source population (Example 13.1), as is the overall odds ratio. No control of the matched confounder is necessary in the analysis, and there is no bias present in the summary table. However, matched cohort data should be analysed using a stratified approach to ensure that the variance estimates are correct.

Observed association between Mh and BRD in a cohort study following matching for BRSV

		Mh			Stratum-specific odds ratios	Crude odds ratio
BRSV	BRD	1	0			
1	1	17	9	26	2	
1	0	408	416	824		
		425	425	850		2
0	1	2	1	3	2	
0	0	73	74	147		
		75	75	150		

In contrast, we later pretend to conduct a case-control study using all 280 cases and 280 controls frequency matched by the confounder BRSV (see Example 13.3).

- if matching is to be conducted on several factors, it can be quite difficult to find controls that have the same distribution of matching factors.

Matching is used frequently in case-control studies to increase the validity and efficiency of the study. For example, Alford et al (2001), in a multicentre case-control study of risk factors for equine laminitis matched on centre, clinician and season of admission.

Veling et al (2002), in a case-control study of risk factors for clinical salmonellosis on dairy farms, matched the control farms to the cases based on region.

However, matching in case-control studies is not without its disadvantages. For example, matching will actually introduce a selection bias into the data. The stronger the exposure-confounder association in the source population, the greater the bias that is introduced. This bias is generally in the direction of the null effect, regardless of the direction of the exposure-confounder association, but is controllable by stratifying on the 'matched' factors in a conditional (*ie* matched data) analysis.

Why does matching have different effects in case-control studies than in cohort studies? In a cohort study, matching makes the exposure independent of the matched extraneous variable so there can be no confounding. Further, because the outcome (*eg* disease) has not happened at the time of matching, the matching process is independent of the outcome. However, in case-control studies, the disease has already occurred when the matching takes place. Hence, if the exposure is related to the matched variable (as it would be if the extraneous variable is a confounder), and if we make the distribution of the matched variable(s) the same in cases and controls, we will alter the distribution of exposure in the controls so that their exposure level is more like that in the cases. An example of this selection bias in a case-control study is presented in Example 13.3. This example also shows that we can prevent the selection bias caused by matching in a case-control study by stratifying on the matched variable(s) in the analysis.

13.6.1 General guidelines for matching

The following guidelines should be considered when contemplating the use of matching (Rothman and Greenland, 1998). First, do not match unless you are certain that the variable is a confounder. This is particularly important in case-control studies if the extraneous variable and exposure are strongly associated. Matching in this situation leads to **overmatching**, because it gives the distribution of the exposure in the cases and controls greater similarity than the corresponding distributions in the base population. This can occur even if the extraneous variable is only related to the exposure and therefore not a confounder in the source population. In addition, with pair-matching (see section 13.6.2), information will be lost because cases and controls with the same value for exposure do not contribute useful data to the analysis (see below), hence effectively reducing the sample size and decreasing precision.

In some situations, however, matching will increase the efficiency of an analysis. For example:

- matching ensures that the dataset contains a control for every case when the matched factor is rare, or if it is a nominal variable with many categories (*eg* farm, sire *etc*). Random sampling in this instance might lead to marginal zeros and the data from such tables is of no value in analysis.
- matching might optimise the amount of information obtained per subject, if exposure information is expensive to obtain.
- matching might be the easiest way to identify controls in a study using a secondary base (*eg* by selecting the next non-case admitted, or listed in the registry). This

Example 13.3 Matching in a case-control study

In our case-control study, we will include all 280 cases from the target population in Example 13.1 as study subjects. This group will have the same exposure and confounder distribution as in the source population. Now, we need to select the controls to match the distribution of BRSV in the cases. Note that 82% (*ie* 230/280) of the cases will be BRSV+, so 230 of the controls will need to be BRSV+. Of these 230, 91.5% (5280/5770) will be Mh+ (n=210). Of the 50 BRSV- controls, 24.8% (980/3950) will be Mh+ (n=12).

Observed association between Mh and BRD in a case control study following matching for BRSV

Case-control structure		Mh			Stratum-specific odds ratio	Crude odds ratio
BRSV	BRD	1	0			
1	1	220	10	230	2.1	
1	0	210	20	230		
						1.6
0	1	20	30	50	2.1	
0	0	12	38	50		

Note The stratum-specific odds ratios are equal to 2 (except for rounding error) but the crude odds ratio is 1.6. The bias induced by matching in a case-control study is a form of selection bias. For example, in the population p(Mh+|BRD+)=86% (240/280) and p(Mh+|BRD-)=64% (6260/9720). In our study population, p(Mh+|BRD+)=86%, as it should, but p(Mh+|BRD-)=79%. The controls no longer represent the level of exposure in the target population. Clearly, analytical control (*eg* stratified analysis) of the matched confounder is necessary to prevent this selection bias in the overall measure of association.

is one of the most common uses of matching and if used only for this purpose, and the frequency of exposure is constant throughout the study period, this matching is often ignored and unmatched analyses performed.

If matching is not needed for one of these reasons, only consider matching in a case-control study if you anticipate a strong association in the population between the outcome and the confounder and a relatively weak association between the exposure and the confounder. In case-control studies, any gains in efficiency from matching are likely to be modest at best.

13.6.2 Frequency and pair matching

In frequency-matching on categorical variables, the overall frequency of the potential confounder(s) is made the same in the two outcome (case control) or exposure (cohort) groups. In pair- or individual-matching, one or more (*eg* m) control is individually matched to each case. Relative to frequency-matching, pair-matching requires a more

complex analysis, is generally less efficient (statistically), and makes it difficult to assess interaction between the exposure and confounder. However, pair-matching might be the only alternative when categories are very refined. Generally, we select between one-to-four controls matched to each case. There is minimal gain in efficiency if the matched control-to-case ratio exceeds 4:1. Although not necessary, it is simplest to use a fixed control-to-case ratio.

13.6.3 Caliper-matching

If the variable to be matched on is continuous, one must specify how close, on the continuous scale, the subject must be in order to be considered matched (called **caliper-matching**). Caliper-matching often produces a problem for analysis in that, if the match must be within, two years of age, for example, then two case (exposed) subjects of the same age could be matched with controls (non-exposed) whose ages differ by almost four years. In this instance, we either have to live with the 'wider' match and chance residual confounding or decide to use strata in our analyses that are no wider than the 'matching' criteria even if that shifts the 'matched' subjects into different strata.

13.6.4 Analysing matched data

In general, frequency-matched data should be analysed using a stratified method to account for the matching. If pair-matching is used, but there are not very many categories of the confounder, and many pairs are present within each category, the data could be analysed by creating a group identifier for the matched set of subjects and analysing the data as for a frequency-matched dataset using the group identifier to form the strata. Interaction between the confounder and exposure should be evaluated in the usual manner.

If the matching is conducted using pair-matching, and there are many categories of the confounder and very few pairs within each category, the data must be analysed using a matched-pair analysis. For these analyses, we use the frequencies of matched sets with every possible exposure and outcome pattern to estimate the odds ratio. In a case-control study, if there is only one control matched to each case, there are four possible exposure patterns: both the case and its matched control were exposed; both non-exposed; case exposed and control non-exposed; case non-exposed and control exposed. The data layout is shown in Table 13.1.

Table 13.1 Data layout for matched-pair case-control analyses

		Control pair		Case totals
		Exposed	Non-exposed	
Case pair	Exposed	t	u	$t+u = a_1$
	Non-exposed	v	w	$v+w = a_0$
	Control totals	$t+v = b_1$	$u+w = b_0$	

The crude odds ratio is $OR = \frac{a_1 b_0}{a_0 b_1}$. **Note** These numbers are available from the

marginal totals of the 'paired' 2X2 table. The Mantel-Haenszel *OR* uses only the data in the discordant cells and is $OR_{MH}=u/v$.

The Mantel-Haenszel χ^2 test (which in the case of 1:1 matching equals McNemar's test), could be used for hypothesis testing. The formula is:

$$\text{McNemar's } \chi^2 = \frac{(u-v)^2}{u+v} \qquad \textit{Eq 13.1}$$

For multivariable analysis, a conditional logistic regression analysis must be used (see section 16.14).

We now move on to describe control of confounding using analytic methods and describe three main ways of implementing that control. The first (which we do not recommend) is to use a statistical criterion, the second is a 'change-in-measure' approach and the third is to supplement the change-in-measure approach through the use of causal diagrams.

13.7 Analytic control of confounding

This approach is widely used and, given sufficient study subjects, might be considered the preferred approach to control confounding (often in conjunction with restricted sampling for control of other confounders). To implement this approach, we need to define and measure the important confounders and then analyse the data appropriately. We will describe the latter methods under stratified analysis (section 13.8) and in more detail based on regression models (see Chapters 14-19). Before discussing stratified analysis, we will review the three general approaches to analytical control of confounding.

13.7.1 Statistical control of confounding

In this approach, we use a statistical algorithm to either select (*ie* forward selection) or eliminate (*ie* backward elimination) variables from a regression model based on their statistical significance. This approach has been used frequently (especially with the advancement of powerful statistical routines to select variables when building models), but it has rapidly lost favour in recent years. The assumptions are that most confounders will be selected as 'significant' by this process thereby preventing confounding. The major problem is that, in using this approach, we cannot (or do not) distinguish between intervening and other types of extraneous variable. Furthermore, the process flies in the face of statements that the extent of confounding bias is a matter of judgement, not a matter of statistical significance. Thus, we do not recommend using this approach for anything other than initial pilot studies of a particular problem, or preliminary analyses of datasets.

Having said that, we need to recognise that when we search for more than one risk factor simultaneously, we will (in fact, we must) break a number of 'rules' about

what variables to control as confounders. With multiple factors under study, the causally prior factor that might need controlling to obtain valid effect estimates of one exposure factor could be an intervening variable for another exposure factor. Hence, the 'adjusted' measures of association we obtain from multivariable models are **direct effects** only, not **total causal effects**. Thus, for estimating the causal association, we will have 'over-controlled' for intervening variables (and perhaps effects of causes). One conservative approach to managing more than one exposure variable in a dataset is to take the set of 'significant' variables and then conduct a separate analysis for each factor as the exposure of interest and use this measure of association as the best estimate of the causal association.

13.7.2 Change in measure of association as an indication of confounding

Suppose we begin our analysis of the study data with an unconditional (crude) association between our exposure and outcome variables and observe a crude odds ratio, OR_c. We then stratify the data based on a potential confounder, or a set of potential confounders. After having ensured that the stratum-specific odds ratios are deemed to be approximately 'equal' to each other, we obtain the adjusted odds ratio OR_a. Almost always OR_a differs somewhat from OR_c, but if we deem the difference to be 'large' (in some practical sense), we say that some or all of the factors we stratified on (or controlled) were confounders. Thus, for example, we use the change in odds ratio between the crude and adjusted values to determine if confounding is present. We need to specify a difference (*eg* >20-30% change in the odds ratio) that would be deemed important given the context of the study. If this difference is exceeded, then we say confounding is present and the adjusted measure is preferred. Conversely, if there is virtually no difference between the crude odds ratio and the adjusted odds ratio, we say that confounding was not present and the crude measure suffices. In part this inference and the 'change-in-estimate' approach to identifying confounders are based on the fact that without confounding, if the stratum-specific measures are equal to X, then when the data are collapsed over that confounder, the crude measure will also be X. If the data meet this criterion, they are called **collapsible**.

Non-collapsibility of odds ratios

The measure of association used can affect our interpretation of confounding. In particular, the odds ratio, which is our most frequently used measure, suffers from the problem that it is not always collapsible. If we are using risk difference or risk ratio measures of association, the crude measure will always be a weighted average of the stratum-specific measures. And, as a result, in the absence of interaction, if no confounding is present, the data can be collapsed (*ie* summed over the levels of the confounder) and the stratum-specific risk ratios will be the same as the crude risk ratio. However, this is not true when the odds ratio is the measure of association. In this instance, the crude odds ratio can be closer to the null than the stratum-specific odds ratios; this is called **non-collapsibility**. This problem usually shows up if the outcome is very common in one or more strata as shown in Example 13.4. Thus, because the crude and adjusted measures differ, it might look as if confounding is present when it really isn't. Be aware of this situation. Despite the problem of non-collapsibility, the change in odds ratio (or other measure of association) has become the standard method

Example 13.4 Non-collapsibility of odds ratios and disease frequency

An example of non-collapsibility of odds ratios between exposure (E) and disease (D) in the presence of a non-confounding extraneous variable (Z)[a]. Disease risk=0.55

	Z+		Z-		Totals	
	E+	E-	E+	E-	E+	E-
D+	870	690	430	200	1300	890
D-	130	310	570	800	700	1110
Totals	1000	1000	1000	1000	2000	2000
Risk	0.87	0.69	0.43	0.20	0.65	0.45
Risk ratio		1.26		2.15		1.44
Risk difference		0.18		0.23		0.20
Odds ratio		3		3		2.3

[a] Example based on Greenland and Morgenstern, 2001

Note that variable Z is not a confounder because it is not associated with exposure; it is however associated with the outcome D. Because the stratum-specific odds ratios are equal to each other, and hence to the OR_{MH}, but differ from the crude odds ratio, we might be tempted to conclude that confounding by Z is present. However, the difference in these odds ratios relates to the use of 'odds' as a measure of outcome frequency; there really is no confounding present in this example. **Note** Both the RD and RR are collapsible in that the crude measure will always lie between the two stratum-specific measures.

Non-collapsibility is a greater problem for interpretation when the outcome frequency is high (55% in this example). In the table below the average risk is 8.3% and the data are 'virtually' collapsible.

An example of near-collapsibility of odds ratios between exposure (E) and disease (D) in the presence of a non-confounding extraneous variable (Z)[a]. Disease risk=0.083

	Z+		Z-		Total	
	E+	E-	E+	E-	E+	E-
D+	211	82	29	10	240	92
D-	789	918	971	990	1760	1908
Totals	1000	1000	1000	1000	2000	2000
Risk	0.21	0.08	0.03	0.01	0.12	0.05
Odds ratio		3		3		2.8

As this example indicates, in practical terms, non-collapsibility is only a problem when the outcome frequency is high.

of identifying confounding. A key to using this approach successfully is to ensure that intervening variables or variables that are affected by the outcome are not included in the model.

13.7.3 Using causal diagrams to identify potential confounding variables

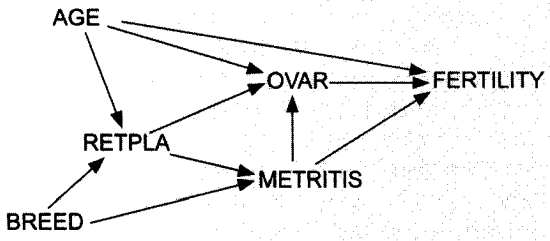
The change in estimate approach is used once the factors to be controlled are selected. In Chapter 1 we introduced the use of causal diagrams. Here we extend this approach to the assessment of potential confounders as a way of determining whether or not a variable should be controlled. First, of course, we need to draw the causal diagram (sometimes referred to as a **directed acyclic graph;** Greenland et al, 1999; Hernan et al, 2002) using the principles explained in Chapter 1. We then identify the exposure factor and the outcome of interest, as specified in the major objective of the study. Now, any factor causally prior to the exposure factor and on a pathway connecting the exposure and outcome is a likely candidate for control as a confounder. Factors that are causally after the exposure variable should not be controlled nor should variables that are causally after the outcome. We formalise the process as follows:

1. Draw the diagram using the guidelines outlined in Chapter 1. Then eliminate all arrows emanating (*ie* leading away) from the exposure factor of interest on the graph.
2. If there are any paths that still connect the exposure and outcome, then the causally prior factors and other non-intervening variables in these paths should be controlled, otherwise these factors will bias the measure of association. In causal terminology, these factors produce **spurious causal effects**.
3. There is a final twist that is needed to complete this process. Suppose that there are two or more factors that 'cause' a third factor that is prior to the exposure factor and that the initial assumption was that these two (or more) factors were unrelated, causally, to each other (*ie* these factors would be marginally independent statistically). AGE and BREED have this structure in Example 13.5 as they both cause RETPLA, but are independent of each other. However, when we control for a factor they cause, this act makes these factors conditionally associated, and we will need to control for at least one of them to prevent bias. To ascertain this, we need to connect all marginally independent factors with a one-headed line. In tracing out pathways between the exposure and outcome we can go either way on this line. In order to 'close' this pathway, we will need to control for one (or more) of these factors in our modelling process. Thus, knowledge of the likely causal structure becomes very important in selecting factors for control, as control of one factor might necessitate control of others.

Now that we have the tools to identify factors needing control, we will explain a process for implementing control – namely, stratified analysis. More elaborate mechanisms of control are explained in the chapters on model-building and regression models (Chapters 14-23).

Example 13.5 Identifying confounders using a causal diagram

We can use the causal diagram from Chapter 1 to demonstrate the application of the criteria for identifying confounders. Recall the example concerned studying the potential impact of selected diseases on infertility in dairy cows. We will add another variable to the diagram, BREED, and we assume that breed effects are transmitted through RETPLA and METRITIS. The causal diagram is:



where RETPLA is retained placenta, OVAR is cystic ovarian disease.

If we were interested in estimating the causal association between METRITIS and FERTILITY:

- omit the arrows leading forward from METRITIS to OVAR and FERTILITY
- this leaves causal paths to FERTILITY from OVAR and AGE
- the spurious causal path from METRITIS back to RETPLA through OVAR means that RETPLA needs to be controlled
- we now connect AGE and BREED with an imaginary non-headed line
- because we can also go backward on the path from RETPLA to AGE and then forward to FERTILITY either directly or via OVAR, AGE as a causally prior variable would also need to be controlled
- once we control for RETPLA we need to control for either or both of AGE and BREED. Although the diagram shows them to be independent, controlling for RETPLA makes them conditionally related
- at this point it appears that BREED should be controlled because it becomes conditionally related to AGE once RETPLA is controlled
- however, as BREED effects go only through RETPLA, we do not need to control for BREED if AGE is controlled. Controlling BREED would not be incorrect but is unnecessary
- note that OVAR is not controlled in the analysis.

Of course, there are more complex causal diagrams (see Hernan et al, 2002) but this example should convey the basics.

13.8 STRATIFIED ANALYSIS TO CONTROL CONFOUNDING

The most widely used stratified analytic approach for dichotomous categorical data is the **Mantel-Haenszel procedure**. The stratification procedure is straightforward, easy to use, and its use can help inform the researcher of details of the data that otherwise might be missed. Indeed, we advise researchers to use this approach in initial analyses even when using more complex analyses such as logistic regression. This method relies on physically stratifying the data according to the combinations of levels of the confounding variables, examining the stratum-specific measures of association (odds ratios for now) and, if these are deemed to be equal (apart from sampling variation), creating a pooled 'weighted' or 'adjusted' estimate of the association. The equality of the stratum-specific measures can be evaluated visually, or statistically using a test for homogeneity. Demonstrating this equality is a prerequisite to calculating an overall measure of association because, if the measure differs across strata, it indicates that interaction is present (see section 13.9). Recall from Chapter 1 the discussion of interaction arising because of the factors being components of the same sufficient cause. Based on this, in Examples 1.1 and 1.2, we demonstrated the relationship between an assumed underlying causal model and the observed risks of disease. In those causal models, our assumption was that two or more causal factors were jointly necessary to complete a sufficient cause. Thus, biological synergism was present. However, whether or not statistical interaction was evident depended on the distribution and frequency of the other sufficient causes and their components.

As stated earlier, from a practical point-of-view, if the adjusted (pooled) measure differs from the crude measure of association (by an amount deemed to be important – a judgement call), then confounding is said to be present. If confounding is deemed to be present, the adjusted measure of association is always preferred to the crude (biased) measure.

In order to describe the Mantel-Haenszel procedure, we will assume that we have dichotomous exposure, outcome variables and a confounder with J levels. Thus, we will have one 2X2 table (*ie* one stratum) for each of the J levels of the confounder, (or each combination of the levels of confounders if there was more than one confounder) (see Table 13.2). We shall assume there are J strata in total; here we denote a specific stratum number by the subscript 'j'):

Table 13.2 Data layout for stratified analyses

	Exposed	Non-exposed	Total
Cases	a_{1j}	a_{0j}	m_{1j}
Non-cases	b_{1j}	b_{0j}	m_{0j}
Total	n_{1j}	n_{0j}	n_j

Recall that the n_j or m_j might not have a population interpretation depending on the study design (*eg* n is not an estimate of population denominators in a case-control study). Nonetheless, the values in the cells might be used for purposes of calculating the measure of association or its variance.

Eqs 13.2 to 13.5 show the necessary formulae for analysing binary data (*ie* risk, not rate, data).

We begin by stratifying the data as shown above and calculating the stratum-specific odds ratio. The odds ratio for the j^{th} stratum is:

$$OR_j = a_{1j} * b_{0j} / a_{0j} * b_{1j} \qquad \textit{Eq 13.2}$$

We also need the expected values and the variance of the exposed-diseased cell. The expected number of exposed cases in the j^{th} stratum is:

$$E_j = m_{1j} * n_{1j} / n_j \qquad \textit{Eq 13.3}$$

and the variance of E_j is:

$$\text{var}(E_j) = V_j = m_{1j} * m_{0j} * n_{1j} * n_{0j} / n_j^2 * (n_j - 1) \qquad \textit{Eq 13.4}$$

The 'adjusted' or Mantel-Haenszel odds ratio is a weighted average across strata:

$$OR_{\text{MH}} = \frac{\sum \left(a_{1j} * b_{0j} / n_j \right)}{\sum \left(a_{0j} * b_{1j} / n_j \right)} \qquad \textit{Eq 13.5}$$

from which we can obtain $\ln OR_{\text{MH}}$ for use in testing homogeneity (Eq 13.6).

Before interpreting the adjusted odds ratio as a valid summary measure of association, we should examine the stratum-specific odds ratios and see if they are 'approximately' equal. Otherwise the adjusted odds ratio oversimplifies the association. Inequality of stratum-specific odds ratios is an indicator of the possible presence of interaction – we say possible presence because confounding by an unknown factor can produce effects that resemble interaction. There is a Wald-type χ^2 test for interaction. This test has low power and so we might benefit from relaxing the P-value for significance to the 10-15% level. The Wald χ^2 test for homogeneity with (J-1) df is:

$$\chi^2_{\text{homo}} \sum \left(\frac{\left[\ln OR_j - \ln OR_{\text{MH}} \right]^2}{\text{var}[\ln OR_j]} \right) \qquad \textit{Eq 13.6}$$

where $\text{var} \ln OR_j = \frac{1}{a_{1j}} + \frac{1}{b_{1j}} + \frac{1}{a_{0j}} + \frac{1}{b_{0j}}$

Whether or not interaction is deemed to be present depends in part on the scale of measurement of association. Here we present only odds ratios but we could use risk difference, relative risk, or rate ratio as measures. The finding of interaction in one scale does not necessarily translate into the presence of interaction in another.

An overall test statistic, with 1 df, for the significance of the summary odds ratio is:

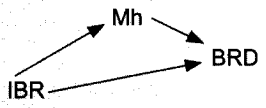
$$\chi^2_{\text{MH}} = \frac{\left(\sum a_{1j} - \sum E_j \right)^2}{\sum V_j} \qquad \textit{Eq 13.7}$$

An example of the use of this approach is given in Examples 13.6 and 13.7. Formulae

Example 13.6 Stratified analysis of respiratory agents and bovine respiratory disease: no confounding

data=feedlot

In this dataset, there are data on the titres to a variety of putative respiratory pathogens in feedlot calves and on the occurrence of bovine respiratory disease (BRD). Experimentally, an interaction has been demonstrated between infectious bovine rhinotracheitis (IBR) virus and *Mannheimia hemolytica* (Mh), and as we have data on these, we can summarise the relationship of each of these agents, alone and together, on the occurrence of BRD. The exposure of interest is Mh, and our proposed causal model is:



We include a direct causal arrow from IBR to BRD because of our belief that IBR could enhance the respiratory pathogenicity of other unmeasured agents, besides Mh, and hence cause BRD. Thus, to ascertain the causal association of Mh with BRD, we need to control for IBR. The relationship of Mh by itself with BRD is:

	Mh+	Mh-	Total
BRD+	167	30	197
BRD-	300	91	391
Total	467	121	588

The *OR* is 1.69 and the χ^2 test is 5.19 with a P-value of 0.023. Hence, when we ignore the effects of IBR, seroconversion to Mh is associated with an increased risk of BRD of about 1.7 times. The joint distribution of Mh and IBR is shown below:

Stratification of BRD by Mh and IBR, prior to Mantel-Haenszel analysis

IBR	BRD	Mh+	Mh-	Total
1	1	83	18	101
1	0	85	48	133
Total		168	66	
0	1	84	12	96
0	0	215	43	258
Total		299	55	

The layout of the essential calculations for the Mantel-Haenszel procedure are:

Stratum	Odds ratio	lnOR	var(lnOR)	a_i	E_i	var(E_i)	$a_{1i}*b_{0i}/n_i$	$a_{0i}*b_{1i}/n_i$
1	2.6	0.96	0.10	83	72.51	11.67	17.03	6.54
2	1.4	0.34	0.12	84	81.08	9.21	10.20	7.29
Totals				167	153.60	20.88	27.23	13.83

(*continued on next page*)

Example 13.6 (*continued*)

The 'adjusted' or Mantel-Haenszel odds ratio is:

$$OR_{MH} = \frac{27.23}{13.83} = 1.97$$

Based on these calculations, it appears that the strength of the association is slightly increased in the presence of IBR virus but perhaps not to the extent of being declared different from the effect when IBR virus is absent. However, we will perform a formal test of equality (or homogeneity) of the stratum-specific *OR*s.

The Wald test for homogeneity is:

$$\chi^2_{\text{homo}} = \sum\left(\frac{(0.96-0.678)^2}{0.100} + \frac{(0.34-0.678)^2}{0.120}\right) = 0.808 + 0.946 = 1.747$$

where 0.678 is the ln1.97.

This test result is reasonably non-significant (P=0.189); thus, we can act as if the stratum-specific *OR*s (*ie* 2.60 and 1.40) do not differ statistically.

An overall test statistic of the null hypothesis that $OR_{MH}=1$ is:

$$\chi^2_{MH} = \frac{(167-153.6)^2}{20.88} = 8.6$$

with 1df P≈0.003 so we can accept that $OR_{MH} > 1$.

Based on this test, because P≈0.003, we can reject the null hypothesis and conclude that there is good evidence that seroconversion to Mh increases the risk of BRD, after controlling the effects of IBR.

Compared with the crude odds ratio of 1.69, the increase in size of OR_{MH} is only about 17% so with our guideline of a change greater than 30%, we might say that serious confounding was not present and we might choose to use the crude *OR* to describe the causal association.

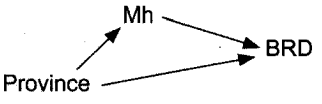
for stratified analyses of risk and rate data from cohort studies are available elsewhere (Rothman and Greenland, 1998, pp79-91).

13.9 Stratified analysis when interaction is present

In Chapter 1 we demonstrated how two or more factors that were members of the same sufficient cause exhibited biological synergism which, in turn, could lead to differences in risk depending on the presence or absence of other component causes. In the section just completed on stratified analysis to control confounding, we indicated that the exposure of interest had to have the same association across all levels of the confounder

Example 13.7 Stratified analysis when confounding is present
data=feedlot

Here we use the same dataset but control for Province (this is a surrogate for location of feedlot, partly for source of calves, and weight of calves on arrival). Our causal diagram is:



The data summary is:

Stratification of BRD by Mh and Province prior to Mantel-Haenszel analysis

Province	BRD	Mh+	Mh-	Odds ratio
1	1	84	21	2.75
1	0	80	55	
2	1	83	9	1.51
2	0	220	36	

The test of homogeneity of the stratum *OR*s had a χ^2 (1 df)=1.47 (P=0.23), so it is legitimate to calculate and interpret a weighted average *OR* as a summary measure. The crude *OR* is 1.69, and the OR_{MH} is 2.19. This is a 30% change in the coefficient and certainly suggestive of moderate confounding by Province being present. The test that the OR_{MH}=1 had a χ^2 (1 df)=11.20 with a P-value of <0.001 so we conclude that Mh and BRD are associated (or that OR_{MH} >1) after controlling for Province.

Thus, based on the crude odds ratio, we might suggest that seroconversion to Mh was associated with an increased risk of BRD. After controlling for province where the feedlot was located, the relationship gets considerably stronger; thus, we would say that confounding by Province was present and the larger OR_{MH} (2.2) is the better indicator of causal association.

in order to support the use of a single summary measure. A test of the equality of the stratum-specific measures of association served to assess this feature. If the stratum-specific measures were declared different, this was an indication that **interaction** was present and that the stratum-specific measures should not be averaged into a single overall measure (technically, unmeasured confounders can produce differences that mimic interaction also) such as the OR_{MH}.

Interaction is a somewhat confusing term. Its presence could provide clues about biological mechanisms or pathways of action, but whether it is deemed to be present or not depends on the statistical model and the scale of measurement. However, regardless of the scale or measure of association, interaction is said to be present when the combined effect of two variables differs from the sum of the individual effects in that scale. For current purposes there are three types of joint effect that two or more

exposure factors can produce: additive, synergistic (if the combined effect is greater than the sum of the individual effects) and antagonistic (if the combined effect is less than the sum of the individual effects).

In order to explain interaction, it will be helpful to return to some basic measures of single and joint exposure factor risks. For this discussion, we will assume that we use the risk of disease as the outcome. Risk will be denoted as

- R_{11} when both exposure factors 1 and 2 are present; as
- R_{10} when only exposure 1 is present; as
- R_{01} when only exposure 2 is present; and, as
- R_{00} when neither exposure factor is present.

Now, the effect of each variable can be measured by either a difference such as the risk difference (*ie* $RD_{10}=R_{10}-R_{00}$) or a relative measure such as the risk ratio (*ie* $RR_{10}=R_{10}/R_{00}$). With these as the basis, we can examine the joint effects of two variables.

13.9.1 Additive scale of association

Using risk difference as the measure of association, additive interaction would be present if

$$RD_{10}+RD_{01}\neq RD_{11} \qquad \textit{Eq 13.8}$$

Generally, if the effects are measured as *RD*, and the effects are additive, this might be taken to indicate that the two factors operate through different biological pathways or mechanisms (*ie* they are not members of the same sufficient causes). However, this 'causal interpretation' might be going beyond what the data can tell us (see section 1.2 and Thompson (1991)). Nonetheless, the risk difference is a common model for assessing the public-health significance of multiple variables as it relates directly to the excess number of cases that an exposure might cause.

13.9.2 Multiplicative scale of association

Using a ratio measure of association, multiplicative interaction would be said to be present if:

$$RR_{10}*RR_{01}\neq RR_{11} \qquad \textit{Eq 13.9}$$

As this involves multiplying the relative measures, it is known as the multiplicative model or scale. **Note** If we take logarithms of Eq 13.9, we have $\ln RR_{10}+\ln RR_{01}\neq\ln RR_{11}$ showing that additive effects on the logarithmic scale are equivalent to multiplicative effects (*ie* interaction) on the original scale. However, as we will point out shortly, the risks of disease in jointly exposed individuals that are consistent with an additive arithmetic-scale model differ greatly from those that are inconsistent with an additive multiplicative-scale model.

When the multiplicative-scale model holds, it can be shown that the *RR* for the primary exposure of interest will be the same in all strata of the extraneous variable(s). Thus, the equality of stratum-specific *RR*s, provides a convenient test for interaction in

the **multiplicative scale**. This is also the basis of the test of homogeneity of *OR*s in the Mantel-Haenszel procedure (Eq 13.6). A significant test result indicates that the stratum-specific ratios are not equal, or equivalently, that the joint effect of the two factors is not what would be predicted based on the singular effects of the two variables (*ie* the effect of one exposure factor **depends on** the level of the other exposure). This phenomenon is referred to as interaction or **effect modification** (Susser, 1973) in the multiplicative scale.

It was noted earlier that whether or not interaction is present depends on the scale of measurement. As noted in Example 13.4, when the stratum-specific odds ratios are equal, the *RR* and *RD* measures will not be, and conversely if the *RD* measures were equal, then *RR* and *OR* would not be. Thus, in large sample-size studies, if the data are consistent with the additive model in one scale, they will be consistent with interaction in another scale. The risks shown in Example 13.4 are consistent with no interaction (*ie* an additive model in the log scale) when the *OR* is the measure of association but, based on these risks, both *RD* and *RR* measures of association show evidence of interaction. Another example, using fictitious data, is shown in Example 13.8. Three different scenarios based on the risk of disease in the jointly exposed subjects are used to indicate the observed risks that would be consistent with a 'no interaction' state in an additive (scenario 'a') versus a multiplicative (scenario 'c') model. The multiplicative model is widely used for assessing associations between dichotomous outcomes and exposures. It is applicable in a variety of contexts and study designs and appears to 'fit' observed data well.

Example 13.8 An example of identifying interaction between exposure factors for BRD dependent on measurement scale

Note the individual effects (bottom half of table) and then the three scenarios for joint effects (top half of table).

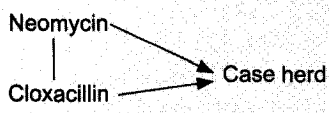
	Mh	BRSV	BRD (cases per 1000)	Risk	RD	Additive scale interaction	RR	Multiplicative scale interaction
Three possible scenarios (ie levels of combined risk) for joint effects								
a	+	+	100	0.100	0.099	synergy	100	antagonism
b	+	+	29	0.029	0.028	none	29	antagonism
c	+	+	200	0.200	0.199	synergy	200	none
Effects of individual factors by scale of measurement								
	+	-	10	0.010	0.009		10	
	-	+	20	0.020	0.019		20	
	-	-	1	0.001				

Note Any joint risk above 29/1000 is considered as indicative of interaction on the additive scale (scenarios a and c) whereas a joint risk of 200/1000 indicates no interaction on the multiplicative (*ie* log) scale (scenario c).

Example 13.9 indicates the result when stratification is used to control confounding in the presence of interaction.

Example 13.9 Controlling for confounding when interaction is present
data=nocardia

The data for this example are from a case-control study of dairy farms with and without Nocardia of mastitis. The exposure of interest was neomycin-containing dry-cow treatments. However, it was believed important to examine other dry-cow treatments also, both as possible risk factors and as potential confounders. Our causal model is:



We use a non-headed line between the two types of dry-cow treatment to indicate a non-causal correlation, likely because of a third common-cause factor such as management style. Even though the association is unlikely causal, using the rules set out in section 13.7.3 we need to control for cloxacillin to determine the causal effect of neomycin-containing treatments.

Stratification of case/control herds by neomycin and cloxacillin

Cloxacillin	Nocardia mastitis	Neomycin+	Neomycin-	Stratum-specific odds ratios
1	1	5	3	1.5
1	0	10	9	
0	1	44	2	29.3
0	0	15	20	

In the herds not using cloxacillin, the *OR* between neomycin use and case status was 29.3, whereas, in those herds using cloxacillin, the *OR* was 1.5. The test of homogeneity had a χ^2 of 6.44 (1 df) with a P-value of 0.011. This is considerable evidence of a difference in *OR* and is consistent with the presence of interaction. Hence, controlling for confounding is mute; we should not compute an adjusted odds ratio because the association between neomycin use and case-control status (Nocardia mastitis) **depends on** the presence or absence of cloxacillin use on the farm. Thus, when interaction is present we should not interpret the summary measure because it depends on the level of other extraneous variables.

13.9.3 Causal structures and interaction

Early on in our examples we demonstrated clear evidence of interaction arising from the sufficient cause model (Chapter 1). The sufficient cause model implies synergism which shows up statistically as interaction. However, we also demonstrated that with

the presence of unknown or unmeasured extraneous variables, interaction is not always detectable (even though the occurrence of synergism is the basis for the causal model). We also know that confounding can produce data that looks as if interaction is present, or conversely hide it. Thus, it is important to control confounding from other factors while trying to identify if interaction is present between two factors of interest.

A biological example of known synergism is the combined effect of viral exposure of the respiratory tract of calves 4-6 days prior to exposure with Mh. Experimentally this is a useful 'model' for reproducing the disease using aerosol challenges (Yates, 1982). Notwithstanding this, when the disease is observed in feedlots, even when a large number of organisms are measured and included in the model, it has not been possible to detect interaction (see Example 13.6).

In this chapter we have been interested in the effect of an extraneous variable given that we know the underlying causal structure. This has hopefully been of use for purposes of understanding the relationship between causal structures and the data we obtain in our studies. We do need to be careful however if, based on our analyses, we try to predict the causal structure. Although a number of researchers have tried to develop a general process for doing this successfully, regrettably, except in limited situations, our ability to infer causal structures from observed data is very limited, largely because we might be missing one or more important extraneous factors in our model (Thompson, 1991).

13.10 External adjustment of odds ratios for unmeasured confounders

Sometimes we might have conducted a study without measuring or otherwise controlling the effects of one or more potentially important extraneous variables. We might have calculated a crude odds ratio between our exposure (E) and disease (D) but wonder what value it would have had if we had measured and controlled a particular confounder (Z). Can we gain some insight into how much bias this unmeasured confounder might produce. The short answer is yes, but we would need to know three things, only one of which can be gleaned from the available data.

1. the prevalence of the exposure variable, E (we can get an estimate of this from the control group in a case-control study)
2. the association between the confounding variable (Z) and disease having adjusted for the exposure ($OR_{ZD|E}$; sometimes we can obtain this value from other studies) and,
3. the prevalence of the confounding variable among the exposed (P_{1Z}) and non-exposed (P_{0Z}) groups. We know these have to differ from each other, or else the factor would not be a confounder, and we might obtain these estimates from other studies, or be able to make educated guesses about their values.

The approach is as follows: first, we will assume the confounding variable is dichotomous, and thus, if we stratify on it, there will be two tables. These tables have the usual risk-based 2X2 structure, the first representing the data when the confounder is absent, and the second the data when the confounder is present. Now if the prevalence

of the confounder is P_{1Z} among the exposed and P_{0Z} among the non-exposed, then within the exposed group, our predicted number of non-cases with the confounder Z will be $b_{11}'=P_{1Z}b_1$. Within the non-exposed the predicted number of non-case subjects with Z is $b_{01}'=P_{0Z}b_0$ (see Example 13.10).

If it is reasonable to assume a common disease-confounding variable odds ratio (OR_{DZ}), we can use these estimates of the number of non-cases to solve for a_{11} and a_{01} (*ie* the number of exposed and non-exposed cases with the confounder). The formulae (see Rothman and Greenland (1998) for details) are:

$$a_{11} = \frac{OR_{DZ}a_1b_{11}'}{(OR_{DZ}b_{11}'+b_1-b_{11}')} \qquad \textit{Eq 13.10}$$

and

$$a_{01} = \frac{OR_{DZ}a_0b_{01}'}{(OR_{DZ}b_{01}'+b_0-b_{01}')} \qquad \textit{Eq 13.11}$$

With these two cell numbers we have complete information for the 2X2 table of subjects with the confounder. The table values for the subjects without the confounder can be obtained by subtracting the values for the subjects with the confounder from the original observed cell values. By substituting a reasonable range of prevalences and confounding-disease odds ratios, we can investigate the likely impact of this unmeasured confounding variable on the exposure-disease association. We recommend programming a spreadsheet to develop these 'what-if' scenarios. One 'what-if' example is shown in Example 13.10.

13.11 Multivariable causation and data analyses

In the discussion that follows, we focus on causal structures and their impact on the disease frequencies that we observe. In reality, there are a number of ways in which factors can combine to produce disease and it is rare that we identify all of the component factors of particular sufficient causes. Thus, if we measure two potentially causal exposures, they might be members of the same or different sufficient causes or they might turn out not to be causes at all. Sometimes, because of the arrangement of some of the underlying causes, we might find **spurious relationships** (*ie* statistical associations when no causal relationship exists). Here we show some of the ways of detecting and understanding these relationships. Not all the relationships we demonstrate relate to confounding factors; however, they are intended to demonstrate the impact that extraneous factors can have on the association between the exposure and outcome of interest. Because of their central value prior to and during analysis, we continue the discussion on causal diagrams that we began in Chapter 1 and elaborated on in section 13.7.3.

13.11.1 Graphical aids to understanding multivariable systems

As a simple biological example we will continue to focus on identifying factors that might be of causal importance for bovine respiratory disease (BRD). We will

Example 13.10 Effects of unmeasured confounders

Suppose we had observed the following data on BRD and Mh in calves. Our interest was to ascertain if calves with Mh were at increased risk of BRD; however, we had not controlled for an important confounder such as BRSV. Our summary 2X2 table data would be:

	Mh+	Mh-	Totals
BRD+	78 (a_1)	11 (a_0)	89
BRD-	86 (b_1)	74 (b_0)	160
	164	85	249

The odds ratio would be 6.11 with a χ^2 statistic of 29.2 (P<0.000); it appears that Mh+ calves were at increased risk of BRD. But, perhaps this relationship was largely explicable by BRSV infection. What effect might this have on our observed association if we had measured it? Suppose there is evidence that BRSV (Z+) doubles (*ie* OR_{EZ} =2) the risk of BRD. We will also suppose that 60% of Mh+ calves and 40% of Mh- calves were infected with BRSV.

Based on this, the predicted number of non-case Mh+ calves that is also infected with BRSV is b_{11}'=0.6*86=51.6 and the predicted number without Mh but with BRSV is b_{10}'=0.4*74=29.6. Hence, solving for the expected number of Mh+ calves with BRD and BRSV we have:

$$a_{11}' = \frac{2*78*5.1.6}{(2*51.6+86-51.6)} = 58.5$$

and for the Mh- cases with BRSV we have:

$$a_{10}' = \frac{2*11*29.6}{(2*29.6+74-29.6)} = 6.3$$

We can now complete the first table for the BRSV-infected subjects (*ie* the Z+ group).

BRSV+	Mh+	Mh-	Totals
BRD+	58.5	6.3	64.8
BRD-	51.6	29.6	81.2

The *OR* between Mh and pneumonia here is 5.3. Now, data for the second table for those without the confounder BRSV (*ie* the Z- group) is obtained by subtraction from the original observed cell values (*eg* $a_{10}' = a_1 - a_{11}'$).

BRSV-	Mh+	Mh-	Totals
BRD+	19.5	4.7	24.1
BRD-	34.4	44.4	78.8

The odds ratio between Mh and pneumonia here is 5.4. The adjusted odds ratio would be close to 5.3; thus, at least with this set of estimates, the presence of BRSV infection in these calves would not explain very much of the observed crude association between Mh and BRD.

suppose that our principal objective is to investigate the association between Mh and the occurrence of BRD. Suppose the additional factor we measure is the presence of BRSV infection (based on titre response). BRSV is only one extraneous factor but we can think of situations where there are numerous factors each with an underlying relationship with the exposure and/or the outcome.

The presumed causal relationship between pairs of variables will be shown using a **causal line diagram**. In this instance, our predictor (or exposure) variables are BRSV and Mh. There are a number of possible causal models involving just two predictors that we will outline subsequently. When describing the causal (structural) relationships between variables using line diagrams, an arrow implies a cause-and-effect relationship, a double-headed arrow indicates causal correlation, a non-headed arrow (*ie* line) non-causal correlation (likely because of an unmeasured factor), and no arrow implies no causal relationship.

We will describe the statistical results we expect, based on the causal structure in the line diagram, both visually using **Venn diagrams** and descriptively in the text. In the Venn diagrams, each circle represents a factor and the amount of overlap in the circles the extent (strength) of their association. If the circles do not overlap, this indicates that the factors are not associated statistically; it does not mean that they are mutually exclusive (*ie* do not occur together). The position (left to right) of each circle represents (where possible) the relative temporal positioning of the variables.

In describing these models we will assume all variables are dichotomous, similar to the factors used in Chapter 1, Example 1.1 where we use a relative measure of association (the risk ratio). We continue to use that approach here except that we will use the odds ratio (*OR*) as our measure of association (see Chapter 6). In the multivariable setting, when examining the Mh–BRD association, any factor that is not the exposure of primary interest is an **extraneous variable**. Susser (1973) named each type of extraneous variable based on their causal relationships with the exposure and outcome; we continue that practice with some revisions from his nomenclature. Hence,

1. *OR* is the unconditional (crude) *OR* between Mh and BRD.

This is the measure we would obtain from a 2X2 table (or by analogy from a simple regression model) when we ignore all other factors. When we 'adjust' or 'control' for other factors, the crude measure of association might change and it is referred to as a conditional measure of association. Hence,

2. *OR*|BRSV is the conditional, or adjusted, *OR* (*eg* OR_{MH}) between Mh and BRD after controlling for the relationships with the extraneous variable BRSV.

We can accomplish that control using either the stratification approach (section 13.7) or a regression approach – these are the subjects of much discussion later in this text (Chapters 14-23). Sometimes, but not always, we prefer the adjusted estimates because, if our analysis is consistent with the proposed causal model, the adjusted estimate should be closer to the 'truth' than a crude estimate of association.

In each of the following sections we will:

1. describe the causal relationships among the exposure, extraneous variable(s) and the outcome of interest,

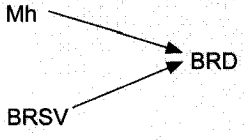
2. draw the causal relationships between the two predictor variables and the outcome to display the underlying causal structure,
3. note the crude statistical association between Mh and BRD that we expect to observe given the causal model, and
4. examine the association (in the absence of any sampling error) between the exposure and outcome after the extraneous variable is 'controlled' (*ie* added to the model).

13.11.2 Exposure-independent variable(s)

See the causal model in Example 13.11. The underlying causal structure is that both Mh and BRSV cause BRD but they are unrelated causally to each other, hence BRSV is called an exposure-independent variable. Because of their lack of causal association with the exposure, unless they are correlated because of the effect of other factors, exposure-independent variables are expected to be uncorrelated with the exposure. In observational studies, exposure-independent variables might arise naturally. In other situations the extraneous variables are causes of the outcome but also are related to the exposure of interest, and might be treated as a confounding variable. However, when matching is used to control these extraneous variables in cohort studies, the matched variables are converted into exposure-independent variables. Thus, they do not bias the measure of association and need not be 'controlled' analytically. In controlled trials (Chapter 11), we rely on randomisation to convert a number of causal extraneous cofactors into treatment-independent variables so they will not bias the measure of effect.

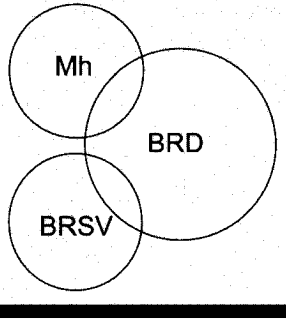
Example 13.11 An exposure-independent variable

Causal model



Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

The two predictor variable circles do not overlap indicating their independence. Both exposure circles overlap with the outcome circle indicating their significant statistical association with BRD.

Exposure-independent variables do not distort the crude measure of association. This is displayed in Example 13.11 by noting that the portion of the outcome explained by BRSV does not overlap with the proportion explained by Mh. Thus, whether BRSV is included in the model or not makes no difference to the *OR*. However, exposure-independent variables account for some of the unexplained variation in BRD, often referred to as the residual variation. Thus, accounting for them in the analysis improves the precision of the estimate of association by reducing the unexplained variability in the outcome.

13.11.3 Simple antecedent variable

See Example 13.12. The underlying causal structure is that BRSV (the simple antecedent) increases susceptibility to Mh which directly causes BRD. A simple antecedent is a variable that occurs temporally before the exposure variable, and is causally related to the outcome only through the exposure variable of interest. In our example, if BRSV is the simple antecedent, adding this variable to our model merely traces the sequence of causation backward in time. (This can be of importance in our understanding of the causal web so simple antecedents should not be dismissed as 'unimportant'.)

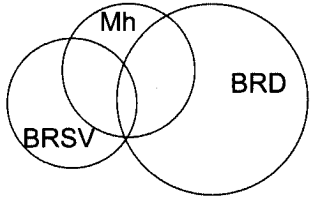
Example 13.12 A simple antecedent variable

Causal model

BRSV → Mh → BRD

Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

Often there is a weak overlap between variables such as BRSV and the outcome, but statistical associations favour direct causes over indirect causes so the strength and significance level of the BRSV association might be low. The Mh–BRD association would not change when BRSV is controlled.

Assuming no sampling error, when BRSV is added to the model (*ie* its effects are controlled) it does not change the Mh–BRD association. By itself, BRSV might or might not be statistically associated with BRD; this depends on how much of Mh susceptibility is caused by BRSV and how much of BRD is attributable to Mh. However, when added to the model containing Mh, BRSV will not be statistically significant; any association it has with the outcome is already contained within the association explained by the exposure factor. Hence, in a forward model-building approach when Mh is in the model, BRSV would not be added and the likely inference

might be that it is causally unimportant. Technically however, it just means it has **no direct effect** on the outcome. The sample statistics are:

Crude: *OR*(Mh) significant
Crude: *OR*(BRSV) might or might not be significant – but *OR*(Mh) >*OR*(BRSV)
Conditional: *OR*(Mh|BRSV)=*OR*(Mh)

Note When describing relative relationships with '>', we assume that the associations are positive, that is producing odds ratios greater than 1. To include the possibility of both associations being negative, the > symbol might be read 'farther from 1' rather than just 'greater than 1'.

The *OR*(BRSV|Mh) is not a valid indicator of the causal association of BRSV with BRD; this *OR* reflects only the direct effect (which in this instance is 0). The crude *OR*(BRSV) is the correct estimate of the **total causal effect** of BRSV on BRD in this example.

13.11.4 Explanatory antecedent variable – complete confounding

See Example 13.13. The underlying causal structure is that BRSV precedes and causes (or predicts) both Mh and BRD, but Mh is not a cause of BRD. Statistically, we expect to observe a significant crude relationship between Mh and BRD because of the common cause BRSV. This association is causally spurious. However, when BRSV is added to the model, the association between Mh and BRD becomes non-significant, because BRSV now 'explains' the original association. Thus, we would infer (correctly) that Mh was not a cause of BRD. Adding BRSV to the model usually reduces the residual variance also. Many extraneous factors function as explanatory antecedents in this manner. The sample statistics are:

Crude: *OR*(Mh) and *OR*(BRSV) are significant, usually with
OR(BRSV) >*OR*(Mh)
Conditional: *OR*(Mh|BRSV)=1, (BRSV biases the *OR* for Mh if it is ignored)
OR(BRSV|Mh) >1

Note The results of the model with both BRSV and Mh included as predictors is not optimal for estimating the BRSV total causal effect. Once we remove all arrows emanating from BRSV, there is no pathway from BRSV through Mh to BRD, hence the model with BRSV only is preferred for estimating this causal effect. Controlling Mh might not change the BRSV coefficient greatly, but it is better NOT to control unnecessary variables as controlling them can necessitate having to control even more variables.

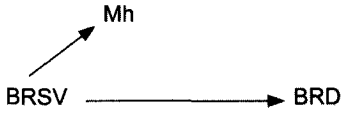
13.11.5 Explanatory antecedent variable – incomplete confounding

See Example 13.14. This is also a very common causal structure. The underlying causal structure is that BRSV causes (or predicts) both Mh and BRD, but Mh is also a cause of BRD. The sample statistics are:

Crude: *OR*(Mh) and *OR*(BRSV) are significant
Conditional: *OR*(Mh|BRSV) <*OR*(Mh) but *OR*(Mh|BRSV) $\neq$ 1

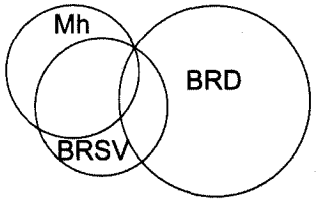
Example 13.13 An explanatory antecedent variable with complete confounding

Causal model



Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

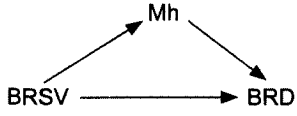
The Mh circle overlaps with the outcome, as they are statistically related until BRSV is added to the model (*ie* controlled). Then, the association becomes non-significant as all of the previous crude association between Mh and BRD is covered by the BRSV-BRD association.

Note The results of the model with both BRSV and Mh included as predictors are inappropriate to estimate the total causal effect of BRSV as only the direct effect would be reflected in the *OR* or regression coefficient. Mh would function as a partial intervening variable and should not be controlled when estimating the BRSV causal association with BRD. Again, the model with only BRSV is preferred for this purpose.

The model with both predictors included is appropriate for estimating the total causal

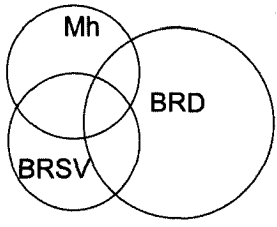
Example 13.14 An explanatory antecedent variable with partial confounding

Causal model



Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

The Mh circle overlaps with the outcome. The association remains statistically significant when BRSV is added to the model (*ie* controlled), but some of the previous association is now attributed to BRSV. Thus, the Mh–BRD association is not as strong when BRSV is controlled as when it was not included. Adding BRSV to the model explains more of the variation in BRD than just knowing Mh status.

effect of Mh. Statistically, as Mh still has an association with BRD after control of BRSV, this is the best estimate of its causal association with BRD. Thus, we would infer that Mh was a cause of BRD, and that the reduced 'strength' was the best estimate of magnitude of causal effect because the spurious causal component (from BRSV) was removed. Again, adding BRSV to the model usually decreases the residual variance of the model.

13.11.6 Intervening variable

See Example 13.15. An intervening variable is one that, in causal or temporal terms, intervenes in the causal or temporal pathway between exposure and disease. Now although unlikely from a biological point of view (humour us on this), the underlying causal structure is that Mh causes (or predicts) BRSV and BRSV causes BRD. The sample statistics are:

Crude: Likely both *OR*(Mh) and *OR*(BRSV) significant
Conditional: *OR*(Mh|BRSV)=0

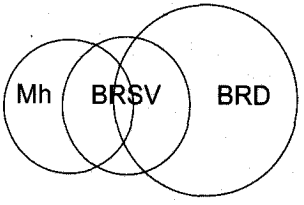
Example 13.15 An intervening variable

Causal model

Mh → BRSV → BRD

Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

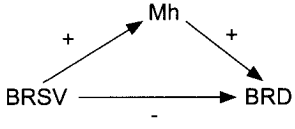
The Mh circle might or might not overlap with the outcome. However, any association of Mh with BRD disappears when BRSV is added to the model (*ie* controlled). It might well be that some or all of the effect of Mh on BRD is mediated through BRSV (and in that sense Mh is still a cause of BRD), but adding BRSV to the model would lead us to conclude that Mh was not associated with BRD and therefore we might infer that Mh was not a cause of BRD. Intervening variables should be identified and should not be controlled when estimating the causal effect of an exposure.

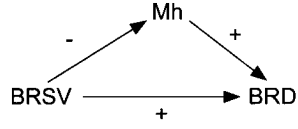
Although this model is improper in the context of ascertaining the causal association of Mh on BRD, the model with both Mh and BRSV would provide a reasonable estimate of the causal association of BRSV with BRD. Nonetheless, the model with only BRSV included would be preferable for estimating the BRSV causal effect.

As noted, we recognise that this is, biologically, a silly example because we have no evidence that Mh would cause increased susceptibility to BRSV in the context of feedlot respiratory disease. However, often it is not so obvious. Thus, it is very important to identify intervening variables and not 'control' them (*ie* do not put them in the model). Intervening variables might be totally or only partly caused by the exposure but should not be 'controlled'. They are not confounders but they cause similar changes in the measure of association to explanatory variables; thus, we must know the likely causal structure and time sequence between variables to differentiate explanatory from intervening variables because they cannot be differentiated analytically. This is a major reason for our stressing the development of explicit causal models before initiating analyses.

13.11.7 Distorter variable

Causally this is the same setup as for explanatory variables except that at least one of the causal actions is of a different sign than the other two (*ie* one of the causal arrows reflects prevention not causation). In our example, there are two possible underlying causal structures, assuming Mh is a cause of BRD. In the model on the left, Mh is a cause of BRD and BRSV prevents BRD but is causally and statistically positively correlated with Mh. In the model on the right, Mh is a cause of BRD and BRSV is also a cause of BRD, but BRSV is causally and statistically negatively correlated with Mh. Thus, the causal structures could be either:





The sample statistics for the left-side model are:

Crude: *OR*(Mh) and *OR*(BRSV) might be <1, =1 or >1

Conditional: *OR*(Mh|BRSV) >1

OR(BRSV|Mh)<1

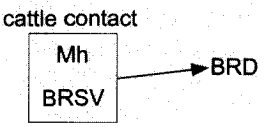
Again, to estimate the causal association of Mh with BRD, we need to control for BRSV. Controlling BRSV will increase the strength of association between Mh and BRD (*eg* a non-significant *OR*(Mh) might become significant when BRSV is controlled). This potential for increasing the *OR* is of significance in model-building. When it occurs it signals an underlying relationship similar to that described here. It is also possible that a significant positive association can become a significant negative association, and only 'distorters' can cause this reversal in the direction of association. The preferred model to estimate the total causal association of BRSV with BRD is the model with only BRSV included. When Mh is included, only the direct effects of BRSV are obtained.

13.11.8 Suppressor variables and refinement of exposure and outcome variables

See Example 13.16. Here the underlying causal structure is that Mh is a cause of BRD

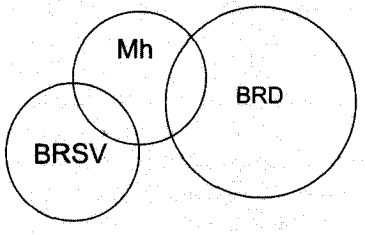
Example 13.16 A suppressor variable

Causal model



Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model



Comment

Before control of BRSV the variable 'cattle contact' is not, or only weakly, associated with BRD. Once BRSV is controlled by refinement (usually) or analysis, the Mh circle overlaps with the outcome indicating an association of 'cattle contact' with BRD. By controlling the non-causal component of our global variable, we increase the strength of the remaining factors' association with the outcome.

and BRSV is not. What distinguishes this from the other examples of relationships with extraneous variables is that both Mh and BRSV are members of the same global variable as defined by the researcher. For example, we might have formed a variable called 'cattle contact' that signified exposure to both these agents. However, because we are assuming that BRSV is not a cause of BRD (in this example), when BRSV is controlled, it will reveal or strengthen the suppressed association between Mh and BRD. BRSV is the (or one of the) irrelevant components of the global variable 'cattle contact'. The refined variable, without BRSV included, would have a stronger association with BRD. Control in situations such as this is usually by **refinement** of the predictor variable(s), but can be accomplished using analytical methods also.

Suppression often occurs with portemanteau-type (global) predictor variables (these are crudely defined or complex variables that contain a number of components). By refinement (stripping away the useless parts), the components of the original variable that are important can be identified. For example, 'ration' might need to be refined to locate which components (if any) of ration (length of roughage, amount of roughage *etc*) are related to abomasal displacement in dairy cows. We had suppression in mind when discussing combining length of exposure with dose of exposure to make a composite variable (in cohort and case-control studies; Chapters 8 and 9). Hence, we stated that it is best to examine the relationship of the components separately before assessing the composite variable for this reason.

Suppression of the dependent variable can also occur. As an example, perhaps only fibrinous pneumonia, not other types of respiratory disease, is related to Mh. Thus, if crude morbidity is the outcome variable, the association between Mh and BRD will be weak. If cause-specific BRD is used as the outcome, the stronger association between Mh and fibrinous pneumonia can be uncovered. Thus, whenever possible, refine the

exposure factors and outcome variables to the point that suppression is unlikely. The extent of refinement used will, however, depend on the objectives of the study as well as practical constraints.

13.11.9 Moderator variable

See Example 13.17. Moderator variables produce statistical interaction. The underlying causal structure is that Mh causes BRD, but only when BRSV is present. Hence, the statistical strength of its association with BRD **depends on** the presence or absence of BRSV. Recall from Chapter 1, that interaction is the statistical result of the joint causal effect of two or more factors on an outcome parameter. Interaction can, but doesn't necessarily, reflect a biological property of the joint effect of variables (*ie* either synergism or antagonism). Moderator variables might or might not be confounders. The sample statistics are

Crude: OR(Mh) and OR(BRSV) usually$\neq$1, but might=1

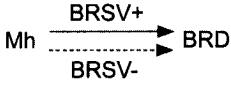
Conditional:OR(Mh|BRSV) might not be meaningful because

OR(Mh|BRSV+)$\neq OR$(Mh|BRSV-)

χ^2_{homo} is significant (section 13.9)

Example 13.17 A moderator variable

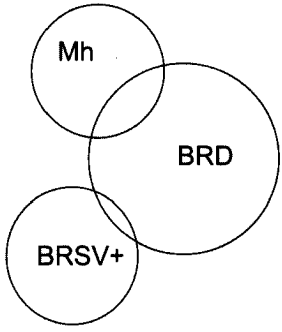
Causal model



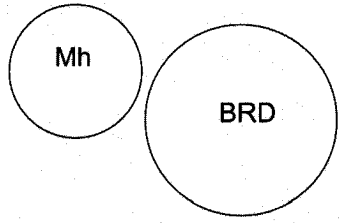
Mh = Mannheimia hemolytica
BRSV = Bovine respiratory syncytial virus
BRD = Bovine respiratory disease

Statistical model

When BRSV is present, the effect of Mh is present:



When BRSV is absent, the effect of Mh is absent:



Comment

The Mh circle overlaps with the outcome only when BRSV is present. This is the exact basis of the causal models shown in Examples 1.1 and 1.2. No disease occurs unless the two factors are present. Interaction is extremely important to identify as it has large implications for disease prevention.

13.12 Summary of effects of extraneous variables

As a summary, in Table 13.3, we indicate the likely impact of adding each type of extraneous variable (*ie* BRSV) to an analysis of the Mh–BRD association on the magnitude (or direction) of the association of Mh with BRD. The association can be measured by an odds ratio or regression coefficient (β_1); the latter denotes the magnitude and direction of association in linear (Chapter 14), logistic (Chapter 16) and Poisson (Chapter 18) regression models, and in survival models (Chapter 19).

Table 13.3 The effect of controlling BRSV on the Mh–BRD association (measured in a regression-type model)

BRSV is an ... variable	Effect	Comments (including impact on regression models)
Exposure-independent	no change	BRSV explains some of BRD incidence, so the residual σ^2 is smaller and the significance of β_1 increases
Simple antecedent	no change	No effect on the analysis but BRSV might be important to know about, from a preventive perspective, if it is easier to modify than Mh
Explanatory antecedent (complete confounding)	 becomes 0	Control of BRSV will remove any Mh association with BRD. The R^2 of the model should increase as the residual variance decreases
Explanatory antecedent (incomplete confounding)		Controlling BRSV will impact on the significance of β_1 depending on the strength of the BRSV effect on Mh and on BRD. The R^2 of the model should increase
Intervening		Because BRSV is more closely related to BRD, it probably has a stronger association and explains more variability. The β_1 is reduced in size and significance. If all of the effect passes through the intervener, it will remove all of the Mh effect on BRD
Distorter		Essentially the same impacts as an explanatory-antecedent variable except the Mh effect is increased, or in the opposite direction, to the crude association
Suppressor		As the global variable containing Mh is refined, it will now have a stronger relationship with BRD, it will probably explain more of the variation in the outcome
Moderator	not applicable	In the presence of interaction, the effect of one variable depends on the level of the other variable, hence separate estimates of effect are required

SELECTED REFERENCES/SUGGESTED READING

1. Alford P, Geller S, Richardson B, Slater M, Honnas C, Foreman J, Robinson J, Messer M, Roberts M, Goble D, Hood D, Chaffin M. A multicenter, matched case-control study of risk factors for equine laminitis. Prev Vet Med 2001; 49: 209-222.
2. Greenland S, Morgenstern H. Confounding in health research. Ann Rev Pub Hlth 2001; 22: 189-212.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48.
4. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as prerequisite for confounding evaluation: An application to birth defects epidemiology. Am J Epidemiol 2002; 155: 176-184.
5. Manske T, Hultgren J, Bergsten C. The effect of claw trimming on the hoof health of Swedish dairy cattle. Prev Vet Med. 2002; 54: 113-29.
6. Rothman KJ, Greenland S. Modern epidemiology 2d ed. Lippincott-Raven, 1998.
7. Susser M. Causal thinking in the health sciences: concepts and strategies of epidemiology, Oxford University Press Toronto (Out of print), 1973.
8. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol. 1991; 44: 221-232.
9. Veling J, Wilpshaar H, Frankena K, Bartels C, Barkema HW. Risk factors for clinical Salmonella enterica subsp. enterica serovar Typhimurium infection on Dutch dairy farms. Prev Vet Med. 2002; 54: 157-68.
10. Walker C, Canfield PJ, Love DN, McNeil DR. A longitudinal study of lymphocyte subsets in a cohort of cats naturally infected with feline immunodeficiency virus. Aust Vet Jour 1996; 73: 218-224.
11. Yates WDG. A review of infectious bovine rhinotracheitis, shipping fever pneumonia and viral-bacterial synergism in respiratory disease of cattle. Can J Comp Med 1982; 46: 225-263.

14

LINEAR REGRESSION

OBJECTIVES

After reading this chapter, you should be able to:

1. Identify if least squares regression is an appropriate analytical tool for meeting your objectives given the characteristics of your data.

2. Construct a linear model to meet your objectives, including control of confounding and identification of interaction.

3. Interpret the regression coefficients from both a technical and a causal perspective.

4. Convert nominal, ordinal or continuous predictor variables into regular or hierarchical variables and interpret the resulting coefficients correctly.

5. Assess the model for linearity between continuous predictors and the outcome, for homoscedasticity, and normality of residuals. You should also be able to identify appropriate transformations of the outcome or predictor variables to help ensure that the model meets these assumptions.

6. Detect and assess individual observations as potential outliers, leverage observations and/or influential observations.

14.1 Introduction

Up to this point, most of our examples in which we relate an outcome to an exposure, have been based on qualitative outcome variables, that is variables that are categorical or dichotomous. Here we will discuss linear regression which is suitable for modelling the outcome when it is measured on a continuous, or near-continuous scale. In regression analysis the relationship is asymmetric in that we think the value of one variable is caused by (or we wish to predict it by) the value or state of another variable. The outcome variable is denoted as the dependent, or outcome, variable, whereas the 'causal' or 'predictor' variables are called the independent or predictor variables. We continue to refer to the **predictor** variable(s) of primary interest as the **exposure** variable(s). The predictor variables can be measured on a continuous, categorical or dichotomous scale.

14.2 Regression analysis

When only one predictor variable is used the model is called a **simple regression model**. The term 'model' is used to denote the formal statistical formula, or equation, that describes the relationship we believe exists between the predictor and the outcome. For example, the model

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \textit{Eq 14.1}$$

is a statistical way of describing how the value of the outcome (variable Y), changes across population groups formed by the values of the predictor variable X_1. More formally it says that the mean value of the outcome for any value of the predictor variable is determined using a starting point, β_0, when X_1 has the value 0 and, for each unit increase in X_1, the outcome changes by β_1 units. β_0 is usually referred to as the **constant** or the **intercept** term whereas β_1 is usually referred to as the **regression coefficient**. The ε component is called the **error** and reflects the fact that the relationship between X_1 and Y is not exact. We will assume that these errors are normally and independently distributed, with zero mean and variance σ^2. We estimate these errors by **residuals**; these are the difference between the observed (actual) value of the observation and the value predicted by the model.

The βs represent population parameters which we estimate based on the observed data and our model; hence, we could write $\hat{\beta}$ but to simplify the notation we will use only β unless otherwise necessary. We will refer to predictor variables as Xs. In general, we will denote the number of observations as n. Thus our predictions for the observed data are:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i}, \quad i = 1,\ldots,n$$

where $\hat{Y}_i$ is the predicted value of the outcome in the i^{th} observation at the observed value (X_{1i}) of the predictor X_1. More generally, for any value X_1 of the predictor variable, the prediction equation is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 \qquad \textit{Eq 14.2}$$

We will denote the specific observation by the subscript *i*; but in most instances, for simplicity we will omit reference to specific observations. Bear in mind that in using *X*-variables to predict *Y* in a regression model there is no necessary underlying assumption of causation; we might just be estimating predictive associations. Nonetheless, we might use terms such as '*X* affects *Y*', or the 'effect of *X* on *Y* is...' when interpreting the results of our models.

Almost without exception, the regression models used by epidemiologists will contain more than one predictor variable. These belong to the family known as multiple regression models, or **multivariable** models (note that 'multivariate' indicates two or more outcome variables). With two predictor variables, the regression model could be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

which suggests that we can predict the value of the outcome *Y* knowing the baseline (intercept or constant) β_0 and the values of the two independent (predictor) variables (*ie* the *X*s). The parameters β_1 and β_2 describe the direction and magnitude of the association of X_1 and X_2 with *Y* – more generally there can be as many *X*-variables as needed, not just two (see Chapter 15 for model-building). A major difference from the simple regression model is that in the above multivariable model, β_1 is an estimate of the effect of X_1 on *Y* after controlling for the effects of X_2, and β_2 is the estimated effect of X_2 on *Y* after controlling for the effects of X_1. As in simple regression, the model suggests that we cannot predict *Y* exactly, so the random error term (ε) takes this into account. Thus, our prediction equation is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where $\hat{Y}$ is the predicted value of the outcome for a specific set X_1 and X_2 of values for the two predictors. In this equation, β_1 describes the number of units change in $\hat{Y}$ as X_1 changes by one unit, X_2 being held constant, while β_2 describes the number of units change in $\hat{Y}$ as X_2 changes by one unit, X_1 being held constant.

In observational studies, incorporating more than one predictor almost always leads to a more complete understanding of how the outcome varies and it also decreases the likelihood that the regression coefficients are biased by confounding variables. Assuming that we have not included intervening variables, or effects of the outcome in our model, the βs are not biased (confounded) by any variable included in the regression equation, but can be biased if confounding variables are omitted from the equation (Robins and Greenland, 1986). From a causal perspective, if intervening variables are included the coefficients, do not estimate the causal effect (see section 14.7). Unfortunately, one can never be sure that there are not other variables that were omitted from the model that also affect *Y* and are related to one or more of the *X*s. These *X*-variables could be unknown, not thought (at least initially) to be important, or (as it often happens) not practical/possible to measure. In other circumstances we might have

numerous potential confounders and need to decide on the important ones to include. As noted in Chapter 15, a major trade-off in model-building is to avoid omitting necessary variables which could confound the relationship described by the βs, while not including variables of little importance in the equation, as this will increase the number of SE(β)s estimated and lead to poor performance of the equation on future datasets. Also, having to measure unnecessary variables increases the cost of future work.

In order to assist with the principles of describing multiple regression we will develop examples from a dataset concerning the impact of diseases on reproductive performance in dairy cows on two farms. In the main, the events in the dataset daisy represent postpartum reproductive events, although other postpartum diseases such as milk fever and mastitis are also recorded. Although these data were obtained from actual studies, they have been changed to allow us to make specific points about correlation analysis and regression models, so the results should not be used as a basis for making biological inferences about reproductive problems. The names of the variables we will use and their labels are shown in Table 14.1; details can be found in Chapter 27. The diseases are listed in order of their average time to occurrence (*eg* milk fever occurs before retained fetal membranes).

Table 14.1 Selected variables from the dataset daisy

Variable	Scale of measurement	Description
farmnum	nominal	denotes the farm identification
age	continuous	age of the cow in years
milkfvr	dichotomous	did the cow have milk fever?
retpla	dichotomous	were fetal membranes retained for 48 hours?
metritis	dichotomous	did the cow have metritis?
ovar	dichotomous	did the cow develop cystic ovarian follicles?
firstbrd	continuous	days from calving to first breeding; 40-day minimum period
firstest	continuous	days from calving to first observed estrus
calvcon	continuous	days from calving to conception (pregnancy)

14.3 Hypothesis testing and effect estimation

14.3.1 The ANOVA table

The idea behind using regression is that we believe that information in X can be used to predict the value of Y. Now, if we have collected the data, we know the observed Y-values and we can describe the distribution of Y using the mean, variance and other statistics. With no further information, the best estimate of the value of Y for a particular subject would be an estimate of central tendency such as the mean value. However, if the X-variable contains information about the Y-variable, we should be able to do a

better job of predicting the value of Y for a given individual (cow) than if we did not have that information. The formal way this is approached in regression is to ascertain how much of the sums of squares of Y (the numerator of the variance of Y) we can explain with knowledge of the X-variable(s). Formally this decomposition of the total sum of squares (SS) is shown in the second column of Table 14.2 (*ie* SST=SSM+SSE; also, dfT=dfM+dfE):

Table 14.2. An analysis of variance (ANOVA) table showing the decomposition of sums of squares in a regression model with k predictor variables

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Model (or regression)	SSM $= \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	dfM = k	MSM = SSM/ dfM	MSM/MSE
Error (or residual)	SSE $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	dfE = n-(k+1)	MSE = SSE/dfE	
Total	SST $= \sum_{i=1}^{n} (Y_i - \overline{Y})^2$	dfT = n-1	MST = SST/dfT	

In the formulae in the table, $\overline{Y}$ is the mean of the Ys, and k is the number of predictor variables in the model (not counting the intercept). When the SS are divided by their degrees of freedom (df), the result is a **mean square**, here denoted as MSM (model), MSE (error) and MST (total) – in other settings we might call these variances but the jargon in regression is to call them mean squares. The MSE is our estimate of the error variance and therefore also denoted as σ^2. Furthermore, σ, the square root of σ^2, is called the **root MSE**, or the **standard error of prediction** (see Example 14.1).

The sums of squares are partitioned by choosing values of the βs that minimise the SSE (or MSE); hence the name 'least squares regression'. There is an explicit formula for doing this, which, in general, involves matrix algebra, but for the simple linear regression model the βs can be determined using:

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}_1 \quad \text{and} \quad \beta_1 = \sum (X_{1i} - \overline{X}_1)(Y_i - \overline{Y}) / \text{SSX}_1 \quad \left(\text{with } \text{SSX}_1 = \sum (X_{1i} - \overline{X}_1)^2\right) \quad \textit{Eq 14.3}$$

For a small dataset, these computations could be done by calculator, but in practice we always use computer software.

14.3.2 Assessing the significance of a multivariable model

To assess whether or not at least some of the variables in the model have a statistically significant relationship with the outcome, we use the F-test of the ANOVA table. The null hypothesis is $H_0 : \beta_1=\beta_2=...\beta_k=0$ (*ie* all regression coefficients except the intercept are zero). The alternative hypothesis is that this is not true, that is at least some (but

not necessarily all) of the βs are non-zero. The distribution of the F-statistic is an F-distribution with the first (*ie* numerator) degrees of freedom equal to dfM and the second (*ie* denominator) degrees of freedom equal to dfE (as given in Table 14.2, the ANOVA table, and shown in Example 14.1).

Example 14.1 Simple linear regression
data=daisy

As an example of a simple regression model, we will develop a model where calving to conception interval (denoted by a variable called -calvcon-) is the outcome variable and -age- (of dairy cow) is the predictor variable. We believe that the outcome is influenced by, or at least predictable by, -age- and each year of -age- will exert an effect on the outcome of size β_1 days for each year increase in age.

Analysis of variance table showing the partitioning of sums of squares in a simple regression model of -calvcon- on - age-:

Number of obs = 145
F(1,143) = 16.50
Prob > F = 0.0001
R-squared = 0.1034
Root MSE = 45.899

Source	SS	df	MS
Model	34758.022	1	34758.022
Residual	301265.744	143	2106.754
Total	336023.766	144	2333.498

The variance (*ie* mean square) of -calvcon- is greater than the mean square residual suggesting that we can do a better job of predicting the calving to conception interval if we know the age of the cow, than if we do not. The root MSE is 45.90 and has the same units (*ie* days) as the SD (standard deviation) of the outcome variable -calvcon- (SD=48.31 days).

Regression coefficients from regressing -calvcon- on -age-:

	Coef	SE	t	P	95% CI	
age	7.816	1.924	4.06	0.000	4.012	11.620
constant	84.904	9.068	9.36	0.000	66.980	102.829

The coefficient for -age- indicates that for each year increase in age, -calvcon- goes up by 7.8 days. The t-statistic (with 143 df) is clearly significant so we reject the null hypothesis in favour of the 7.8-day value. Often our preference is to estimate the likely values of the effect measure by examining the confidence intervals for the coefficient. Here the likely interval is from 4.0 to 11.6 days increase in -calvcon- for each year increase in age.

We usually do not test the intercept (or constant) value. β_0 reflects the value of the outcome (-calvcon-) when the X-variable (-age-) has the value 0, but no cows calve when they are 0 years of age. We shall comment subsequently on how to modify the predictor variable so that the constant term gets a sensible interpretation.

However, some care is necessary when interpreting the model F-statistic as its meaning changes with the method of model-building. The F-test probably only has a straightforward meaning when the Xs are manipulated treatments in a controlled experiment, and all comparisons are appropriately planned *a priori*. In observational studies, the F-statistic is influenced by the number of variables available for entry, their correlations with each other, and the total number of subjects (sampling units). Most variable selection methods (Chapter 15) choose variables in a manner that tends to maximise F; hence the observed F overestimates the actual significance of the model. On the other hand, if useless variables are forced into the model with the hope of controlling all confounding, the F-statistic might be biased downwards. Sometimes with highly correlated variables in the model the F-test might be significant yet the test of the individual coefficients might suggest that none of them differ significantly from zero (see section 14.5).

14.3.3 Testing the significance of a regression coefficient

A t-test with n-(k+1) degrees of freedom (dfE) is used to evaluate the significance of any of the regression coefficients, *eg* the j^{th} coefficient. The usual null hypothesis is $H_0 : \beta_j=0$ but any value of β^* other than 0 can be used in $H_0 : \beta_j=\beta^*$ depending on the context. The t-test formula is:

$$t = \frac{\beta_j - \beta^*}{SE(\beta_j)} \qquad \textit{Eq 14.4}$$

where $SE(\beta_j)$ is the standard error of the estimated coefficient. This standard error is always computed as the root MSE times some constant depending on the formula for the estimated coefficient and the values of the X-variables in the model. Except for the simplest situations it is not easily computable; however, it is always given in the computer output from the estimation of the model. For a model with only one predictor (X_1), the standard error (SE) of the regression coefficient is:

$$SE(\beta_1) = \sqrt{MSE/SSX_1} \qquad \textit{Eq 14.5}$$

As the formula indicates, both the variance of X_1 and the MSE affect the standard error. Examples of individual t-tests of coefficients are shown in Examples 14.1 and 14.2.

Similar to the F-statistic, the inference to be made based on the probability associated with the calculated t-statistic is often difficult to assess in non-experimental studies. In experiments, the Xs are manipulated treatments, or blocking factors, and the observed t-value can be referred to tables (of the t-distribution) for a P-value (observed level of significance). The same is probably true if the variable being tested in an observational study was of *a priori* interest (*eg* if the observational study was conducted to determine the effect of a specific X on Y, given control of a set of other variables). However, if a variable selection programme was used to sort through a list of variables, selecting those with large t-values in the absence of specific *a priori* hypothesis, then the actual level of significance is higher than the nominal level of significance (usually termed α) that you specify for a variable to enter/stay in the equation. Nonetheless, using the

Example 14.2 Multiple linear regression
data=daisy

A multiple regression model of -calvcon- on -age-, -metritis- and -ovar- is shown below. We code the variable -metritis- (denoting whether or not a cow had -metritis-) as 1 if the cow had -metritis-, and as 0 for cows without -metritis-. The variable -ovar- is similarly coded, -age- is in years.

Number of obs = 145
F(1,143) = 16.79
Prob > F = 0.0000
R-squared = 0.2632
Adj R-squared = 0.2475
Root MSE = 41.904

Source	SS	df	MS
Model	88440.458	3	29480.153
Residual	247583.307	141	1755.901
Total	336023.766	144	2333.498

	Coef	SE	t	P	95% CI	
age	6.342	1.777	3.57	0.000	2.829	9.855
metritis	18.550	8.522	2.18	0.031	1.703	35.397
ovar	53.341	10.212	5.22	0.000	33.153	73.530
constant	79.884	8.410	9.50	0.000	63.258	96.511

When we include data on -age-, -ovar- and -metritis- the overall model is highly significant with an explained proportion of the total variation in the data (R^2=SSM/SST) of 26.3%. Controlling for -age- and -ovar-, cows with -metritis- have an 18.6-day delay in calving to conception interval, and controlling for -age- and -metritis-, cows with -ovar- have a 53.3-day delay. The constant (intercept) represents the value of -calvcon- for cows of 0 years of age without -ovar- and -metritis-. We will leave the 'sensible' interpretation of this to section 14.4.1.

P-value as a guideline is a convenient and accepted way of identifying potentially useful predictors of the outcome.

14.3.4 Estimates and intervals for prediction

Calculating the point estimate for predictions in regression is straightforward. The complex component is determining the appropriate variance associated with the estimate, because there are two types of variation in play. One derives from the estimation of the parameters of the regression equation; this is our usual SE. The other is the variation associated with a new observation, *ie* the variation about the regression equation for the mean. The **prediction error** (PE) of a new observation involves both of these variations.

For example, in a simple linear regression model, the fitted (*ie* **forecast**) value $\hat{Y}$ for individuals with X_1=x* has an SE of:

$$\mathrm{SE}\left(\hat{Y}\middle|x^*\right)=\sigma\sqrt{\frac{1}{n}+\frac{\left(x^*-\overline{X}_1\right)^2}{\mathrm{SSX}_1}} \qquad \textit{Eq 14.6}$$

which can be interpreted as the variation associated with the expectation (*ie* mean) of a new observation, or an average of a large number of new observations, with the particular value x* chosen for prediction.

The PE for a new single observation with predictor value x* is increased because we must account for the additional σ^2 because the individual predicted value is unlikely to equal its expectation (*ie* unlikely to exactly equal the average value for all individuals with X=x*):

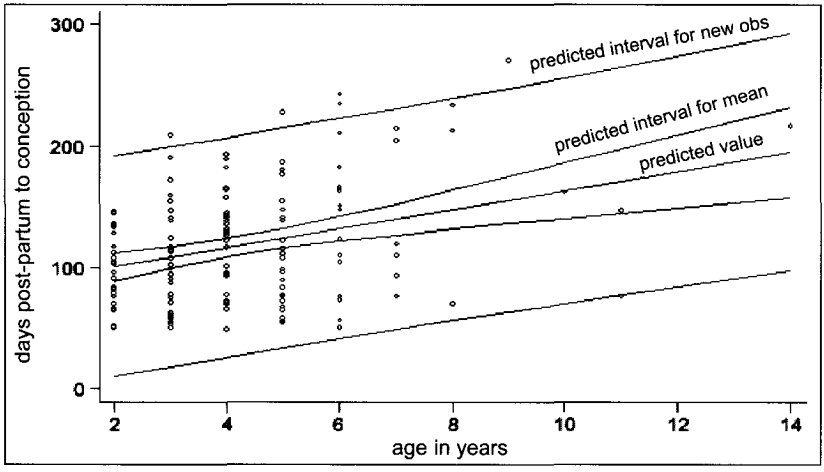
$$\mathrm{PE}\left(\hat{Y}\middle|x^*\right)=\sigma\sqrt{1+\frac{1}{n}+\frac{\left(x^*-\overline{X}_1\right)^2}{\mathrm{SSX}_1}} \qquad \textit{Eq 14.7}$$

Three points can be made here. First, the variation associated with the mean is much less, and forecast intervals much more narrow than those for a specific instance or subject (a single new observation). Second, the further that x* is from the mean value of X_1, the greater the variability in the prediction. We show the 95% confidence bounds (or forecast intervals) for the mean (using SE) and for a specific subject (using PE) in Fig. 14.1. Finally, the general formula for computing prediction errors from standard errors is:

$$\mathrm{PE}=\sqrt{\mathrm{SE}^2+\sigma^2} \qquad \textit{Eq 14.8}$$

where σ^2=MSE.

Fig. 14.1 Prediction intervals for means and observations



14.3.5 Interpreting R^2 and adjusted R^2

R^2 is sometimes called the **coefficient of determination** of the model; it describes the amount of variance in the outcome variable that is 'explained' or 'accounted for' by the predictor variables (see Example 14.2). It also is the squared correlation coefficient between the predicted and observed Y-values (and in simple linear regression as well the squared correlation between outcome and predictor).

Unfortunately, R^2 always increases as variables are added to a multiple regression model which makes R^2 useless for variable selection and potentially misleading. Hence, R^2 can be adjusted for the number of variables in the equation (k), and this adjusted value will tend to decline if the variables added contain little additional information about the outcome. The formula for the adjusted R^2 is: adjusted R^2=1-(MSE/MST); notice the similarity with the formula R^2=SSM/SST=1-(SSE/SST).

Note in Example 14.2 that the adjusted R^2 is slightly lower than the R^2. The adjusted R^2 is also useful for comparing the relative predictive abilities of equations, with different numbers of variables in them. For example, if one equation has seven variables and another equation three, the R^2 for the model with seven might exceed that for the model with three (and it always will if the smaller model is a submodel of the larger one), but its adjusted R^2 might be less. Adjusted R^2 is sometimes used as a basis for selecting potentially good models, but this approach is not without its drawbacks (see Chapter 15 about variable selection methods).

When assessing R^2 we should be aware that non-random sampling can have a pronounced effect on its value. For example, if you select subjects on the basis of extreme X-values, as in a cohort study, you might artificially increase the R^2. It would be okay to use regression to estimate the effect of X on Y, but the R^2 *per se* would be of little value. In a similar manner, if the X-values are limited to a narrow range, the R^2 might be very low.

14.3.6 Assessing the significance of groups of predictor variables

Sometimes (often) it is necessary to simultaneously evaluate the significance of a group of X-variables, as opposed to just one variable. For example, this approach should be used when a set of indicator variables has been created from a nominal variable (section 14.4.2), or if it is desired to add or remove more than one variable at a time (*eg* a set of variables relating to housing or feeding practices).

In order to assess the impact of the set of variables, we note the change in the error (residual) sum of squares (SSE) before and after entering (or deleting) the set of variables. (Alternatively, one might use the model sum of squares, as indicated below.) That is, note SSE_{full} with the variable set of interest in the model (called the 'full model'), then remove the set of variables (*eg* X_j and $X_{j'}$) and note the SSE_{red} (for the 'reduced model'). If variables X_j and $X_{j'}$ are important, then $SSE_{full} \ll SSE_{red}$ (and $SSM_{full} \gg SSM_{red}$).

The increase in SSE (or reduction in SSM) by deleting the variables from the model is divided by the number of variables in the set (which equals dfE_{red} - dfE_{full}) to give us the MSM from these variables. Dividing this MS by the MSE_{full} provides an F-test of the significance of X_j and $X_{j'}$ conditional on the other variables in the model. In summary, the formula to assess a set of variables is:

$$F_{group} = \frac{\left(\frac{SSE_{red} - SSE_{full}}{dfE_{red} - dfE_{full}}\right)}{MSE_{full}} \sim F(dfE_{red} - dfE_{full}, dfE_{full}) \text{ under } H_0 \qquad \textit{Eq 14.9}$$

where the null hypothesis H_0 is that the reduced model gives an adequate description of the data, and large values of the F-test are considered as evidence against H_0. The numerator of the formula might alternatively be calculated from differences of MS- and df-values for the model (instead of error); as SSM_{full}-SSM_{red}=SSE_{red}-SSE_{full}, it gives the same result. As will be seen subsequently, there is a similar test for use in linear, logistic and Poisson models. Most software contains specific procedures to automate this process. Example 14.3 shows the calculation of an F-test for three categorical variables that were added to the simple linear model shown in Example 14.1.

Example 14.3 Testing the significance of multiple variables
data=daisy

For example, suppose we had a model containing only -age- (see Example 14.1) and wondered, if as a set, the variables -retpla-, -metritis- and -ovar- added significantly to the model. The ANOVA table from the full model is shown below:

Source	SS	df	MS
Model	98924.110	4	24731.027
Residual	237099.656	140	1693.569
Total	336023.766	144	2333.498

In the simple model with only -age- as a predictor, we had SSE_{red}=301265.744. Hence the F-test is:

$$F = \frac{(301265.744 - 237099.656)/(143 - 140)}{1693.569} = 12.6$$

This F-statistic is very significant with 3 and 140 df ($P<0.001$) so we can safely infer that the three variables collectively add significant information to the model containing only -age-. This test does not say that all three are significant, only that the amount of information in the set of three variables adds significantly to the model.

14.4 NATURE OF THE X-VARIABLES

The X-variables can be continuous or categorical with the latter being either nominal (meaning that the variable's values constitute 'levels' (or categories) with no meaningful numerical representation) or ordinal (in which case the values represent ordered levels of the variable, *eg* high, medium, low). Examples of nominal variables include: farm identification, categories representing different ways of feeding colostrum, categories representing different breeds of dogs *etc*. Nominal and ordinal variables with more than two levels should not be used as predictors in their numerical form, they need to be converted to indicator variables. This is because the corresponding βs would be meaningless (*eg* because herd 4 is not twice something in herd 2, or breed 5 is not three units more than breed 2 *etc*), and would not achieve the desired effect (*eg* of removing herd-to-herd variation when examining the relationship between disease and production in cows or calves).

However, a nominal predictor with only two levels (a dichotomous variable) can be used directly when it is coded as 1 or 0 (*eg* the variables -metritis- and -ovar- in Example 14.2). Such variables often serve as answers to questions about present/absent, alive/dead, sick/well *etc*. The regression coefficient represents the difference in the outcome between the levels (*ie* level 1 minus level 0).

For categorical (nominal or ordinal) variables with multiple levels, we use indicator variables (also called dummy variables) to code the information into a set of dichotomous variables. Sections 14.4.2 and 14.4.3 review **regular indicator variables** that can be used for both nominal and ordinal variables, and **hierarchical indicator variables** applicable only to ordinal or quantitative variables. However, we first examine how to improve the interpretability of regression parameters.

14.4.1 Improving interpretability of the regression parameters

Often, the predictor variables have a limited range of possible, or sensible values. For example, many cannot be interpreted, sensibly, at the value 0 (*ie* if age, weight, or days to breeding were predictor variables, they have no meaningful interpretation at the value 0). Yet, the intercept reflects the value of the outcome when the predictor has the value 0. Thus, it is often useful to 'rescale' these variables by subtracting the lowest possible sensible value from each observed value before entering the variable into the model. Then, the intercept coefficient β_0 will be the value of the outcome at this lowest possible value of the X-variable instead of at zero. As an example this could be two years (age of first calving) for -age-. Rescaling has no effect on the regression coefficient or its SE, but it does change the value of the intercept (constant) (see Example 14.4). The rescaling can also be done with values other than the lowest possible value, for example a centre value (mean or median) of distribution of X.

Another use of rescaling is when the X-variable is measured with much greater accuracy than needed (*eg* measuring -age- in days in our example). Hence, in its original form, even if the variable has a large 'effect' on the outcome, its coefficient will be small reflecting the change in -calvcon- for each additional day of -age-. This problem can be

Example 14.4 Rescaling predictor variables
data=daisy

Here we rescale -age- by subtracting two years from the actual age, so our new variable is: age_sc=age^2. The effect of an increase of one year in age is the same as in the unscaled Example 14.1:

	Coef	SE	t	P	95% CI	
age_sc	7.816	1.924	4.06	0.000	4.012	11.620
constant	100.536	5.806	17.32	0.000	89.060	112.012

In the original scale (Example 14.1), days to conception was predicted to be 89.2 days for a 0-year-old cow. Here it is 100.5 days for a two-year-old cow, hopefully a much more sensible number.

circumvented by dividing the value of X by a suitable constant (such as age/365). For example, if one was predicting badger numbers using the area of pasture, if the latter was measured in square metres, it might be more practical to divide this by 100^2 so that the X-variable is now measuring hectares. Now the coefficient reflects the change in the number of badgers as pasture is increased by one hectare. In a similar manner, depending on the context, distances might be more practically expressed as km than metres, and weights by kg or 100 kg instead of grams.

14.4.2 Coding regular indicator variables

Indicator variables are created variables whose values have no direct physical relationship to the characteristic being described. For example, suppose there is a variable called -herdnum- that identifies what herd the animals in your study came from. Further, suppose there are three herds coded as -herdnum-=1, 2, or 3 (or A, B, C) and we wish to control for 'herd effect' when examining the potential effect of calfhood disease(s) on growth rate in calves. To do this, we create two regular **indicator** (sometimes called **disjoint**) variables (X_1 and X_2) as logical answers to the following questions: Is this calf from herd 1?; if yes, then X_1=1 else X_1=0. For the next indicator variable we ask: is this calf from herd 2?; if yes, then X_2=1 else X_2=0. With respect to these variables, the following values would be present in the dataset:

herdnum	X_1	X_2
1	1	0
2	0	1
3	0	0

Thus, herd 3 is identified as the herd with both indicator variables equal to 0, and will be the **referent** (or comparison level or reference category) for assessing the effect of herds 1 and 2 on the outcome. So, in general to code j levels of a nominal variable, j-1 indicator variables are required, and the j^{th} herd takes the value 0 for all the

indicators (see Example 14.5). As the third herd has become the referent level (when all the indicator variables are in the equation), β_1 (the coefficient of X_1) estimates the difference in the outcome between herds 1 and 3, whereas β_2 estimates the difference in the outcome between herds 2 and 3.

Example 14.5 Coding indicator (regular dummy) variables

We will demonstrate forming regular (*ie* disjoint) indicator variables from a nominal variable. For example, when conducting a study in which one predictor is method of colostrum feeding we might have coded the answers in the variable -colfeed- as 1=suckling, 2=nipple pail, 3=open pail, and 4=intubation. Let's assume that 'nipple pail' is a sensible referent and has sufficient sample size. The coding of the three disjoint variables could be completed by writing logical code to answer the following:

If colfeed=1	then suckle=1	else suckle=0
If colfeed=3	then openpail=1	else openpail=0
If colfeed=4	then tube=1	else tube=0

The effect and significance of each new variable (-suckle-, -openpail- and -tube-) would be in relation to nipple-pail feeding. Whether or not the information in the original variable -colfeed- added significantly to the model should be assessed by an F-test (see Example 14.3).

Because one of the levels of the nominal variable will be the referent, often there is merit in considering which level it should be. In terms of the information provided to the model, it does not matter which level is the referent, but careful consideration can enhance the interpretation of the coefficients. In essence, considerations about biological interpretation and the precision of estimates in each level of the nominal variable should be weighed in choosing a referent (*eg* if body temperature is recorded as below normal, normal or above normal, it might make sense to use 'normal' as the referent value). The referent should have a sufficiently large sample size so that the contrasts (comparisons with the referent) have reasonable precision. Sometimes the level of the nominal variable that has an 'average' response (*eg* close to the mean of the dependent variable) is the desired referent; this can lead to a situation where no design variables are significant, as the extreme categories might differ from each other but not from the outcome in the middle (mean) indicator. However the significance of the indicator variables as a set (section 14.3.6) is unaffected by the choice of reference category. In other instances, the choice of the referent can be arbitrary, as for example when the indicators are herd indicators and the herd effects are not of primary interest but they must be controlled to prevent confounding.

Most software programs have automated procedures to create indicator variables, and the coding can be more flexible than shown here. By default, some use the first category as the referent, others use the last category as the referent. All allow the user to set the referent using the contextual considerations just mentioned.

As noted earlier, all indicator variables (of each nominal variable) usually should be entered or excluded from the model as a set using the F-test in section 14.3.6. Once the set has been deemed to be important in a statistical sense or from the perspective of confounding control, it is often desirable to allow only some (*eg* the statistically significant or the 'important' indicators) to remain in the model. Removal of unnecessary indicators can aid the development of a more parsimonious model. The decision about removing some of the indicators can be assisted by testing the equality of selected indicator coefficients. (**Note** To select indicators in a statistically stringent sense, multiple comparison procedures must be applied.) One should also be aware that removal of some indicators changes the interpretation of the coefficients for the remaining indicators. For example, if only indicator X_1 is in the model, the referent (in the above example) will be both of the remaining herds; the referent will be the weighted average of the outcome in herds 2 and 3 and the coefficient associated with X_1 will represent the difference in response between herd 1 and this average. Any effects from indicators not included in the model are present in the constant term.

14.4.3 Coding hierarchical indicator variables

If the predictor variables are ordinal in type, (*eg* reflect relative changes in an underlying characteristic, *eg* severity of milk fever), it is sometimes difficult to associate the levels of severity with specific numerical values that would make it meaningful to use the variable as a continuous predictor. As an example, when coding a variable representing severity (*eg* using 1, 2, or 3 representing stage 1, stage 2 or stage 3 milk fever), there might be concern when using it as a continuous predictor based on the actual codes (*eg* is the biological effect of the difference between stage 1 and stage 2 milk fever the same as between stage 2 and stage 3?). It is always possible to use regular indicator variables, but they do not reflect the ordering of levels. Therefore, the use of **hierarchical** (or **incremental**) indicator variables is often the preferred approach, in order to maintain the ordering inherent in the original variable. This approach can also be used to recode a continuous variable based on using appropriate cutpoints. Hierarchical indicator variables contrast the outcome in the categorised version of the original variable against the level just preceding it (assuming all incremental variables are in the model). As with regular indicator variables it is possible to just include a subset of the indicators. One such situation occurs if we are interested in identifying cutpoints of an ordinal or continuous variable where the relationship with the outcome changes. In this setting we can select the most significant incremental variable for entry. The corresponding coefficient contrasts the outcome in levels of the categorised X-variable at or above the specific increment selected against the average of the outcome in the levels below it (Walter et al, 1987). Example 14.6 compares regular and hierarchical indicator variables.

14.4.4 Errors in the X-variables

In the regression model, the X-variables are 'fixed' (*ie* constant). In reality, they might be fixed because they are set by the experimenter in a controlled trial (*eg* treatment or dose) or because they represent values that *are* constant (*eg* site or year). However, when the X-variables are measured quantities (*eg* in observational studies), these measurements

Example 14.6 Coding hierarchical dummies
data=daisy

In this example, we will assume that relationship of -age- to -calvcon- is neither linear, nor curvilinear, and, wanting to avoid the use of a more complicated function, we want to use all of the information in -age- to prevent it from being a confounder of the effects of postpartum diseases. The average -calvcon- in each age is shown in Table 14.3 (all cows eight or more years of age are combined into one category for this example). Note the general increase in -calvcon- with age except for the five-year old and seven-year old groups. In order to create hierarchical variables our code is a logical answer to the following statements:

If age >2	then X_1=1	else X_1=0
If age >3	then X_2=1	else X_2=0
	...	
If age >6	then X_5=1	else X_5=0
If age >7	then X_6=1	else X_6=0

If we enter these into a multiple regression we obtain the coefficients shown in Table 14.3 (for purposes of comparison we also include the coefficients for a set of regular indicator variables, with age category '2 years' as referent).

Table 14.3 Summary of days postpartum to conception by age, and coefficients from regressing -calvcon- on -age- coded by regular or hierarchical indicator variables

	Number of	Mean	Regression	Coding	
Age	cows	calvcon	coefficient	Regular	Hierarchical
2	24	101.58	β_0	101.58	101.58
3	34	110.68	β_1	9.09	9.09
4	32	116.94	β_2	15.35	6.26
5	25	109.88	β_3	8.30	-7.06
6	16	141.69	β_4	40.10	31.81
7	6	136.00	β_5	34.42	-5.68
>7	8	173.00	β_6	71.42	37.00

Note that for both codings of -age- the intercept (β_0) is the -calvcon- value for two-year-old cows. With regular indicators, cows of three years of age have a 9.1 day longer calving to conception period than two year olds, and cows of four years of age have a 15.4 day longer calving to conception interval than two-year-old cows *etc.* With hierarchical (incremental) coding, cows of four years of age have a 6.26 longer calving to conception than three-year-old cows and cows of five years of age have a 7.06 day shorter interval than cows of age four *etc.* You might verify that both give the correct mean -calvcon- for each -age- interval.

might contain error: either a natural variation related to the measurements, or error in the sense of misrecordings. The consequence of this error is that relations between the outcome and the observed X-values are not the same as those with the true X values. The regression model estimates the relationship between the observed X-values and the outcome, and this is the appropriate relationship for prediction. A causal relationship between the X-variables and the outcome would usually rather involve the true values of the X-variables. Special models exist for taking error in the X-variables into account, so-called measurement error models, but they are beyond the scope of this book. However, as indicated in Chapter 12, if the magnitude of the measurement error is small relative to the range of the X-values in the model, we are not unduly worried when using the ordinary regression model. To ignore measurement errors generally tends to bias the parameters towards the null (*ie* effects will be (numerically) smaller than if the completely accurate information about the X-values was present). It can also be said that if the errors are large relative to the range of X-values serious consideration of the need for validation studies is in order.

14.5 Modeling highly correlated (collinear) variables

Despite the fact that multiple regression is used to adjust for correlations among predictor variables in the model, if the variables are too highly correlated then a number of problems might arise. Before discussing these, recall that in a multivariable regression model the estimated effect of each variable generally depends on the other variables in the model. On one hand, this is the advantage of a multivariable analysis – that variables are studied while taking the others into account and thereby avoiding duplication of effects. On the other hand, this means that the effect of any variable might change when other variables are added to or removed from the model. If, for a particular variable, such changes are large (*eg* involving a shift of sign), its interpretation becomes difficult. Only in the special case that all the X-variables are uncorrelated are the effects of different variables estimated completely independently of each other. Thus, the first problem arising from highly correlated (or collinear) predictors is that estimated effects will depend strongly on the other predictors present in the model. As a consequence, it might be difficult to statistically select the 'important' predictors from a larger group of predictors. Both of these concerns are less serious when the purpose of the analysis is prediction than when interpretation of causal effects is the objective. On a more technical note, the standard errors of regression coefficients might become very large in a highly collinear model (section 14.5.1), and hence we become less certain of the likely magnitude of the association.

To avoid this, a single X-variable should not be a perfect mapping of another X-variable or be perfectly predictable by a combination of the other X-variables in the regression model. However, even before the correlations become 'perfect' as a general rule, if two (or more) variables are highly correlated (collinear, $|\rho| > 0.8$–0.9), it will be difficult to select between (among) them for inclusion in the regression equation. When two variables are highly and positively correlated, the resulting coefficients (βs) will be highly and negatively correlated. (This means that if, as a result of sampling variation, one coefficient gets larger, then the other gets smaller. To illustrate, if $\beta_1=3$ and $\beta_2=2$,

then removing one of the variables from the model will increase the other coefficient. One well-known example is the negative correlation between the slope and intercept in a simple regression model.) In extreme situations none of the coefficients of the highly correlated variables will be declared significantly different from zero, despite the fact that the *F*-test of the model might indicate that the variable(s) contributes significantly to the model.

The best way of eliminating collinearity problems is through considered exclusion of one of the variables, or by making a new combination of the variables on substantive grounds. In extreme situations specialised regression approaches, such as ridge regression, might be needed.

Most software provides some warnings about possible collinearity using a variance inflation factor (section 14.5.1) or its reciprocal **tolerance**. Unfortunately, the methods we use for including interaction terms (section 14.6) and power terms (section 14.9.3) in models sometimes leads to a high collinearity between the variables. It is less serious than the collinearity between variables that are not constructed from each other, but will nevertheless affect the variance inflation factors. Thus we describe a general method for circumventing high correlations between these variables, known as **centring the variables** (section 14.5.2).

14.5.1 Variance inflation factors

The effect of entering a new variable into the model, on the variance of the coefficients for variables currently in the model can be assessed with a statistic known as the **variance inflation factor** (*VIF*). The formula for *VIF* is:

$$VIF = \frac{1}{1 - R_X^2} \qquad \textit{Eq 14.10}$$

where R_X^2 is the coefficient of determination from regressing the variable that is about to enter the model on the other variables in the model. As this coefficient gets larger so does the *VIF*. We illustrate the importance of the *VIF* in a simple linear regression model, in which the variance of the regression coefficient β_1 for X_1 is from Eq 14.4.

$$\text{var}^{(1)}(\beta_1) = \frac{\text{MSE}^{(1)}}{\text{SSX}_1} \qquad \textit{Eq 14.11}$$

where the superscript (1) refers to the simple linear regression model. When we place X_2 in the model, if it is related to X_1, three things will happen:

- the coefficient β_1 will change because we account for the correlation of X_1 with X_2,
- the residual sum of squares (and in most cases also the $\text{MSE}^{(2)}$) will become smaller because X_1 and X_2 together can predict Y better than X_1 on its own,

and

- the standard error of β_1 might increase by an amount roughly equal to $\sqrt{VIF}$; specifically, its variance in the combined model (2) with both X_1 and X_2 is:

$$\text{var}^{(1)}(\beta_1) = \frac{\text{MSE}^{(1)}}{\text{SSX}_1} \qquad \textit{Eq 14.12}$$

It is seen that the variance increases unless the reduction in $\text{MSE}^{(2)}$ from $\text{MSE}^{(1)}$ by adding X_2 more than offsets the increase due to the *VIF*. It is also true that adding variable X_2 can cause the variance of β_1 to decrease. This is most likely to happen if X_2 is a good predictor of the outcome and X_1 and X_2 are nearly (or totally) independent of each other.

The role of the *VIF* in multiple regression models is similar. A (conservative) rule of thumb for interpreting *VIF*s is that values above 10 indicate serious collinearity. As discussed above, this does not necessarily mean that the model is useless or that one is obliged to remove one or more *X*-variables from the model; it should however always be taken as a sign of warning for the interpretation of regression coefficients and the increase in standard errors. **Note** The *VIF*s reported in most software do not take into account if *X*-variables are related by construction, *eg* quadratic terms or indicator variables of a categorical variable – this limits the usefulness of *VIF*s in these situations.

14.5.2 Centring variables to reduce collinearity

Centring a continuous variable is performed by subtracting the mean value (or some other central value) from each observed *X*-value, similarly to the rescaling discussed in section 14.4.1. Centring the main effect variable and the power term (or an interaction term between two continuous variables) reduces the correlation between the variables to a low level (provided the variables are symmetrically distributed about their mean). If the distribution is not symmetric, then larger (or smaller) values than the mean might need to be subtracted. It should be stressed that centring only affects correlations between variables constructed from each other, and centring does not change the predictions or the fit of the model, only the values and interpretation of its coefficients. See Example 14.7 for a discussion of *VIF*s and centring.

14.6 Detecting and Modeling Interaction

In Chapter 1 we developed the view that given the component cause model we might expect to see interaction when two factors act synergistically or antagonistically. Whereas within limits this might be true, the significance of an interaction term need not indicate anything about the causal model; it might merely describe the nature of the relationship being modelled. In the previous section the model contained only 'main effects' of the *X*s, hence it assumes that the association of X_1 to Y is the same at all levels of X_2 and the association of X_2 to Y is the same at all levels of X_1. A test of this assumption is to examine if an 'interaction term' adds significantly to the regression model (see Example 14.8).

Example 14.7 Detecting and resolving collinearity by centring variables
data=daisy

In this example we regress -calvcon- on -age- and then on -age- and age-squared (-age_sq-). The latter term can be used to explore if the -age- versus -calvcon- relationship is curvilinear, and if it is, a convenient way of 'linearising' the relationship of -age- to -calvcon- is to insert a power term for -age-, in this case -age_sq-. However, the correlation between -age_sq- and -age- is 0.95. We also show the *VIF*s resulting from the two related variables being in the model.

First we regress -calvcon- on -age- only, then on -age- and -age_sq-; the results are shown below:

	Model 1			Model 2		
Variable	β	SE(β)	VIF	β	SE(β)	VIF
age	7.816	1.924	1	7.072	6.277	10.57
age_sq				0.063	0.507	10.57

This result clearly suggests that we do not need the -age_sq- term in the model. But humour us for this example. Notice the large *VIF*s and the corresponding increase in the SE(β_1). Below we show the results of the same model using centred variables. The mean -age- is 4.34 years so we subtract that from the cow's age and denote the variable as -age_ct-. Then we regress -calvcon- on -age_ct- alone, and on both -age_ct- and -age_ct_sq-:

	Model 1			Model 2		
Variable	β	SE(β)	VIF	β	SE(β)	VIF
age_ct	7.816	1.924	1	7.620	2.489	1.66
age_ct_sq				0.063	0.507	1.66

Notice that the coefficient and SE change only for the linear term in Model 2. In the multivariable model the SE of the linear effect of -age- is greatly reduced from that in the uncentred model. You should also note that centring is not necessary to determine if the power term is needed, only to reduce the *VIF* if it is needed. In the multivariable model, the coefficient for -age_ct- changes only a little from the earlier model, due to the almost negligible effect of -age_sq-. Note that the *VIF*s are now very small, because -age_ct- and -age_ct_sq- have a correlation of only 0.62. (In this example, because -age- is skewed to the right, to further reduce the correlation between -age_ct- and -age_ct_sq-, it would be better to use a value somewhat larger than the mean.)

Before leaving this example, a little reflection on what has been achieved by centring. The *VIF* and the SE(β) are lowered, but we don't intrepret the linear term independently from the quadratic term anyway, because that would contradict our curvilinear model. In situations with extreme correlation between a linear and quadratic term, the estimation might run into numerical trouble and centring will often improve matters considerably, but this is not the case here. The interpretation of the intercept has most likely improved, as discussed in section 14.4.1. If further variables were to be added to the model, the usefulness of *VIF*s has been improved by eliminating the intrinsic collinearity between linear and quadratic terms. Taken as a whole, these improvements might or might not warrant the effort.

Example 14.8 Testing for interaction
data=daisy

To test for interaction, we generate a product variable (ovarmet=ovar*metritis) and assess its significance:

	Coef	SE	t	P	95% CI	
metritis	23.602	9.352	2.52	0.013	5.112	42.089
ovar	63.109	11.462	5.51	0.000	40.449	85.768
ovarmet	-29.561	28.857	-1.02	0.307	-86.609	27.487
constant	105.186	4.426	23.76	0.000	96.435	113.936

Here, because the interaction term was not significant it appears that we can act as if the effects of -metritis- on -calvcon- are the same in cows with and without ovarian disease. Had the product term been significant it would imply that the effect of -ovar- on -calvcon- was different in cows with metritis than in cows without metritis, or the effect of -metritis- depends on -ovar-.

In the situation where X-variables are not indicator variables of a categorical variable, the interaction term is formed by the product X_1*X_2 which can be tested in the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

by assessing if β_3=0 (see Example 14.8). If interaction is absent (β_3 is deemed to be not different from 0), the main effects (or 'additive') model is deemed to describe the effects adequately. It is not necessary to centre variables (X_1 and X_2) to see if an interaction term is needed, because β_3 and its standard error will be unaffected. However, if the interaction is needed, centring might be useful because it allows us to interpret β_1 and β_2 as linear effects when the interaction cancels (*eg* β_1 applies to the situation when (the centred version of) X_2 is zero). Higher order interactions can be investigated by extending this process to an interaction term that is the product of three (or more) variables (see Chapter 15).

Interactions involving categorical variables (with more than two levels) are modelled by including products between all indicator variables needed in the main effects model. For example, the interaction between a 3-level and a 4-level categorical variable requires (3-1)*(4-1)=6 product variables. These six variables should be tested and explored as a group (section 14.3.6).

In many multivariable analyses the number of possibilities for interaction is large and there is no one correct way of selecting the model to assess if interaction is present. One strategy is to identify variables that make a significant contribution to the linear model without interaction terms present; this is the main effects model. Then, construct all possible two-way interactions for the significant main effect variables. Force all significant main effect variables into the equation and allow the program to select

significant interaction terms using a forward stepwise method (or eliminate them using a backwards approach). Another strategy is to only form biologically sensible interactions, perhaps based on what is already known in the literature. In cases with a large number of dummy variables (*eg* to control herd effects), one might have to assume that interaction is absent between the factors represented by the dummies and the other variables in the model unless the number of observations is sufficient to give reasonable power for assessing the numerous interactions. Three- and four-way interactions can be assessed in a similar manner; however their interpretation becomes more difficult as the number of terms increases. Thus, we recommend that these interactions only be investigated when there are good, biologically sound, reasons for doing so (see Chapter 15 for a further discussion of interaction terms). Example 14.9 demonstrates interaction between two dichotomous variables, 14.10 between a dichotomous and a continuous predictor, and 14.11 between two continuous predictors. Hopefully this approach to demonstrating and understanding interaction will provide a useful platform for your own analyses.

14.7 Causal interpretation of a multivariable linear model

So far in this chapter we have focused on the technical interpretation of regression coefficients. Example 14.12 is presented to focus on the causal interpretation of a multivariable linear model and in this circumstance, care is needed to ensure that only the appropriate variables are controlled in the analysis (see section 13.3).

14.8 Evaluating the least squares model

Valid regression analyses are based on a set of assumptions, and once our initial model is built we need to evaluate whether the model meets these (we say initial because after checking whether the model meets the key assumptions we might have to alter it). We will use the model from Example 14.12 containing five predictor variables for the purposes of this evaluation.

The key assumptions of the model are:

- **homoscedasticity** - the variance of the outcome is the same at all levels of the predictor variables (*ie* the variance in -calvcon- in cows that have a -firstbrd- of 63 days should be the same as the variance for those that have a -firstbrd- of 126 days, *etc*) and within all combinations of the values of the predictor variables. If this is true, then the error variance will be constant. This is an important assumption, perhaps more so than having a normal distribution of residuals.
- **normal distribution** - the errors should be normally distributed at all levels of the predictors, or at all combinations of predictors in the model (*ie* -calvcon- values for cows that have a -firstbrd- of 84 should be normally distributed as they should for all other values of -firstbrd-). We often try to get a quick assessment of this before starting the regression analysis by looking at the distribution of the outcome for normality. The errors from very non-normally distributed outcomes are unlikely to be 'normalised' by regression on the predictor variables unless the R^2 of

Example 14.9 Interaction between dichotomous variables
data=daisy

If we regress -calvcon- on -retpla-, -metritis-, and their interaction (denoted as -rpmet-) we observe the following estimates:

	Coef	SE	t	P	95% CI	
retpla	72.731	19.171	3.79	0.000	34.831	110.632
metritis	10.996	11.926	0.92	0.358	-12.581	34.574
rpmet	-49.996	25.292	-1.98	0.050	-99.996	0.004
constant	110.769	4.398	25.18	0.000	102.074	119.464

These results indicate that in the absence of either disease, the -calvcon- value is 110.8 days. By itself -retpla- adds almost 73 days to this, whereas by itself -metritis- adds only 11 days and this is not deemed to be significant. However when both are present instead of adding 84 days (73+11) to the baseline -calvcon-, only 84-50=34 days are added. This might represent a type of antagonism between these two diseases, the biology of which we will not attempt to explain. An easy way to see the results of interaction involving two dichotomous variables is to calculate the mean calving to conception interval for each of the covariate patterns formed by the combination of -metritis- and -retpla-, as shown in the table below (also giving the number of observations).

	metritis=0		metritis=1		total	
retpla	mean calvcon	no. obs.	mean calvcon	no. obs.	mean calvcon	no. obs.
0	110.8	108	121.7	17	112.3	125
1	183.5	6	144.5	14	156.2	20
total	114.6	114	132.0	31	118.3	145

You can also see from this table why the detection of interaction has low power (note the small cell sizes). Thus if detecting interaction is a major feature of the work it is often necessary to increase the sample size accordingly (see section 2.10.8).

the model is very high. On the other hand, as a simple example, if a strong dichotomous predictor for the outcome exists, then the raw distribution of the outcome will show as bimodal and therefore non-normal, but the residuals from the model might be normally distributed.

- **linearity** - because the relationship between the outcome and continuous or ordinal predictors (modelled as continuous) is described by a single coefficient, this assumes that the association is a straight-line relationship (*ie* a 21-day increase in -firstbrd- from 42 to 63 days affects -calvcon- by the same amount as a 21-day increase from a -firstbrd- of 105 to 126). There is no assumption involved for dichotomous variables as two points can always be connected by a straight line.
- **independence** - the values of the dependent variable are statistically independent from one another (*ie* the -calvcon- value of one cow does not depend on the -calvcon- value of other cows in the dataset). Usually we do not worry about

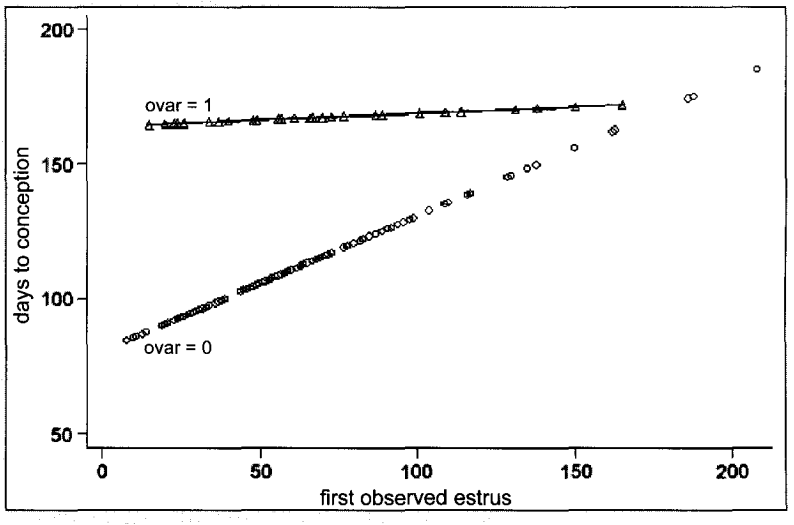
Example 14.10 Interaction between a dichotomous and a continuous variable
data=daisy

Here we regress -calvcon- on -ovar-, -firstest- and their interaction (denoted as -ovarest-) to see if the effect of timing of first estrus on calving to conception interval depends on whether or not the cow has cystic ovaries, or if the effect of cystic ovaries depends on when estrus is first observed.

	Coef	SE	t	P	95% CI	
ovar	83.635	18.659	4.48	0.000	46.747	120.522
firstest	0.504	0.093	5.41	0.000	0.320	0.687
ovarest	-0.456	0.218	-2.09	0.039	-0.887	-0.024
constant	80.162	6.670	12.02	0.000	66.976	93.347

The results of this model indicate that -ovar- by itself has a major impact on delaying conception, but this impact depends on the values of -firstest-; -ovar- has a big effect on cows with smaller values of -firstest- and a smaller effect on cows with a larger value of -firstest-. In cows that do not have -ovar-, a one-day delay in -firstest- delays conception by about 0.5 days. However, in cows with -ovar- the effect of -firstest- is virtually non-existent (as the +0.5 and the -0.5 cancel each other). Thus the effect of -firstest- depends on the level of -ovar-. Again in this example biological antagonism seems to be at work. In a situation such as this a graph might make the interaction effects more apparent. This is easily accomplished by obtaining the predicted -calvcon- from the model and graphing it against the continuous predictor (-firstest-) in cows with and without ovarian disease to produce Fig. 14.2.

Fig. 14.2 Interaction between -ovar- and -firstest- on days to conception



Here we can see the difference in effect of -firstest- between when -ovar- is absent (the sloped line) and when it is present (the near horizontal line). The interaction in the model corresponds to **different slopes** for the regression on -firstest- at the two levels of -ovar-. If interaction was absent, the regression lines on -firstest- would be **parallel**.

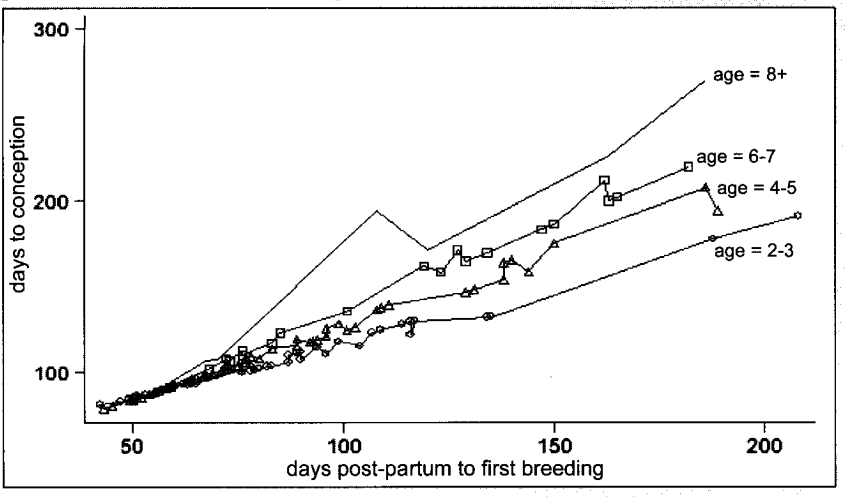
Example 14.11 Interaction between two continuous variables
data=daisy

Here we regress -calvcon- on -age- (ignoring that -age- is recorded to the nearest year), -firstbrd- and their interaction term (denoted -agebrd-). The results are:

	Coef	SE	t	P	95% CI	
age	-6.610	3.729	-1.77	0.078	-13.982	0.763
firstbrd	0.309	0.197	1.57	0.119	-0.081	0.700
agebrd	0.120	0.038	3.13	0.002	0.044	0.196
constant	71.564	17.890	4.00	0.000	36.197	106.930

In this model, the interaction contributes significantly to the prediction, but it is not very easy to tell what the effects of either variable are because we cannot examine an effect when the other variable is not present (as neither -age- nor -firstbrd- has a sensible interpretation at the value 0). Centring of both variables might therefore be helpful here. Generally, when trying to understand interaction between two or more continuous variables, a graph can almost be indispensable. The general approach we use is to obtain the predicted values of -calvcon- from the model, then divide one of the continuous variables into categories, and then plot the predicted values against the values of the continuous variable in each of the categories just created. Here we created four age categories beginning at 2, 4, 6, and 8 years and determined the predicted -calvcon- in each, then plotted these against -firstbrd- to produce Fig. 14.3.

Fig. 14.3 Interaction between -age- and -firstest- on days to conception



We see that the lines representing the effect of -firstbrd- on -calvcon- are not parallel, but rather they diverge, depending on the age of the cow. The non-linearity reflects the effect of different ages within each age category. The trend is that as the cow gets older, the impact of delaying -firstbrd- on -calvcon- is increased. Here we might have biological synergism at work. **Note** In small datasets the graph obtained in this manner might be very noisy for subgroups with only a few observations. In this instance, it is recommended to create the data for the graph directly from the regression model, by computing predicted -calvcon- values for a range of -firstbrd- values while fixing -age- at values 2, 4, 6 and 8 years. This graph will contain four straight (and non-parallel) lines.

Example 14.12 Causal interpretation of regression coefficients
data=daisy

As a concrete model to discuss we will regress -calvcon- on -farm1-, -age-, -metritis-, -ovar- and -firstbrd-. Here, -farm1- is an indicator variable representing the farm identification and is coded as 1 if the cow is from farm 1 and as 0 if the cow is from farm 2; you can imagine extending this process if we had a larger number of herds in our dataset, but this dataset only has observations from two herds. As a whole, this model is clearly significant, and gives the following coefficients with their SE, significance and confidence intervals:

	Coef	SE	t	P	95% CI	
farm1	6.108	5.551	1.10	0.273	-4.868	17.084
age	3.615	1.353	2.67	0.008	0.939	6.291
metritis	14.368	6.416	2.24	0.027	1.683	27.053
ovar	34.007	7.835	4.34	0.000	18.515	49.499
firstbrd	0.777	0.078	9.98	0.000	0.623	0.931
constant	23.585	8.199	2.88	0.005	7.375	39.795

All of the coefficients are significant at $P<0.05$ except for -farm1-; however we will leave this variable in the model on the assumption that we had strong beliefs that it would be a confounder and it will help correct for a lack of independence of outcome (*ie* -calvcon-) within each farm.

From a technical perspective the interpretation of the coefficients is as follows: 'having adjusted the coefficient for the other variables in the model', it holds that

- cows on farm 1 take 6.1 days longer to conceive than cows on farm 2,
- each year a cow gets older results in a 3.6-day delay in conception,
- cows with metritis have a 14.4-day delay in conception relative to cows without metritis,
- cows with -ovar- have a 34-day delay in conception relative to cows without -ovar-,
- each day's delay in -firstbrd- results in a 0.8-day delay in conception.

All these assertions are interpreted as 'all other things being equal' (*eg* cows compared with and without metritis should be on the same farm, of the same age, and have the same values of -ovar- and -firstbrd-). The confidence limits give a 'feel' for the upper and lower limits of these associations.

(*continued on next page*)

Example 14.12 (*continued*)

It is worthwhile to remind ourselves that the coefficients represent associations. If we are reasonably certain that we have controlled for the major sources of confounding then we can act as if these coefficients are reflecting the causal effects of each variable on -calvcon-. However, before that, there is one major item to consider; specifically what is the objective of the model? If it is to estimate the impact of delayed time to first breeding then the previous coefficient for -firstbrd- can be interpreted as the causal effect (0.8 days) of -firstbrd-. The coefficients for the remaining variables represent direct effects only, and thus they do not represent causal coefficients for these three variables.

If our goal was to estimate the causal effect of metritis, then neither -ovar-, nor -firstbrd- should be in the model because they are intervening variables for the relationship between -ovar- and -firstbrd-. Thus, the correct estimate for the causal effect of -metritis- would be 12 days (based on the model omitting -ovar- and -firstbrd-) and it is no longer significant. In this instance, inference about any causal effect different from zero would be based on the confidence limits (Robins and Greenland, 1986).

	Coef	SE	t	P	95% CI	
metritis	11.984	9.054	1.32	0.188	-5.915	29.883

If we have inadvertently, or by commission, placed an intervening variable in our model, then the coefficients for the other variables do not represent the causal effects as at least part of their effect is removed by the intervening variable. This feature of causal interpretation is very important to remember; it also points out that we can usually only estimate the causal effect of one or two variables in a given model. The coefficients for the remaining variables, if significant, represent only direct effects.

independence unless the context is such that the assumption is likely to be broken. For example, the structure of the data might signal a lack of independence when there are multiple observations from a single animal, or multiple animals from a herd. Methods for dealing with clustered data of this sort are presented in Chapters 20-23. Another situation where errors are likely correlated is in time series analyses because the value of the outcome on one day is likely correlated with the value on the previous day.

Each of the first three assumptions will now be discussed in more detail, and we can learn much about them by examining residuals, often using graphical methods, although formal tests are also available. Once we are satisfied that these three major assumptions have been met, we should pursue a more detailed search for specific observations that might be outliers, leverage points, and/or influential points. Because of the importance of residuals in these assessments we begin by describing different types of residual.

14.8.1 Residuals

The **raw residual** (r_i) is the difference between the observed and predicted value for the i^{th} observation and has the same units as the outcome variable,

$$r_i = Y_i - \hat{Y}_i \qquad \textit{Eq 14.13}$$

where the subscript i denotes the particular observation from 1 to n. The raw residual r_i is our 'estimate' of the error for observation i, by subtracting its predicted mean from the observation itself.

The mean of all residuals is zero, and the variance of each residual is:

$$\text{var}(r_i)=\sigma^2(1-h_i) \qquad \textit{Eq 14.14}$$

where h_i is the weight of the i^{th} observation in determining $\hat{Y}_i$. The h_i is called the **leverage** of that observation and indicates the potential for this observation to have a major impact on the model. In a simple regression model h_i has the following formula:

$$h_i = \frac{1}{n} + \frac{(X_{1i} - \overline{X}_1)^2}{\text{SSX}_1} \qquad \textit{Eq 14.15}$$

indicating that as the value of the predictor becomes farther from its mean, the leverage of the observation increases. Note that this 'potential' impact depends only on the predictor, not on the value of the outcome. Leverage has a more obvious meaning when the predictor is measured on the continuous scale. We return to the subject of leverage in section 14.10.2.

The raw residuals can be scaled by dividing them by their SE. If all observations are used to estimate σ^2 this produces what are called **standardised residuals** (these are also called internally studentised residuals by other authors):

$$r_{si} = \frac{r_i}{\sigma\sqrt{1-h_i}} \qquad \textit{Eq 14.16}$$

The reference distribution for standardised residuals is a t(dfE), so for sample sizes with n>30, based on the Gaussian distribution, there should be only about 5% of values outside of the interval (-2, 2). The major advantage of standardised residuals relative to raw residuals is that we have this absolute scale for what constitutes a large residual.

The raw and standardised residuals are computed from the prediction for the i^{th} observation from the regression equation based on all observations. That is, the observation itself contributes to the prediction. An influential observation might not show a large residual because of its impact on the prediction. To 'truly' examine whether the i^{th} observation is in agreement with the model based on the others, we should compare it with the prediction based on the other n-1 observations. Such (standardised) residuals are called **studentised** residuals or externally studentised residuals (others denote them as **deleted** residuals, or **jackknife** residuals):

$$r_{ti} = \frac{r_{-i}}{\sigma_{-i}\sqrt{1-h_i}} \qquad \textit{Eq 14.17}$$

where the '-*i*' notation indicates that observation *i* is not included in the prediction or the model's variance. These residuals are distributed as a *t*-distribution (with dfE-1; Table 14.2), assuming the model is correct.

To summarise, standardised residuals might yield a large value if:

- the observation is an outlier in the response (Y) variable (*ie* r_i is large), or
- the observation is an outlier in the predictor variable(s) (*ie* h_i is large),

and studentised residuals might be large if either of the above are true, or if

- the observation strongly affects the fit of the model (*ie* the model changes considerably when the observation is removed).

We now proceed to use data on the residuals to assess the overall fit of the model. Although we separate the study of homoscedasticity from normality, in practice one should look at both, as well as linearity before deciding on modifications (*eg* transformations) to the variables.

14.9 Evaluating the major assumptions

In general, evaluating the model assumptions relies heavily on graphical methods, although a large battery of statistical tests exists for evaluating different assumptions. However, we recommend the tests to be used only as a supplement to the graphical methods, and that caution should be exercised when tests and graphics lead to different conclusions.

14.9.1 Homoscedasticity

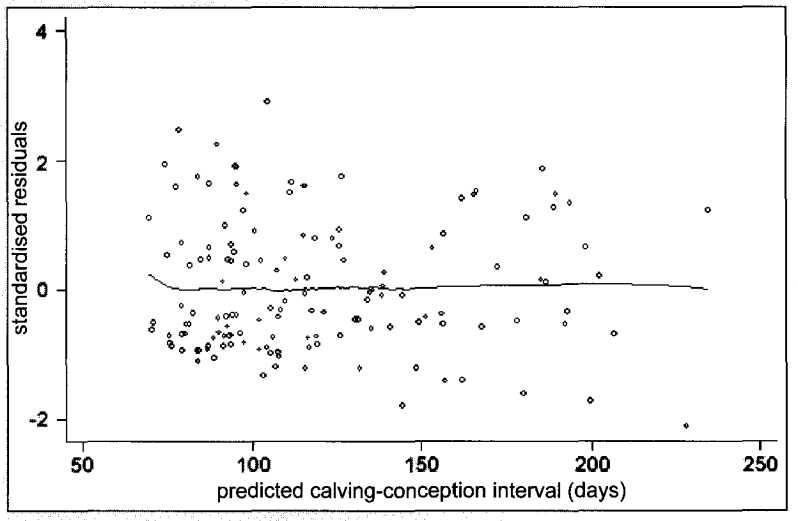
A constant variance is an important assumption in linear regression. Without equality of variance (a situation called heteroscedasticity), the significance tests are at best only approximate because the standard error is too small for some values and too large for others. One can examine the homoscedasticity assumption, by plotting the standardised residuals against the predicted values. If the variance is constant across the range of predicted Y-values, then a scatter of points resembling a horizontal band will result. If the variance is not constant, a pattern such as fanning (increased variance with larger $\hat{Y}$), or coning (decreased variance with larger $\hat{Y}$) might result. These patterns suggest that the dependent variable might need to be transformed (or a weighted regression used). It might also be useful to plot standardised residuals against individual (continuous) predictors and look for similar patterns, and to compare the residual variances in the groups formed by levels of categorical variables (Example 14.13). The plot of standardised residuals against the predicted -calvcon- in our model is shown in Fig. 14.4.

14.9.2 Normality of residuals

To examine for normality one can plot the residuals in the form of a histogram (Fig. 14.5, Example 14.14). An alternative, and more sensitive display, is a normal probability plot (sometimes called Q-Q (quantile-quantile) plot) for the residuals. If

Example 14.13 Examining residuals for equality of variances
data=daisy

Fig. 14.4 Plot of standardised residuals vs predicted -calvcon- values



The approximately equal-width band of points suggests that our model likely meets the assumption of equal variances. A formal test of equal variances (homoscedasticity) is the Cook-Weisberg test for heteroscedasticity (*ie* the null hypothesis is H_0: homoscedasticity). In the example we get a value of 0.22 for a χ^2-statistic with 1 df. This non-significant result (P>0.5) is consistent with a constant variance.

they are normally distributed, the resulting plot will be (approximately) a straight line at 45° to the horizontal (see Fig. 14.5).

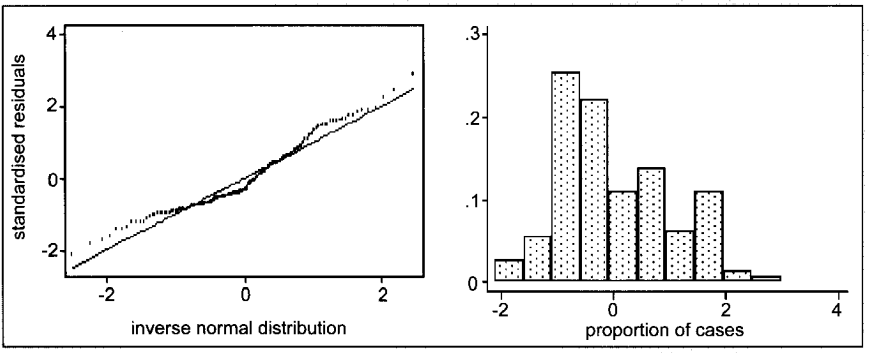
If the residuals are skewed to the right, the normal plot will curve below the 45° line (the curve is convex), whereas if the residuals are left skewed the normal plot will curve above the 45° line (the curve is concave). If the residuals are too peaked (platykurtic), the normal plot will be sigmoid curved. Whether such departures from normality are most easily visualised in the normal plot or the histogram is largely a matter of taste. As an aid for the interpretation, the skewness and kurtosis of the standardised residuals can also be computed.

14.9.3 Linearity of predictor-outcome association

In a regression model we assume that the relationship between the continuous predictor and the outcome is linear. With multiple continuous variables in the model, one approach to detecting non-linearity is to plot the residuals against each of the continuous predictor variables (see Example 14.15). The sensitivity of this process can be increased by using a kernel smoothing function to help you visualise any pattern

Example 14.14 Examining the assumption of normality
data=daisy

Fig. 14.5 Q-Q plot and histogram of residuals



The Q-Q plot displays the quantiles of the residuals versus the quantiles of the normal probability distribution. Here we have a dish-shaped (convex) plot suggesting a right skewed distribution of residuals. This is verified in the graph displaying a histogram of residuals.

Further evidence of a lack of normality can be obtained from a test for a normal distribution. One standard test is the Shapiro-Wilk's statistic, which for this example gives a value of W=0.954 (small values are critical for H_0: normal distribution) and P<0.001. **Note** Due to the dependence among the residuals, P-values from tests of normality are strictly speaking not valid (because they refer to testing an independent sample); however, the tests can be used as a rough guideline.

The residuals are clearly not normally distributed in this example, so we need to consider improving this aspect of our model, as is discussed in subsequent examples and summarised in section 14.11.

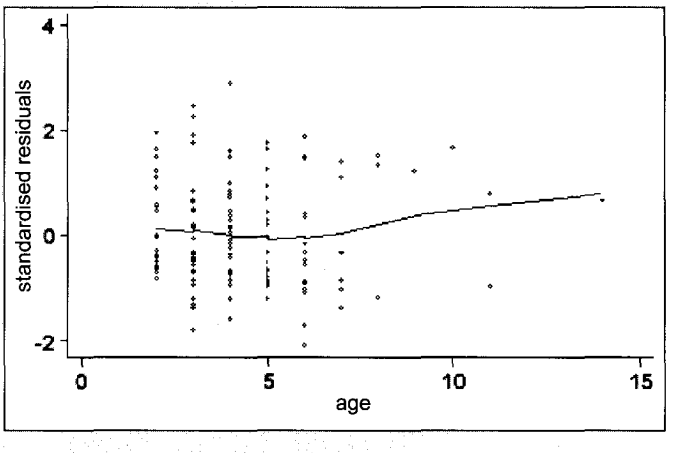
that might be present, but be careful of patterns in areas where the data are sparse. Methods for assessing linearity and dealing with non-linearity are discussed much more fully in section 16.10. However, three possible approaches to solving the problem will be mentioned here. The first is to add a power term (*eg* quadratic and higher order polynomials, spline polynomials, fractional polynomials) to the X-variables. This has the drawback of involving several parameters for the effect of the original X-variable. The second approach is to try to transform the Y-variable (as discussed below). Many software programs have helpful routines to guide the selection of an appropriate transformation. The third is to categorise the continuous predictor and include either regular or hierarchical indicator variables in the model in place of the continuous predictor variable.

Suggestions for correcting a lack of linearity by transformation

In order to correct a lack of linearity, we can transform the outcome or the predictor(s) or both. As will become apparent, we often have to use transformations to correct for

Example 14.15 Examining the linearity of an association
data=daisy

Fig. 14.6 Scatter plot of standardised residuals vs -age- with lowess smoother



The relationship, while not totally linear, is sufficiently close for our purposes. This is confirmed by using -age- and -age_sq- terms in the model (see Example 14.7); the coefficient for the latter variable was not close to significant.

heteroscedasticity and lack of normality also. Sometimes correcting for one problem solves others, but sometimes correcting one problem makes a new problem on the other fronts. If we transform the outcome variable to improve linearity, this will definitely affect the variance and normality of residuals so these must be rechecked after transforming the outcome variable. Indeed we might have to rebuild the model. If we transform the offending predictor(s), then the variance and normality of residuals are likely to remain relatively stable. Thus, often the route of choice for improving linearity is to test quadratic, or other power transformations of the predictor(s) within a power of ±2 to assess their significance. The following are guidelines:

- if the outcome increases, at a decreasing rate with X, then try a $\ln X$ or a $X^{1/2}$ transformation
- if the outcome increases, at an increasing rate with X, then try X^2 or e^X
- if the outcome decreases, at a decreasing rate with X, then try X^{-1} or e^{-X}.

If the relationship is more complex, it is often helpful (as noted above) to use hierarchical indicators instead of the continuous scaled variable. We can choose the important cutpoints for the indicators by selecting them in a forward manner (section 14.4.3). We have previously shown the use of hierarchical indicators for age as a predictor of -calvcon- (Example 14.6).

14.9.4 Specification bias

If the model is correct, the residuals are uncorrelated with the predicted outcome ($\hat{Y}$). However, if an important variable is missing from the equation, the model suffers from specification bias. This might reflect itself in a linear pattern of the standardised residuals with the predicted values of Y. For example, small (negative) residuals might be associated with lower values of $\hat{Y}$ and large (positive) residuals with large values of $\hat{Y}$ suggesting that one or more important predictor variables are missing. Specifically, the sampling units with positive residuals have something in common that also gives them large observed values of Y, and this feature might help identify the missing variable. Unfortunately, with a 'weak' (low R^2) model, it is difficult to discern some of these patterns because of the relatively large variability in r_i. There are formal tests for specification bias, but they are beyond the scope of this text.

14.9.5 Correcting error distribution problems: transformations of the outcome

There are a number of possible transformations but only the more frequently used ones are mentioned here. Most programs provide a variety of easily accessed transformations so that we can readily try different approaches. The selection of the correct transformation also is aided by knowledge of what has worked in similar situations in the past, although formal assessment of the appropriate transformation remains useful. Some general rules:

- if the variance of the residuals increases (mildly) with the outcome, and the underlying relationship is not necessarily linear, then a square-root transform of Y is often useful (it is the variance-stabilising transform for variances proportional to the mean),
- if the 'fanning' is strong and increases with the outcome, a logarithmic transformation of Y often works (the variance-stabilising transform for variances proportional to the mean squared),
- if the 'fanning' decreases with the outcome and the relationship of X and Y is nearly linear, a reciprocal transformation of Y could prove helpful,
- if Y is a proportion (p) (or more generally, an outcome in a bounded interval but without a binomial denominator) the variance-stabilising transformation for proportions is $\arcsin(\sqrt{p})$.

Often one of these suggested solutions will solve the problem. However, sometimes a more formal approach to deciding the optimal transformation is needed. In this regard, if we are concerned about a lack of normality, there is a family of transformations called **Box-Cox transformations**. The intent here is to determine the power transformation Y^{λ} (except for λ=0, see below) which will make the distribution of the errors as close to an independent Gaussian sample as possible. The Box-Cox analysis, available in most software, computes the value of λ which best 'normalises' the errors using an iterative maximum likelihood procedure. The Box-Cox transforms can only be used on positive numbers (*ie* >0), but they can be applied to the outcome variable, the predictor(s) or both. Some examples of Box-Cox transformations (where Y^* is the transformed value of Y) are:

- if λ=1, we use $Y^*=Y$

- if λ=1/2, we use $Y^*=\sqrt{Y}$
- if λ=0, we use $Y^*=\ln Y$,
- if λ=-1, we use $Y^*=-1/Y$.

Usually it is sufficient to round the estimated λ to the nearest 1/4 unit (*ie* λ=0.45 would be λ=1/2), or to pick a 'nice' value within the 95% confidence interval for λ.

Note that the analysis should be based on the residuals (from an appropriate linear model), not on the distribution of the outcome itself. It should also be noted that Box-Cox transforms is only one (and commonly used) type of transformation; there is no guarantee that the optimal λ works well (only that it is the best among the power transforms), and many other transformations might be relevant. For example, if the distributional problem with the residuals is mainly one of skewness, an alternative transform is of the form $Y^*=\ln(Y-c)$, where c is a value to be selected to help correct the skewness. An advantage of this transform is that it is not constrained to transforming only positive numbers. The application of these methods to our example data is shown in Example 14.16.

Example 14.16 Box-Cox and skewness transformations
data=daisy

In Example 14.14, the residuals were clearly non-normal. No transformations were found that correct this without these same transformations leading to heteroscedasticity of residuals. The suggested Box-Cox power transform was $(\text{calvcon})^{0.460}$ and the suggested skewness correction transform was ln(calvcon-13.5). **Note** The Box-Cox transformation was computed after adjustment for other predictors in the model (and hence adjusts the residuals) while the skewness transformation only works on the original outcome variable -calvcon-. However, the application of these transformations improved but did not solve the lack of normality and they both led to heteroscedasticity in the residuals.

A major reason for the non-normality in this data set is 'structural' in that -calvcon- has a constrained lower limit and although the observed -calvcon- can be less than the predicted value, given that -calvcon- cannot be less than -firstbrd-, it is more likely that there will be more large positive residuals than large negative residuals. Also, because applying the transformations did not change any inferences about the significance of the variables in the model (results not shown), in order to enhance interpretation we chose to leave the variables in their original scales.

14.9.6 Interpreting transformed models

One problem with transformations is that they change the structure of the model and interpretation can become more difficult. Among transformations of the outcome, only the log transformation allows for back-transformation of regression coefficients (to give multiplicative effects on original scale). In general, rather than trying to explain

the model in a mathematical sense, we suggest that you make extensive use of graphical techniques, compute the predicted values and plot the back-transformed outcomes. The key is to obtain the predicted outcome (and any confidence limits) in the transformed scale and then use the back-transform to determine the outcome in the original scale – on the assumption that explanations of effect are much easier in the original scale. Sometimes it is advantageous to leave the model in its transformed format. For example, it has now become standard practice to use log transformed somatic cell counts in models of factors that affect cell counts.

When applying transformations to multivariable models we need to be careful when making predictions because additive and linear models in one scale become (possibly strongly) non-linear and non-additive (*ie* showing interaction) in another scale. Thus the outcome depends on the values of all of the variables in the model even though there is no actual interaction. A recommended practice here is to use the mean values for variables not of direct interest and a range of values for those variables of primary interest. Again, all confidence limits *etc* are determined in the transformed scale and then back-transformed into the original scale as necessary.

14.9.7 Correcting distribution problems using robust standard errors

A number of distributional problems can be dealt with using robust standard errors. These are discussed in more detail in section 23.2.3 as they might also play a role in dealing with clustered data. Robust SEs are generally larger than regular SEs and hence, the CIs for the coefficients are wider (see Example 14.17). If robust errors are used, be careful not to use the F-test to assess the model as it is no longer valid. Also, the MSE no longer estimates σ^2 as there is no single parametric value.

14.10 Assessment of each observation

Our previous efforts were directed toward evaluating the major assumptions on which linear regression models are based. Here we assess the fit of the model on an observation by observation basis. Specifically we look for:

- cases that are not well fit by the model and hence have large residuals; some of these might be deemed **outliers**.
- cases with unusual X-values; these are called **leverage** observations.
- cases that have an unduly large impact on the model; these are called **influential** observations.

Our rationale for pursuing this observation-by-observation analysis is that we want to be sure the model is correct for the great majority of the study subjects, and if we can identify specific instances of observations that do not fit, or have a big influence on, our model, it can help us identify the reason(s) for that impact.

There are two general approaches to assist in this task, one is to use graphical techniques to detect observations with an unusual value (*ie* atypical relative to the others) on the test statistic, and the other is based on identifying observations that exceed a specific cutpoint. Both have their advantages, the key is to try a variety of approaches and see

Example 14.17 Robust standard errors
data=daisy

In the model shown below we requested that robust standard errors be calculated.

	Coef	'Usual' SE	Robust SE	t	P	95% CI	
farm1	-6.108	5.551	5.663	1.08	0.283	-5.089	17.305
age	3.615	1.353	1.279	2.83	0.005	1.086	6.144
metritis	14.368	6.416	6.956	2.07	0.041	0.616	28.120
ovar	34.007	7.835	8.023	4.24	0.000	18.144	49.870
firstbrd	0.777	0.078	0.064	12.08	0.000	0.650	0.904
constant	23.585	8.199	8.317	2.84	0.005	7.140	40.029

The coefficients are identical to those in the original regression (Example 14.12). The SE for metritis, as an example, has increased by about 8%, but its contribution is still statistically significant. Our preference is to use transformations whenever possible, rather than relying on robust standard errors to 'solve' the distributional problem(s).

which you prefer, but there is no need to use all possible approaches in a given dataset. Although we use graphical techniques regularly, here we present only tabular results.

14.10.1 Outliers

In general, an outlier is an observation in a dataset which is far removed in value from the others in the dataset. In multivariable datasets, we need to make precise the meaning of 'far removed in value', because it may be only in the combination of several variables that an observation becomes outlying. In regression analysis, we distinguish between outliers in the outcome variable and outliers among the predictor variables (not involving the outcome).

An outlier in the outcome is detected by a (numerically) large residual, where 'large' is viewed relative to both the other observations and to what would be expected for a dataset of the same size.

It is important to note that, although we are interested in identifying outliers, we do so largely to try and explain/understand why they fit poorly, not to remove them without reason. Outliers inflate the standard error of the estimate and hence, reduce the power of statistical tests. They might arise because of transcription or data entry errors, or they might signal that we are missing important covariates that could 'explain' the poor fitting points. In most instances, one should not be unduly concerned about outliers unless their standardised value is greater than 3, although values between 2 and 3 might be having an impact on the model. Recall that in a normal distribution, a small percentage of observations would be expected to lie outside of 3 SDs.

If an observation is suspected to be an outlier it can be assessed with a *t*-test based on the studentised residual. However, the probability associated with this test depends on whether the observation was suspected of being an outlier *a priori* or not. If an observation was suspected before hand, then the P-value is found by relating the studentised residual to a *t*-distribution with dfE-1 degrees of freedom. If we are testing subsequent to observing the residuals, we should multiply the above probability by the number of observations (n) (equivalent to using the Bonferroni adjustment).

Some general rules in managing outlier observations include:

- identify observations with large standardised residuals;
- try and find an obvious explanation for them, such as a recording error or erroneous test result (*ie* equipment or operator problem);
- if there is no recording error then think about what factors the outliers might have in common that, if measured, could explain their lack of fit;
- try refitting the model without the outliers to see the effect on the model; and
- if the observations are to be deleted (which they rarely are), be sure to explicitly record this for yourself and those who read your research report. (It is hard to justify the deletion of observations.)

Although deleting outliers will improve the fit of the model to the sample data, it might actually decrease the model's validity as a predictor for future observations. In Example 14.18, we have presented the five largest positive and negative residuals from our model along with the values of the predictor variables; this presentation often helps you understand the reason for the departures from expectation.

14.10.2 Detecting 'unusual' observations: leverage

This activity focuses on identifying subjects with unusual values in the Xs and is particularly applicable when many continuous variables are present in the model. For this purpose, we use the leverage from Eqs 14.14 and 14.15 which indicates the potential for the i^{th} observation to have a major impact on the model.

In general, observations with a value of (at least) one of the predictors that is far from the mean will tend to have a large leverage; note that we always have $1/n \leq h_i \leq 1$. Observations with a very high leverage may have a large influence on the regression model; whether they do or not depends on the observed Y-values. A common rule is to examine observations that have leverage values $>2(k+1)/n$, where k is the number of predictors in the model (or the number of regression parameters, excluding the intercept). There is a fair bit of arbitrariness in this cutpoint (another commonly used value is $3(k+1)/n$), and hence one should initially look for observations with relatively extreme leverage values regardless of the cutpoints. Any observation with a leverage above 0.9 can be considered as extreme in its predictor values.

Note that observations with a large leverage often have a very small residual (as is apparent from Eq 14.14), and thus, they do not show up when searching for large residuals. Having identified potentially influential observations, we proceed to identify their actual influence on the model.

Example 14.18 Individual observation assessment using residuals

The results below are based on the previously discussed multiple regression model for -calvcon- with the five predictors -farm1-, -age-, -metritis-, -ovar-, and -firstbrd-.

Potential outliers

The largest five negative residual (observed < predicted) observations:

cownum	calvcon	age	firstbrd	raw residual	standardised residual	studentised residual
114	165	6	165	-62.950	-2.115	-2.142
68	90	3	90	-54.468	-1.805	-1.820
48	147	6	147	-52.598	-1.738	-1.751
40	131	4	131	-48.937	-1.615	-1.625
158	114	3	114	-43.006	-1.423	-1.428

The largest five positive residual (observed > predicted) observations:

cownum	calvcon	age	firstbrd	raw residual	standardised residual	studentised residual
122	154	3	78	58.970	1.900	1.919
96	134	2	56	59.678	1.930	1.949
75	159	3	71	69.409	2.237	2.270
133	154	3	49	75.393	2.455	2.501
1	193	4	67	88.533	2.891	2.971

None of these residuals are very extreme but there are two or three cows with relatively large positive residuals. If we were to note the observation with the largest studentised residual (cow number 1), the P-value associated with a value of 2.971 from a $t(138)$-distribution is 0.0035; when this is multiplied by the number of observations (n=145), it is clearly non-significant (P=0.51).

Leverage observations

Here we list the five observations with the largest leverage values:

cownum	calvcon	age	metritis	ovar	firstbrd	leverage	standardised residual
84	76	11	0	0	54	0.114	-0.995
15	270	9	0	1	186	0.118	1.204
117	140	3	1	1	49	0.120	0.444
92	146	11	1	0	59	0.131	0.770
58	216	14	0	1	108	0.219	0.643

The two most commonly used cutpoints for leverages $(2(k+1)/n$ and $3(k+1)/n)$ are 2*6/145=0.083 and 3*6/145=0.124.

(*continued on next page*)

Example 14.18 *(continued)*

With the exception of one old (14 years) cow, none of the cows has really extreme values of the predictor variables, although all exceed the lower cutpoint. Three (perhaps four) of the cows were rather old, while one (#117) was young, had both diseases and was bred very early at day 49 postpartum. With this warning, we might accept that removing the few older cows (*eg* >8) in our dataset might be a more effective way to model factors affecting -calvcon- for the majority of cows. However, we examine the actual impact of these observations below.

Influential observations

Here we display the data for subjects with the five largest negative DFITS:

cownum	calvcon	age	metritis	ovar	firstbrd	Cook's D	DFITS
114	165	6	1	1	165	.0778	-.692
68	90	3	0	1	90	.0403	-.496
48	147	6	0	1	147	.0346	-.460
40	131	4	0	1	131	.0286	-.417
158	114	3	0	1	114	.0239	-.380

and the data for subjects having the five largest positive DFITS:

cownum	calvcon	age	metritis	ovar	firstbrd	Cook's D	DFITS
15	270	9	0	1	186	.0324	.441
135	242	6	0	1	129	.0376	.479
133	154	3	0	0	49	.0379	.486
174	161	10	0	0	67	.0404	.496
1	193	4	1	0	67	.0601	.617

Note that only one (#15) of the potential leverage subjects was unduly influential so removing the older cows' data would not change the model substantially. However, all of these observations had DFITS values that exceed the critical value of ±0.41 ($2*\sqrt{6/145}$). The model changes when they are omitted, but not drastically, and because we have no explicit rationale for removing them, they should stay in the model.

14.10.3 Detecting influential observations: Cook's distance and DFITS

An intuitive test of an observation's overall influence is to omit it from the equation, recalculate the model and note the amount of change in the predicted outcome. If an observation is influential the change will be large; if not, the change will be small (see Example 14.18). This approach forms the basis of **Cook's Distance** D_i which is the sum of squared differences in fitted values with and without observation i (summed over all other observations and scaled suitably). A more direct interpretation of Cook's distance derives from the formula:

$$D_i = \frac{r_{si}^2}{(k+1)} \frac{h_i}{(1-h_i)} \qquad \textit{Eq 14.18}$$

emphasising that a large standardised residual, a large leverage, or both can lead to undue influence.

A commonly suggested cutpoint is to compare the Cook's value with the $F(k+1, n-k-1)$ distribution. If it exceeds the 50% percentile (not 5%), which is essentially 1, then the observation should be investigated. However, in our practical experience, the values of D_i rarely exceed this cutpoint, so it is recommended to look instead for values that are extreme relative to the others in the data.

A similar approach is used with a statistic known as **DFITS** (or DFFITS). It is an acronym that stands for 'difference in fit' between when the observation is in the model versus when it is out. DFITS indicates the number of standard errors change to the model when that observation is deleted. The following formula for DFITS shows its strong similarity to Cook's distance:

$$\text{DFITS}_i = r_{ti}\sqrt{\frac{h_i}{(1-h_i)}} \qquad \textit{Eq 14.19}$$

Thus, the DFITS statistic is based on the studentised residual and retains its sign. Again, if the DFITS numerically exceeds a value of, for example, 1 for $n<120$ or $2\sqrt{(k+1)/n}$ in a larger dataset, it means that if that observation was deleted, the model would change by a relatively large amount (recall that k is the number of predictor variables in the model). As with outliers, be hesitant to remove influential observations without good reason. In general, we do not remove influential observations unless the data are known to be incorrect. If observations are removed, this, and the reason(s) for their removal, must be drawn to the attention of those reading your research results.

14.10.4 Detecting influential values of specific predictors

Given an exposure variable of interest, one can assess the impact of deleting a specific observation on the value of the regression coefficient for that variable. The statistic used for this is known as a delta-beta (DB) and reflects the number of standard errors by which the specific regression coefficient changes when that observation is deleted. Thus it helps identify if a particular variable has a large influence on the β for that variable. Critical values for $n<120$ are 1 and for larger datasets $2/\sqrt{n}$. Again, this value might be too sensitive and initially one should just focus on observations with very extreme DB values.

14.11 Comments on the model deficiencies

In our examples we have taken you through the basic steps of assessing a linear regression model. We did have a few problem cows (subjects) that were minimally influential, or poor fitting, but we had a nasty problem relating to the lack of normality with more positive residuals than negative residuals. The reality is that if we correct the normality assumption, we create unequal variances. Having tried a number of transformations and seeing the model having similar coefficients we will content ourselves with these efforts. Because the outcome in these examples is time to an event, we might consider reanalysing these data using survival methods discussed in Chapter 19.

SELECTED REFERENCES/SUGGESTED READING

1. Belsley DA, Kuh E, Welsch RE. Regression diagnostics. New York: Wiley, 1980.
2. Dorfman A, Kimball AW, Friedman LA. Regression modeling of consumption or exposure variables classified by type. Am J Epidemiol 1985;122: 1096-1107.
3. Draper NR, Smith H. Applied Regression Analysis 3d ed. Toronto: John Wiley and Sons, 1998.
4. Kutner MH, Nachtschiem CJ, Wasserman W, Neter J. Applied linear statistical models. 4th ed. Boston: McGraw-Hill/Irwin, 1996.
5. Robins JM, Greenland S. The role of model selection in causal inferences from non-experimental data. Am J Epidemiol 1986; 123: 392-402.
6. Walter SD, Feinstein AR, Wells CK. Coding ordinal independent variables in multiple regression analyses. Am J Epidemiol 1987; 125: 319-323.
7. Weisberg S. Applied linear regression 2d ed. New York: Wiley, 1985.

SAMPLE PROBLEMS

Use the dataset pig to evaluate the effects of various diseases (and other factors) on the average daily gain of pigs. These are a subset of data on the growth performance and abattoir findings of pigs; the dataset is described further in Chapter 27. The variables we will use are explained below.

Field	Description	Codes/units
farm	Farm identification code	1-15
sex	Sex of the pig	0 = female
		1 = castrated male
worms	Count of nematodes in small intestine at time of slaughter	Continuous
lu	Lung score for enzootic pneumonia	0 = negative (no lesions)
		1 = mild (<10% affected)
		2 = moderate (10-20% affected)
		3 = severe (>20% affected)
ar	Atrophic rhinitis score (0-5 in half-point increments	0 = no lesions
		5 = complete erosion of turbinates
adg	Average daily gain in kg (based on an assumed birth weight of 0 kg and a live weight at slaughter estimated form the carcass weight	Continuous
pn	Enzootic pneumonia (present or absent) – not in the dataset but you will create this variable	

Exercise 1

1. Create -pn- (a variable indicating the presence/absence of lung lesions).
2. Compute descriptive statistics for all continuous variables and frequency distributions for all categorical variables. Do all of the values look reasonable?
3. What are the simple pairwise correlations between -worms-, -pn-, -ar- and -adg-? Create some scatterplots to examine the data. What do the plots suggest about the nature of the relationships?
4. Carry out simple linear regressions for each of the factors -worms- and -ar-. Interpret the results.
5. What is the 95% CI for the mean value of -adg- for a pig with an average worm burden? What is the forecast interval (95% CI for -adg- for an individual pig) for a pig with an average worm burden?
6. What is the 95% CI for the mean value of -adg- for a pig with no worms? What is the forecast interval (95% confidence interval for -adg- for an individual pig) for a pig with no worms?

Exercise 2

Use the dataset that you created in the previous exercise to evaluate the effects of various diseases (and other factors) on the average daily gain of pigs.

1. Fit a regression model for -adg- using -worms-, -ar-, -sex- and -pn- as predictor variables. Interpret the results.
2. Include a series of indicator (dummy) variables representing 'farms' and refit the regression.
3. Perform a multiple partial F-test to determine if the dummy variables representing farms are a significant 'group' of predictors when added to the model containing -worms-, -ar-, -sex- and -pn-.
4. Convert -ar- to a four-level categorical variable called (ar_c4) coded as follows

ar_c4	Range of values in ar
0	ar = 0
1	0<ar≤2.0
2	2.0<ar≤4.0
3	4.0<ar

5. Create a set of indicator variables for farms called -f01- to -f14-.
6. Assess the assumption that the relationship between -ar- and -adg- is linear (hint: fit a model with ar_c4 along with -worms-, -pn- and f01..f14).
7. Compute a partial F-test to assess the significance of -pn- in the above model (this effectively gives the same answer as the t-test of the coefficient).
8. Create a new variable representing severe -ar- (ar≥4.5). Call it -ar_sev-. Investigate possible interactions between the disease variables (-worms-, -pn- and -ar_sev-) and -sex- to see if the disease effects depend on the sex of the pigs.
9. Is -sex- a confounding variable for the relationships between the disease variables and -adg-?
10. Is -farm- a confounding variable for the relationships between the disease variables and -adg-?
11. Build an appropriate model for -adg-. Keep in mind that the primary objective is to determine how diseases affect -adg-.

Exercise 3

1. Evaluate the model built in the previous exercise by looking at various regression diagnostics. Before you start this, we would suggest that you sort the data by 'farm, sex and adg' and then generate a unique pig identification number
2. Compute each of the following and become familiar with them by scanning the computed values:

   predicted values
   raw residuals
   standardised residuals
   studentised residuals
   leverage values
   Cook's distance values

DFITS
delta-betas for the variable -pn-.

3. Use the above to evaluate each of the following assumptions:
   homoscedasticity
   normality of residuals
4. Did you identify any problems in question 3. See what happens if you make a natural log transform of -adg-. Also evaluate Box-Cox transformations to see if there is a more appropriate transformation scale.
5. Refit the model with -adg- (not log transformed). Identify if there are any of the following, and if so, determine why they have arisen and evaluate what effect they are having on the model:
   outliers
   high leverage points
   influential observations.

Exercise 4

1. Using the dataset daisy, build a model using only the 'disease variables' to see their relationship with the days to conception (-calvcon-). Interpret the results.
2. What model is appropriate if your intention is to identify the causal effect of -pyomet- on -calvcon-?
3. Now add -firstest- to the model. Interpret the results.
4. Repeat Exercise 3 for assessing the model.
5. Log transform -calvcon- to create -calvcon_ln- and interpret the effect of -pyomet- on this model compared with the effect obtained from the non-transformed model.

15

MODEL-BUILDING STRATEGIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Develop a 'full' (maximal) model which incorporates your biological understanding of the system being investigated.

2. Carry out procedures to reduce a large number of predictors to a more manageable subset.

3. Build regression-type models while considering statistical and non-statistical criteria.

4. Evaluate the reliability of a regression-type model.

5. Present the results from your analysis in a meaningful way.

15.1 INTRODUCTION

When building a regression model, we need to balance the desire to get the model which 'best fits' the data with the desire for parsimony (simplicity in the model). As will become apparent, the definition of 'best fit' depends on the goal of the analysis.

One goal might be to come up with the best model for predicting future observations. In this case, the details of the model (*eg* the effects of specific predictors) might be of little consequence but we want to keep any variables whose relationship with the dependent variable is questionable out of the model. If the latter are allowed in and a future observation has a relatively extreme value for one of those variables, the prediction might be inaccurate.

A second goal could be to obtain the most precise estimates possible of coefficients for selected variables of interest. This is often our goal when trying to elucidate causal associations. In this strategy, careful attention must be paid to possible interaction and confounding effects (see section 15.2.2).

The steps involved in building a regression model are:

1. specify the maximum model to be considered (*ie* identify the outcome and the full set of predictors that you want to consider)
2. specify the criterion (criteria) to be used in selecting the variables to be included in the model
3. specify the strategy for applying the criterion (criteria)
4. conduct the analyses
5. evaluate the reliability of the model chosen
6. present the results.

Unless otherwise specified, the discussions that follow relate to all types of regression model, not just linear regression models.

15.2 SPECIFYING THE MAXIMUM MODEL

The first step in specifying the maximum model is to identify the outcome variable and determine whether it is likely to need transformation (*eg* natural log transformation) or other form of manipulation (*eg* recategorisation of a categorical outcome). Discussion of issues related to the outcome variable is presented in chapters dealing with specific modelling techniques (*eg* Chapter 14 for linear regression models).

The maximum model is the model with all possible predictors of interest included. There are pros and cons to making the maximum model very large. On one hand, it will prevent you from overlooking some potentially important predictors. However, on the other, adding a lot of predictors increases the chances of:

a. collinearity among predictor variables (if two or more independent variables are highly correlated, the estimates of their coefficients in a regression model will be unstable), and

b. including variables that are not important 'in the real world' but happen to be significant in your dataset. (Interpretation of these results might be difficult and the risk of identifying spurious associations is high.)

Bear in mind that building the maximum model is as much a scientific/clinical task as it is a statistical one. In general, the desire for parsimony should be your guiding light but do not exclude variables that you have good reason to believe (*ie* for biological reasons) should be in the model. Remember, the goal of most statistical analyses is to extract meaningful results from a complex dataset. If the final results are almost as complex as the original data, nothing has been gained. (Statistically speaking this would happen if the number of regression coefficients equalled the number of observations in the dataset).

When specifying the maximum model, you need to identify which variables should be included in the model-building process, how many should be included and whether or not interaction terms need to be considered.

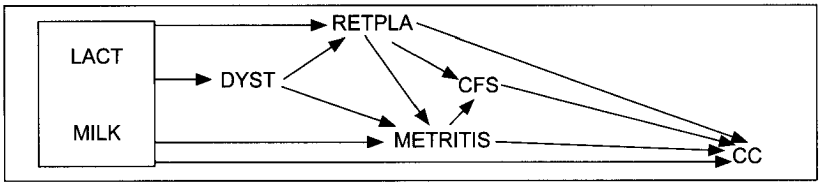
15.2.1 Building a causal model

It is imperative that you have a causal model in place before you begin the model-building process. This will identify potential causal relationships among the predictors and the outcome of interest. For example, if you were interested in evaluating the effects of retained placenta (RETPLA) on reproductive performance (as measured by the calving-to-conception interval) in multiparous dairy cows and had recorded data on:

- the lactation number (surrogate measure for cow's age) (LACT)
- previous lactation milk production (kg) (MILK)
- dystocia (DYST)
- retained placenta (RETPLA)
- metritis (METRITIS)
- days from calving to first service (CFS)
- days from calving to conception (CC),

then a putative causal diagram might look like Fig. 15.1.

Fig. 15.1 Putative causal diagram for effects of RETPLA on reproductive performance



If the objective of the study was to quantify the effects of RETPLA on the calving to conception interval, you would NOT include any intervening variables (metritis, days to first service) in the regression model. Inclusion of these variables would remove any

of the effect from RETPLA that was mediated through the intervening variables. On the other hand, if lactation number is suspected to be an important confounder, it might be designated to remain in the model regardless of whether or not it is statistically significant. (See Chapter 13 for a more detailed discussion of confounding.)

Even if a study has a very large number of predictors, it is essential to start with a causal structure in mind and this can often be drawn by grouping variables into logical clusters (*eg* all farm management variables together, all measures of disease levels together).

15.2.2 Reducing the number of predictors

One is often faced with building regression models using datasets with a large number of predictor variables. One rule of thumb suggests that there must be at least 10 observations for each predictor considered for inclusion in the model. There are a variety of ways of reducing the number of variables that need to be considered for inclusion in a regression model. These include:

- screening variables based on descriptive statistics
- correlation analysis of independent variables
- creation of indices
- screening variables based on unconditional associations
- principle components analysis/factor analysis
- correspondence analysis.

These will each be reviewed briefly and more detail can be found in Dohoo et al (1997). However, before any reduction in the number of independent variables is undertaken, it is essential to identify the primary variables of interest and any other variables for which there is already evidence that they might be confounders or interacting variables. These should always be retained for consideration in the model.

Before proceeding with an overview of the approaches for reducing the number of variables, we must point out that, in many cases, the most appropriate procedure would be to design a study which was much more focused and which collected high-quality data on far fewer predictors. This will greatly reduce the risk of identifying associations for which making a causal inference is very precarious.

Screening variables based on descriptive statistics

Descriptive statistics (means, variances, percentiles *etc* for continuous variables and frequency tabulations for categorical variables) can be very helpful in identifying variables which might be of little value in your model. Keep in mind that, in general, you want to keep variables that you are confident have been measured accurately and precisely, and which are relatively complete. Some specific guidelines to consider are:

- Avoid variables with large numbers of missing observations.
- Select only variables with substantial variability (*eg* if almost all of the animals in a study are males, adding sex as a predictor is not likely to be helpful).
- If a categorical variable has many categories with small numbers of observations in each, consider combining categories (if this makes biological sense), or eliminating the variable.

Correlation analysis

Examination of all pairwise correlations among predictor variables will identify pairs of variables that contain essentially the same information. Inclusion of highly correlated variables will result in multicollinearity in the model, producing unstable estimates of coefficients and incorrect standard errors. Collinearity will certainly be a problem with correlation coefficients greater than 0.9, but could occur at lower levels. If pairs of highly correlated variables are found, one of them should be selected for inclusion in the model based on criteria such as: biological plausibility, fewer missing observations, ease and/or reliability of measurement. **Note** Examining correlations among variables in a pairwise manner will not necessarily prevent multicollinearity because the problem can also arise from correlations among linear combinations of predictors. However, screening based on pairwise correlations will remove one potential source of the problem.

Creation of indices

It might be possible to combine a number of predictor variables that are related into a single predictor that represents some overall level of a factor. This might be done subjectively based on the perceived importance of the contribution of a number of factors. For example, an index representing the level of hygiene in stalls for dairy cows might be created as a linear combination of scores for factors such as quantity of bedding present, wetness of the bedding, amount of manure present and amount of fecal soiling of the udder and flanks of the cows. The weights assigned to each factor might be subjectively assigned although, if possible, they should be based on evidence from previous research. Alternatively, data on a number of factors can be combined in an objective manner if procedures to do so exist. For example, data on fan capacity, size and shape of air inlets and barn size might be used to compute the number of air changes per hour in a swine barn. This might then be expressed as the proportion of a recommended ventilation level. One drawback to the creation of indices is that it precludes the evaluation of the effects of individual factors which were used to create the index (see discussion of suppressor variables in section 13.11.8).

Screening variables based on unconditional associations

One of the most commonly used approaches to reducing the number of predictor variables is to select only those that have unconditional associations with the outcome that are significant at some very liberal P-value (*eg* 0.15 or 0.2). The types of test used to evaluate these associations will depend on the form of the outcome and predictor variables. However, simple forms of a regression model (*eg* a linear or logistic regression model with a single predictor) will always be appropriate for this investigation.

One drawback to this approach is that an important predictor might be excluded if its effect is masked by another variable (*ie* the effect of a predictor only becomes evident once a confounder is controlled) (see distorter variables, section 13.11.7). Using a liberal P-value helps prevent this problem. Another approach is to add all eliminated predictors, one at a time, back into the final model. If the confounder was included in the final model, then the eliminated predictor might then turn out to have a statistically significant association and be added back into the model.

This process of screening predictors individually can be extended to include building multivariable models using mutually exclusive logical subsets of predictors to identify the key predictors in each subset, which are then retained for consideration in a final multivariable model. For example, Lofstedt et al (1999), when evaluating a wide range of possible predictors of septicemia in diarrheic calves, built separate models using demographic and physical examination data, clinical chemistry data and hematology data. The important predictors from each of these three models were then evaluated in an overall model.

Principle components analysis/factor analysis

Principle components analysis and factor analysis are two closely related techniques that can be used to consolidate the information contained in a set of predictor variables into a new set of uncorrelated (*ie* orthogonal) predictor variables. A detailed discussion of the techniques is beyond the scope of this book but they will be summarised briefly. Both are designed primarily to work with quantitative (continuous) predictors, but techniques are available to allow categorical predictors to be included.

Principle components analysis is used to convert a set of k predictor variables into a set of k principle components with each successive component containing a decreasing proportion of the total variation among the original predictor variables. Because most of the variation is often contained in the first few principle components, this subset is often selected for use as predictors in the regression model. The composition of the principle components does not vary depending on the number of components selected for retention. Once the regression model has been built with this subset of the principle components, the resulting coefficients can be back-transformed to obtain coefficients for the full set of original predictors. This resulting set of coefficients will be more stable than those from a model built directly from the original predictors because the problem of multicollinearity has been eliminated. However, there will be no evaluation of the statistical significance of each of the predictors and hence, no identification of which ones are most 'important'.

Factor analysis is a closely related technique, but is based on the assumption that a set of factors that have inherent meaning can be created from the original variables. For example, Hurnik et al (1994a,b) used factor analysis to create six factors that they claimed represented specific types of swine farms, in a study of risk factors for respiratory disease in swine. Unlike principle components, the composition of the factors does vary as the number of factors selected for creation varies. The strength of a factor analysis rests with the plausibility of the assumption that the factors are truly measuring an underlying latent structure (*eg* swine farm type). If this assumption is valid, then knowing which of those underlying structures (*eg* farm type) are associated with the outcome (*eg* respiratory disease) might be as important as information about individual predictor variables. Determining which of the original predictors are important determinants of the outcome is a subjective process based on determining which predictors are highly correlated (or have high 'factor loadings') with factors found to be significant predictors of the outcome. As with principle components analysis, there is no statistical testing of individual predictors.

Correspondence analysis

Correspondence analysis is a form of exploratory data analysis designed to analyse the relationships among a set of categorical variables. One of the main objectives of correspondence analysis is to produce a visual summary (usually two-dimensional) of the complex relationships that exist among a set of categorical variables (both predictors and the outcome). The two axes are factorial axes which reflect the most 'inertia' (variability) in the original predictor variables. The result is a scatterplot which identifies clusters of predictors that are closely associated, with clusters farther from the intersection of the axes having stronger associations. The values of the outcome variable (also categorical) can also be projected on the same axes to determine which clusters of predictor variable values are associated with the outcome(s) of interest. A correspondence analysis of a subset of the risk factors for swine respiratory disease is presented in Example 15.1.

While principle components analysis, factor analysis and correspondence analysis can be used to deal with the problem of large numbers of independent variables, they are perhaps better viewed as complementary techniques to model-building procedures. They provide insight into how predictor variables are related to each other and ultimately, into how groups of predictors are related to the outcome of interest.

15.2.3 Identifying interaction terms of interest

It is important to consider including interaction terms when specifying the maximum model. There are five general strategies for creating and evaluating two-way interactions.

1. Create and evaluate all possible two-way interaction terms. This will only be feasible if the total number of predictors is small (*eg* ≤ 8).
2. Create two-way interactions among all predictors that are significant in the final main effects model.
3. Create two-way interactions among all predictors found to have a significant unconditional association with the outcome.
4. Create two-way interactions only among pairs of variables which you suspect (based on evidence from the literature *etc*) might interact. This will probably focus on interactions involving the primary predictor(s) of interest and important confounders.
5. Create two-way interactions that involve the exposure variable (predictor) of interest.

Regardless of how the set of interaction terms is created, you could subject them to the same sort of screening processes described above to reduce the number included in the model-building process. If an interaction term is to be included in the model, then it is recommended that the main effects that make up that interaction term also be included. Evaluation of a large number of two-way interactions could identify spurious associations, due to the fact that a large number of associations are being evaluated. Two-way interactions between continuous predictors are difficult to interpret, and, whenever significant, should be evaluated by fitting a range of possible values for both predictors (see Example 14.11).

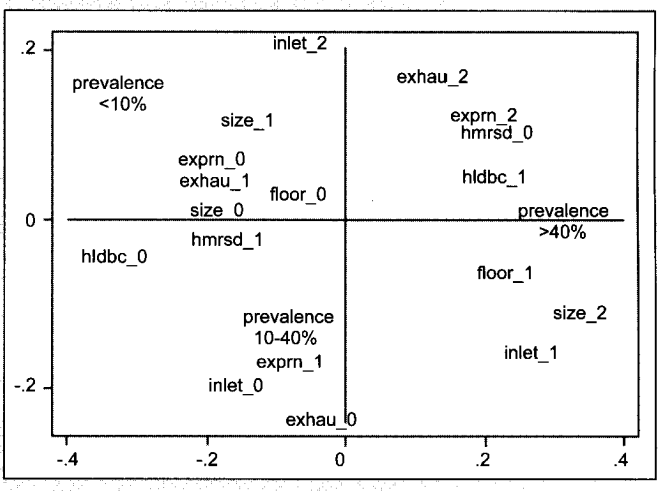
Example 15.1 Correspondence analysis of pig respiratory disease risk factors
data=pig_farm

In a study aimed at identifying risk factors for enzootic pneumonia in swine, a large number of herd characteristics were evaluated (details of the dataset are presented in Chapter 27). For the purpose of this example, only factors selected by automated model-building procedures (Example 15.2) have been included. They include:

pncode	pneumonia prevalence (the outcome variable) (labelled as prevalence in Fig. 15.2)
inlet	air inlet size (as proportion of recommendation) (recoded 0,1,2)
exhau	exhaust fan capacity (as proportion of recommendation) (recoded 0,1,2)
hldbc	holding back slow-growing pigs (coded 0,1)
size	herd size ('000s of pigs) (recoded 0,1,2)
exprn	farmer experience (recoded 0,1,2)
floor	type of flooring (coded 0,1)
hmrsd	only home-raised pigs (coded 0,1)

Correspondence analysis was used to visually evaluate the relationships among these variables with the results presented in Fig. 15.2. **Note** This correspondence analysis is provided for pedagogical purposes only; a more complete analysis is described in Dohoo et al (1997).

Fig. 15.2 Multiple correspondence analysis of risk factors for enzootic pneumonia in pigs



As can be seen, high levels of pneumonia were associated with holding back slow-growing pigs (hldbc_1), barns with slatted floors (floor_1), and large herd sizes (size_2). Low levels of pneumonia were associated with moderate-sized herds (size_1), high air inlet capacity (inlet_2) and young (low level of experience) farmers (exprn_0).

Three-way interactions might be considered, but they are usually difficult to interpret. They should only be included if you have good reason (*a priori*) to suspect the existence of such an effect or if they are made up of variables with significant main effects and two-way interactions. Three-way interactions might also unnecessarily complicate the model because all of the main effects and two-way interactions among the predictors making up the three-way interaction need to be included in the model.

15.3 Specify the selection criteria

Once a maximum model has been specified, you need to decide how you will determine which predictors need to be retained in the model. Criteria for retention can be based on non-statistical considerations or the statistical significance of the predictor. It is essential that both be considered and the non-statistical considerations will be discussed first.

15.3.1 Non-statistical considerations

Variables should be retained in the model if they meet any of the following criteria.

- They are a primary predictor of interest.
- They are known, *a priori*, to be potential confounders for the primary predictor of interest.
- They show evidence of being a confounder in this dataset because their removal results in a substantial change in the coefficient for one of the primary predictors of interest. **Note** Building an appropriate causal model before starting the model-building process will help ensure that the variable is not an intervening variable (see section 13.11.6).
- They are a component of an interaction term which is included in the model.

15.3.2 Statistical criteria – nested models

Nested models are models based on the same set of observations in which the predictors in one model are a subset of the predictors in the other model. By far the most common approach to evaluating the statistical significance of individual predictors is to use tests based on nested models. For a linear regression model this would involve carrying out a partial F-test for the predictor, while in other types of regression model (*eg* logistic, Poisson) a Wald test or likelihood-ratio test could be used. When evaluating the significance of a categorical variable (included in the model as a set of indicator variables), the overall significance of all the indicator variables in the model should be used, not the statistical significance of individual indicator variables.

15.3.3 Statistical considerations – non-nested models

A number of **information measures** have been developed for use in comparing models that are not nested. The most common are the **AIC** (Akaike's Information Criteria) and the **BIC** (Bayesian Information Criteria – also known as the Schwartz Bayesian Criteria). They are based on an overall assessment of the model and can be

used to compare different models, regardless of whether they are nested. Although the following formulae are presented in terms of likelihoods (see Chapter 16), they are equally applicable to linear regression models. However, some words of caution are in order. First, these tests should not be used to compare nested models – test- based comparisons (*eg* partial *F*-tests or likelihood-ratio tests) are superior. Second, these tests cannot be used to compare models which are based on different numbers of observations.

The AIC is computed as:

$$AIC = -2\ln L + 2(k + 1) \qquad \textit{Eq 15.1}$$

where L is the log likelihood and k is the number of predictors in the model. The smaller the value of the AIC, the better the model. If two models have comparable log likelihoods, the more parsimonious (*ie* fewer parameters) will have the smaller AIC.

The BIC is computed as:

$$BIC = -2\ln L - (N - k - 1)\ln(N) \qquad \textit{Eq 15.2}$$

where N is the number of observations in the dataset.

A 'better' model will be more negative than the 'poorer' model and Table 15.1 provides guidelines for assessing the evidence of superiority of one model over another (Raftery, 1996). In general, the BIC leads to more parsimonious models than the AIC does.

Table 15.1 Guidelines for interpreting BIC values from non-nested models

Absolute difference in BIC	Evidence for superiority of the better model
0 - <2	weak
2 - <6	positive
6 - <10	strong
≥ 10	very strong

Two additional approaches, applicable to linear regression models, are based on the adjusted R^2 or a statistic called Mallow's *Cp*. The model which maximises the adjusted R^2 (see section 14.3.5) is, in effect, maximising the amount of variance explained by the model, while precluding the incorporation of predictors which explain only a very small amount of the variance. This approach is equivalent to finding the model which minimises the mean square error (MSE). **Note** Adding more and more unimportant terms to the model will actually increase the MSE because the df on which it is based becomes smaller.

Mallow's *Cp* is computed as follows. If k predictors are selected from a complete set of p predictors, then Mallow's *Cp* for that model is:

$$Cp = \sum \frac{(Y - \hat{Y})^2}{\sigma^2} - n + 2k \qquad \textit{Eq 15.3}$$

where Y and $\hat{Y}$ are the observed and expected values of Y for a model based on the k predictors, σ^2 is the MSE from a model based on all (p) predictors and n is the sample size. Mallow's Cp is a special case of the AIC. Models with the lowest Cp are generally considered the best. In general, Mallow's Cp is not negative, but might be so in cases in which the dataset is small and there are a large number of predictors (*eg* Example 15.2).

15.4 Specifying the selection strategy

Once the criteria (both statistical and non-statistical) to be used in the selection process have been specified, there are a number of ways to actually carry out the selection.

15.4.1 All possible/best subset regressions

If the number of predictors in the maximum model is small, then it is possible to examine all possible combinations of predictors. Once all of the models have been fit, it is relatively easy to apply both the non-statistical and statistical criteria described above in order to select the 'best' model. This approach is best applied in a context that a researcher is searching for a number of good models, such as early in an investigation on a topic.

This process is modified slightly with best subset regression. In this procedure, the computer identifies the 'best' model (according to one of the criteria outlined above), with a given number of predictors. For example, it will identify the single-term model with the largest R^2, the two-term model with the largest R^2, the three-term model with the largest R^2 *etc*. The investigator can then identify the point at which increasing the number of predictors in the model is of little value in terms of improving the predictive ability of the model. Both nested and non-nested models can be compared using 'all possible' or 'best subset' selection procedures.

15.4.2 Forward selection/backward elimination/stepwise

When a **forward selection** process is used, the computer first fits a model with only the intercept and then selectively adds terms that meet a specified criterion. The usual criterion for inclusion is a partial F-test statistic (Wald test in logistic, Poisson *etc* regression) over a specified value (equivalent to a P-value below a specified value such as 0.05). The term with the largest partial F is added first and then the process is repeated. This continues until no term meets the entry criterion. If there is a very large number of potential predictors, forward selection might be the only feasible approach because it might be impossible to fit the maximum model and obtain reasonable estimates (due to problems with collinearity).

With **backward elimination**, the process is reversed. The maximum model is fit and then terms are removed sequentially until none of the terms remaining in the model has a partial F statistic under the specified criterion. An advantage of backward elimination is that the statistical significance of terms is assessed after adjustment for the potential

confounding effect of other variables in the model. With forward selection, this happens to a much more limited extent (only after confounders have been selected and incorporated into the model). Therefore, forward selection often identifies a smaller model than backward elimination, and might be considered the inferior procedure for datasets with a reasonably small number of predictors.

Stepwise regression is simply a combination of forward selection and backward elimination. If it starts with forward selection, after the addition of each variable, the criterion for backward elimination is applied to each variable in the model to see if it should remain. If it starts with backward elimination, after the removal of each variable, all removed variables are checked to see if any of them would meet the forward selection criterion for inclusion.

Different selection procedures will often result in different final models, as can be seen in Example 15.2.

15.4.3 Cautions in using any automated selection procedures

While the automated selection procedures described above are convenient, easy to apply and quickly reduce a large complex dataset to a succinct regression model, they must be applied judiciously and should be considered methods of data exploration rather than definitive approaches to building a model. Some scientific journals will no longer accept regression models which have been built solely using automated selection criteria.

Some of the problems with automated model-building procedures are that they:

- yield R^2 values which are too high (see more on validation in section 15.6)
- are based on methods (*eg* partial F-tests) which were designed to test specific hypotheses in the data (as opposed to evaluating all possible relationships) so they produce P-values which are too small and confidence intervals for parameters which are too narrow (more on this below)
- can have severe problems in the face of collinearity
- cannot incorporate any of the non-statistical considerations identified above
- make the predictive ability of the model look better than it really is
- waste a lot of paper.

However, the most serious drawback in their use is that they allow the investigator to avoid thinking about their data and the questions to be asked. By turning the model-building procedure over to an automated process, the investigator abdicates all responsibility for the results of their analysis. To quote Ronan Conroy (Sribney et al, 1998): "Personally, I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase."

However, when faced with a large number of predictor variables, using a variety of automated selection procedures might be helpful in identifying all of the predictors which potentially have statistically significant associations with the outcome.

Example 15.2 Automated model selection for risk factors for pneumonia in swine
data=pig_farm

Starting with a full set of 43 predictors in this dataset and using the natural log of the prevalence of pneumonia (proportion of hogs with typical lung lesions at slaughter) as the outcome (n=66 observations), both forward and backward selection procedures were applied using a selection threshold of P=0.05. The predictors selected by each approach (and their coefficients) were:

Variable	Forward selection	Backward elimination
inlet	-0.04	
hldbc	0.50	0.67
size	0.43	0.67
exhau	-0.37	-0.46
exprn	0.03	0.02
floor (slatted)		-0.51
hmrsd	-0.50	
constant	-2.10	-2.62
Model parameters		
SS_{tot}	62.9	62.9
SSE	28.2	31.6
square root MSE	0.69	0.73
-2lnL	131	139
adjusted R^2	0.51	0.46
AIC	145	151
BIC	-116	-113
Cp	-11.1	-8.1

The description of the selected variables is presented in Example 15.1. A full description of the dataset can be found in Chapter 27. The two procedures arrived at different final models, which was not surprising given the low variable to number of cases ratio in this dataset. The forward selection procedure has produced a superior model which explains more of the variation in the log-prevalence of pneumonia, has lower AIC and BIC scores and a lower Mallow's *Cp*. A best subset approach might be useful to identify a number of good models in situations such as this. However, variables that were selected in both procedures were consistent in their direction, although there were substantial differences in the coefficients. The model which gave the lowest Mallow's *Cp* (-11.72) was very similar to the forward selection model but had one additional term (floor feeding). The model which maximised the adjusted R^2 contained 19 predictors and would have been totally unsuitable (results not shown).

Note This example is provided for pedagogical purposes only, not as a recommended approach to model-building.

Three additional points must be kept in mind when using any automated procedure. First, groups of indicator variables formed by breaking down a categorical variable must all be added or removed together. Second, if any interaction term is included, the main effects of both variables that make up the interaction term must be kept in the model. Third, the analysis will only be based on those observations for which all variables are not missing. If there are many missing observations in the dataset, the data used to estimate the model might be a very small subset of the full dataset.

P-values and automated selection procedures

It is important to note that if you allow an automated selection procedure to sift through all of your predictors and select a group that are significant, the actual level of significance of the selected predictors is less than the α level that you set (*eg* 0.05). Specifically, if you select 'significant' predictors from a list of 10 variables (with α=0.05), then the probability of finding at least one predictor significant due to chance alone is:

$$\alpha^* = 1 - (1 - 0.05)^{10} = 0.40$$

There is a 40% chance that at least one predictor will be significant, even if none of them has any association with the outcome. This value (40%) is called the experiment-wise error rate.

15.5 Conduct the analysis

Once the issues described in the preceding sections have been addressed, the analysis should be relatively straightforward. However, it is inevitably an iterative process. As models are built and evaluated, the investigator gains insight into the complex relationships that exist among the variable in the dataset which allows for more refined, and biologically reasonable models to be built. In the process, investigators must incorporate their biological knowledge of the system being studied along with the results of the statistical analyses.

15.6 Evaluate the reliability of the model

Evaluating any regression model is a two-step process. The first step is to thoroughly evaluate the model using regression 'diagnostics' (*eg* evaluating the normality of residuals from a linear regression model). This assesses the **validity** of the model and procedures for doing this are described in each chapter describing specific model types. The second step is to evaluate the **reliability** of the model. That is, to address the question of 'how well will the model predict observations in subsequent samples?'

Note The term reliability is used differently by various authors, but we will use it to describe how well the conclusions from a regression model can be generalised – *ie* make future predictions (Kleinbaum et al, 1988). Simply reporting the R^2 of a linear model or computing the '% correctly classified' by a logistic model does not evaluate reliability as these estimates will always overstate the true reliability of the model.

The two most common approaches to assessing reliability are a split-sample analysis and leave-one-out analysis. A **split-sample analysis** involves dividing the data into two groups. In a longitudinal study, the data might be divided by time period (*eg* early versus late) and in a cross-sectional study they might be divided randomly. A regression model is built using the data from one of the two groups and the model is then applied to the second group to obtain predicted values for the remaining observations. For linear regression models, the correlation between the predicted and observed values in the second group is called the **cross-validation correlation**. The difference between the R^2 obtained from the analysis of the first group's data and the square of the cross-correlation validation is called the **shrinkage on cross-validation**. If it is small (a subjective decision, although 0.1 would be considered small by most people), then the model is considered reliable. For non-linear regression models (*eg* logistic models), the same general approach can be used but some other measure of predictive ability (*eg* replace R^2 with % correctly classified) needs to be used to compare the two sets of results.

If only a small dataset is available, it might be desirable to put more than 50% of the observations in the first group (the one used to build the prediction model). Alternatively, a 10-fold cross-validation can be carried out in which the data are divided into 10 subsets with 9 being used to estimate the model and that model used to generate predicted values for the 10th subset. This process is repeated with each subset being left out of the model estimation procedure. Split-sample validation of the swine pneumonia model from Example 15.2 (based on the backwards elimination model) is shown in Example 15.3.

Example 15.3 Validating the swine respiratory disease model
data=pig_farm

A forward selection procedure was used to build a regression model (as in Example 15.2) but only a randomly selected 75% subset of the data (n=47 herds) was used to build the model. The resulting model was as shown. This model was then used to compute the predicted pneumonia prevalence for the remaining 19 herds and the correlation between those values and the actual prevalences was computed.

Variable	Coef	P-value
hldbc	0.694	0.005
size (x1000)	0.383	0.020
exprn	0.026	0.005
inlet	-0.048	0.020
constant	-2.733	<0.001

The correlation between the predicted and actual values for the estimation subset (n=47) was 0.703 (R^2=0.495) while for the remaining 19 herds it was 0.633 (R^2=0.400). The shrinkage on cross-validation was <0.1 suggesting that the selected predictors make up a reasonably reliable model.

An alternative approach to split-sample validation involves building separate regression models for each of the sub-groups and subjectively comparing the regression coefficient. **Note** This can be done for any type of regression model. If the coefficients are substantially different in the two models, then the model is not reliable.

A **leave-one-out** approach to validation is based on fitting the model many times, with one observation left out each time (until all have been omitted). The residuals for the omitted observations are summed to provide an estimate of the prediction error which can then be compared with the prediction error from the model based on all observations. If the two values are close, it suggests that the model will predict future observations well.

15.7 Presenting the results

The standard method of presenting results from a regression model is to present the coefficients (don't forget to include the intercept), their standard errors and/or their confidence interval. Assuming the observed effects are causal, the coefficients represent the change that would be expected in the outcome for a unit change in the predictor. For dichotomous predictors (or categorical variables that have been converted to a set of dichotomous predictors), the coefficient represents the effect of the factor being present compared with when it is absent. However, for continuous variables, assessing their impact is more difficult because they are all measured on different scales (and hence, a 'unit change' might represent either a small or large change in the predictor). Consequently, it is difficult to determine the magnitude of the impact of each predictor on the outcome. In order to obtain a better understanding of the effect of a predictor, it would be helpful to have an idea of what constitutes a reasonable change in any predictor measured on a continuous scale. Two approaches to presenting results in order that the relative impact of different predictors can be compared are to

a. use standardised coefficients or
b. compute predicted effects as a continuous predictor changes over its interquartile range.

Each of these will be discussed briefly.

15.7.1 Standardised coefficients

Standardised coefficients represent the effect on the outcome that results from a change of 1 standard deviation (SD) in the predictor. They can be computed by rescaling the coefficient by multiplying it by the ratio of the SD of the predictor to the SD of the outcome [$\beta^*=\beta(\sigma_X/\sigma_Y)$]. In the past, they have not only been used to evaluate the relative magnitude of effects for various predictors in a model, but to compare results across studies. However, there are two problems with this approach. First, the SD might not be a good measure of the variability of a continuous predictor variable. If the distribution is skewed to the right, a few large values might unduly inflate the estimate of the SD. More importantly, the SD of the predictor or the outcome might vary from population to population. If standardised coefficients are used to compare results across studies, identical results from two studies can appear different due to differences in the scaling factor. Consequently, standardised coefficients are no longer recommended for general use.

15.7.2 Interquartile ranges

Rather than computing standardised coefficients, the effect of a predictor can be represented by computing the change in the outcome that would be expected to accompany a change in the predictor across its interquartile range (IQR) (*ie* from its 25th to 75th percentile). This avoids the problem of outlying observations having a big impact on the standard deviation. Although the IQR might also vary across populations (as the SD does), the problem of comparability across studies can be avoided by supplementing the ordinary coefficients with the estimates of effect based on the IQR, rather than replacing the ordinary coefficients with standardised ones. Example 15.4 shows the effects of various variables on the log-prevalence of respiratory disease in swine.

Example 15.4 Effects of predictors
data=pig_farm

Based on the model selected using backward elimination (Example 15.2), the effects of the various predictors was evaluated by computing the expected change in the log-prevalence of pneumonia for defined changes in each of the predictors.

Variable	Coef	Basis	Estimated effect change	Effect
hldbc	0.666	dichotomous	0 - 1	0.666
size ('000s)	0.669	IQR	0.600 - 1.600	0.669
exhau	-0.458	IQR	0.141 - 1.407	-1.342
exprn	0.023	IQR	9 - 26	0.391
floor	-0.509	dichotomous	0 - 1	-0.509

It appears that the capacity of the exhaust fans is one of the largest determinants of the prevalence of respiratory disease in this study population.

15.7.3 Predictors eliminated from a model

When presenting results from a multivariable model, you might also want to discuss the potential effects of predictors not included in the model. Unless the P-value is very large, it is unwise to assume that the effect is zero. Some investigators will discuss unconditional associations between those predictors and the outcome. An alternative, if a backward elimination procedure has been used in the model-building process, is to use the coefficient of the predictor at the last step before it was removed from the model. A third approach is to force the predictor back into the final model and use its coefficient from that model as an estimate of its effect (adjusted for other predictors in the model).

15.7.4 Scale of results

All of the models presented in the examples in this chapter have shown the effects of a predictor on the log-prevalence of pneumonia. The log transformation was necessary to ensure that the residuals from the regression model had an approximately normal distribution. However, it makes the interpretation of the results more difficult and it is often desirable to present results on a different scale than was used in the analysis. Back-transformations following linear regressions are discussed in section 14.9.6. Converting results from the logit scale to the probability scale after logistic regression is discussed in section 16.8.5.

In Example 15.4, the effect of each predictor is assumed to be linear on the log scale, which is equivalent to having a multiplicative effect on the original scale. For example, holding back slow-growing pigs (β=0.666) increases the prevalence of pneumonia by a factor of 1.95 times ($e^{0.666}$=1.95). Consequently, the effect of holding back pigs will depend on the values of other factors in the model, because they will determine the prevalence of pneumonia that is multiplied by 1.95. It is often useful to compute the expected effects of key predictors on the original scale at various levels of other factors in the model.

SELECTED REFERENCES/SUGGESTED READING

1. Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med 1997; 29: 221-239.
2. Kleinbaum DG, Kupper LL, Mullen KE. Applied regression analysis and other multivariable methods. 2d ed. Boston: PWS-Kent Publishing Co, 1988.
3. Hurnik D, Dohoo IR, Donald AW, Robinson NP. Factor analysis of swine farm management practices on Prince Edward Island. Prev Vet Med 1994; 20: 135-146.
4. Hurnik D, Dohoo IR, Bate LA. Types of farm management as risk factors for swine respiratory disease. Prev Vet Med 1994; 20: 147-157.
5. Lofstedt J, Dohoo IR, Duizer G. Model to predict septicemia in diarrheic calves. J Vet Int Med 1999; 13: 81-88.
6. Raftery AE. Bayesian model selection in social research. In: Marsden PV, editor. Sociological Methodology. Oxford: Basil Blackwell, 1996. pp 111-163.
7. Sribney B, Harrell F, Conroy R. Problems with stepwise regression. Stata, Frequently Asked Questions, 1998.

16

LOGISTIC REGRESSION

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand logistic regression
   a. Understand log odds as a measure of disease and how it relates to a linear combination of predictors.

2. Build and interpret logistic regression models
   a. Compute and interpret odds ratios derived from a logistic regression model.
   b. Evaluate the effects of predictors on the outcome of interest on a probability scale.
   c. Statistically compare different logistic models using both Wald tests and likelihood ratio tests.
   d. Determine if the relationship between a continuous predictor variable and the log odds of disease is linear.

3. Evaluate logistic regression models
   a. Understand covariate patterns and how they impact the computation of residuals for logistic regression models.
   b. Compute residuals on the basis of one per covariate pattern and one per observation.
   c. Select and use the appropriate test(s) to evaluate the goodness of fit of a logistic model.
   d. Determine the effect of changing the threshold ('cutpoint') on the sensitivity and specificity of the model.
   e. Generate ROC curves as a method of evaluating the goodness of fit.
   f. Identify and determine the impact of influential observations on a logistic model.

16.1 INTRODUCTION

In veterinary epidemiology, we are often in the situation where the outcome in our study is dichotomous (*ie* Y=0 or 1). Most commonly, this variable represents either the presence or absence of disease or mortality. We can't use linear regression techniques to analyse these data as a function of a set of linear predictors $X=(X_j)$ for the following reasons.

1. The error terms (ε) are not normally (Gaussian) distributed. In fact, they can only take on two values.

$$\text{if } Y = 1 \text{ then } \varepsilon = 1 - (\beta_0 + \sum \beta_j X_j)$$
$$\text{if } Y = 0 \text{ then } \varepsilon = -(\beta_0 + \sum \beta_j X_j) \qquad \textit{Eq 16.1}$$

2. The probability of the outcome occurring (*ie* $p(Y=1)$) depends on the values of the predictor variables (*ie* X). Since the variance of a binomial distribution is a function of the probability (p), the error variance will also vary with the level of X and consequently, the assumption of homoscedasticity will be violated.

3. The mean responses should be constrained as:

$$0 \leq \mathrm{E}(Y) = p \leq 1$$

However, with a linear regression model, the predicted values might fall outside of these constraints.

In this chapter, we will explore the use of logistic regression to avoid the problems identified above. The primary dataset used in the examples in this chapter is one derived from a case-control study of *Nocardia spp.* mastitis that was carried out during an outbreak of this disease in dairy herds in Nova Scotia, Canada. The data consist of observations from 54 case herds and 54 control herds. The predictors of interest were primarily related to the management of the cows during the dry period and, in particular, the use of specific types of dry cow mastitis treatment. The variables used in this chapter are presented in Table 16.1.

Table 16.1 Selected variables from the Nocardia dataset

Variable	Description
casecont	case or control status of the herd (the outcome)
dcpct	percentage of cows treated with dry cow treatments
dneo	use of neomycin-based dry cow products in the last year (yes/no)
dclox	use of cloxacillin-based dry cow products in the last year (yes/no)
dbarn	categorical variable for barn type (1=freestall, 2=tiestall, 3=other)

Details of the dataset can be found in Chapter 27.

16.2 THE LOGIT TRANSFORM

One way of getting around the problems described in section 16.1 is to use a logit transform of the probability of the outcome and model this as a linear function of a set of predictor variables.

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \sum \beta_j X_j \qquad \textit{Eq 16.2}$$

where $\ln\left[\frac{p}{1-p}\right]$ is the logit transform. This value is the log of the odds of the outcome (because odds=$p/(1-p)$), so a logistic regression model is sometimes referred to as a log odds model.

Fig. 16.1 p vs logit of p

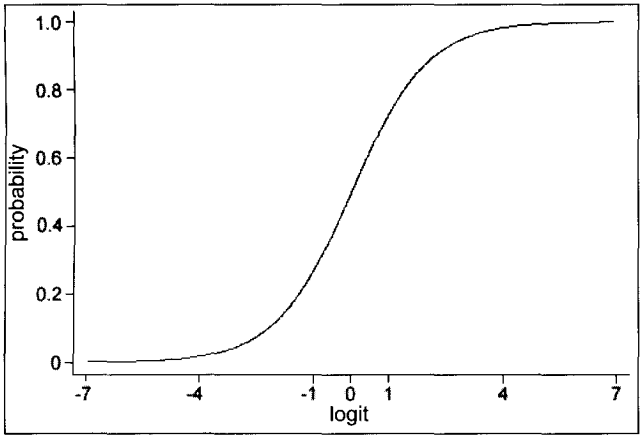


Fig. 16.1 shows that while the logit of p might become very large or very small, p does not go beyond the bounds of 0 and 1. In fact, logit values tend to remain between -7 and +7 as these are associated with very small (<0.001) and very large (>0.999) probabilities, respectively.

This transformation leads to the logistic model in which the probability of the outcome can be expressed in one of the two following ways (they are equivalent).

$$p = \frac{1}{1+e^{-(\beta_0+\sum \beta_j X_j)}} = \frac{e^{(\beta_0+\sum \beta_j X_j)}}{1+e^{(\beta_0+\sum \beta_j X_j)}} \qquad \textit{Eq 16.3}$$

16.3 ODDS AND ODDS RATIOS

Let's look at the simple situation in which the occurrence of disease is the event of interest (Y=0 or 1) and we have a single dichotomous predictor variable (*ie* X=0 or 1). The probability of disease becomes:

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad \textit{Eq 16.4}$$

From this, we can compute the odds of disease (*ie* $p/1\text{-}p$). To simplify calculating the odds of disease:

$$\text{let } \alpha = e^{\beta_0+\beta_1 X} \text{ so } p = \frac{\alpha}{1+\alpha}$$

Then it follows that:

$$\begin{aligned}\text{odds} &= \frac{p}{1-p} = \frac{\alpha}{1+\alpha} \Big/ \left(1-\frac{\alpha}{1+\alpha}\right) \\ &= \frac{\alpha}{1+\alpha} \Big/ \frac{1+\alpha-\alpha}{1+\alpha} \\ &= \alpha = e^{\beta_0+\beta_1 X} \end{aligned} \qquad \textit{Eq 16.5}$$

From this it is a relatively simple process to determine the odds ratio (*OR*) for disease that is associated with the presence of factor '*X*'.

$$\text{if } X = 1 \quad \text{odds} = e^{\beta_0+\beta_1}$$

$$\text{if } X = 0 \quad \text{odds} = e^{\beta_0}$$

The odds ratio is then:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad \textit{Eq 16.6}$$

This can be extended to the situation in which there are multiple predictors and the *OR* for the k^{th} variable will be e^{β_k}.

16.4 Fitting a logistic regression model

In linear regression, we used least squares techniques to estimate the regression coefficients (or at least the computer did this for us). Because the error term has a Gaussian distribution, this approach produces maximum likelihood estimates of the coefficients. In a logistic model, we use a different maximum likelihood estimation procedure to estimate the coefficients.

The key feature of maximum likelihood estimation is that it estimates values for parameters (the βs) which are most likely to have produced the data that have been observed. Rather than starting with the observed data and computing parameter estimates (as is done with least squares estimates), one determines the likelihood (probability) of the observed data for various combinations of parameter values. The set of parameter values that was most likely to have produced the observed data are the maximum likelihood (ML) estimates.

The following is a very simple example which demonstrates the maximum likelihood estimation process. Assume that you have a set of serologic results from a sample of 10 cows from a dairy herd and the parameter you want to estimate is the prevalence of the disease. Three of the 10 samples are positive (these are the observed data).

The likelihood (L) of getting three positive results from 10 cows if the true prevalence is P:

$$L(P) = \binom{10}{3} P^3 (1-P)^7$$

The log likelihood (lnL) is:

$$\ln L(P) = \ln\left\{\binom{10}{3}\right\} + 3\ln(P) + 7\ln(1-P)$$

In this situation, the maximum value of the lnL can be determined directly, but in many cases an iterative approach is required. If such a procedure was being followed, the steps would be:

a. Pick a value for the prevalence (perhaps your first guess is 0.2). The probability of observing three positive cows out of 10, if the true prevalence (P) is 0.2, is:

$$L(0.2) = \binom{n}{x} P^x (1-P)^{n-x} = \binom{10}{3} 0.2^3 (1-0.2)^{10-3} = 0.201 \qquad \textit{Eq 16.7}$$

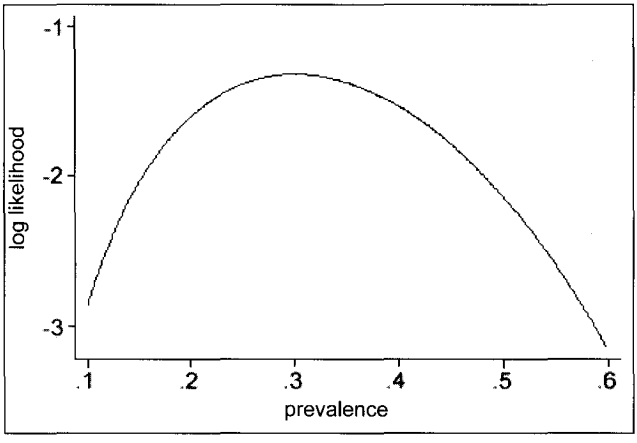
The lnL is -1.60.

b. Pick another prevalence (perhaps your next guess is 0.35) and recompute the likelihood. This turns out to be 0.252 (lnL=-1.38).

c. Keep repeating this process until you have the estimate of the parameter that gives you the highest likelihood (*ie* the maximum likelihood). This would occur at P=0.3 (but you already knew that, didn't you?).

A graph of the relationship between lnL and prevalence (Fig. 16.2) shows the maximum value at P=0.3.

Fig. 16.2 Log likelihood versus prevalence



Of course, the computer doesn't just pick values of parameters at random; there are ways of estimating what the parameter is likely to be and then refining that estimate.

Since it is possible to keep refining the estimates to more and more decimal places, you have to specify the **convergence criterion**. Once the estimates change by less than the convergence criterion, the process of refining the estimates is stopped (*ie* convergence has been achieved).

16.5 Assumptions in logistic regression

As with linear regression, there are a number of assumptions inherent in fitting a logistic model. In a logistic model, the outcome Y is dichotomous:

$$Y_i \begin{cases} 1 & p(Y_i = 1) = p_i = 1 - p(Y_i = 0) \\ 0 & \end{cases} \qquad \textit{Eq 16.8}$$

and two important assumptions are independence and linearity.

Independence It is assumed that the observations are independent from each other (the same assumption was made in linear regression). If animals are maintained in groups or, if multiple measurements are being made on the same individual, this assumption has probably been violated. For example, if animals are kept in herds, variation between animals in the study population results from the usual variation between animals plus the variation that is due to differences between herds. This often results in 'over-dispersion' or 'extra-binomial variation' in the data. Some methods of checking this assumption will be presented in section 16.11.3 and methods of dealing with the problem are discussed in Chapters 20-23.

Linearity As with linear regression, any predictor that is measured on a continuous scale is assumed to have a linear (straight-line) relationship with the outcome. Techniques for evaluating this assumption are presented in section 16.10.

16.6 Likelihood ratio statistics

Although the maximum likelihood estimation process produces the largest possible (*ie* maximum) likelihood value, these values are always very, very small because they are describing the probability of an exact set of observations given the parameter estimates selected. Because of this (and the fact that the estimation process is simpler), computer programs usually work with the log likelihood which will be a moderately sized negative number. Most computer programs print out the log likelihood of the model that has been fit to the data. It is a key component in testing logistic regression models.

16.6.1 Significance of the full model

The test used to determine the overall significance of a logistic model is called the **likelihood ratio test** (*LRT*) as it compares the likelihood of the 'full' model (*ie* with all the predictors included) with the likelihood of the

'null' model (*ie* a model which contains only the intercept). Consequently, it is analogous to the overall *F*-test of the model in linear regressions. The formula for the likelihood ratio test statistic $\left(G_0^2\right)$ is:

$$G_0^2 = 2\ln\frac{L}{L_0} = 2(\ln L - \ln L_0) \qquad \textit{Eq 16.9}$$

where L is the likelihood of the full model and L_0 is the likelihood of the null model. The statistic $\left(G_0^2\right)$ has an approximate χ^2 distribution with k degrees of freedom (df) (k=number of predictors in the full model). If significant, it suggests that, taken together, the predictors contribute significantly to the prediction of the outcome.

Note When computing an *LRT* statistic, two conditions must be met.

1. Both models must be fit using exactly the same observations. If a dataset contains missing values for some predictors in the full model, then these would be omitted from the full model but included when the null model is computed. This must be avoided.
2. The models must be **nested**. This means that the predictors in the simpler model must be a subset of those in the full model. This will not be a problem when the smaller model is the null model, but might be a problem in other situations.

In Example 16.1, a logistic regression model from the case-control study of *Nocardia spp* mastitis has been fit with three predictor variables (-dneo-, -dclox--dcpct-). The likelihood ratio test evaluating the three predictors as a group is highly statistically significant (G_0^2=41.72, df=3, P <0.001).

16.6.2 Comparing full and reduced models

In the preceding section, the *LRT* was used to compare the full and null models but an *LRT* can also be used to test the contribution of any subset of parameters in much the same way as a multiple partial *F*-test is used in linear regression. The formula is:

$$G_0^2 = 2\ln\frac{L_{full}}{L_{red}} = 2\left(\ln L_{full} - \ln L_{red}\right) \qquad \textit{Eq 16.10}$$

where L_{full} and L_{red} refer to the likelihood of the full and reduced models, respectively. As can be seen in Example 16.1, the two antibiotic specific predictors (-dneo-, -dclox-) are highly significant predictors of case-control status. This test is sometimes referred to as the 'improvement χ^2'.

16.6.3 Comparing full and saturated models (deviance)

A special case of the likelihood ratio test is the comparison of the likelihood of the model under investigation to the likelihood of a fully saturated model (one in which there would be one parameter fit for each data point). Since a fully saturated model should perfectly predict the data, the likelihood of the observed data, given this model,

Example 16.1 Comparing logistic regression models
data=Nocardia

The log likelihoods from four different models were:

Model	Predictors	# of predictors	Log likelihood
null	intercept β_0	1	-74.86
full	intercept, dcpct, dneo, dclox $\beta_0, \beta_1, \beta_2, \beta_3$	4	-54.00
reduced	intercept, dcpct β_0, β_1	2	-69.07
saturated	108 'hypothetical' predictors $\beta_0, \beta_1 \ldots \beta_{n-1}$	108	0

Overall likelihood ratio test of the full model:
G_0^2=2(-54.00 - (-74.86))=41.72 with 3 df (P <0.001)
Taken together, the three predictors are highly significant predictors of case-control status.

Likelihood ratio test comparing the full and reduced models:
G_0^2=2(-54.00 - (-69.07))=30.16 with 2 df (P <0.001)
The two antibiotic specific predictors (-dneo- and -dclox-) are highly significant predictors.

Likelihood ratio test comparing the saturated and full models:
G_0^2=2(0 - (-54.00))=108.00 with 104 df.
Note This does not have a χ^2 distribution.

should be 1 (or $\ln L_{sat}=0$). This comparison yields a statistic called the **deviance** which is analogous to the **error sum of squares** (SSE) in linear regression. The deviance is a measure of the unexplained variation in the data.

$$D = 2 \ln \frac{L_{sat}}{L_{full}} = 2(\ln L_{sat} - \ln L_{full}) = -2(\ln L_{full}) \qquad \textit{Eq 16.11}$$

Note The deviance computed in this manner does not have a χ^2 distribution. (See section 16.11.2 for more discussion of deviance.)

16.7 Wald tests

An alternative approach to evaluating the significance of a single coefficient is to use a test that relates the coefficient to its SE. A Wald test is the ratio of the coefficient to its SE and it follows (asymptotically) a standard normal (Z) distribution. This tests

whether the coefficient is significantly different from zero. It is routinely computed by most computer programs and is the most widely used test of the significance of coefficients. However, the estimates of the coefficient and its SE are only estimates and consequently, the normal approximation of its distribution might not be reliable particularly if the sample size is small. Consequently, to evaluate the significance of variables with a P-value close to the rejection region, it is best to use a likelihood ratio test.

Just as with multiple partial F-tests in linear regression, multiple parameters in a logistic model can be tested with a multiple Wald test. For example, comparing the full and reduced models in Example 16.1 would be equivalent to testing the null hypothesis:

$$H_0: \beta_2 = \beta_3 = 0$$

In this case, the test statistic is compared to a χ^2 distribution with the df equal to the number of predictors being tested. In Example 16.1, the Wald χ^2 for comparing the full and reduced models has a value of 21.4 and 2 df. This is a more conservative test statistic (although this is not generally the case) than the likelihood ratio test (G_0^2=30.16), but it is still highly significant.

16.8 INTERPRETATION OF COEFFICIENTS

The coefficients in a logistic regression model represent the amount the logit of the probability of the outcome changes with a unit increase in the predictor. Unfortunately, this is hard to interpret so we usually convert the coefficients into odds ratios. The following sections are based on the model shown in Example 16.2.

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1(\text{dcpct}) + \beta_2(\text{dneo}) + \beta_3(\text{dclox}) + \beta_4(\text{dbarn\_2}) + \beta_5(\text{dbarn\_3})$$

16.8.1 Dichotomous predictor

Coefficients for a dichotomous predictor represent the amount that the log odds of disease increase (or decrease) when the factor is present. These can be easily converted into *OR* by exponentiating the coefficient. For example, the *OR* for -dneo- in Example 16.2 is:

$$OR = e^{\beta_2} = e^{2.685} = 14.7$$

If the outcome of interest is relatively rare, the *OR* provides a good approximation of the risk ratio (*RR*). If the data come from a case-control study in which incidence density sampling was employed, the *OR* is a good estimate of the incidence rate ratio (*IR*) in the original population (see Chapter 6).

16.8.2 Continuous predictor

For a continuous predictor, the coefficient (β_1) represents the change in the log odds of

Example 16.2 Interpreting logistic regression coefficients
data=Nocardia

The tables below present results from a logistic regression of -casecont- on -dcpct- -dneo- -dclox- and two levels of -dbarn-. The first table presents the effects of the predictors on the logit of the outcome (case-control status), while the second shows the same results expressed as odds ratios.

Number of obs = 108
LR chi2 (5) = 47.40
Prob > chi2 = 0.000
Log likelihood = -51.168

Predictor	Coef	SE	Z	P	95% CI	
dcpct	0.022	0.008	2.82	0.005	0.006	0.037
dneo	2.685	0.677	3.96	0.000	1.358	4.013
dclox	-1.235	0.581	-2.13	0.033	-2.374	-0.096
dbarn_2	-1.334	0.632	-2.11	0.035	-2.572	-0.095
dbarn_3	-0.218	1.154	-0.19	0.850	-2.481	2.044
constant	-2.446	0.854	-2.86	0.004	-4.120	-0.771

Predictor	OR	SE	95% CI	
dcpct	1.022	0.008	1.007	1.037
dneo	14.662	9.931	3.888	55.296
dclox	0.291	0.169	0.093	0.908
dbarn_2	0.263	0.166	0.076	0.909
dbarn_3	0.804	0.928	0.084	7.722

Effect of -dneo- Use of neomycin-based products in the herd increased the log odds of Nocardia mastitis by 2.685 units. Alternatively, one can say that using neomycin-based products increased the odds 14.7 times. Since Nocardia mastitis is a relatively rare condition, it would be reasonable to interpret the odds ratio as a risk ratio and state that use of neomycin-based products increased the risk of Nocardia mastitis by approximately 15 times.

Effect of -dcpct- Changing the percentage of dry cows treated from 50% to 75% increases the log odds of disease by: (75-50)*0.022=0.55 units. Alternatively, it increases the odds of disease by: $(1.022)^{(75-50)}=1.73$. An increase of 25% in the percentage of cows dry-treated increases the risk of disease by about 73% (*ie* 1.73 times).

Effect of -dbarn- Tiestall barns (-dbarn_2-) and other barn types (-dbarn_3-) both had lower risks of Nocardia mastitis (*ie* OR <1) than did freestall barns (-dbarn_1- was the omitted baseline). However, the multiple Wald test and the likelihood ratio tests of the two included categories were 0.08 and 0.06, respectively, suggesting that barn type was only borderline significant (0.1 >P >0.05).

disease for a one-unit change in the predictor. Similarly, the computed *OR* represents the factor by which the odds of disease are increased (or decreased) for each one-unit change in the predictor. However, we are often interested in changes of multiple units of the exposure variable(s), such as from x_1 to x_2. For example, for a change from 50% to 75% of cows dry-treated, the log odds of disease changes by:

$$\log \text{odds}\,(x_1, x_2) = (x_2 - x_1) * \beta_1 = (75 - 50) * 0.022 = 0.55 \qquad \textit{Eq 16.12}$$

For this 25% change in -dcpct-, the odds of disease change by:

$$e^{0.55} = 1.73, \quad \text{or} \quad OR\,(x_1, x_2) = OR^{(x_2 - x_1)} = 1.022^{(75-50)} = 1.73 \qquad \textit{Eq 16.13}$$

16.8.3 Categorical predictor

As in linear regression, predictors with multiple categories (*eg* '*j*' categories) must be converted to a series of indicator variables (also called 'dummy' variables) with *j*-1 variables put into the model. The coefficient for each indicator variable represents the effect of that level compared to the category (*ie* the 'baseline') not included in the model. The coefficients are interpreted in the same manner as for any other dichotomous predictor.

Note There are other ways of coding categorical variables, such as hierarchical indicator variables, and these are used in the same way as described in Chapter 14.

When creating indicator variables, the choice of the baseline might be important. In general, we choose one that makes biological sense (*ie* makes some sense as a reference level) and one that has a reasonable number of observations so we are not comparing everything with a category for which the effect can only be estimated very imprecisely. When evaluating the statistical significance of coefficients for categorical variables, it is important NOT to pay much attention to the P-values of individual coefficients. This P-value indicates whether or not the chosen level is statistically different from the baseline level. However, because the choice of the baseline is arbitrary, any category has a range of possible P-values that could be computed. Instead, you should evaluate the statistical significance of all of the categories together with a multiple Wald test or a likelihood ratio test.

In Example 16.2, the variable -dbarn- was converted to a series of three dummy variables and two of these (-dbarn_2-, -dbarn_3-) were included in the model. These represented tiestall and 'other' types of housing, respectively and, consequently, the coefficients represent the effects of these types of housing on the risk of Nocardia mastitis compared with freestall barns (the category that was omitted).

16.8.4 Interpretation of the intercept

Interpretation of the intercept (constant) in the regression model depends on how the data were collected. The intercept represents the logit of the probability of disease if all of the 'risk factors' are absent (*ie* equal to zero). This can be expressed as:

$$\ln\left(\frac{p_0}{1-p_0}\right) = \beta_0 \qquad \textit{Eq 16.14}$$

where p_0 equals the probability of disease in this 'non-exposed group'. In a cross-sectional or cohort study, p_0 has real meaning because it represents the frequency of disease in the non-exposed group. However, in a case-control study, p_0 will vary depending on how many cases and controls are selected for inclusion in the study. We don't really know what the frequency of disease is in the non-exposed group because we didn't take a sample from that group. Consequently, the value of the intercept cannot be meaningfully interpreted if the data came from a case-control study.

16.8.5 Presenting effects of factors on the probability scale

As has been presented above, the coefficients from a logistic model represent the change in the log odds of disease that is associated with a unit change in the factor of interest. These can be relatively easily converted to an odds ratio (by exponentiating the coefficient) but there is a limitation to the usefulness of this parameter.

We normally think about the probability of disease (rather than the odds) and the probability of disease is not linearly related to the factor of interest. Consequently, the effect of a unit increase in the factor usually does not increase the probability of disease by a fixed amount. The amount that a unit increase in the factor changes the probability of disease depends on the level of the factor and the levels of other factors in the model.

In Example 16.3, you can see that the effect of a 10% increase in the percentage of cows dry treated depends heavily on whether it occurs in a herd that uses neomycin or one that uses cloxacillin. It also depends on whether the change is from 10-20% or 80-90%. It is very helpful to generate some graphs of predicted probabilities to get a full understanding of the effects of key variables in your model.

As can be seen, a 10% increase in the level of -dcpct- has a greater effect on the probability of Nocardia mastitis in herds using neomycin; furthermore, in the cloxacillin herds, there is a bigger increase in the predicted probability of mastitis going from 80% to 90% than in going from 0 to 10%.

16.9 Assessing interaction and confounding

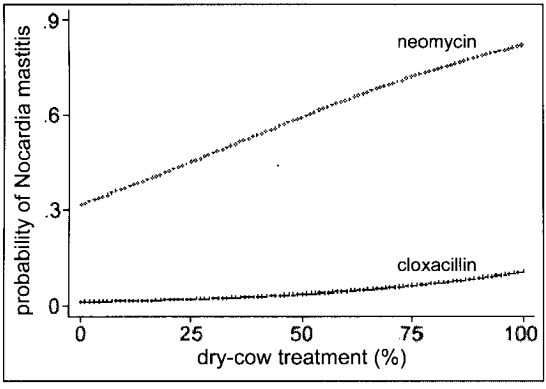
Assessment of interaction and confounding in logistic regression models is similar to the process used in linear regression. **Confounding** is assessed by adding the potential confounding variable to the model and making a subjective decision as to whether or not the coefficient of the variable of interest has changed 'substantially'. In Example 16.4, it appears there is some degree of confounding between -dcpct- and -dclox-.

Example 16.3 Effects of factors on the probability scale
dataset=Nocardia

In this example, a model containing -dcpct-, -dneo- and -dclox- was fit and the predicted probability of Nocardia mastitis computed as -dcpct- rose from 0 to 100%. Predicted probabilities were computed separately for neomycin-using herds and cloxacillin-using herds.

Predictor	Coef	SE	Z	P	95% CI	
dcpct	.023	.007	3.15	0.002	.008	.037
dneo	2.212	.578	3.83	0.000	1.080	3.345
dclox	-1.412	.557	-2.53	0.011	-2.505	-.320
constant	-2.984	.772	-3.86	0.000	-4.498	-1.471

Fig. 16.3 Effect of dry-cow treatment



The effect of a 10% increase in -dcpct- depends on whether the herd is a neomycin-using or a cloxacillin-using herd (*ie* the effect is much greater in neomycin-using herds). It also depends on where on the scale of -dcpct- the increase occurs (going from 10-20% in a cloxacillin-using herd has a smaller effect than going from 80-90%).

Interaction is assessed by adding the cross-product term ($X_1 * X_2$) and determining if the coefficient for the term is statistically significant. Estimation of *OR*s in the presence of interaction deserves some attention though. If interaction is present, the *OR* for the variable of interest has to be determined at a predefined level of the interacting variable because it will vary with the level of the interacting variable.

If the interaction is between two dichotomous predictors, the coefficients for the main effects and the interaction term have straightforward interpretations. The coefficient for each main effect represents the effect of that variable in observations in which the other variable is absent. In Example 16.5, the coefficient for -dneo- (3.184) is a measure of the effect of neomycin used in herds that don't use cloxacillin. The interaction term represents the additional effect of having both factors present, over the sum of the two individual effects. The results shown in Example 16.5 are summarised in Table 16.2.

Example 16.4 Assessment of confounding

dataset=Nocardia

First a 'full' model containing -dcpct- -dneo- and -dclox- was fit, and then -dcpct- was dropped from the model.

	Full model		Reduced model	
Predictor	Coef	SE	Coef	SE
dcpct	0.023	0.007	n/a	n/a
dneo	2.212	0.578	2.377	0.550
dclox	-1.412	0.557	-1.010	0.532
constant	-2.984	0.772	-1.480	0.501

When -dcpct- was removed from the model, the coefficient for -dneo- changes very little (~7%), but the coefficient for -dclox- changes by almost 30% suggesting that -dcpct- and -dclox- might be related (and acting as confounders for each other).

Example 16.5 Assessment of interaction

dataset=Nocardia

Interaction between -dneo- and -dclox- was evaluated by adding their cross-product term:

Predictor	Coef	SE	Z	P	95% CI	
dcpct	0.023	0.008	2.93	0.003	0.007	0.038
dneo	3.184	0.837	3.80	0.000	1.543	4.825
dclox	0.446	1.026	0.43	0.664	-1.565	2.457
neoclox	-2.552	1.205	-2.12	0.034	-4.914	-0.190
constant	-3.777	0.993	-3.80	0.000	-5.724	-1.830

The effect of neomycin and cloxacillin use can be summarised as follows:

neomycin only	log odds goes up by: 3.18 units
cloxacillin only	log odds goes up by: 0.45 units
using both	log odds goes up by: 3.18 + 0.45 -2.55 = 1.08 units

Consequently, using neomycin-based products is much more harmful (increase of 3.18 units in log odds of Nocardia mastitis) in herds using neomycin exclusively. If the herd uses cloxacillin as well, the effect of neomycin is only an increase of 0.63 units (1.08-0.45). Alternatively, cloxacillin seems to have a small (insignificant) detrimental effect when used in herds that don't use neomycin (increase of 0.45 units), but in herds that use neomycin, it is highly beneficial (reduces log odds by 2.1 units (3.18-1.08).

Table 16.2 Effect of neomycin and cloxacillin use on the log odds of Nocardia mastitis compared with using neither (from Example 16.5)

		cloxacillin	
		0	1
neomycin	0	0	0.446
	1	3.184	1.078

Note 1.078=3.184+0.446–2.552

Higher-order interactions (*eg* three-way interactions) might also be evaluated (see section 15.2.3).

16.10 Model-building

In general, the process of building a logistic model is very similar to that of building a linear regression model. It might involve any of the following steps.

- laying out a tentative causal diagram to guide your thinking
- unconditional analyses of relationships between predictors and the outcome of interest using a 'liberal' P-value
- evaluation of relationships (correlations) among predictor variables
- automated model-building processes (used with caution)
  - forward selection
  - backward elimination
  - stepwise selection
  - best subset regression
- manual model-building guided by a causal diagram (preferred method) including:
  - evaluation of confounding
  - evaluation of interaction.

However, there is one fundamental difference and it relates to the process of evaluating the shape of the relationship between a continuous predictor variable and the outcome of interest. In linear regression, you might be able to get a reasonable estimate of the relationship between a continuous predictor and the outcome of interest by looking at a simple scatterplot of the two variables. In a logistic model, the assumption behind a linear modelling of any continuous predictor is that the log odds of disease increases linearly with the predictor. Unfortunately, a simple scatterplot of the two variables is of little use with a dichotomous outcome because it produces points that make two horizontal lines across the graph.

Some of the options that you have for evaluating the shape of this relationship include:

1. Plotting the residuals from the model, with the predictor included, against the values of the predictor.

2. Categorising the continuous predictor and:
   a. inserting the indicator variables into the model, or
   b. computing and plotting the log odds of the outcome against the category means.
3. Adding higher order terms to the model:
   a. quadratic and possibly cubic terms, or
   b. orthogonal polynomials, or
   c. fractional polynomials.
4. Generating a smoothed scatterplot of the log odds of the outcome against the predictor.
5. Creating several linear splines to use instead of the original variable.

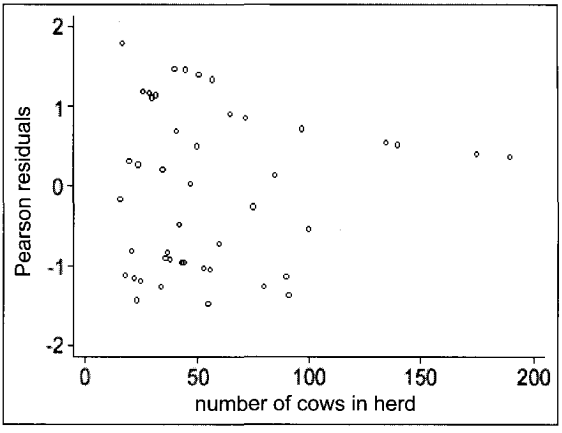
We will discuss each of these in turn using the effect of herd size (-numcow-) on the odds of Nocardia mastitis. Another approach, based on generalised additive models, will not be discussed but has been reviewed elsewhere (Hastie and Tibshirani, 1995).

16.10.1 Plotting residuals

Fig. 16.4 in Example 16.6 shows the Pearson residuals (see section 16.11.2) from a model containing only -numcow- as a predictor of Nocardia mastitis. Any departure from linearity is not obvious. However, in general, this method of evaluating the functional form of the relationship between a continuous predictor and the log odds of the outcome is often not very informative. The plot of residuals will only show a distinct non-linear pattern if the predictor is strongly associated with the outcome and has a distinctly non-linear relationship.

Example 16.6 Evaluating continuous predictors (Part 1) – plots of residuals

Fig. 16.4 Pearson residuals



The graph shows a plot of Pearson residuals from a logistic model with -numcow- as the only predictor against herd size (-numcow-). No indication of non-linearity is evident. -numcow- is borderline significant in this model (β=0.015, P=0.04)

16.10.2 Categorising a continuous predictor variable

If the relationship between a continuous predictor and the log odds of the outcome is not linear, one simple approach is to chop the predictor up into categories and fit a set of indicator variables in the model. The two drawbacks to this approach are that you throw away information by categorising the continuous variable and, if many categories are required to capture the effect of the predictor, then you will have to include a lot of indicator variables in the model (*ie* estimate many parameters). The decision on where to divide the categories should be based, if possible, on what would be biologically meaningful. For example, if rectal temperature had been measured in a study of horses with colic, it might be reasonable to categorise the temperature as below normal, normal range or elevated. However, in many cases, there is no obvious choice for cutpoints. In these situations, choosing cutpoints that create roughly equal-sized groups is an appropriate strategy.

If ordinary indicator variables are created, then a general increase (or decrease) in the coefficients as you go up through the categories of the predictor will be indicative of an approximately linear relationship. The coefficients for the indicator variables shown in Example 16.7 suggest there is little difference between the two largest herd-size groups. If you create hierarchical indicator variables, and the categories of the predictor are equally spaced, then you would expect all of the coefficients to be approximately the same size (because hierarchical variables reflect the effect of going from one category to the next highest category).

Once the continuous variable has been divided into categories, you can compute the log odds of the outcome among the observations that fall in each of the categories and plot those values against the midpoints of the categories. This provides a very good visual assessment of the linearity of the relationship provided you have reasonable sample sizes in each of the groups. Fig. 16.5 suggests that the odds of the outcome increases quickly at small herd sizes but flattens off at about 50 cows.

16.10.3 Polynomials

Polynomials are power terms (*eg* X_1^2) that are added to a model to allow the regression line to follow a curve through the data rather than a straight line. The complexity of the curve (*ie* number of bends) depends on the number of power terms included in the polynomial. Three ways of creating polynomials will be discussed: quadratic, orthogonal and fractional.

16.10.4 Quadratic polynomials

The most common way to fit a curve (rather than a straight line) through the data is to add a quadratic term (the predictor squared). This fits a simple curve which bends in only one direction. The significance of the quadratic term can be used as a check of whether the assumption of linearity is acceptable (provided the data do not follow a more complex pattern than suggested by the single curve of a quadratic model). One issue to keep in mind is that the original value is often highly correlated with its

Example 16.7 Evaluating continuous predictors (Part 2) – categorising the data

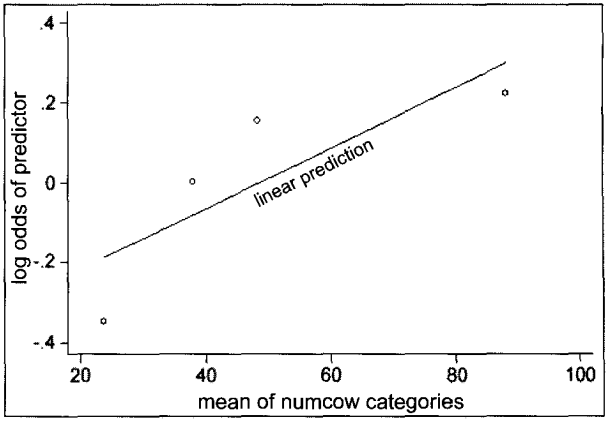
The table below presents a logistic model with the original variable -numcow- replaced by three indicator variables representing herds from 32-41 cows, 42-54 cows and ≥55 cows. All are compared to the smallest herds (≤31 cows).

Number of obs = 108
LR chi2 (3) = 1.35
Prob > chi2 = 0.716
Log likelihood = -74.183

Predictor	Coef	SE	Group size	Mean herd size in group
numcow_0	–	–	29	23.6
numcow_32	0.348	0.544	26	37.9
numcow_42	0.502	0.545	26	48.2
numcow_55	0.571	0.541	27	88.1
constant	-0.348	0.371	NA	NA

It appears that there is a relatively large jump in the log odds of the outcome between the smallest herds (baseline) and the next size group (log odds increases 0.35 units), but as herd size continues to increase, there is little additional increase. Note, however, that when categorised using these cutpoints, -numcow- is not a statistically significant predictor of -casecont- (P=0.72).

Fig. 16.5 Assessing linearity assumption - log odds



A graph of log odds of the outcome predictor against the midpoints of the four categories of herd size (divided at the quartiles) suggests that the risk of Nocardia mastitis increases substantially going from the smallest herds to those in the second quartile. Increases beyond this are successively smaller.

squared term and collinearity might be a problem in the model. The usual way to avoid this problem is to centre the original variable before squaring it (see section 14.6.2). Example 16.8 shows that the quadratic term is not significant, suggesting that a linear relationship might be acceptable. Caution must be used when interpreting results from polynomial models. They might be heavily influenced by points at the ends of the range of values for the predictor. It is also very dangerous to make any predictions outside the range of observed values.

Example 16.8 Evaluating continuous predictors (Part 3) - polynomials

Quadratic term: -numcow- was centered by subtracting the value 75 from all observations and then squared. The correlation between these two new variables was 0.27 which should not cause any problem with collinearity.

Number of obs = 108
LR chi2 (2) = 4.67
Prob > chi2 = 0.097
Log likelihood = -72.525

Predictor	Coef	SE	Z	P	95% CI	
numcow_ct	0.016	0.009	1.74	0.081	-0.002	0.033
numcow_ct_sq (x 10,000)	0.981	1.827	0.54	0.591	-2.600	4.563
constant	0.284	0.352	0.81	0.420	-0.407	0.975

The complete lack of significance for the quadratic term suggests that using a linear term might be adequate.

Orthogonal polynomials: two new terms (-numcow_op1- -numcow_op2-) were created as orthogonal polynomials from -numcow-. These variables are on a new scale with each having a mean of 0 and a standard deviation of 1 and they are totally uncorrelated.

Variable	Obs	Mean	SD	Min	Max
numcow	108	49.092	30.020	16	190
numcow_op1	108	0	1.005	-1.107	4.716
numcow_op2	108	0	1.005	-1.209	5.697

The logistic model was refit using the orthogonal polynomials.

Predictor	Coef	SE	Z	P	95% CI	
numcow_op1	0.515	0.309	1.66	0.096	-0.091	1.121
numcow_op2	0.164	0.305	0.54	0.591	0.433	0.761
constant	0.030	0.202	0.15	0.881	-0.367	0.427

(continued on next page)

Example 16.8 (*continued*)

Fractional polynomials: two new variables (-fp_nc1- and -fp_nc2-) have been created:

numcow_fp1=(numcow/100)$^{-2}$ - 4.149

numcow_fp2=(numcow/100)$^{+3}$ - 0.1183

Unlike orthogonal polynomials, these two variables have some degree of correlation (ρ=-0.28). These have been used in a logistic model of -casecont-.

Number of obs = 108
LR chi2 (2) = 5.53
Prob > chi2 = 0.063
Log likelihood = -72.097

Predictor	Coef	SE	Z	P	95% CI	
numcow_fp1	-0.018	0.024	-0.77	0.442	-0.066	0.029
numcow_fp2	0.682	0.548	1.24	0.213	-0.392	1.757
constant	0.001	0.235	0.01	0.995	-0.459	0.462

The log likelihood of the model has been reduced to -72.1 (from -72.5), but neither of the terms is significant individually, although they are jointly borderline (P=0.063) significant.

16.10.5 Orthogonal polynomials

One way to ensure the new variables that are replacing the original variable are uncorrelated is to create orthogonal polynomials. These are variables that are constructed from the original data but are on a new scale with each variable having a mean of 0 and a SD of 1. The correlation between any pair of these variables is 0. These new variables can be used in place of the original variables in the logistic model. The model based on orthogonal polynomials in Example 16.8 shows exactly the same result as the quadratic model described above. Although the coefficients are different (because the new variables are on a different scale), the P-value of the squared (quadratic) term is exactly the same. Removal of the collinearity makes it possible to interpret the lower order terms, but the fact that they are not in the original scale makes this difficult.

16.10.6 Fractional polynomials

While any set of variables might be orthogonalised, orthogonal polynomials are usually limited to power terms that have positive integer values (*eg* X_1^2, X_1^3). One way of exploring more flexible functional forms is to use fractional polynomials. Fractional polynomials are power terms that can take on both positive and negative integer values and fractional values (*eg* $X^{0.5}$, X^{-2}, X^3). In addition, they can include the natural log of the original value. The combination of fractional polynomials that best fits the data (*ie* the model with the smallest log likelihood) can be determined. Because there are

an infinite number of possibilities to choose from, it is usual to restrict the search to a range of power terms (*eg* between -2 and +3 and including -0.5 and +0.5). In Example 16.8, the power terms -2 and +3 have been selected as the best fit. The drawback of fractional polynomials is that, while they might fit the data very well, the coefficients are almost impossible to interpret. However, if you want to merely control for the effect of a factor (*ie* a potential confounder) in a logistic regression model, then fitting fractional polynomials can be a useful approach.

16.10.7 Smoothed scatterplots

A simple graph of the outcome against the predictor is uninformative because it consists of two horizontal lines of points. However, a smoothed scatterplot of the probability of the outcome against the predictor variable will show you how the mean probability of the outcome changes as the predictor increases. While this can be generated with the original data (*ie* on the probability scale), it is more informative to generate a graph on the log odds scale. An advantage of smoothed scatterplots over fitting some form of polynomial function is that local weighting in the smoothing process makes it easier to pick up departures from global curves. However, while a smoothed scatterplot might identify the functional form of the relationship, it does not generate the variables necessary for inclusion in the logistic model. It is therefore a descriptive rather than a modelling tool.

Smoothed scatterplots can be computed as a function of the original data, or from predicted values based on a regression function. In the former, the running mean at any given point is computed as a weighted average of the points around the point of interest. Weights are applied so that data points close by are weighted more heavily than those at a distance. In the latter, instead of weighting the original data points, a weighted linear regression is run on the points surrounding the one of interest and the predicted value is obtained. These predicted values are then joined to form the smoothed line, referred to as a **lowess curve**. This latter approach tends to produce smoother curves, but might generate unusual results at the extreme values of the predictor. In both cases, a parameter called the **bandwidth** determines how large a subset of the data is used in the computation at each point. For example, if the bandwidth is 0.5, then 50% of the observations will be included in each calculation (except towards the ends of the distribution where smaller uncentred subsets are used). However, points towards the ends of the bandwidth receive relatively little weight in the calculation.

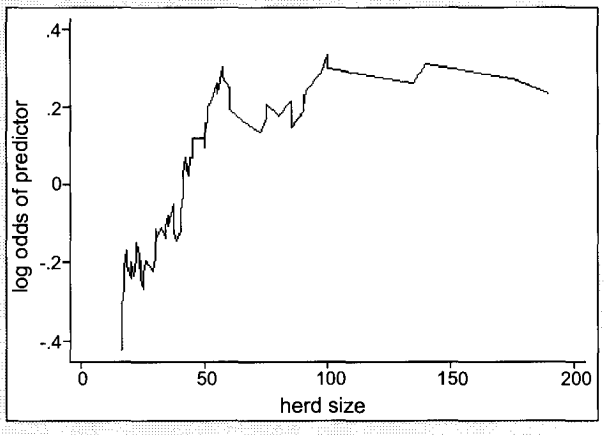
A smoothed (running mean) scatterplot of the log odds of -casecont- against -numcow- is shown in Fig. 16.6 in Example 16.9. The log odds of -casecont- increases in a curved fashion up to about 70 cows and then levels off.

16.10.8 Piecewise linear functions (splines)

In situations when the functional form of the relationships appears to consist of a set of straight lines, you can replace the original variable with a set of variables that will fit straight lines through subsets of the data. This set of piecewise linear functions is called

Example 16.9 Evaluating continuous predictors (Part 4) – smoothed scatterplots

Fig. 16.6 Smoothed scatterplot



A smoothed scatterplot of the log odds of -casecont- on -numcow-. Smoothing was done using Cleveland's tricube weighting function (Cleveland, 1979) with the bandwidth set to 80%.

This graph shows an increase in the log odds (approximately linear) up to about 70 cows and then a relatively flat relationship after that.

a **spline**. Based on the scatterplot in Fig 16.6, it appears that a straight line up to about 70 cows, followed by a second straight line up to 190 cows might fit the data reasonably well. The point at which one piecewise linear function stops and the next begins is known as a **knot point**. In a simple situation with two linear splines, the first will take the values of the original variable up to the knot point while the second is set to zero. Beyond the knot point, the first variable has the value of the knot point while the second has the original value minus the knot point. This can be seen in Example 16.10.

Splines are not necessarily restricted to fitting linear functions to subsets of the data (*ie* polynomial functions can be fit as well and the way these are pieced together at the knot points becomes part of the spline specification), but the computation of these is beyond the scope of this book.

At this point, you have conflicting evidence about the nature of the relationship between herd size (-numcow-) and the log odds of -casecont-. The smoothed scatterplot clearly suggests that the functional form is not linear, but none of the methods that add additional terms (*eg* polynomials, splines) provide any statistical support for the addition of these terms. This lack of statistical significance is, in part, a function of the fact that herd size is only a weak predictor of whether or not the herd was a case or a control. If you were exploring herd size as a risk factor, there would be little justification for using anything other than a linear term. If you had reasonable evidence to believe that herd size was a potential confounder and you wanted to do the best job possible of removing that confounding effect when evaluating other risk factors, then fitting a polynomial model of some sort would be appropriate.

Example 16.10 Evaluating continuous predictors (Part 5) – splines

Two new variables (-numcow_spl- and -numcow_sp2-) were computed to fit straight lines through the herd sizes ≤70 and >70. These were then used in a logistic model of -casecont-. Selected -numcow- data are used to demonstrate the corresponding spline value.

Herd	numcow	numcow_spl	numcow-sp2
32	42	42	0
89	50	50	0
97	65	65	0
98	72	70	2
103	97	70	27
106	140	70	70

Number of obs = 108
LR chi2 (2) = 4.52
Prob > chi2 = 0.105
Log likelihood = -72.601

Predictor	Coef	SE	Z	P	95% CI	
numcow_sp1	0.010	0.014	0.68	0.497	-0.018	0.037
numcow_sp2	0.022	0.019	1.16	0.247	0.015	0.058
constant	-0.513	0.610	-0.84	0.400	-1.709	0.682

This model doesn't fit the data quite as well as the quadratic polynomial (log likelihood is -72.6 versus -72.5) and even less well than the fractional polynomial.

16.11 Evaluating logistic regression models

There are two steps in assessing the fit of the model. The first is to determine if the model fits, in general, using summary measures of goodness of fit or by assessing the predictive ability of the model. The second is to determine whether there are any specific observations (or groups of observations) that do not fit the model or that are having an undue influence on the model. However, before proceeding with either of these two areas, it is important to understand the distinction between residuals computed on the basis of 'covariate patterns' and those computed on the basis of 'observations'.

16.11.1 Covariate patterns

Most of the summary measures of goodness of fit depend on an understanding of the term **covariate pattern.** A covariate pattern is a unique combination of values

of predictor variables. For example, if the model contains only two dichotomous predictors, there will be four covariate patterns: (1,1) (1,0) (0,1) (0,0). Data of this form are called binomial data because observations within the groups formed by the covariate pattern are modelled as replications and the number of positives within a group is binomially distributed. On the other hand, if the model contains many continuous variables, there might very well be as many covariate patterns as there are data points (*ie* each covariate pattern will have only one observation in it) and these data are referred to as binary data.

Residuals from logistic models can be computed on the basis of one residual per observation or one residual per covariate pattern. To get a feeling for the difference between these two approaches, imagine a covariate pattern 'A' with two observations, 1 disease '+' and one disease '-'. Further assume that the predicted value for the probability of disease in animals with this covariate pattern is 0.5 (Table 16.3).

Table 16.3 Residuals computed on the basis of one per observation and one per covariate pattern

				Residuals	
Observation	Covariate pattern	Disease	Predicted value	1 per observation	1 per covariate pattern
1	A	1	0.5	positive	0
2	A	0	0.5	negative	

With one residual per observation, we have two residuals, of which one will be positive and one will be negative. With residuals computed on the basis of covariate patterns, the predicted value (0.5) exactly equals the observed value (0.5) so the residual is zero. For logistic models, residuals are normally computed on the basis of one per covariate pattern and some of the desirable properties of the residuals only apply if there is a reasonable number of observations in each covariate pattern.

In the following discussion, we will use j to represent the number of covariate patterns, m_j to represent the number of data points in the j^{th} covariate pattern, k to represent the number of predictors in the model (not including the constant) and n is the number of data points in the dataset.

All of the examples in this section are based on the model shown in Example 16.4. The values of the predictors in this model make up 30 distinct covariate patterns.

16.11.2 Pearson and deviance residuals

Computing residuals for a logistic model is not as straightforward as it is following a linear regression model (*ie* observed value-expected value). A number of different types of residual have been proposed, but the two most commonly used are Pearson residuals and deviance residuals.

Pearson residuals are roughly analogous to standardised residuals in linear regression. They are based on the difference between the observed and expected values for a given covariate pattern, but are adjusted based on the precision of the estimate of the observed value (*ie* covariate patterns with a large number of observations will have a more precise estimate than those in which there are few observations). **Pearson residuals** are computed as:

$$r_j = \frac{y_j - m_j p_j}{\sqrt{m_j p_j (1 - p_j)}}$$

where y_j=the number of positive outcomes in the j^{th} covariate pattern and p_j=the predicted probability for the j^{th} covariate pattern. Pearson residuals might also be standardised to have a mean of zero and unit variance if the data are binomial. These are called standardised Pearson residuals. Pearson residuals computed on the basis of one per covariate pattern and one per observation are presented in Example 16.11.

Example 16.11 Residuals and covariate patterns

data=Nocardia

Logistic regression model of -casecont- on -dcpct-, -dneo-, and -dclox- was fit (see Example 16.4).

It turns out that there were 30 distinct covariate patterns represented in this model. The data for covariate pattern #9 (herds that dry-treated 20% of their cows, and used neomycin-based products but not cloxacillin-based products) are shown below.

id	case-control	dcpct	dneo	dclox	cov. pattern	pred. value	Pearson residual (covariate)	Pearson residual (observ.)
86	1	20	1	0	9	0.421	0.226	1.173
22	0	20	1	0	9	0.421	0.226	-0.853

There were two observations in covariate pattern #9 and an observed probability of a positive outcome of 0.5 (1 of the 2 herds was positive). The predicted probability was 0.421 and the Pearson residual computed on the basis of one residual per covariate pattern was a small positive value (0.226). However, when residuals were computed for each observation individually, there was one moderately large positive residual value (1.173 for the case herd) and a negative residual value of a similar magnitude (-0.853) for the control herd.

Deviance residuals represent the contribution of each observation to the overall deviance. The sum of deviance residuals computed on the basis of individual observations (rather than covariate patterns) is the deviance (-2*log likelihood) that was observed when comparing the full and saturated models (section 16.6.3).

16.11.3 Goodness-of-fit tests

A variety of tests are available to provide an overall assessment of how well the model fits the observed data. All of these tests are based on the premise that the data will be divided into subsets and within each subset, the predicted number of outcome events will be computed and this will be compared with the observed number of outcome events. Two tests (the Pearson χ^2 and the deviance χ^2) are based on dividing the data up into the natural covariate patterns. A third test (Hosmer-Lemeshow test) is based on a more arbitrary division of the data. Other measures of fit are also described.

Pearson and deviance χ^2 tests
The sum of Pearson residuals squared is known as the Pearson χ^2 statistic. When computed on the basis of one per covariate pattern, this statistic has a χ^2 distribution with (j-k-1) df provided that j is much smaller than n (*ie* on average, the m_j are large). j being much smaller than n ensures that the observed probability of the outcome in each covariate pattern is based on a reasonable sample size. If j=n (*ie* binary data), or almost so, the statistic does not follow a χ^2 distribution, so this goodness-of-fit statistic cannot be used.

The Pearson χ^2 indicates whether or not there is sufficient evidence that the observed data do not fit the model (*ie* H_0 is that the model fits the data). If it is not significant, it suggests that there is no reason to assume that the model is not correct (*ie* we accept that the model generally fits the data). **Note** In general, goodness-of-fit tests do not have a lot of power to detect inadequacies in the model.

The sum of the squared deviance residuals computed on the basis of 1 per covariate pattern (*ie* only applicable to binomial data) is called the deviance χ^2. **Note** The term deviance χ^2 is used to differentiate this deviance from that computed on the basis of 1 per observation (discussed in section 16.6.3). As with the Pearson χ^2, it has a χ^2 distribution with (j-k-1) df. If either the Pearson χ^2 or the deviance χ^2 are significant, you should be suspicious that the data do not fit the model. Example 16.12 shows the Pearson χ^2 and deviance χ^2 for the model presented in Example 16.4.

Hosmer-Lemeshow goodness-of-fit test
If you have binary data (or any situation where j is not much less than n), you can't rely on covariate patterns to divide your data into subsets of sufficient size for a valid goodness-of-fit test. One way to get around this problem is to group the data using some method other than covariate patterns and compare the observed and predicted probabilities of disease (if that is the outcome of interest) in each group. This is the basis of the Hosmer-Lemeshow test.

There are two ways to group the data. The first is on the basis of percentiles of estimated probability and the second is on fixed values of estimated probability. For example, if you want 10 groups, the first method would take the 10% of the data points with the lowest predicted probabilities of disease and put them in group 1, the next 10% in group 2 *etc*. The second approach would take all data points for which the predicted probability of disease was less than 0.1 and put them in a group (regardless of how

many data points fell into that group). In general, the first approach is preferable because it avoids the problem of some groups having very small sample sizes.

Once the data are grouped, a 2*g table is set up (g is the number of groups and should not be <6) with the observed and expected number of cases included in each cell. The expected number of cases in the g=1 row of the table is simply the sum of the estimated probabilities for all subjects in the group. The observed number of cases is simply the number of observations with Y=1. The observed and expected values are compared using a χ^2 statistic with g-2 df. A visual comparison of the observed and expected values will also identify areas where the model might not fit well. Example 16.12 shows the Hosmer-Lemeshow χ^2 along with the observed and expected values.

Example 16.12 Goodness-of-fit tests

Test	χ^2	df	P
Pearson χ^2	45.58	26	0.010
Deviance χ^2	22.98	26	0.136
Hosmer-Lemeshow	6.06	2	0.048

As can be seen from the P values, there is quite a range of estimates. Since goodness-of-fit tests generally have low power for detecting inabilities of models to adequately fit the data, the general guideline is that if any goodness-of-fit test is statistically significant, you should assume there is a problem with the model and try to correct it. It is also worth noting that with 108 observations and 30 covariate patterns, the average number of observations per covariate pattern is quite low, so the Hosmer-Lemeshow test might provide the most reliable evaluation.

A table of the observed and expected values from the Hosmer-Lemeshow test provides some insight into where the model does not fit the data very well.

Group	p(D+)	Cases observed	Expected	# of herds
1	0.106	4	1.9	24
2	0.408	3	6.2	20
3	0.717	12	13.2	24
4	0.817	35	32.7	40

Proportionally, the largest difference between the observed and expected number of cases is in the first group (lowest predicted probabilities). One possible explanation of this is that some cases might have arisen from mechanisms not included in the model.

R^2 (pseudo-R^2)
A number of pseudo R^2-type measures for estimating the amount of variation explained by a logistic regression model have been proposed and recently reviewed (Mittlbock and Schemper, 1996). In general, Hosmer and Lemeshow (2000) argue that the pseudo-R^2 is equivalent to the likelihood ratio test for all of the parameters in the model (*ie* comparing the likelihood of the full model to one with only the intercept). It does not compare the fit of the model with the observed values and consequently is better suited for comparing models than for assessing the goodness of fit of a selected model.

16.11.4 Predictive ability of model

A second general approach to assessing the overall usefulness of the model is to assess its predictive ability (*ie* how good a job does it do in predicting the outcome?). This can involve computing the sensitivity and specificity of the model at various probability thresholds and/or generating a receiver operating characteristic (ROC) curve.

Sensitivity and specificity
The ability of the model to correctly classify individuals (or in this example, herds) can be assessed by computing the classification statistics after fitting a model. By default, these are computed by classifying every observation that has a predicted probability ≥ 0.5 as positive and those with values <0.5 as negative. However, this cutpoint can be lowered (to increase the sensitivity of the model) or raised (to increase the specificity) similar to the discussion of cutpoints for tests (section 5.6.3). A graph of the sensitivity and specificity vs the potential cutpoint values (two-graph ROC curve) is helpful in selecting an appropriate cutpoint (Example 16.13).

Receiver operating characteristic curves
An ROC curve for the model can also be generated to evaluate the performance of the model at all possible cutpoints. The closer the curve comes to the upper left corner of the graph, the better the predictive ability of the model. If the ROC curve is close to the diagonal line, it indicates that the model has very little predictive ability. The maximum area under an ROC curve is 1.0 (*ie* sensitivity=100% and specificity=100%) while the area will be 0.5 if the curve falls on the diagonal line (*ie* has no predictive ability at all). (See section 5.6.4 for a more complete discussion of ROC curves.) The predictive ability of the model for Nocardia mastitis is shown in Example 16.13.

16.11.5 Identifying important observations

Detecting observations which either do not fit the model well, or which might have an undue influence on the model is an important component of evaluating a logistic regression model, particularly if any of the goodness-of-fit statistics indicate problems with the model.

Outliers
Pearson residuals and deviance residuals represent the square root of the contribution of the covariate pattern to the Pearson and deviance χ^2 statistics, respectively. As with standardised residuals from linear regression, large positive or negative standardised

residuals identify points which are not well fit by the model. If outliers are observed, it is important to try to determine:

1. Why they are outliers (what are the characteristics of the observations that make them outliers?).
2. If the data are found to be erroneous, they should be corrected, or failing that, deleted.
3. If the data are correct, determine if they are having an undue effect on the model.

This last point can be evaluated by looking at other diagnostic parameters (leverage,

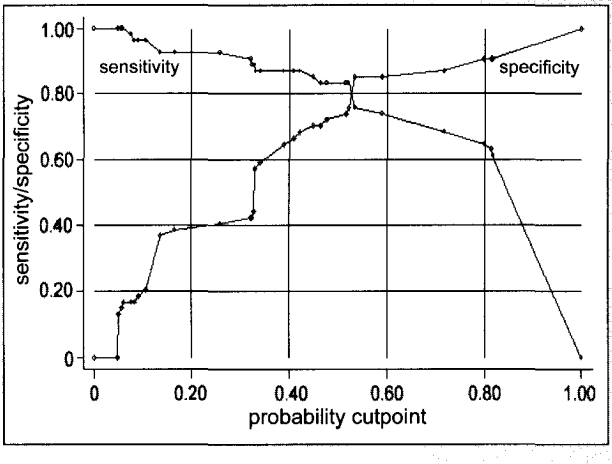
Example 16.13 Predictive ability of a model

data=Nocardia

For the model presented in Example 16.10, the classification statistics are:

	Classified (predicted) status		
True status	T+ p(D+)≥0.5	T- p(D+)<0.5	Total
D+	45	9	54
D-	14	40	54
Total	59	49	108
Sensitivity		pr (T+|D+)	83.3%
Specificity		pr(T-|D-)	74.1%
Positive predictive value		pr(D+|T+)	76.3%
Negative predictive value		pr(D-|T-)	81.6%

Fig. 16.7 Use of a two-graph ROC to show the effect of changing cutpoint



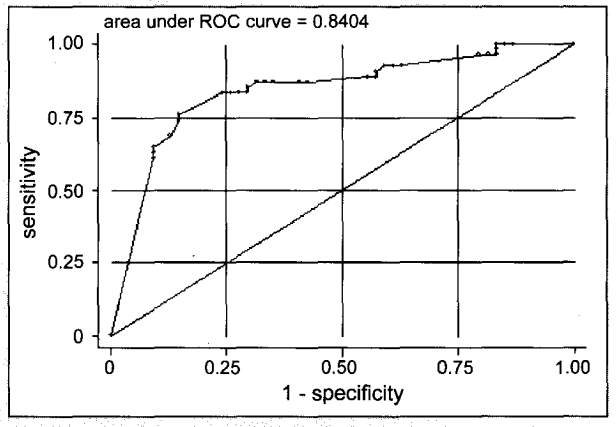
At a cutpoint of 0.5, the sensitivity and specificity of the model are roughly balanced. The effect of changing the cutpoint can be evaluated visually in the graph.

In this situation, reducing the cutpoint would reduce specificity quite dramatically and raising it beyond about 0.75 would seriously affect sensitivity.

(continued on next page)

Example 16.13 (*continued*)

Fig. 16.8 ROC curve



An ROC curve for this model is presented in Fig. 16.8.

The ROC curve extends reasonably well into the upper left-hand corner of the graph and the area under the curve is 0.84. Both of these indicate that the model has a moderate predictive ability.

delta-betas, *etc* (see below) or by refitting the model with the outliers omitted. (Deleting the outliers should only be done for the purpose of evaluating their impact on the model and they must be put back in the dataset.) In general, outliers contribute to the lack of fit of a model but often do not have an undue influence on it. An index plot of standardised Pearson residuals (1 per covariate pattern) is shown in Example 16.14.

Hat matrix and leverage

Another quantity central to the discussion of logistic regression diagnostics is the hat matrix. It is used to calculate leverage values and other diagnostic parameters. The hat matrix is a square matrix of dimension $j * j$ (j=number of covariate patterns) or $n * n$ (n=number of data points) depending on whether the data are binomial or binary. The diagonal elements of the hat matrix are the logistic regression leverage values (h_j) (see Hosmer and Lemeshow, 2000 for details).

As in linear regression, leverage measures the potential impact of an observation (or covariate pattern) on the model. Points with high leverage certainly deserve evaluation given their potential impact.

Unlike leverage values in linear regression models, the leverage of a data point in a logistic model is not exclusively a function of the values of the predictors. Data points that have extreme values of predictor variables (which would have high leverage in linear regression) might, in fact, have low leverage in logistic regression if the predicted value is very large or very small. Observations with extreme values of the predictor(s) will have leverage values that are: highest if the predicted probability lies between 0.1 and 0.3 or 0.7 and 0.9, moderate between 0.3 and 0.7, and low if the predicted probability is <0.1 or >0.9. The covariate patterns with the highest leverage are shown in Example 16.15.

Delta-betas

Values of delta-beta provide an estimate of the effect of the j^{th} covariate pattern on the logistic regression coefficients. These values are analogous to Cook's distance in linear regression models.

A single set of values of delta-beta can be calculated – one value for each covariate pattern – and this represents the overall effect of the covariate pattern on the regression

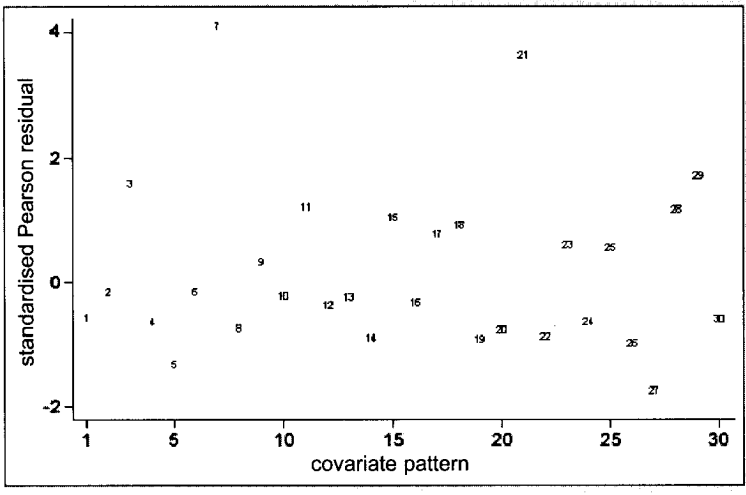
Example 16.14 Identifying important observations
data=Nocardia

Closer examination of the two large standardised residuals identifies why they are so large.

id	covariate pattern	case-control	dcpct	dneo	dclox	predicted value	standardised residual
84	21	1	83	0	1	0.075	3.569
77	7	1	10	0	0	0.060	4.029

These two covariate patterns each consisted of a single-case herd which had a very low predicted probability of being a case herd (< 8%). This suggests that Nocardia mastitis might have arisen in these herds from some mechanism other than those covered by the predictors in the model, although the possibility of misclassification bias (*ie* false positive cases) cannot be ruled out.

Fig. 16.9 Index plot of standardised residuals



From the model fit in Exercise 16.8, an index plot of standardised residuals with the covariate pattern identification number used as the plotting symbol identifies two outliers. There were no covariate patterns with particularly large negative residuals.

(continued on next page)

Example 16.14 (*continued*)

If the model is refit with those two points omitted, the resulting model is:

Number of obs = 106
LR chi2 (5) = 51.24
Prob > chi2 = 0.000
Log likelihood = -47.836

Predictor	Coef	SE	Z	P	95% CI	
dcpct	.027	.008	3.44	0.001	.012	.043
dneo	2.871	.702	4.09	0.000	1.495	4.247
dclox	-1.642	.600	-2.73	0.006	-2.819	-.465
constant	-3.956	.943	-4.20	0.000	-5.805	-2.108

The effect of removing the two outliers is that the coefficients for all three predictors have moved away from the null (*ie* either larger positive or negative values). This suggests that the model based on the full dataset might provide slightly conservative estimates of the effects of these three factors (but the full model should be used).

model. It is a measure of the distance between the observed set of regression coefficients and a similar set that would be obtained if the observations in the covariate pattern of interest were omitted when building the model. Alternatively, separate sets of delta-betas could be determined for each predictor variable to measure the effect of the covariate pattern on each coefficient in the model.

Values of delta-beta will depend on the leverage that the covariate pattern has, the predicted value, whether or not the model fits the data point well (*ie* is it an outlier?) and also on the number of observations in the covariate pattern. Covariate patterns with a large number of observations will naturally tend to have a large influence on the model, so we want to identify covariate patterns with a large influence but a small m_j, for further investigation.

If a particular pattern has a large delta-beta, it is important to determine why that is. As noted in our example (16.15), when m_j is large, that covariate pattern will likely have a big impact on the model. This is as it should be and need not concern us. However, if it is a covariate pattern with relatively few observations, then it is important to verify that the data are correct and determine if there is a logical explanation for the influence it is exerting.

Other parameters

Two other parameters which measure the overall influence of a covariate pattern on the model are the delta-χ^2 and the delta-deviance. The delta-χ^2 provides an overall estimate of the effect of the j^{th} covariate on the Pearson χ^2 statistic. The delta-deviance provides an overall estimate of the effect of the j^{th} covariate on the deviance χ^2. These

Example 16.15 Identifying influential observations
data=Nocardia

Based on the model fit in Example 16.8, the covariate patterns with the largest leverage values are:

covariate pattern	# of herds	p(D+)	dcpct	dneo	dclox	predicted value	leverage
28	11	0.18	100	0	1	.106	.427
30	9	0.44	100	1	1	.521	.540
27	8	0.13	100	0	0	.328	.547
29	38	0.87	100	1	0	.817	.747

None of the covariate patterns have a particularly large leverage value and nor were either of the outliers a covariate pattern with high leverage. The covariate patterns with the largest overall delta-betas were determined:

covariate pattern	# of herds	p(D+)	dcpct	dneo	dclox	predicted value	delta-beta
5	3	0.00	5	1	0	.341	.552
28	11	0.18	100	0	1	.106	.860
27	8	0.13	100	0	0	.328	3.98
29	38	0.87	100	1	0	.817	7.89

The covariate pattern with the largest delta-beta is pattern #29. This is not surprising since this covariate pattern contained 38 observations (approximately 1/3 of the data). Neither the evaluation of the leverage values nor the delta-betas cause particular concern for this model.

The two observations that were previously identified as outliers are also the covariate patterns with the largest delta-chi-square and delta-deviance values (data not shown).

two measures are overall evaluations of the fit of the model (*ie* they are based on the unexplained variation) so points that are outliers will tend to have large values for the delta-χ^2 and delta-deviance. However, as noted, these observations can only be deleted if you are certain that the data are erroneous.

16.12 Sample size considerations

There are two important issues related to sample size in logistic regression analyses. The first relates to the power of the study to detect effects of interest. For a simple logistic regression model with a single dichotomous predictor, the formula for comparing two proportions in Eq 2.6 will provide a reasonable estimate of the sample

size. For multivariable models, the sample size adjustment shown in Eq 2.10 or 2.11 can be used. The simulation approach described in section 2.10.8 provides a very flexible method of addressing all sample size issues.

The second issue relates to the adequacy of the obtained sample to support the fitting of a logistic model. In addition to considering the total sample size, the number of positive and negative outcomes in the observed data influence the precision of the estimates of the coefficients in the model. If positive outcomes are rare, then variances might be over- or underestimated and hence parameter estimates and test statistics might be affected. It has been suggested that the dataset should contain a minimum of $10(k+1)$ positive outcomes where k is the number of predictors in the model (not counting the intercept) in order to adequately fit the model (Hosmer and Lemeshow, 2000). The same rationale applies if negative outcomes are rare: there should be $10(k+1)$ negative outcomes in the dataset.

16.13 LOGISTIC REGRESSION USING DATA FROM COMPLEX SAMPLE SURVEYS

Data used in logistic regression analyses often come from complex sample surveys in which one or more of the following features might apply to the data.

- The population from which the data are obtained could be divided into **strata** and the sampling might have been done within each stratum.
- The **primary sampling unit** might not have been the unit of observation. For example, a sample of dairy herds might have been chosen from a sampling frame. Subsequently, data might have been collected on all cows (cluster sample) or a sample of cows (multistage sampling) in each herd. Alternative approaches to dealing with this problem of clustering are discussed in Chapters 20-23.
- The probability of each individual in the population being selected for inclusion in the study might have varied (across strata, sampling units *etc*). In this case, each observation should have a **sampling weight** applied to it.

These features and their effects on parameter estimates are discussed in Chapter 2.

Specialised software is required to fit logistic models to this type of data and to take the nature of the sampling process into account. As examples, these are available in the -svy- commands in Stata or in SUDAAN (a specialised SAS-compatible program for the analysis of survey data). Three main points which need to be considered are:

1. Analytic procedures which take into account the sampling strategy should be used to obtain appropriate parameter and variance estimates.
2. Likelihood ratio tests cannot be used to compare nested models because the likelihood function which is used to determine the estimates is only an approximation.
3. Specialised diagnostic procedures for evaluating these logistic models are not yet available so methods applicable to ordinary logistic models have to be used.

See section 6.4 of Hosmer and Lemeshow (2000) for a more complete discussion of these issues.

16.14 CONDITIONAL LOGISTIC REGRESSION FOR MATCHED STUDIES

In our discussions of procedures to control confounding, we discussed the technique of matching. The most common application of this technique is in matched case-control studies in which a case is matched with one or more controls on the basis of some factor such as age, breed, herd of origin *etc*. Because there might be one case and a variable number of controls, this is often referred to as 1-M matching, of which 1-1 matching is a special case.

We could analyse the data using regular logistic regression procedures by simply including dummy variables to represent the j strata, where a case and its control(s) make up a stratum. Unfortunately, the generally desirable properties of maximum likelihood estimation of a logistic regression model only hold if the sample size is large relative to the number of parameters estimated and this wouldn't be true in a matched study with j-1 dummy variables to indicate the strata in addition to the predictors of interest. With matched-pair data (*ie* one case and one control in each matched set), an unconditional logistic regression model including j-1 dummy variables produces estimates of the coefficients of interest that are the square of their true value (*eg* 4 vs 2). This is clearly undesirable.

As we don't really care about the coefficients for the j strata variables, we can use a technique known as conditional logistic regression to analyse matched data. (The conditional likelihood for the j^{th} stratum is simply the probability of the observed data conditional on the number of observations in the stratum and the total number of cases in the study).

While logistic regression models from studies which employed 1-1 matching can be fit using ordinary logistic regression programs, provided the data are appropriately reformatted, a conditional logistic regression model needs to be employed in the case of 1-M matching. **Note** Conditional logistic regression models can also be fit using a Cox proportional hazards model (described in Chapter 19).

When analysing matched data using conditional logistic regression, only predictors that vary within a matched set can be evaluated. Coefficients cannot be estimated for variables that are constant within all matched sets, even if they vary between sets. It is also important to note that conditional models do not estimate an intercept.

If data that were collected in a matched-design study are analysed using an unconditional logistic regression model, one of two effects can occur. If the matching was done on variables that are confounders (*ie* matching was required to prevent bias) then the estimates from the unconditional analysis will be biased towards the null (*ie* a conservative estimate). If the matching was not necessary to avoid bias, then the coefficients from the unconditional analysis will not be biased, but will be less efficient (*ie* will have wider confidence intervals). Consequently, matching should be accounted for in the analysis if it was incorporated into the design of the study (Breslow and Day, 1980).

The evaluation of these models (*ie* regression diagnostics) is not as straightforward

as it is for ordinary logistic models (*eg* the Hosmer-Lemeshow goodness-of-fit test is inappropriate). In the absence of readily available software to compute diagnostic parameters specific for matched-study designs, we recommend that you fit an unconditional model and use the standard regression diagnostics described in section 16.11.5. The reader is referred to Chapter 7 Hosmer and Lemeshow (2000) for a more complete discussion of some of the issues related to conditional logistic regression. Example 16.16 provides an example of the analysis of data from a matched case-control study.

Example 16.16 Conditional logistic regression
data=sal_outbrk

An outbreak of Salmonella in Funen County of Denmark in 1996 was investigated (see Chapter 27 for description of dataset). The data consisted of 39 cases of *Salmonella typhimurium* phage type 12 and 73 controls matched for age, sex and municipality of residence. Data on numerous food exposures were recorded and a small subset of those data are included in the dataset -sal_outbrk-.

Only one food exposure (consuming pork produced by slaughterhouse A) was significantly associated with the outcome. Conditional logistic regression of the outcome on this factor produced an *OR* of 4.42 (95% CI: 1.60-12.19). In comparison, an ordinary logistic regression analysis of these data produced an (incorrect) *OR* of 3.21 (95% CI: 1.42-7.27).

SELECTED REFERENCES/SUGGESTED READING

1. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1: the analysis of case-control studies. Lyon: Intl. Agency for Research on Cancer, 1980.
2. Collett D. Modelling binary data 2d ed. New York: Chapman and Hall, 2002.
3. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 1979 74: 829-836.
4. Hastie T, Tibshirani R. Generalized additive models for medical research. Stat Methods Med Res 1995; 4: 187-196.
5. Hosmer DW, Lemeshow S. Applied logistic regression 2d ed. Toronto: John Wiley and Sons, 2000.
6. Lofstedt J, Dohoo IR, Duizer G. Model to predict septicemia in diarrheic calves. J Vet Int Med 1999; 13: 81-88.
7. Long JS. Regression models for categorical and limited outcomes. Thousand Oaks CA, Sage Publications, 1997
8 Mittlbock M, Schemper M. Explained variation for logistic regression. Stat Med 1996; 15: 1987-1997.
9. Pregibon D. Logistic regression diagnostics. The Annals of Statistics 1981 9: 705-724.

SAMPLE PROBLEMS

The data for the logistic regression exercises come from a retrospective analysis of the medical records from all diarrheic calves which were presented to the Atlantic Veterinary College between 1989 and 1993. The ultimate objective of the study was to develop a logistic model which would predict whether or not the calf was septic at the time of admission (septic calves have a much poorer prognosis than non-septic calves and are not usually worth treating from an economic point of view). The principal investigator in this study was Dr Jeanne Lofstedt (Lofstedt et al, 1999.)

There are 254 observations (records) and 14 variables in the dataset (calf). The original dataset had far more variables (including a lot of laboratory data) but we will restrict ourselves to using a subset of the demographic data and the physical examination data collected. These predictor variables would be readily available to any practitioner and they are virtually complete (*ie* few missing values). A description of the variables in the dataset can be found in Chapter 27.

Exercise 1

1. Familiarise yourself with the data. Look at descriptive statistics to check that all the values look reasonable. How many calves were ultimately diagnosed as septic?
2. Next, look into unconditional associations between the predictor variables and the outcome for sepsis (-sepsis-). Use simple statistics (*t*-tests, χ^2 as appropriate) to do this.
3. Identify all variables with a significant ($P \leq 0.1$) association with sepsis.
4. Build a simple logistic model using only posture and swollen umbilicus as predictors. Remember, that posture is not a dichotomous variable so you will have to convert it to a series of dummy variables.
5. Based on the model in 4. explain the relationship between a swollen umbilicus and the risk of being septic?
   a. How does the predicted probability of sepsis change as posture changes from standing to sternal to lateral?
   b. What is the predicted probability of sepsis in calf #1294?

Exercise 2

We want to build a logistic model for -sepsis- using, as a starting point, the following predictors which were found to have significant ($P \leq 0.1$) association with sepsis.

Categorical variables	Continuous variables
attd	age
eye	dehy
jnts	resp
post	temp
umb	

1. First, consider what type of causal structure might exist among the predictors and with the outcome.
2. One of the assumptions in a logistic regression model is that the relationship between the log odds of the outcome and a continuous predictor variable is linear. Evaluate this assumption for each of the continuous predictor variables using the following two approaches:
   a. Categorise the predictor and 'eyeball' the coefficients in a logistic regression model including dummy variables for each level, and
   b. Categorise the predictor and plot the log odds of disease against the predictor.

   Create quadratic or ordinal variables from continuous variables which do not have an approximate linear relationship with the log odds of sepsis.
3. Build a logistic model to predict sepsis using $P \leq 0.05$ as the criterion for statistical significance when considering terms to keep in the model. Approach this in two ways.
   a. First build one 'manually' by looking at various combinations of terms to include in the model. Use likelihood ratio tests to evaluate the significance of groups of terms (*eg* categorical variables). Also, subjectively assess the impact of term addition/removal on the coefficients (and SEs) of other terms in the model.
   b. Once you have settled on what you feel is a reasonable model, try using a stepwise selection procedure to build the model. Do you get the same result?
4. Using a model which includes only -age- and -age-squared-, -posture-, and -umbilicus-.
   a. Investigate the association between swollen umbilicus and posture further by seeing if there is evidence of confounding or interaction between those two variables.
   b. Is -age- a confounder for -umbilicus- or -posture-?

Exercise 3

Evaluate the model specified in Exercise 2, 4. (*ie* including -age-, -age-squared-, -posture- and -umbilicus-).

Specifically:

1. Assess the fit of the model based on the Hosmer-Lemeshow goodness-of-fit test and the estimation of the sensitivity and specificity of the model.
2. Examine residuals, leverages, and delta-betas. Are there any individual calves that have an unduly large influence on the model?
3. How well does the model predict sepsis? Evaluate an ROC curve for the model.
4. What would be an appropriate value to use as a cutpoint or threshold for the model if we wanted to predict sepsis? What factors should you consider in making that choice?

17

MODELLING MULTINOMIAL DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Select an appropriate model from the following based upon the objectives of your study and the nature of your data
   - multinomial logistic model
   - adjacent-category model
   - continuation-ratio model
   - proportional-odds model.

2. Fit all of the models listed above.

3. Evaluate the assumptions on which the models are based and use one or more tests to compare different models.

4. Interpret *OR* estimates from each of the models.

5. Compute predicted probabilities from each of the models.

17.1 INTRODUCTION

In some studies, the outcome of interest might be categorical but have more than two categories (*ie* multinomial). These data could be recorded on either a nominal or ordinal scale. Nominal data arise when the outcome categories have no specific ordering (*eg* reason for culling might be classified as due to low production, reproduction, mastitis or other). Ordinal data arise when the outcome categories have a distinct order to them (*eg* severity of disease might be classified as absent, mild, moderate or severe).

Nominal data can be analysed using log-linear models or multinomial logistic regression models. Log-linear models can simultaneously evaluate the effects of multiple predictors on multiple outcomes but are limited to the evaluation of categorical variables (predictors and outcomes). Log-linear models are used less frequently than regression-type models in veterinary epidemiology so they will not be discussed further.

An overview of a variety of regression models applicable to nominal and ordinal data is presented in section 17.2. Each of the four models introduced in this section is described in more detail in sections 17.3 to 17.7. All of the examples used in this chapter are based on data derived from a study designed to determine if ultrasound evaluation of beef cattle at the start of the feeding (fattening) period could be used to predict whether the carcass from the animal would eventually be graded as 1=AAA (highest grade), 2=AA, or 3=A (lowest grade in terms of price) (Keefe et al, 2003). This classification is based on the amount of 'marbling' (intramuscular fat in the loin region) present in the carcass after slaughter with grade AAA selling for the highest price. The dataset (beef_ultra) is described more fully in Chapter 27, but the main variables used in this chapter are shown in Table 17.1.

Table 17.1 Variable from beef ultrasound dataset

farm	farm id
id	animal id
grade	carcass grade 1=AAA 2=AA 3=A
sex	0=heifer (female)
	1=steer (castrated male)
backfat	backfat thickness (mm)
ribeye	area of ribeye muscle (sq cm)
imfat	intramuscular fat score (%)
carc_wt	carcass weight (kg)

17.2 OVERVIEW OF MODELS

An overview of the four models to be discussed in this chapter is presented here. In each case we will assume that the outcome has J categories with j being used to designate the

categories from 2 to J (*ie* j=2,...,J). For the sake of simplicity, we will assume that there is a single dichotomous predictor in the model, but these models can easily be extended to multiple predictors. A simple example, based on the data in Table 17.2, will be used to demonstrate most of the models.

Table 17.2 Cross-tabulation of grade and sex from the dataset beef_ultra

Category	Grade	Steer	Female	Totals
1	AAA	100	64	164
2	AA	185	92	277
3	A	29	17	46
		314	173	487

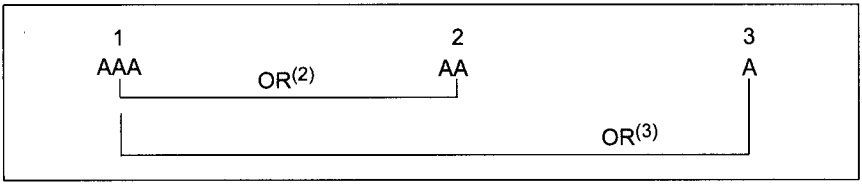
17.2.1 Multinomial logistic model

Nominal data can be analysed using a **multinomial logistic model** which relates the probability of being in category j to the probability of being in a baseline category (which we will refer to as category 1). The model can be written as follows.

$$\ln \frac{p(Y=j)}{p(Y=1)} = \beta_0^{(j)} + \beta_1^{(j)} X \qquad \textit{Eq 17.1}$$

A complete set of coefficients (β_0 and β_1) are estimated for each of the J-1 levels being compared with the baseline (these are designated as $\beta^{(j)}$). Graphically, the effect of the predictor can be seen in Fig. 17.1.

Fig. 17.1 Multinomial logistic model



Based on the data in Table 17.2, the odds ratio (OR) associated with being a steer for category 2 (AA) (compared with category 1) is:

$$OR^{(2)} = \frac{64*185}{100*92} = 1.29$$

Similarly, the OR for category 3 (A) compared with category 1 (AAA) is:

$$OR^{(3)} = \frac{64*29}{100*17} = 1.09$$

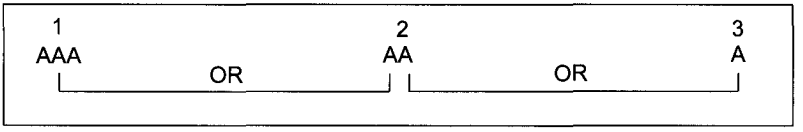
17.2.2 Adjacent-category model

If the categories are ordered, and in some sense 'equidistant', then a constrained multinomial model, or **adjacent-category model** can be fit to the data. This model is based on the assumption that the predictor increases (or decreases) the log odds of a category occurring by a fixed amount as you go up through the categories. Consequently, the model can be written as follows.

$$\ln\frac{p(Y=j)}{p(Y=1)} = \beta_0^{(j)} + (J-1)\beta_1 X \qquad \textit{Eq 17.2}$$

Fitting this model requires that J-1 intercepts (β_0) be estimated, but only a single β_1. Graphically, the effects of the predictor can be seen in Fig. 17.2.

Fig. 17.2 Adjacent-category model



The estimate of β cannot be derived easily from the data in Table 17.2, but the *OR* for AA vs AAA is 1.13 while that for A vs AAA is $(1.13)^2$=1.28.

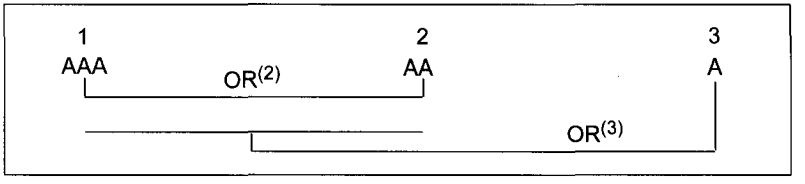
17.2.3 Continuation-ratio model

An alternative for analysing ordinal data is to use a continuation-ratio model which relates the probability of being in a category to the probability of being in any lower category. The model can be written as follows.

$$\ln\frac{p(Y=j)}{p(Y<j)} = \beta_0^{(j)} + \beta_1^{(j)} X \qquad \textit{Eq 17.3}$$

A complete set of coefficients (β_0 and β_1) is estimated for each of the J-1 categories above the baseline. Graphically, the effect of the predictor can be seen in Fig. 17.3.

Fig. 17.3 Continuation-ratio model



Based on the data in Table 17.2, the *OR* associated with being a steer for category 2 (compared with category 1) is:

$$OR^{(2)} = \frac{64*185}{100*92} = 1.29$$

while the OR associated with being a steer for category 3 (compared with being <3) (*ie* A vs AA or AAA) is:

$$OR^{(3)} = \frac{(64+92)*29}{17*(100+185)} = 0.93$$

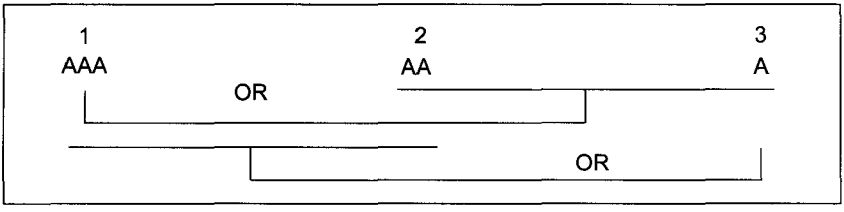
17.2.4 Proportional odds model

A third approach for analysing ordinal data is to use a proportional odds model which relates the probability of being at or above a category to the probability of being in any lower category. The model assumes that this relationship is the same at each of the categories. The model can be written as follows.

$$\ln \frac{p(Y \geq j)}{p(Y < j)} = \beta_0^{(j)} + \beta_1 X \qquad \textit{Eq 17.4}$$

Fitting this model requires that J-1 intercepts (β_0) be estimated, but only a single β_1. Graphically, the effects of the predictor can be seen in Fig. 17.4.

Fig. 17.4 Proportional odds model



Based on the data in Table 17.2, the *OR* associated with being a steer (castrated male) for category 2 or 3 (compared with category 1) is:

$$OR^{(2)} = \frac{64*(185+29)}{(92+17)*100} = 1.26$$

while the *OR* associated with being a steer for category 3 (compared with being <3) (*ie* A vs AA or AAA) is:

$$OR^{(3)} = \frac{(64+92)*29}{17*(100+185)} = 0.93$$

Because the two *OR*s are not the same (or close), the assumption of proportional odds seems to have been violated, so this might not be an appropriate model.

17.3 Multinomial logistic regression

In multinomial logistic regression of an outcome that has J categories, the probability of membership in each of the outcome categories is computed by simultaneously fitting *J-1* separate logistic models (with one category serving as the baseline or reference category). Consequently, for a dependent variable with four levels (leaving the first

level as the baseline category), we estimate three sets of coefficients ($\beta^{(2)}$, $\beta^{(3)}$, $\beta^{(4)}$) corresponding to the remaining outcome categories. Because $\beta^{(1)}=0$, the predicted probability that an observation is in category 1 is:

$$p(Y=1)=\frac{1}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}+e^{X\beta^{(4)}}} \qquad \textit{Eq 17.5}$$

while the probability of being in category 2 is:

$$p(Y=2)=\frac{e^{X\beta^{(2)}}}{1+e^{X\beta^{(2)}}+e^{X\beta^{(3)}}+e^{X\beta^{(4)}}} \qquad \textit{Eq 17.6}$$

and similarly for categories 3 and 4.

17.3.1 Odds ratios

For any given predictor (*eg* sex=steer), there is a separate estimate of the effect of that predictor on each outcome (relative to the base level). Exponentiation of the coefficients from a multinomial regression model produces odds ratios as a measure of effect. **Note** Strictly speaking, these effect measures are not odds ratios. They are actually the ratio of two relative risks (or risk ratios) with each relative risk describing the probability of the outcome in the category of interest relative to the baseline category. Consequently, it would be more appropriate to refer to them as relative risk ratios and some computer programs do so. However, the term odds ratio is more commonly applied and will be used in this chapter.

Example 17.1 shows a very simple model for carcass classification with -sex- as the single predictor. The odds ratios are exactly the same as were found in section 17.2. They indicate that a steer was 1.29 times as likely to grade AA (compared with AAA) as a heifer was. Similarly, a steer was 1.09 times as likely as a heifer to grade A.

Both of the *OR*s in Example 17.1 suggest that being a steer increases the risk of a lower carcass grade. However, the effect is not statistically significant (see section 17.3.3 for assessment of significance).

As with ordinary logistic regression, multinomial logistic regression can be extended to model the effects of multiple predictors that might be categorical or continuous in nature. Example 17.2 shows a model for carcass grade including several predictors with results presented as coefficients.

17.3.2 Interpretation of estimates

Estimates (coefficients or *OR*s) from multinomial logistic regression models are interpreted in a manner similar to those from ordinary logistic regression. The *OR* for the predictor -imfat- in Example 17.2 suggests that, for a unit increase in the intramuscular fat reading from the ultrasound examination, the odds of being graded AA goes down by a factor of $e^{-0.488}=0.61$ (39% reduction) while the odds of being graded A goes down by a factor of $e^{-1.114}=0.33$ (67% reduction). In Example 17.2, all of the predictors have more pronounced effects on the A vs AAA comparison compared

Example 17.1 Simple multinomial logistic regression
data=beef_ultra

A simple multinomial logistic regression of carcass grade (3 levels) was carried out with -sex- as the sole predictor. Carcass grade AAA was the baseline (referent) level.

The first table presents the results in terms of coefficients of the logistic models.

Number of obs = 487
LR chi2 (2) = 1.56
Prob > chi2 = 0.459
Log likelihood = -442.56

		Coef	SE	Z	P	95% CI	
AA							
	sex=steer	0.252	0.205	1.23	0.218	-0.149	0.653
	constant	0.363	0.163	2.23	0.026	0.044	0.682
A							
	sex=steer	0.088	0.345	0.25	0.799	-0.588	0.764
	constant	-1.326	0.273	-4.86	<0.001	-1.860	-0.791

Being a steer compared with being a female, increased the logit of the probability of grading AA or A by 0.25 and 0.09 units, respectively.

The second table presents the results in terms of odds ratios.

		OR	SE	95% CI	
AA					
	sex=steer	1.287	0.263	0.862	1.922
A					
	sex=steer	1.092	0.377	0.555	2.146

Steers were 1.29 and 1.09 times as likely to be downgraded to AA or A compared with females.

with the A vs AA comparison. This was expected given the ordinal nature of the data, but nothing in the model guarantees this (it was not true in Example 17.1). This pattern would not be expected if unordered nominal data were being analysed.

17.3.3 Testing significance of predictors

The significance of predictors can be assessed using either a Wald test or a likelihood ratio test (*LRT*). Overall tests of significance can be carried out (for all logistic models fit) as well as tests for coefficients within individual models. Note however, that tests of significance for a predictor within a given logistic model (*eg* for grade=A) will change

Example 17.2 Multiple multinomial logistic regression
data=beef_ultra

Prediction of carcass grade based on the sex and weight of the animal and three ultrasound measurements determined at the start of the feeding period.

Number of obs = 487
LR chi2 (10) = 125.63
Prob > chi2 < 0.001
Log likelihood = -380.522

		Coef	SE	Z	P	95% CI	
AA							
	sex=steer	0.912	0.262	3.49	0.000	0.399	1.425
	backfat	-0.278	0.116	-2.40	0.017	-0.506	-0.051
	ribeye	0.316	0.077	4.07	0.000	0.163	0.468
	imfat	-0.488	0.121	-4.02	0.000	-0.726	-0.250
	carc_wt	-0.022	0.003	-6.54	0.000	-0.028	-0.015
	constant	7.449	1.223	6.09	0.000	5.052	9.846
A							
	sex=steer	1.533	0.449	3.42	0.001	0.653	2.412
	backfat	-0.708	0.251	-2.82	0.005	-1.200	-0.215
	ribeye	0.382	0.147	2.59	0.010	0.093	0.671
	imfat	-1.114	0.233	-4.77	0.000	-1.572	-0.656
	carc_wt	-0.046	0.006	-7.24	0.000	-0.058	-0.033
	constant	15.745	2.123	7.42	0.000	11.584	19.905

Carcass grade=AAA was the baseline (referent) level.

Sex=steer is now highly significant and is a strong risk factor for lower grades. -carc_wt-, -backfat-, -ribeye- and -imfat- are intervening variables between -sex- and -grade- suggesting that the direct effect of -sex- is much stronger than the total effect. See Chapter 13 for a discussion of intervening variables.

if the baseline category is changed. Consequently, overall tests of significance provide a better estimate of the significance of the predictor. In Example 17.1 neither the Wald nor the likelihood ratio test produced a significant result (P=0.46 for both). However, based on the model in Example 17.2, while the Wald likelihood ratio tests for -sex- give slightly different values (χ^2 of 17.4 and 18.3, respectively on 2 df); both were highly significant (P<0.001). Control of other factors (intervening variables), has made sex an important predictor of carcass grade.

17.3.4 Obtaining predicted probabilities

Predicted probabilities of the occurrence of each outcome category can be computed from the multinomial logistic regression (see Eqs 17.5 and 17.6). These will, of course, vary with the values of the predictors for the animal. Table 17.3 shows those values for a small subset of the cattle based on the model shown in Example 17.2.

Table 17.3 Predicted probabilities from a multinomial logistic regression model

			observed				Probability of grade		
id	grade	sex	backfat	ribeye	imfat	carc_wt	AAA	AA	A
1	AA	steer	2.51	8.94	4.46	357.7	.035	.588	.377
2	AA	steer	5.86	11.77	5.24	323.2	.016	.675	.309
3	AAA	steer	3.14	9.68	3.50	360.0	.055	.676	.269
4	AA	female	2.47	7.46	5.18	307.3	.027	.446	.526
5	AAA	steer	1.85	8.03	4.89	354.5	.032	.532	.435

The sum of the probabilities for each animal equals 1.

17.3.5 Regression diagnostics

Specialised diagnostics for multinomial logistic regression are not as readily available as they are for ordinary logistic regression. One approach is to fit ordinary logistic models for pairs of comparisons (*eg* grade=A vs AAA and AA vs AAA) and evaluate the regression diagnostics for those models.

17.4 Modelling ordinal data

Ordinal data can arise in a variety of ways. For example, an observed continuous variable might be divided into categories. Alternatively, levels of an ordinal variable could represent categories of an unobserved (but hypothesised) continuous variable (*eg* opinions ranging from strongly agree to strongly disagree, or disease severity ranging from absent to severe). Finally, categories might represent total values of a composite variable made up of a series of scored variables (*eg* a hygiene score that represents the sum of scores from several questions about hygiene in a barn).

While the multinomial models described above can also be used to analyse ordinal data, they ignore the fact that the categories fall in a logical, ordered sequence. There are a number of different ways of fitting ordinal models. We will consider three of them: adjacent-category models (section 17.5), continuation-ratio models (section 17.6) and proportional-odds models (section 17.7).

17.5 Adjacent-category model

In an adjacent-category logistic regression model, each coefficient measures the effect of a factor on the logit of the probability of being in a specified level compared with the probability of being in the level below. For any given predictor, this results in the estimation of a single effect that expresses how the predictor influences the log odds of the outcome moving up to the next (adjacent) category. This model is also known as a constrained multinomial model because it is estimated as a multinomial model with the constraint that the coefficient for categories n levels apart be n times the coefficient for adjacent categories. (Alternatively, the *OR* for categories n levels apart will be the *OR* for adjacent levels raised to the power n.) This model is based on the assumption that, as you go from one level to the next, the *OR* is constant. A graphic representation is shown in Fig. 17.2.

Example 17.3 presents an adjacent-category model based on the multinomial model fit in Example 17.2. A likelihood ratio test can be used to compare this 'constrained multinomial model' with the usual multinomial model. If the test is significant, it suggests that the multinomial model is superior. The *LRT* for the model in Example 17.3 had a χ^2 of 7.73 with 5 df (because five fewer coefficients were estimated) with a P-value of 0.19, suggesting that there is little evidence that the unconstrained model fits the data better than the adjacent-category model. In this case, for the sake of simplicity, the adjacent-category model is preferable.

17.6 Continuation-ratio model

In continuation-ratio models, the log *OR* measures the effect of a factor on the odds of being in a specified level compared with the odds of being in any of the lower levels. This type of model is useful in situations where the dependent variable represents the number of attempts required to achieve an outcome (*eg* number of breedings required for conception in dairy cows). A graphic representation is shown in Fig. 17.3.

This model can be fit as a series of simple logistic models in which the dependent variable (Y) is recoded to be 1 for the level of interest, 0 for all lower levels and missing for all higher levels. For example, a continuation-ratio model evaluating the effects of predictors on the probability of conception for up to four breedings would require three separate logistic regressions. The data would be recoded as shown in Table 17.4.

Table 17.4 Coding of data for a continuation-ratio model of the effect of predictors on the number of services required for conception

	Breeding			
	1	2	3	4
Y1	0	1	missing	missing
Y2	0	0	1	missing
Y3	0	0	0	1

Example 17.3 Adjacent-category model
data=beef_ultra

An adjacent-category model was fit using the same predictors presented in Example 17.2. The constraint that coefficients for categories two levels apart be twice those of the adjacent categories reduces the number of parameters which need to be estimated.

Number of obs = 487
LR chi2 (5) = 117.90
Prob > chi2 < 0.001
Log likelihood = -384.385

		Coef	SE	Z	P	95% CI	
AA							
	sex=steer	.811	.206	3.94	0.000	.407	1.214
	backfat	-.305	.095	-3.21	0.001	-.491	-.118
	ribeye	.237	.060	3.94	0.000	.119	.356
	imfat	-.520	.099	-5.25	0.000	-.714	-.323
	carc_wt	-.023	.003	-8.22	0.000	-.028	-.017
	constant	8.713	1.010	8.63	0.000	6.733	10.692
A							
	sex=steer	1.621	.412	3.94	0.000	.814	2.428
	backfat	-.610	.190	-3.21	0.001	-.983	-.237
	ribeye	.475	.121	3.94	0.000	.238	.711
	imfat	-1.039	.198	-5.25	0.000	-1.427	-.651
	carc_wt	-.046	.006	-8.22	0.000	-.057	-.035
	constant	14.393	1.906	7.55	0.000	10.658	18.129

Outcome grade=AAA is the comparison group.

Note The coefficient for each predictor for grade=A is twice that for grade=AA because it is two categories away from AAA. For example, being a steer increases the log odds of being graded A by 1.62 units but the log odds of being graded AA by only 0.81 units.

In this example, the coefficient for a predictor represents the effect of the factor on the log odds of conceiving on the j^{th} breeding, conditional on not conceiving on any previous breedings.

The model contains the same number of parameters as the multinomial model presented in section 17.3. Consequently, the model is no more 'parsimonious', but it results in estimates of the *OR* which have different interpretations than those from a multinomial logistic regression model. A constrained continuation-ratio model can be fit with the *OR* for each predictor constrained to be equal for each increment in the outcome. A likelihood ratio test, comparing the constrained and unconstrained models, can be used to evaluate the assumption of equal *OR*.

The *OR* from the separate logistic models for the beef ultrasound data are not presented because it does not make biological sense to fit these data with a continuation-ratio model (*ie* movements between grades are not sequential events).

17.7 Proportional odds model (constrained cumulative logit model)

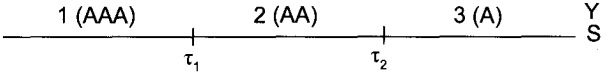
This is the most commonly encountered type of ordinal logistic model. In a proportional odds model, the coefficients measure the effect of a predictor on the log odds of being above a specified level compared with the log odds of being at or below the specified level. It is based on the assumption that the coefficients do not depend upon the outcome level, so only a single coefficient for each predictor is estimated. A graphic representation of this model is presented in Fig. 17.4.

A proportional-odds model assumes that the ordinal outcome variable represents categories of an underlying continuous latent (unobserved) variable. Assume that the value of the underlying latent variable (or 'score') (S_i) is a linear combination of predictor variables.

$$S_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \textit{Eq 17.7}$$

where ε_i is a random error term from a continuous distribution.

The latent variable (S) is divided by cutpoints (τ_j) so that the i^{th} individual is classified as category 1 (AAA) if $S_i \leq \tau_1$ and is classified as category 2 (AA) if $\tau_1 < S_i \leq \tau_2$, and so on.



The probability of observing outcome j in the i^{th} individual is:

$$\text{p(outcome}_i = j) = \text{p}(\tau_{j-1} < S_i < \tau_j) \qquad \textit{Eq 17.8}$$

If the random term (ε_i) is assumed to have a logistic distribution (with a mean of 0 and a variance of $\pi^2/3$), then

$$\text{p}(S_i < \tau_j) = \frac{1}{1 + e^{S_i - \tau_j}} \qquad \textit{Eq 17.9}$$

Note Assuming the latent variable has a normal distribution gives rise to an ordinal probit model, but those are not discussed in this chapter.

The model fit by assuming a logistically distributed latent variable can also be written as (presented with a single predictor X for simplicity):

$$\text{logit}\,(\text{p}(Y \leq j)) = \beta_{0j} + \beta X \qquad \textit{Eq 17.10}$$

where the β_{0j} are intercepts and β is the effect (slope) of the predictor. Thus, the model is one in which the log odds of the outcome can be viewed as being represented by a series of parallel lines with different intercepts.

Example 17.4 presents a proportional-odds model for the carcass grade data.

Example 17.4 Proportional-odds model
data=beef_ultra

A proportional-odds model was fit to the carcass grade data with the same predictors as used in Example 17.2 and 17.3.

Number of obs = 487
LR chi2 (5) = 115.97
Prob > chi2 < 0.001
Log likelihood = -385.349

	Coef	SE	Z	P	95% CI	
sex=steer	0.919	0.229	4.02	0.000	0.471	1.368
backfat	-0.326	0.105	-3.09	0.002	-0.533	-0.120
ribeye	0.274	0.067	4.08	0.000	0.142	0.405
imfat	-0.569	0.108	-5.26	0.000	-0.781	-0.357
carc_wt	-0.026	0.003	-8.56	0.000	-0.032	-0.020
cutpoint 1	-9.678	1.078				
cutpoint 2	-6.151	1.004				

The odds ratio associated with being a steer, compared with being a female is:

$$e^{0.919}=2.51$$

This suggests that being a steer increases the odds of being at or above any given carcass grade compared with being below that grade by 2.51 times. (Remember that the data are coded so that A is grade 3 – *ie* greatest economic loss). As such it measures the overall increased chance of a poor (higher) grade that is associated with being a steer.

17.7.1 Predicted probabilities

The first observation in the dataset is a steer (sex=1) with a backfat=2.51, a ribeye=8.94, an imfat=4.46 and a carc_wt=357.7. For this animal, the latent variable (S_i) is:

$$S_i=-9.202$$

Consequently, the probability of this animal being in category 1 (AAA) (from Eq 17.9) is:

$$p(Y=1)=\frac{1}{1+e^{-9.202-(-9.678)}}=0.38$$

Similarly, the probability of this animal being graded AA is 0.57 and A is 0.05.

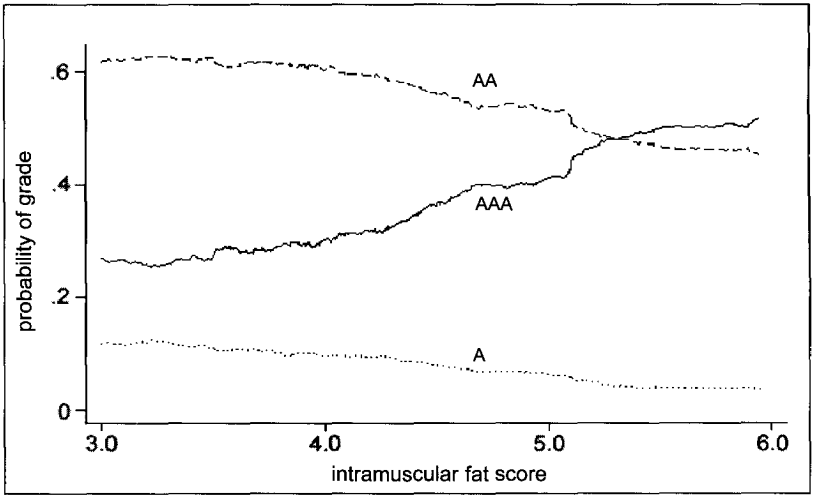
The probabilities of each grade outcome for the first five animals from this dataset (and the values of the predictor variables for those animals) are shown in Table 17.5.

Table 17.5 Values of predictor variables, latent variables (S_i) and predicted probabilities of each of the carcass grades from the proportional-odds model

							Probability of grade		
id	sex	backfat	ribeye	imfat	carc_wt	S	AAA	AA	A
1	steer	2.51	8.94	4.46	357.7	-9.202	.38	.57	.05
2	steer	5.86	11.77	5.24	323.2	-9.076	.35	.60	.05
3	steer	3.14	9.68	3.50	360.0	-8.718	.28	.65	.07
4	female	2.47	7.46	5.18	307.3	-9.625	.49	.48	.03
5	steer	1.85	8.03	4.89	354.5	-9.398	.43	.53	.04

The effect of a single predictor (-imfat-) on the predicted probability can best be viewed by generating smoothed curves of the probability of each grade against -imfat-. Fig. 17.5 shows a graph of kernel smoothed mean probabilities (smoothed with a bandwidth of 30%) of each grade against the intramuscular fat level (over the range of 3.0 to 6.0 – the range in which most -imfat- values fall). **Note** As the probability of each outcome depends on the value of all predictors in the model, the smoothed curves shown in Fig. 17.5 represent average probabilities of the grade as -imfat- changes.

Fig. 17.5 Smoothed mean probabilities of grades



As can be seen, the probability of a carcass being graded AA or A goes down as the intramuscular fat level at the start of the feeding period goes up. On the other hand, the probability of a grade of AAA goes up substantially.

17.7.2 Proportional-odds assumption

A rough assessment of the assumption of proportional odds can be obtained by comparing the log likelihood of the ordered logit model (L_1) with one obtained from the multinomial logit model (L_0) using a likelihood ratio test. If there are k predictors (not counting the intercept) and J categories of outcome, the multinomial model will fit $(k+1)(J-1)$ parameters, while the proportional-odds model will fit $k+(J-1)$ so the difference in degrees of freedom is $k(J-2)$. Consequently, $-2(L_1-L_0)$ should have an approximate χ^2 distribution with $k(J-2)$ degrees of freedom. **Note** This is only an approximate test because the proportional-odds model is not nested within the multinomial model. However, it gives a rough assessment of the assumption of the proportional-odds assumption.

In our example, the log likelihoods of the two models are -380.5 and -385.3, respectively so the *LRT* is:

$$LRT=-2(-380.5-[-385.3])=9.6$$

The χ^2 statistic has $k(J-2)=5$ df which yields a P-value of 0.09. Consequently, this provides marginal evidence that the proportional-odds assumption does not hold.

An alternative approximate *LRT* based on fitting J-1 separate binary models has been developed (Wolfe and Gould, 1998). The models are fit first assuming the βs are constant across all models (proportional-odds assumption) and the sum of these log likelihoods are compared with the sum of those obtained by fitting the models without the assumption of constant βs. For the beef ultrasound model, this test produces a χ^2 value of 7.47 (P=0.19).

The likelihood ratio tests described above are omnibus tests which evaluate the assumption of proportional odds over all predictors. A Wald test which will provide an overall assessment as well as an evaluation of the assumption for each predictor separately is available (Brant, 1990). The results of this test for the model fit in Example 17.4 are presented in Table 17.6.

Table 17.6 Brant (Wald) test of proportional-odds assumption

Variable	χ^2	P	df
all	7.81	0.167	5
sex	0.08	0.777	1
backfat	0.58	0.448	1
ribeye	1.53	0.216	1
imfat	0.77	0.380	1
carc_wt	0.64	0.425	1

The P-value of the overall Wald test is comparable to the second approximate likelihood ratio test described above. None of the individual predictors have significant test results suggesting that the proportional-odds assumption is valid. Other tests of

the proportional-odds assumption are available but there are no clear guidelines for choosing one test over another. In general, if any of the tests discussed above yields a significant result, the assumption should be investigated further.

SELECTED REFERENCES/SUGGESTED READING

1. Agresti A. Categorical data analysis. New York: John Wiley and Sons, 1990.
2. Agresti A. Modelling ordered categorical data: recent advances and future challenges. Stat Med 1999; 18: 2191-2207.
3. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol 1997; 26: 1323-1333.
4. Brant R. Assessing proportionality in the proportional-odds model for ordinary logistic regression. Biometrics 1990; 46: 1171-1178.
5. Greenland S. Alternative models for ordinal logistic regression. Stat Med 1994; 13: 1665-1677.
6. Hosmer DW, Lemeshow S. Applied logistic regression, Chapter 8. 2d ed. Toronto: John Wiley and Sons, 2000.
7. Keefe GP, Dohoo IR, Valcour J, Milton RL. Assessment of ultrasonic imaging of marbling at entry into the feedlot as a predictor of carcass traits at slaughter. J Anim Sci 2003; submitted.
8. Long JS. Regression models for categorical and limited dependent variables, chapters 5 and 6. London: Sage Publications, 1997.
9. Long JS, Freese J. Regression models for categorical dependent variables using Stata, chapters 5 and 6. College Station: Stata Press, 2001.
10. Wolfe R, Gould W. An approximate likelihood-ratio test for ordinal response models. Stata Tech Bull 1998; 42: 24-27.

SAMPLE PROBLEMS

The data for this exercise are included in the dataset 'colostrum'. They represent a small subset of the data collected during a study looking at the effects on calf health of bacterial levels in colostrum. It had been postulated that high bacterial loads in colostrum might adversely affect calf health either directly, or by reducing immunoglobulin absorption by the calf. The variables in this dataset are:

Variable	Description
health	a 3-level outcome variable describing the health of the calf over the first four weeks of life
	0=completely healthy
	2=mild illness
	3=serious illness (including a few calves that died)
herd	the 6 herds in the study have been divided into 2 groups
	0=the single large herd
	1=a collection of the five smaller herds
qty	quantity of colostrum fed (litres)
log_tot	natural logarithm of the total bacterial load (bact/ml) in the colostrum (sample taken just before first feeding)

1. Fit multinomial, adjacent-category and proportional-odds models for -health- with -herd-, -qty- and -log_tot- as predictors.
2. Is there evidence that the multinomial model fits the data better than the adjacent-category model?
3. Evaluate the assumption of 'proportional odds' for the overall proportional-odds model using both a likelihood ratio test and a Wald test. Use the Wald test to determine if any individual predictors violate the assumption.
4. Test the overall significance of the variables -herd- and -qty-. Although the variable -qty- is not statistically significant, examine what effects removing it has on the apparent effect of -log_tot-.
5. What is your overall conclusion about the effect of bacterial load in colostrum on calf health?

18

MODELLING COUNT AND RATE DATA

Objectives

After reading this chapter, you should be able to:

1. Understand the relationship between counts of disease events and incidence rates.

2. Fit, evaluate and interpret Poisson regression models.

3. Be able to determine when a negative binomial model is likely to be more appropriate than a Poisson model and to quantify and statistically assess overdispersion.

4. Fit, evaluate and interpret negative binomial regression models.

5. Decide when a zero-inflated model might be more appropriate than a Poisson or negative binomial model.

6. Fit a zero-inflated model and interpret the model and the Vuong statistic.

18.1 INTRODUCTION

In previous chapters, we have looked at methods of analysing data measured on a continuous scale (Chapter 14) and two types of discrete data: binary/binomial data (Chapter 16) and multinomial data (Chapter 17). Here, we are introduced to the situation in which the outcome we are measuring represents a count of the number of times an event occurs in an individual or group of individuals.

a. It might be a simple count of events, such as the number of breedings required for a dairy cow to conceive. A recently published manuscript used Poisson regression to evaluate the effect of peripartum treatment with an anthelmintic on the number of services per conception in dairy cows (Sanchez et al, 2002).
b. It might be a count of cases of disease over a period of time with the amount of animal-time at risk having to be taken into consideration (*eg* total number of cases of clinical mastitis in a dairy herd over a year with the number of cow-months contributed by lactating cows as the amount of animal-time). Hence, this is a measure of the incidence rate (I) of disease. The examples used in this chapter will focus on this kind of data: the incidence rate of clinical mastitis in a hypothetical trial of 'pre-dipping' in a dairy herd and the incidence rate of new *Mycobacterium bovis* infections in cattle and cervid herds after the introduction of the agent to the herd.
c. It might be a count of cases of disease with the size of the population at risk being taken into consideration (*eg* cases of lymphosarcoma in slaughtered cattle seen at various abattoirs with the number of cattle slaughtered as the population at risk). This is an estimate of the (lifetime) incidence risk of lymphosarcoma in cattle.
d. It might be a count of an outcome that is measured over a geographical area. For example, Poisson models are also used to investigate factors related to the number of events per unit area. Hammond et al (2001) investigated whether land use was predictive of the number of badgers in 500 m^2 grids in an area in Ireland. The study area was overlaid by a 500 m^2 grid and the number of badgers caught in each grid was recorded. Land use within each cell of the grid was described by a set of categorical variables. The mean number of badgers per grid was 0.6 and the variance was 1.5. A major finding was that as the area of high-quality pasture within a grid increased, the number of badgers also increased.

18.1.1 Approaches to analysis

We might want to evaluate the effect of 'pre-dipping' (disinfection of teat ends prior to milking) on the incidence of clinical mastitis. We will assume that a controlled trial can be carried out in three large dairy herds and cows will be individually assigned to be pre-dipped or not. The outcome of interest might be the total number of cases of clinical mastitis occurring in each cow over a full lactation. Other factors that will have to be taken into consideration are the age of the cow (it is generally accepted that the risk of clinical mastitis increases with age) and which herd the cow is in (because

incidence rates of mastitis vary among herds). While random assignment of cows to treatment groups should balance the age and herd factors across the study groups, you might still want to consider them in the analysis. Given the clinical trial design, we can assume that the population is closed, but the time at risk will vary among cows. **Note** In this example, we are interested in the total number of cases of mastitis. If we were interested only in first cases, we could create a binary variable for each cow and fit a logistic model.

There are a number of ways to approach the analysis of the data generated by this study.

a. The incidence rate of clinical mastitis could be computed for each study group and the difference between the two groups tested using an unconditional *Z*-test that was discussed in Chapter 6. This approach does not allow for the control of other potential confounding variables (*ie* age of cow and herd), so it would rely totally on random assignment to control for these effects.
b. Alternatively, you could determine the incidence rate of clinical mastitis within each cow and use that value as the dependent variable in a linear regression with pre-dipping as the primary exposure (predictor) of interest and age and herd as extraneous variables. However, most cows would have an *I* of zero, so it is very unlikely that the error terms would have anything close to a normal distribution. Consequently, one of the fundamental assumptions of linear regression would be violated. It is also possible that some combination of predictor variables could be found that predicted a negative *I* for the cow. This approach looks worse than the first.
c. The preferred approach is to use Poisson regression to model the incidence of new cases while adjusting for the amount of time each cow was at risk.

18.2 The Poisson distribution

The Poisson distribution is often used to model counts of 'rare' events:

$$p(Y = y) = \frac{\mu^y e^{-\mu}}{y!} \qquad \textit{Eq 18.1}$$

where y is the observed count of events and μ is the mean number of events. An interesting characteristic of the Poisson distribution is that the mean and the variance are equal (*ie* μ).

The Poisson distribution can be thought of in two ways.

a. If the times between events (*eg* cases of mastitis) are independent and follow an exponential distribution with a mean value of t, then the number of cases of mastitis (Y) in a defined time period (T) will follow a Poisson distribution with $\mu=T/t$. For example, if the mean time between cases of mastitis is 150 days, then the expected number of cases in a 300-day lactation will be 300/150=2 cases. The time between events is sometimes referred to as the 'waiting time'. Using this formulation of the Poisson distribution, there is a natural connection

between the analysis of counts of events (Poisson regression) and the analysis of time to event occurrence (survival analysis - Chapter 19).

b. The Poisson distribution approximates the binomial distribution if the population (n) is large, consists of independent units and the binomial proportion (p) is small (*ie* the two characteristics that would make an event 'rare'). In this case, $\mu=np$. For example, if the probability of the occurrence of mastitis on any given day is 1/150=0.0067, then the expected number of cases in a 300-day lactation will be 300*0.0067=2.

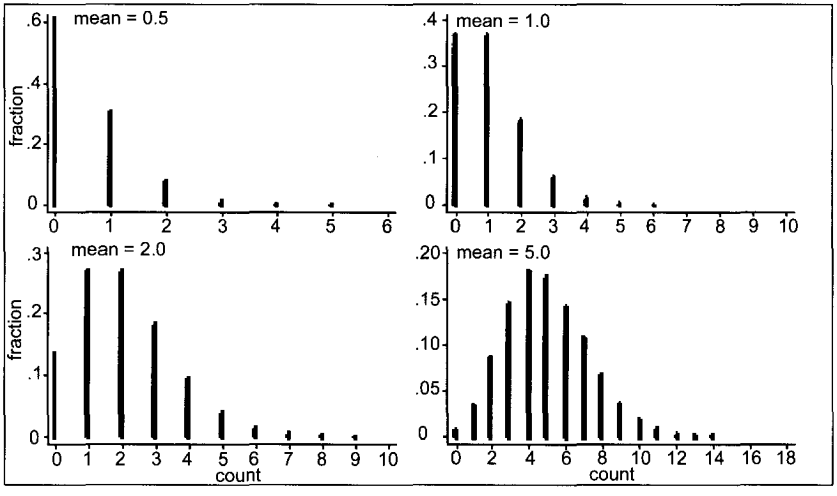
If the outcome follows a Poisson distribution and the mean is known, you can calculate the probability of a specific number of events occurring. For example, if the average number of milk fever cases in a dairy herd is 5 per year, the probability of getting 10 cases in a year is:

$$p(Y=10)=\frac{5^{10}e^{-5}}{10!}=0.018$$

This indicates that there is approximately a 2% chance of having exactly 10 cases in a year (provided the mean for the population is not changing over time).

Poisson distributions with means of 0.5, 1.0, 2.0 and 5.0 are shown in Fig. 18.1. As this figure indicates, as the mean increases, the Poisson distribution approaches a normal distribution.

Fig. 18.1 Poisson distributions



18.3 Poisson regression model

The usual form of the Poisson regression model is:

$$E(Y) = \mu = n\lambda \qquad \textit{Eq 18.2}$$

where $E(Y)$ = the expected number of cases of disease
n = exposure (*eg* animal-time units at risk)
λ = represents a function which defines the disease incidence rate.

The exposure (n) adjusts for different amounts of time at risk (or alternatively different sizes of populations at risk) for the various study subjects (animals or groups of animals). (**Note** Throughout this text, the letter n is most commonly used to denote the number of animals in a population. Here we are also using it to denote the amount of animal-time at risk.) n could be recorded either on the original scale (*ie* the amount of animal-time at risk, referred to as the **exposure**), or on a log-scale (*ie* the log (animal-time at risk), referred to as an **offset**). However, if n is equal for all subjects, it can be omitted but you must remember that predicted counts will refer to the expected number of cases in n animal-time units at risk. For example, in the badger study referred to, each count related to the same 500 m^2 grid size so no offset or exposure was required. However, the predicted counts were counts per 500 m^2.

One of the ways that λ can be related to the predictor(s) is:

$$\lambda = e^{\beta_0+\beta_1 X} \quad \text{or} \quad \ln(\lambda) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.3}$$

Consequently, the Poisson regression model is:

$$E(Y) = ne^{\beta_0+\beta_1 X} \quad \text{or} \quad \ln E(Y) = \ln n + \beta_0 + \beta_1 X$$

$$\text{or} \quad \ln E(I) = \ln \frac{E(Y)}{n} = \beta_0 + \beta_1 X \qquad \textit{Eq 18.4}$$

where $\ln E(I)$ is the log of the expected value of the incidence rate (I) of disease which is being modelled as a linear combination of predictors. **Note** This example assumes that there is a single predictor variable (X), but the model can be extended to include multiple predictors.

As with logistic regression, Poisson regression models are fit using an iterative maximum likelihood estimation procedure. The statistical significance of the contribution of individual predictors (or groups of predictors) to the model can be tested using either Wald tests or likelihood ratio tests. An example of a Poisson regression analysis, based on tuberculosis data from cattle and cervids is presented in Example 18.1.

Example 18.1 Poisson regression model
data=tb_real

The incidence rates of new tuberculosis (TB) infections in cattle, cervid (deer and elk) and bison herds in 9 TB outbreaks in Canada between 1985 and 1994 were estimated (see Chapter 27 for explanation of method of estimation). These incidence rates were modelled as a function of several characteristics of the animals in the herd (type, sex and age). The key predictors were:

type: 1=dairy, 2=beef, 3=cervid, 4=other
sex: 1=female, 2=male
age: 0=0-12 mo, 1=12-24 mo, 2=24+ mo

A more complete description of the dataset is in Chapter 27.

A Poisson regression model with the three predictor variables and the time at risk included as an exposure variable (or offset) produced the following results.

Number of obs = 134
Log likelihood = -238.7

					95% CI for IR	
Variable	Coef	SE	P	IR	Lower	Upper
type=beef	0.442	0.236	0.061	1.56	0.98	2.47
type=cervid	1.066	0.233	<0.001	2.90	1.84	4.59
type=other	0.438	0.615	0.476	1.55	1.46	5.17
sex=male	-0.362	0.195	0.064	0.70	0.47	1.02
age=12-24 mo	2.673	0.722	<0.001	14.49	3.52	59.62
age=24+ mo	2.601	0.714	<0.001	13.48	3.33	54.59
constant	-11.328	0.771	<0.001	NA	NA	NA

Herd type was a significant predictor (P <0.001) with incidence rates in beef and cervid herds higher than in dairy herds although the coefficient for beef herds was only borderline significant. Males appeared to have a lower incidence rate (again borderline significance) and animals over 12 months of age definitely had higher incidence rates.

The deviance and Pearson goodness-of-fit test statistics were:

	df	χ^2	P
Deviance	127	348.4	<0.001
Pearson	127	1105.7	<0.001

These suggest that there are serious problems with the model (*ie* strong evidence of lack of fit).

18.4 Interpretation of coefficients

The coefficients from a Poisson regression model represent the amount the log of I ($\ln I$) is expected to change with a unit change in the predictor. Assuming that there are two exposure groups (X=0 and X=1), then the incidence rate ratio (IR) associated with belonging to group X=1 is:

$$IR_{(X=1)} = \frac{\lambda_1}{\lambda_0} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad \textit{Eq 18.5}$$

so the coefficients from a Poisson regression can easily be converted into IR estimates. In general, the IR represents the proportional increase in I for a unit change in the predictor. For example, if the IR for lactation number in a study of clinical mastitis cases was 1.5, that would suggest that the incidence rate of clinical mastitis went up by 50% for each additional lactation that a cow had (*ie* that it was 1.5 times higher than the rate in the previous lactation). **Note** In general $e^{\beta 1}$ represents the ratio between mean counts in two groups. However, because Poisson regression is most commonly used for incidence rate data in epidemiologic studies, this specific use will be emphasised throughout this chapter.

The effect of a predictor on the absolute number of cases of disease (or other outcome event) depends on the values for other predictors in the model. For example, the IR for type=cervid in Example 18.1 was $e^{1.066}$=2.9 (compared with dairy herds). The predicted I for young (0-12 mo), females in a dairy herd was $e^{-11.328}$=0.12 cases/10,000 animal-days at risk. For cervids, the predicted rate would be 0.12*2.9=0.35 cases/10,000 animal-days, or an extra 0.23 cases per 10,000 animal-days. In animals aged 12-24 mo, the predicted rate for females in dairy herds was $e^{-8.655}$=1.7 cases/10,000 animal-days. For cervids of this age, the rate would be 1.7*2.9=4.9 cases/10,000 animal-days, or an extra 3.2 cases per 10,000 animal-days.

18.5 Evaluating Poisson regression models

18.5.1 Residuals

Raw residuals can be computed for each observation as the observed number of cases (*obs*) minus the expected number of cases (*exp*) predicted from the model. Residuals are computed on the basis of 1 per observation.

Pearson residuals can be computed as:

$$res_p = \frac{\text{obs} - \text{exp}}{\sqrt{\text{var}}} \qquad \textit{Eq 18.6}$$

where var is the estimated variance of the observations. For a Poisson model, the estimated variance is equal to the expected number of cases (μ). For negative binomial models (discussed below), the variance is $\mu+\alpha\mu^2$.

Deviance residuals (formula not shown) can also be computed. The sum of the squared deviance residuals gives the deviance for the model which is defined as minus twice the difference between the log likelihood of the model and the maximum log likelihood achievable. **Anscombe residuals** are very similar to deviance residuals but are adjusted so that they will most closely follow a normal distribution if the Poisson model is appropriate.

18.5.2 Assessing overall fit

As with logistic regression, χ^2 goodness-of-fit tests can be computed as the sum of the squared deviance or Pearson residuals. The resulting test statistic has an approximate χ^2 distribution if there are multiple observations within each covariate pattern defined by the predictors in the model. However, the values of the two test statistics could be quite different and, if either is indicative of a lack of fit, the model should be investigated thoroughly. As with all overall goodness-of-fit statistics, a significant result (indicating lack of fit) provides no information about what the cause of the lack of fit is. However, with Poisson models, a common cause is overdispersion (*ie* the variance of the counts is much larger than the mean).

18.5.3 Evaluating overdispersion

The assumption behind a Poisson model is that the mean and the variance are equal (conditional upon the predictors in the model); that is, the mean and the variance of the number of events are equal after the effects of the predictors in the model have been taken into account. Including predictors in the model does not change the mean number of events, but will reduce the variance. Consequently, one could have a variance greater than the mean in the raw data (*ie* unadjusted estimates), but still meet the assumption of equidispersion. However, as a rule, if the unadjusted variance is greater than twice the unadjusted mean, one must be suspicious that overdispersion will be present.

Having a variance much larger than the mean is a common problem with count data. This is called **extra-Poisson variation** or **overdispersion**. It often arises when the data are clustered (*eg* animals within a herd) and the clustering has not been adequately taken into account.

For example, in the TB data described in Examples 18.1 and 18.2, each herd contributed multiple observations to the dataset. (While most herds had only one type of animal, they would have had multiple age classes and perhaps, males and females). The incidence rate of TB cases is more alike among groups of animals within a herd than across different herds. Consequently, part of the variation between groups of animals is due to the variation between herds, and this has not been taken into account. Thus the model does not fit the data well.

Overdispersion can be dealt with by fitting a model which allows the variance to be larger than the mean. This can be done in one of two ways. (Actually, there are lots of ways it can be modified, but these two are the most common). The simplest modification

Example 18.2 Poisson regression – examining the model
data=tb-real

Based on the model fit in Example 18.1, the observations with the five largest and smallest deviance residuals are:

farm	type	sex	age (mo)	time at risk	reactors observed	reactors predicted	deviance residual	Cook's distance
89	cervid	male	> 24	27410	0	3.261	-3.539	0.307
54	cervid	female	> 24	26182	1	8.588	-3.298	0.174
53	cervid	female	12-24	12420	0	4.379	-2.959	0.097
27	dairy	female	> 24	57176	1	6.457	-2.680	0.311
108	beef	female	> 24	19146	0	3.365	-2.594	0.030
120	cervid	male	12-24	15921	11	3.909	2.929	0.389
25	dairy	female	0-12	389	1	0.003	3.076	0.078
119	cervid	female	> 24	12269	12	4.024	3.205	0.164
45	cervid	female	> 24	21848	29	7.167	6.116	1.371
133	beef	female	> 24	6418	20	1.128	8.790	0.875

Large negative residuals were associated with groups of animals where many cases were expected, but few observed. Large positive residuals were found in groups of animals where many more cases were observed than were expected. Although they only accounted for 38% of the observations in the dataset, groups of cervids accounted for 5 of the 8 most extreme residuals suggesting that the model did not fit as well for cervids as it did for cattle.

The four observations with the largest Cook's distance were as follows.

farm	type	sex	age (mo)	pop. at risk	reactors observed	reactors predicted	deviance residual	Cook's distance
118	cervid	female	12-24	21660	17	7.637	2.912	0.552
92	other	female	> 24	9360	0	1.639	-1.810	0.656
133	beef	female	> 24	6418	20	1.128	8.790	0.875
45	cervid	female	> 24	21848	29	7.167	6.116	1.371

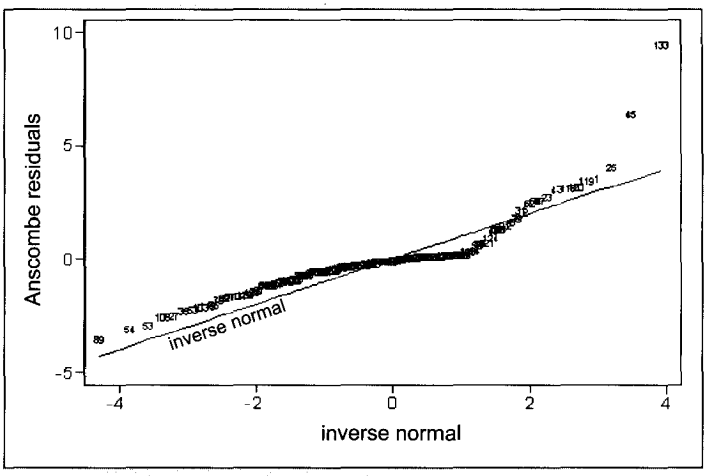
Large Cook's distances were associated with observations that had moderate or large residuals and contributed greatly to the overall time at risk. Observations with small amounts of time at risk, tended not to have a large impact on the model (small Cook's distance) (*eg* observation 25 – in the table of large residuals).

(*continued on next page*)

Example 18.2 *(continued)*
data=tb-real

A normal probability plot of Anscombe residuals shows an approximate normal distribution but highlights the large positive residuals.

Fig. 18.2 Normal probability of Anscombe residuals



is to assume that the variance is a constant multiple of the mean instead of equal to the mean (this allows the variance to be larger than the mean):

$$\text{var} = (1+\alpha)\mu = \mu + \alpha\mu \qquad \textit{Eq 18.7}$$

where α is referred to as the dispersion parameter.

An alternative formulation for the variance is to assume that the variance is a function of the mean:

$$\text{var} = (1+\alpha\mu)\mu = \mu + \alpha\mu^2 \qquad \textit{Eq 18.8}$$

This formulation of the variance gives rise to a negative binomial model which is discussed further in section 18.6.

Note In either of the above two formulations, if α=0, then the variance once again equals μ and the model is a simple Poisson model.

18.5.4 Influential points and outliers

Outliers can be identified by looking for large values of either the Pearson or deviance residuals. Influential points can be identified by looking for large values of Cook's

distance (see Chapter 14 for introduction to Cook's distance). Examples of these are shown in Example 18.2. As with other forms of regression models, ill-fitting points must be checked thoroughly. If the data are incorrect, they must be fixed or excluded. If the data are correct, evaluation of poorly fitting points could provide insight into reasons why the model does not fit well.

18.6 NEGATIVE BINOMIAL REGRESSION

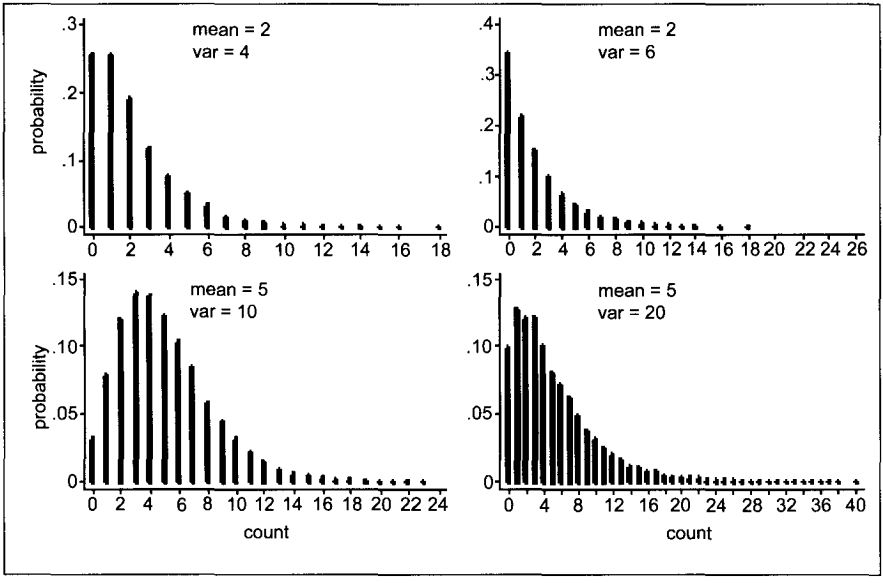
The negative binomial distribution is a two-parameter distribution for counts, and as noted in section 18.5.3, its variance can be expressed as a function of the mean as shown in Eq 18.8 and repeated here:

$$\text{var} = \mu + \alpha\mu^2$$

If α=0, then the negative binomial distribution reduces to the usual Poisson distribution with a variance equal to the mean. Fig. 18.3 shows four negative binomial distributions with various combinations of means and variances. Comparing these distributions to the Poisson distributions with means of 2 and 5, you can see the more prominent right tails on the negative binomial distributions.

The interpretation of a negative binomial distribution as a Poisson distribution with extra dispersion corresponds to a random effects model where the distribution of Poisson means are subjected to additional variation which has a gamma distribution (see section 22.4.3).

Fig. 18.3 Negative binomial distributions



As with the Poisson distribution, the usual form of a negative binomial regression model is:

$$E(Y) = n\lambda \quad \text{or} \quad \frac{E(Y)}{n} = \lambda$$

except in this case $E(Y)$ has a negative binomial distribution. As in a Poisson model, n is a measure of exposure and λ is a function of the predictors, with the most usual form for λ being:

$$\lambda = e^{\beta_0+\beta_1 X} \quad \text{or} \quad \ln(\lambda) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.9}$$

Consequently, exponentiated regression coefficients from a negative binomial model in which the exposure was a measure of time at risk are interpreted as incidence rate ratios.

As presented above, the most common form of the negative binomial model is based on the assumption that α is a constant value for all observations. More complex models might allow α to vary as a function of one or more predictors (which might, or might not, be the same predictors as in the negative binomial model). A more detailed discussion of these models is beyond the scope of this text and the reader is referred to Irvine et al (2000).

18.6.1 Evaluating overdispersion

A likelihood ratio test which compares the usual Poisson model to the negative binomial model is equivalent to a test of α=0. This provides a formal test for the presence of overdispersion in the model. Because α cannot be negative, this is a 1-tailed test.

As the additional variance is now a function of both α and μ [var=$(1+\alpha\mu)\mu$], the amount of overdispersion is a function of both values. If $\alpha\mu > 1$, then $(1+\alpha\mu) > 2$, which would indicate substantial overdispersion. For example, if α=0.5 and most counts are 0, 1 or 2 with a mean of 1.0, then $(1+\alpha\mu)$=1.5, so there is only slight evidence of overdispersion. However, if α=0.5 and most counts range from 0 to 15, with a mean of 5.0, then $(1+\alpha\mu)$=3.5 which is indicative of substantial overdispersion. Example 18.3 provides an example of a negative binomial model and an assessment of overdispersion.

18.7 ZERO-INFLATED MODELS

One occasionally encounters situations in which the distribution of outcome events might follow a Poisson (or negative binomial) distribution, except that there is an excess of zero counts in the data. This might be because there are two processes by which zero counts might arise. For example, assume your outcome of interest is the number of slaughter hogs with lesions of enzootic pneumonia (caused by *Mycoplasma hyopneumonia*), in samples taken from many herds. In herds in which *Mycoplasma hyopneumonia* is endemic, the count of lesioned animals might follow a Poisson

Example 18.3 Negative binomial regression
data=tb_real

Fitting the same data as were used in Examples 18.1 and 18.2 with a negative binomial model yields the following results.

Log likelihood = -157.7

Variable	Coef	SE	Z	P	95% CI	
type=beef	0.605	0.675	0.90	0.370	-0.718	1.927
type=cervid	0.666	0.684	0.97	0.330	-0.675	2.006
type=other	0.800	1.119	0.71	0.475	-1.393	2.993
sex=male	-0.057	0.405	-0.14	0.887	-0.851	0.736
age=12-24 mos	2.253	0.903	2.49	0.013	0.483	4.023
age=24+ mos	2.481	0.882	2.81	0.005	0.753	4.209
constant	-11.124	1.171	-9.50	0.000	-13.418	-8.829
alpha	1.740	0.441			1.059	2.860

The likelihood ratio test of $\alpha=0$ is highly significant ($P<0.001$) suggesting that the variance in the data is higher than would be expected for a Poisson regression. Since the overall mean number of reactors in these data was 1.46, the value of $(1+\alpha\mu)=1+(1.74*1.46)=3.54$ (*ie* substantial overdispersion).

The deviance χ^2 goodness-of-fit test was not significant ($\chi^2=99.4$ on 127 df, $P=0.97$) indicating the model fit the data well. An examination of the residuals did not identify any very large negative or positive deviance residuals. However, a listing of the observations with the five largest Cook's distance values showed that the model was heavily influenced by one group of animals (observation #133). Clearly this group of beef cows with 20 reactors had a strong influence in the model.

			age	pop	reactors		deviance	Cook's
obs	type	sex	(mo)	at risk	observed	predicted	residual	distance
117	cervid	male	12-24	6224	0	1.516	-1.218	6.593
49	dairy	female	12-24	6588	0	0.873	-1.031	7.557
26	dairy	female	12-24	6526	0	0.865	-1.027	11.398
133	beef	female	> 24	6418	20	1.956	2.600	133.366

As was suggested in Chapter 15, you might omit this group of animals and refit the model. However, this should only be done to further evaluate the impact of this observation on the model.

distribution, and some herds could still have zero counts (*ie* no animals with lesions at slaughter). However, other herds will have zero counts because *Mycoplasma hyopneumonia* is not present in the herd. Consequently, a count of zero might arise from either of the two situations.

Zero-inflated models deal with an excess of zero counts by simultaneously fitting a binary model (usually a logistic regression model) and a Poisson (or negative binomial) model. The two models might have the same, or different, sets of predictors. The outcome in the binary model is the probability of a zero count so coefficients have an opposite sign than they would in a usual logistic regression (and if the same predictor is in the Poisson model, they often have opposite signs in the two models).

Whether or not a zero-inflated model fits the data better than the usual Poisson or negative binomial model can be assessed using a Vuong test (V). This test compares two non-nested models and is asymptotically distributed as normal. If the value of V is >1.96, one model (*eg* the usual Poisson or negative binomial model) is favoured. If V <-1.96, the second model (*ie* the zero-inflated model) is favoured. If V lies between -1.96 and 1.96, neither model is preferred.

A zero-inflated negative binomial model was recently used to model factors affecting fecal egg counts in adult dairy cattle (Nødtvedt et al, 2002). Fecal egg counts are generally low in adult cattle and might arise because the animal is uninfected or is infected but shedding eggs in numbers too low to be detected. The zero-inflated negative binomial model identified a number of factors which influenced the number of eggs among animals shedding, but only a single factor (lactation number) was associated with any shedding at all in the logistic component of the model (older cows were more likely to shed no eggs). Example 18.4 shows the application of a zero-inflated negative binomial model to these data.

Example 18.4 Zero-inflated negative-binomial model
data=fec

Fecal egg counts (n=2250) from 304 cows in 38 dairy herds in four regions of Canada were determined over a one-year period in conjunction with a clinical trial of eprinomectin as a treatment for gastro-intestinal nematodes in Canadian dairy cattle. A more detailed description of the dataset can be found in Chapter 27. Although the mean fecal egg count was 8.6, almost one-half the observations had zero counts.

Counts obtained from control cows and treated cows prior to treatment were analysed using a zero-inflated negative binomial model. Lactation (two groups), season (four groups), province (four groups) and several herd management variables were included in the negative binomial portion of the model. Age and herd of origin (38 groups) were included in the logistic portion of the model. Clustering of observations within cows was accounted for by using robust standard error estimates. The resulting model follows.

Number of obs = 1840
Non-zero obs = 983
Zero obs = 857
Log likelihood = -4428.953

Variable	Coef	SE	Z	P	95% CI	
Negative binomial portion						
lact=2+	-0.942	0.229	-4.11	0.000	-1.391	-0.493
season=Jan-Mar	-0.705	0.176	-4.00	0.000	-1.050	-0.359
season=Apr-Jun	0.361	0.224	1.61	0.107	-0.078	0.800
season=Jul-Sep	0.076	0.256	0.30	0.766	-0.426	0.578
prov=Quebec	-0.556	0.274	-2.03	0.042	-1.092	-0.020
prov=Ontario	-0.333	0.535	-0.62	0.533	-1.382	0.716
prov=Sask	0.477	0.809	0.59	0.555	-1.108	2.063
past_lact	0.925	0.334	2.77	0.006	0.271	1.579
man_heif	-0.878	0.244	-3.60	0.000	-1.356	-0.399
man_lact	0.601	0.285	2.11	0.035	0.043	1.159
constant	2.474	0.306	8.08	0.000	1.874	3.074
Logistic portion						
lact=2+	1.495	0.530	2.82	0.005	0.456	2.535
herds 2-38	coefficients not shown					
constant	-3.132	1.431	-2.19	0.029	-5.936	-0.327
alpha	2.889	0.133			2.640	3.163

The Vuong statistic was 7.55 suggesting that the zero-inflated model was clearly superior to the regular negative binomial model. The value of α (2.89, 95% CI of 2.64 to 3.16) suggests that a negative binomial model is preferable to an ordinary Poisson model. The coefficients for -lact- in the negative binomial portion (-0.94) and the logistic portion (1.50) of the model indicated that multiparous cows generally had lower fecal egg counts and were more likely to have zero egg counts.

SELECTED REFERENCES/SUGGESTED READING

1. Cameron AC, Trivedie PK. Regression analysis of count data. Cambridge: Cambridge University Press, 1998.
2. Hammond RF, McGrath G, Martin SW. Irish soil and land-use classifications as predictors of numbers of badgers and badger setts. Prev Vet Med 2001; 51: 137-148.
3. Irvine RJ, Stein A, Halvorsen O, Langvatn R, Albon SD. Life-history strategies and population dynamics of abomasal nematodes in Svalbard reindeer (Rangifer tarandus platyrhunchus). Parasitology 2000; 120: 297-311.
4. Long JS. Regression models for categorical and limited dependent variables. London: Sage Publications, 1997.
5. Long JS, Freese J. Regression models for categorical dependent variables using Stata. College Station: Stata Press, 2001.
6. Nødtvedt A, Dohoo IR, Sanchez J, Conboy G, DesCôteaux L, Keefe GP et al. The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. Can J Vet Res 2002; 66: 249-257.
7. Sanchez J, Nødtvedt A, Dohoo IR, DesCôteaux L. The effect of eprinomectin at calving on reproduction parameters in adult dairy cows in Canada. Prev Vet Med 2002; 56: 165-177.

SAMPLE PROBLEMS

1. You are interested in modelling the spread of tuberculosis (TB) within cattle herds so you decide to retrospectively collect data from a number of herds in which there was good evidence that TB had been introduced. For those herds which you know the most likely date of entry of the organism into the herd, you collect the following data:

Variable	Description
dairy	dairy=1, beef=0
bunk	cattle eat from a common bunk: yes=1 no=0
confine	animals confined in some manner (barn, dry lot) all year round=1 cattle on pasture for part of the year=0
herdsz	number of animals in the herd
time	number of months from the time the infection was introduced to the time the infection was discovered and the herd completely tested
tb	number of animals that tested positive for TB when the herd was tested

The data from 60 herds are included in the dataset -tb_fake-. **Note** Some of these herds had no positive animals when tested but you are fairly certain that they were exposed to TB.

a. Compute a variable -cowmo- that represents the cow-months at risk between the time of introduction of the infection and the time of testing. (Don't worry about adjusting the estimate based on the number of positive cases found at testing.)
b. Use Poisson regression to determine the effects of -dairy-, -bunk- and -confine- on the rate of spread of TB. Make sure you adjust for the effect of different sample sizes.
c. What does each of the parameters in the final model mean (including the coefficient for the constant)?
d. Does the Poisson model fit the data well? Are there any large (negative or positive) residuals or herds which have a large influence on the model?
e. Would you expect a negative binomial model to fit better? Does it?

19

MODELLING SURVIVAL DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Distinguish between non-parametric, semi-parametric and parametric analyses of survival time data.

2. Carry out non-parametric analyses using either actuarial or Kaplan-Meier lifetables and compare the survival experiences of groups of animals using a variety of statistical tests.

3. Generate survivor and cumulative hazard function graphs to display survival data.

4. Understand the relationships among survivor functions $S(t)$, failure functions $F(t)$, probability density functions $f(t)$, hazard functions $h(t)$ and cumulative hazard functions $H(t)$.

5. Carry out a semi-parametric analysis of survival data using a Cox proportional hazards model.
   a. Evaluate the model to:
      i. assess the validity of the assumption of proportional hazards,
      ii. assess the validity of the assumption of independent censoring,
      iii. check the overall fit of the model,
      iv. evaluate the functional form of the predictors in the model, and
      v. check for outliers and influential points.
   b. Incorporate time-varying covariates into the model to evaluate or account for non-proportional hazards.

6. Carry out a parametric analysis of survival data based on an assumption that the survival times have an exponential, Weibull or log-normal distribution.

7. Incorporate frailty effects into a model to account for unmeasured covariates at the individual or group level.

8. Analyse multiple failure (recurrence) type data.

19.1 INTRODUCTION

In previous chapters we have looked at statistical models for evaluating how much of an outcome occurred (linear regression), whether or not an event occurred (logistic regression), which category of event occurred (multinomial models) and the number of events that occurred (or the rate of event occurrence) (Poisson regression). However, we are often interested in how long it took for an event to occur (time-to-event data). These data are often referred to as 'survival' data because the outcome of interest is often the time until death (*ie* the survival time). However, the analytical approaches discussed in this chapter apply equally to any time-to-event data (*eg* interval from calving to conception in dairy cows, or time to reoccurrence of Mycoplasma infections in swine barns after an eradication program). As these examples suggest, the unit of analysis could be an animal or a group of animals (litter, pen, herd) although in general we will present the discussion in terms of animals. The occurrence of the event of interest is often referred to as a 'failure' even though in some cases the outcome is desirable (*eg* time to conception after calving in dairy cows). Some relatively recently published texts which cover the analysis of survival data include (Collett, 1994, Hosmer and Lemeshow, 1999, Cleves et al, 2002).

There are specific issues that affect how we quantify and express time to occurrence of an event and how we evaluate the effects of factors (predictors) on that time. However, before discussing these issues, let's look at a simple hypothetical example (Example 19.1).

19.1.1 Quantifying survival time

How should the time to recurrence (*ie* time after initial diagnosis) of lymphosarcoma in dogs that have been treated for lymphosarcoma be quantified and expressed (Example 19.1)? For many dogs, we do not know what the time to recurrence was. All we know is that the disease did not occur in the time period for which the dogs were followed. These 'non-failures' are called **censored** observations and are a unique feature of time-to-event data.

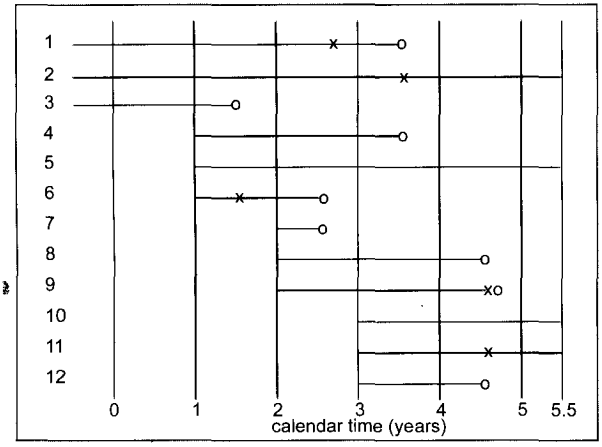
Some possible ways of quantifying and expressing the time to recurrence are as follows (using data from Example 19.1).

1. Mean time to recurrence - The mean time to recurrence can only be computed using data from the dogs in which recurrence has been observed. Consequently, we can only use data from five dogs (mean survival=2.1 years). The estimate will have a downward bias because recurrence in dogs which had a long time to recurrence are less likely to be observed. On the other hand, if the follow-up observation period is long, the mean suffers from the problem that it might be heavily influenced by a few animals with long survival times. Time-to-event data often have an asymmetrical distribution with a long right tail (*ie* right skew).
2. Median time to recurrence - This can only be computed if at least 50% of the animals are observed to have the event of interest and if none of the censored observations were censored before the failure of 50th percentile individual (*ie* if they were going

Example 19.1 Hypothetical survival data

Fig. 19.1 shows the time from treatment of 12 dogs with lymphosarcoma to the recurrence of the cancer. The study was carried out over a 5.5 year period with dogs entering the study as they were diagnosed and treated for the first occurrence. Once enrolled, not all dogs were followed for the rest of the study period because some died (from other diseases) or the owner moved away from the study location. For convenience, all dogs were assumed to have had the initial diagnosis and treatment at the start of a year and events (recurrence or loss to study) occurred at the mid-point of a year. In reality, this would not normally be the case.

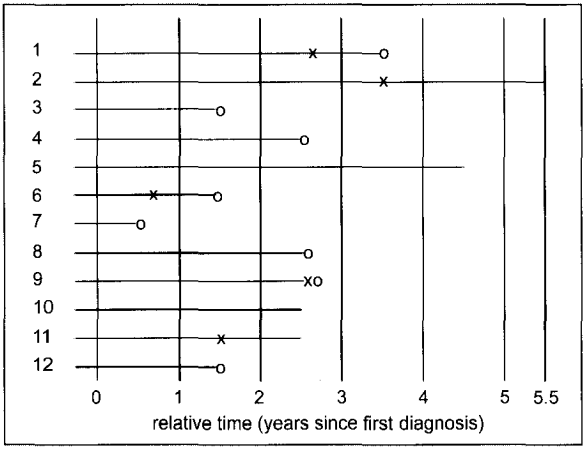
Fig. 19.1 Time from treatment to recurrence of cancer



x=recurrence of lymphosarcoma
o=death (due to other disease) or loss to follow-up

One way to simplify the graphic representation of these 12 dogs would be to express all times as being relative to the time of first diagnosis (Fig. 19.2).

Fig. 19.2 All times relative to time of first diagnosis



x=recurrence of lymphosarcoma
o=death (due to other disease) or loss to follow-up

to fail, they had a failure time at least as large as the median). It could not be computed for the data in Example 19.1. However, if it can be computed, the median is not influenced by a few animals having long times to recurrence in the same way that the mean is.

3. Overall probability of recurrence - The proportion of dogs having a recurrence of the tumour could be computed, but it is not at all clear what constitutes a 'negative' dog (*ie* one which does not have a recurrence). Should the dog be required to have some minimum number of years of follow-up to be considered eligible to contribute to the denominator of the proportion?
4. n-year survival risk - This expresses the number of dogs which have not had a recurrence by the n^{th} year. For each year (*eg* first, second) it can be computed based on the dogs that were observed for that number of years. This approach is often used in human epidemiology to quantify survival of people diagnosed with various forms of cancer (*eg* five-year survival of breast cancer patients). The two-year 'survival' for dogs in Example 19.1 is 0.78 (two recurrences among nine dogs that had either two complete years of follow-up or a failure at <2 years).
5. Incidence rate - The number of recurrences relative to the accumulated dog-years at risk would be one way to use all of the available data. In some cases, the average time to recurrence could be estimated from the incidence rate (see section 4.5). However, this approach assumes that the incidence rate of recurrence remains constant after first diagnosis and this is often not the case with time-to event data. The incidence rate in Example 19.1 is 0.19 cases per dog-year (five cases in 26 dog-years of follow-up time – dogs no longer contribute to the pool of dog-years once they have experienced a recurrence).

The approaches outlined above identify two key problems to be considered when analysing time-to-event data. First, many observations are censored; that is, the individual is not followed for a sufficiently long period of time to observe the event of interest if it were to occur. Second, the distribution of survival times is often not symmetrical, and might not even be unimodal. For example, tumour recurrences might be common in the first year after first diagnosis and then relatively rare for several years before becoming more common again as the dog ages. These issues are also important when evaluating the effect of predictors on the time-to-event occurrence.

19.1.2 Evaluating the effect of factors on survival times

Because time-to-event data are continuous, it would be tempting to evaluate the effects of factors on the time to the occurrence of an event using linear regression models. In some cases, this would be appropriate. However, as noted above, time-to-event data are often not symmetrical and might not even be unimodal. The assumption of normally distributed residuals required for a linear regression model would be violated in these cases. (In extreme cases, a linear model might predict negative survival times which are impossible). However, linear models have been successfully used to analyse time-to-event data. We recently evaluated calving-to-conception intervals in dairy cattle using this approach (Dohoo et al, 2001). The data were either log or Box-Cox transformed to deal with the distribution of residuals being skewed to the right. Numerous examples in Chapter 14 were also based on calving-to-conception intervals.

Even if the distribution of the residuals is (or can be made) approximately normal, the problem of censored observations remains. Censoring is discussed further in section 19.1.4. In the case of calving-to-conception interval data, because most cows are not culled until the end of the lactation, the follow-up period for most cows is adequate. However, many cows are bred unsuccessfully several times and then the producer stops trying. The data from these cows are lost to the analyses so the effects of factors which reduce conception might be underestimated.

19.1.3 General approaches to analysing survival data

There are three general approaches to analysing survival data:

- non-parametric analyses
- semi-parametric models
- parametric models.

These are discussed in much more detail later, but the essential features of each approach is summarised here.

In a **non-parametric** analysis, we make no assumptions about the distribution of survival times, nor about the functional form of the relationship between a factor (predictor) and the survival time. Consequently, they are only appropriate for evaluating the effect of qualitative (categorical) predictors.

In a **semi-parametric** analysis, we make no assumption about the distribution of the survival time, but merely use the survival time to order the observations in terms of time of occurrence of the event. We then evaluate the probability of the event occurring at each of those time points as a function of the predictors of interest. Because the time variable is only used to order the observations, it makes no difference if there was a large time interval or a small time interval between successive events.

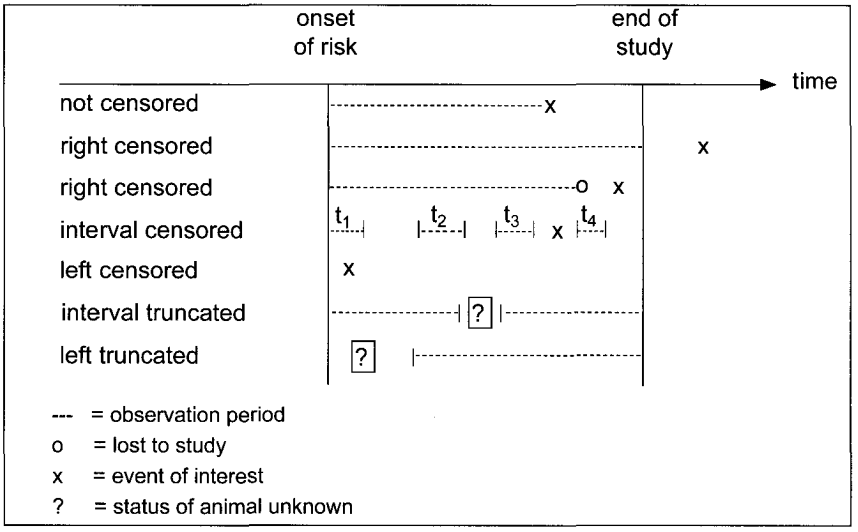
In a **parametric** analysis, we replace the distributional assumption that the residuals are normally distributed (as required in a linear model) by a more appropriate distribution that reflects the pattern of survival times. Because we specify a distribution for the survival times, the length of the interval between events is relevant for the analysis. Consequently, if the assumed distribution is correct, a parametric model is more efficient than a semi-parametric model (*ie* it makes better use of the available data).

19.1.4 Censoring

Censoring is defined as the occurrence (or possible occurrence) of a failure when the animal is not under observation. Censoring can arise in a variety of ways and these are summarised in Fig. 19.3.

Right censoring occurs when an animal is lost to a study, before the outcome of interest has occurred. This might arise because the study ends before the event occurs or because it is lost to follow-up during the study (*eg* the owner moves to another city). Right censoring is the most common form of censoring that needs to be dealt with in survival analyses.

Fig. 19.3 Summary of censoring



Interval censoring might arise when an animal is only observed periodically throughout a study period. If examinations are conducted every six months and at one examination (t_4 in Fig. 19.3) it is determined that the event had happened in the preceding six months, all that is known is that the event occurred sometime between t_3 and t_4. The precise time the event occurred is not known.

Left censoring is similar to interval censoring except that the 'interval' occurs at the start of the study (*ie* the event occurred in the animal before the animal was observed). Consequently, the animal is not put in the study. Left censoring usually arises if the onset of risk occurs before the start of the study. For example, if a study of calving to conception intervals started following cows at 45 days post-partum, a cow which conceived to a breeding at 42 days would be left censored. (**Note** If multiple failures are possible, the animal might be put on the study and the left censoring then becomes left truncation (see below)).

A related concept is that of **truncation**. While censoring relates to the possible occurrence of events during periods when the animal was not observed, truncation refers to periods of time in which nothing is known about the animal in terms of whether or not the event occurred or what the values of the predictors were. These periods of time might be referred to as **gaps**. In cases where multiple events are possible (*eg* cases of mastitis), you have no knowledge of how many cases occurred during the gap. For outbreaks which can only occur once (*eg* death), all that is known is that the event did not occur during the gap (or the animal would not have come back into the study). Truncation can occur throughout a study (**interval truncation**) or at the start of a study (**left truncation**). Right truncation is the same as right censoring.

As noted above, the most common problem is with right censoring and it will be the only type of censoring or truncation that we deal with in examples in this chapter. A

more complete discussion of censoring and how the various forms are dealt with can be found in Chapter 4 of Cleves et al (2002).

19.2 NON-PARAMETRIC ANALYSIS

In a non-parametric analysis of survival data, we make no assumption about either the distribution of survival times or the functional form of the relationship between a predictor and survival. Hence, they can be used to compare survival experiences of groups of animals, but not to evaluate the effect of a continuous predictor on survival times. We will look at three non-parametric methods for analysing survival data:

- actuarial life tables
- Kaplan-Meier estimator of the survivor function
- Nelson-Aalen estimator of the cumulative hazard function

In this section, we introduce the concepts of survivor and hazard functions. These will be described more formally in section 19.7.

19.3 ACTUARIAL LIFE TABLES

Life tables were originally developed to summarise long-term human-survival data by dividing the lifespan into short intervals in which the probability of dying was reasonably constant over the time interval. (It certainly isn't constant over an entire lifespan).

The requirements to create an actuarial life table are as follows.

- A clearly demarcated starting point to the period of risk (*eg* birth, calving, first diagnosis, first exposure *etc*)
- A well-defined study outcome (death, seroconversion, pregnancy diagnosis, calving)
- Only one episode or event per individual animal (not multiple remissions or relapses)
- Losses to follow-up should be independent of the study outcome (another way of saying this is that the animals lost from the study should have the same future experience as those that remain under observation)
- The risk of the outcome remains constant over calendar time (no secular changes in risk). This does not imply that risk stays the same in an individual over time. Secular changes in survival rates for cancers (*eg* due to better therapies), might for example, affect validity of studies of survivorship
- The risk of outcome must remain constant within the intervals used for constructing a life table. Intervals of any length could be calculated to meet this requirement. In fact, the intervals need not be of the same length.

19.3.1 Steps in constructing the actuarial life table

Table 19.1 shows the columns required to build an actuarial life table, based on the data from Example 19.1.

Table 19.1 Actuarial life table

j	t_{j-1},t_j	l_j	w_j	r_j	d_j	q_j	p_j	S_j
1	0 < 1	12	1	11.5	1	0.087	0.913	0.913
2	1 < 2	10	2	9.0	1	0.111	0.889	0.812
3	2 < 3	7	3	5.5	2	0.364	0.636	0.516
4	3 < 4	2	0	2.0	1	0.500	0.500	0.258
5	4 < 5	1	1	0.5	0	0.000	1.000	0.258

where ...

j listing of time intervals (time intervals should be established *a priori*).

t_{j-1},t_j time span covered in the interval

l_j subjects at risk of failure at the start of the time interval
$l_j = l_{(j-1)} - (w_{(j-1)} + d_{(j-1)})$

w_j subjects withdrawn during interval (censored observations)
These are animals who died of causes other than the condition under study or were otherwise lost to follow up during that interval. Animals who were still free of the outcome when the study ended are counted as withdrawals in that interval

r_j average number of subjects at risk during the current time interval
$r_j = l_j - (w_j/2)$
This calculation is based on the assumption that the censored observations were withdrawn, on average, at the midpoint of the interval

d_j outcomes (failures) during the interval
This is the number experiencing the outcome during the time interval (death, seroconversion, relapse *etc*)

q_j risk of event during interval
$q_j = (d_j)/(r_j)$
This is the probability that the subject will develop the study outcome during the given interval, conditional upon surviving without the outcome up to the beginning of the time interval

p_j probability of surviving the interval
$p_j = 1 - q_j$
The conditional probability of surviving the time interval, given survival to the beginning of the interval

S_j cumulative survival probability to the end of the interval
$S_j = (p_1)(p_2)(p_3)....(p_j)$
The probability of surviving without experiencing the event of interest from the start of follow-up through the end of the current interval in the life table.

The risk of an animal experiencing the event of interest during the interval (q_j) divided by the length of the interval is also known as the **hazard**. The cumulative survival probability (S_j) is also known as the **survivor function**. These two quantities are key elements of all survival analyses.

19.4 KAPLAN-MEIER ESTIMATE OF SURVIVOR FUNCTION

19.4.1 Overview and comparison to actuarial method

The Kaplan-Meier (K-M) (Kaplan and Meier, 1958) estimate of the survivor function is also known as the **product-limit estimate**. It has two important differences from the actuarial estimate described above.

1. The K-M method does not depend on discrete time intervals constructed by the investigator. Each row in the table (hence, each time interval) is defined by the time at which the next subject (or subjects, in the case of two events happening at the same time) experiences the event of interest.
2. Censored observations (losses to follow up *etc*) between two events are counted as animals at risk only up to the time of the earlier of the two events.

The K-M method has the advantage that it avoids the assumption that withdrawals occurred uniformly throughout the interval (*ie* the actuarial assumption) and that the risk is constant over the arbitrarily selected interval. (The only remaining assumption about withdrawals is that they have the same future experiences as those remaining under observation). However, because it creates an 'interval' for each unique time to the event of interest, it is best suited for small sample sizes (or you end up with a very large number of 'intervals').

19.4.2 Construction of the K-M life table

An ordered list of the event times is constructed from the sample, with patients ranked in ascending order of the time of the event of interest. Based on these, Table 19.2 can be filled out (using the data from Example 19.1)

Table 19.2 Kaplan-Meier life table

j	t_j	r_j	d_j	w_j	q_j	p_j	S_j
1	0.5	12	1	1	0.083	0.917	0.917
2	1.5	10	1	2	0.100	0.900	0.825
3	2.5	7	2	3	0.286	0.714	0.589
4	3.5	2	1	1	0.500	0.500	0.295
5	-	1	0	0	0.000	1.000	0.295

where:

j listing of time points

t_j time of event

r_j number at risk of event at time t_j

$r_j = r_{j-1} - (d_{j-1} + w_{j-1})$

Includes all subjects known to be alive and in the study (not censored) at the time of the event at time t, plus the number experiencing the event at time t. When censored times are tied with event times, the event is usually assumed to have occurred first

d_j number of events at time t_j

w_j number of censored observations at time t_j
Censoring between time t_j and t_{j+1} is assumed to have happened at t_j so the animals will not be considered at risk at time t_{j+1}.

q_j risk of event at time t_j
$q_j = (d_j)/(r_j)$
Also known as the instantaneous hazard, this is the individual probability of the event at time t_j, conditional upon survival to time t_j

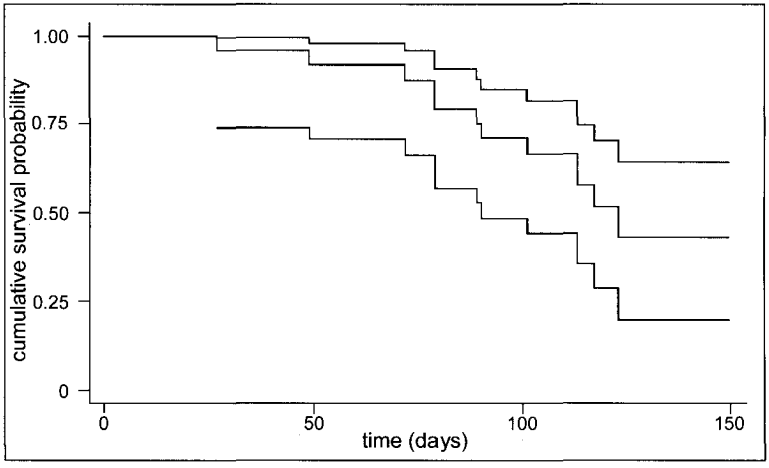
p_j probability of survival at time t_j
$p_j = 1 - q_j$

S_j cumulative probability of surviving up to and including time t_j
$S_j = (p_1)(p_2)(p_3)....(p_j)$

Survivor functions are usually presented graphically as step functions of the cumulative survival over time. They start at 1 and monotonically descend (*ie* they never go up) as time proceeds. The Kaplan-Meier survivor function is shown in Fig. 19.4 along with 95% confidence intervals of the survivor function.

Example 19.2 shows an actuarial life table and a Kaplan-Meier estimate of a survivor function using some published data on calf pneumonia (Thysen, 1988).

Fig. 19.4 Kaplan-Meier survivor function (with 95% confidence interval)



Example 19.2 Actuarial and Kaplan-Meier estimates of survivor functions
data=calf_pneu

Data on the occurrence of calf pneumonia in calves raised in two different housing systems were published (Thysen, 1988). Calves surviving to 150 days without experiencing pneumonia were considered censored at that time.

The table below presents an actuarial life table estimate of the cumulative survivor function.

Interval		Beg. total	Deaths	Lost	Survival	SE	95% CI	
15	30	24	1	0	0.958	0.041	0.739	0.994
45	60	23	1	0	0.917	0.056	0.706	0.979
60	75	22	1	0	0.875	0.068	0.661	0.958
75	90	21	3	0	0.750	0.088	0.526	0.879
90	105	18	2	1	0.664	0.097	0.439	0.816
105	120	15	3	6	0.498	0.110	0.273	0.688
120	135	6	1	0	0.415	0.119	0.189	0.629
150	165	5	0	5	0.415	0.119	0.189	0.629

Note that survival estimates are only presented for intervals in which at least one event or censoring occurred. Thus, the cumulative survival at the end of the 30-45 day interval would be exactly the same as at the end of the 15-30 day interval (0.958).

The table below presents the results of a Kaplan-Meier estimate of the survivor function.

Time	Beg. total	Fail	Net lost	Survivor function	SE	95% CI	
27	24	1	0	0.958	0.041	0.739	0.994
49	23	1	0	0.917	0.056	0.706	0.979
72	22	1	0	0.875	0.068	0.661	0.958
79	21	2	0	0.792	0.083	0.570	0.908
89	19	1	0	0.750	0.088	0.526	0.879
90	18	1	0	0.708	0.093	0.484	0.849
101	17	1	1	0.667	0.096	0.443	0.817
113	15	2	4	0.578	0.102	0.357	0.747
117	9	1	2	0.514	0.109	0.288	0.700
123	6	1	0	0.428	0.120	0.200	0.641
150	5	0	5	0.428	0.120	0.200	0.641

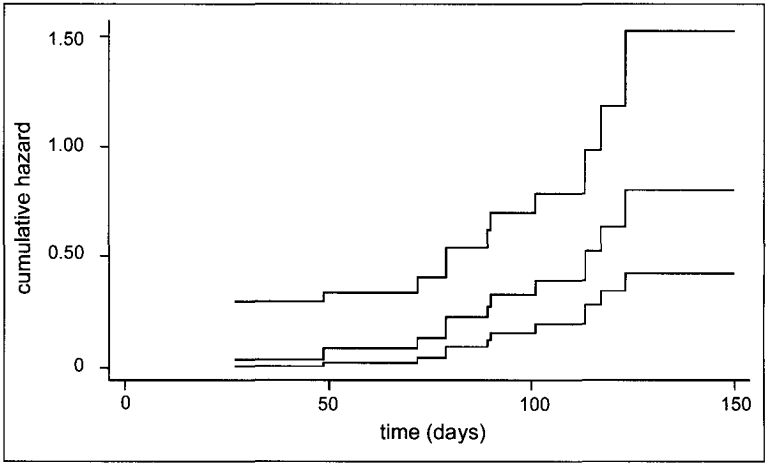
The two estimates of the probability of survival up to day 150 are very close (41.5% and 42.8%).

19.5 Nelson-Aalen estimate of cumulative hazard

In the above two sections, we introduced the concept of 'hazard', being the probability of failure at a point in time, given that the animal had survived up to that time point. This is discussed more formally in section 19.7, but for now, we note that a **cumulative hazard** (Nelson-Aalen estimate) can also be computed. The cumulative hazard is the expected number of outcomes occurring up to a point in time (assuming that the outcome could occur multiple times in an individual). For example, in the calf pneumonia data, the cumulative hazard at day 60 would be the sum of all the individual hazards (computed at failure times), up to day 60.

The cumulative hazard can range from 0 to infinity (as the time period gets longer, the expected number of outcomes keeps going up with no upper bound). A graph of the cumulative hazard is, like a graph of the survivor function, a way of expressing the overall failure (survival) experience of the population. Fig 19.5 shows the cumulative hazard (and 95% CI) for the calf-pneumonia data.

Fig. 19.5 Nelson-Aalen cumulative hazard estimate (95% confidence interval)



19.6 Statistical inference in non-parametric analyses

19.6.1 Confidence intervals and 'point-in-time' comparisons

Although the formulae have not been shown, standard errors (SE) of the cumulative survival estimates can be computed from actuarial or Kaplan-Meier survivor functions at any point in time. These SE can be used to compute confidence intervals around the functions (see Example 19.2 and Figs. 19.4 and 19.5). They can also be used to test the difference between survivor functions for two (or more) populations at any point in time using a standard normal Z-test.

$$Z = \frac{S_{j1} - S_{j2}}{\sqrt{[\mathrm{SE}(S_{j1})]^2 + [\mathrm{SE}(S_{j2})]^2}} \qquad \textit{Eq 19.1}$$

where S_{j1}=cumulative probability of survival in population 1 at time=j and SE(S_{j1}) is the standard error of the estimate. S_{j2} and SE(S_{j2}) are the same estimates from population 2. There are potentially, an infinite number of points at which the cumulative survival probabilities could be computed. This could lead to a serious problem of 'data snooping' or multiple comparisons and consequently, 'point-in-time' comparisons are only valid if it is possible to identify specific times at which the comparison of survival probabilities is warranted. These should be specified *a priori* (*ie* before the data are collected) and if multiple time points are evaluated, some adjustment for multiple comparisons must be made.

19.6.2 Tests of the overall survival curve

There are several tests that can be used to test whether the overall survivor functions in two (or more) groups are equal. They are all based on a series of contingency tables of observed and expected events for each group at each time point at which an event occurred (assuming the test is based on a Kaplan-Meier survivor function). The observed number of events at each time point is compared to the expected number (under the H_0 that there is no difference between the two groups) and a χ^2 statistic computed. Consequently, the tests can be viewed as the survival analysis equivalent of the Mantel-Haenszel test for stratified data.

All of the tests assume that the ratio of risks of the event of interest for the two groups is constant across all strata (equivalent to the no-interaction assumption in a Mantel-Haenszel test). This assumption is known as the 'proportional hazards' assumption (you will see more of this later). If the survivor functions cross over, then it is clear that this assumption is violated. The differences among the tests depend on the weights used to combine the estimates derived at each point in time.

Log-rank test
The log-rank test is the simplest test as it assigns equal weight to each point estimate (weights=1). Consequently, it is equivalent to doing a standard Mantel-Haenszel procedure to combine the estimates.

Wilcoxon test
This test weights the intervals according to the sample size ($w(t_j)=n_j$) . Consequently, it is more sensitive to differences early in the time period when the sample size is larger. Some people advocate using both Wilcoxon and the log-rank test to see if differences in the survival curves occur early or late in the time period studied. The Wilcoxon test is less sensitive than the log-rank test to the assumption of proportional hazards, but will be unreliable if the censoring patterns vary across the groups being compared.

Tarone-Ware test
The Tarone-Ware test is intermediate between the log-rank and Wilcoxon tests in that

it weights the estimates by the square root of the population at risk at each time point ($w(t_j) = \sqrt{nj}$).

Peto-Peto-Prentice and Harrington-Flemming tests
All of the previous tests weight the estimates according to the total population at risk. Consequently, if two groups have markedly different censoring patterns, this can have a substantial influence on the test statistic. The Peto-Peto-Prentice and Harrington-Flemming tests weight the estimates on the overall survival experience (estimated just before the time point of interest). These tests are preferred if the groups have markedly different censoring patterns.

Example 19.3 shows separate survivor functions for 'batch' and 'continuous' stocked calves and the results from several of the tests for the overall equality of the survivor functions.

19.7 Survivor, failure and hazard functions

The concepts of survivor, and hazard functions were introduced when we looked at non-parametric methods of analysis of survival data. Before proceeding with semi-parametric and parametric analyses, we need to develop a more complete understanding of these and related functions.

19.7.1 Survivor function

The survivor function ($S(t)$) is the probability that an individual's survival time (T) (or more generally, their time to event occurrence) will exceed some specified time t. It can be written as:

$$S(t) = p(T \geq t) \qquad \textit{Eq 19.2}$$

Survivor functions are 'non-increasing' (*ie* they are flat or head downwards). They start at 1 and drop to 0 if all individuals ultimately experience the event of interest. **Note** By convention, cumulative functions will be designated by upper-case letters and density functions by lower-case letters. The survivor function is a cumulative function in that it represents the cumulative probability of surviving up to a point in time t.

19.7.2 Failure function

The failure function ($F(t)$) is the probability of not surviving past time t. Consequently, it is:

$$F(t) = 1\text{-}S(t) \qquad \textit{Eq 19.3}$$

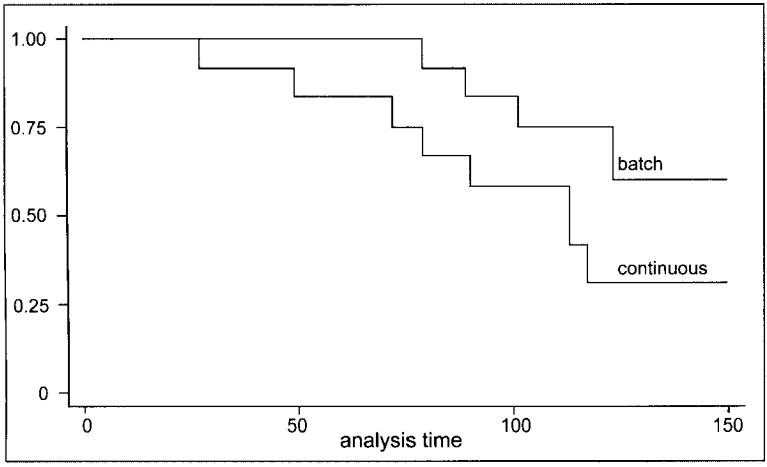
19.7.3 Probability density function

The probability density function ($f(t)$) is the slope of the failure function. Consequently, it represents the instantaneous rate at which failures are occurring in the study

Example 19.3 Comparing survivor functions
data=calf_pneu

Fig. 19.6 shows the Kaplan-Meier survival functions for batch and continuous-stocked calves.

Fig. 19.6 Kaplan-Meier survivor function



Continuous-stocked calves were at greater risk of having pneumonia than batch-stocked calves. The statistical significance of the test results for the difference between these two survivor functions were as follows:

Test	P-value
log-rank test	0.084
Wilcoxon	0.083
Tarone-Ware	0.081
Peto-Peto-Prentice	0.078

All tests provided comparable results (borderline significant).

population at a point in time. It is determined by taking the derivative of the failure function with respect to time.

19.7.4 Hazard function

The hazard function ($h(t)$) is the probability of an event occurring at time t given that it had not occurred up to time t. With time divided into discrete intervals (as in a life table) it can be expressed as:

$$h(t) = p\,(T = t \mid T \geq t) \qquad \textit{Eq 19.4}$$

With time on a continuous scale, the hazard function describes the instantaneous probability of an event occurring at a point in time given that it did not occur previously. The hazard function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

The hazard function can also be computed as the ratio of the probability density function (which represents the rate at which failures are occurring at a point in time) and the survivor function (which represents the probability of surviving up to that point in time).

$$h(t) = \frac{f(t)}{S(t)} = -\left[\frac{\frac{d\,S(t)}{d\,t}}{S\,(t)}\right] \qquad \textit{Eq 19.5}$$

Hazard functions are always non-negative (*ie* greater than or equal to zero) and have no upper bound (their value will change with the time scale used).

19.7.5 Cumulative hazard function

The cumulative hazard $H(t)$, also known as the integrated hazard, represents the accumulation of hazard over time. It can be computed as the integral of the hazard function but is more conveniently found using the following equation.

$$H(t) = -\ln S(t) \qquad \textit{Eq 19.6}$$

As noted before, the cumulative hazard represents the expected number of outcomes of interest that would occur (assuming that repeat occurrences were possible in an individual). For example, if you were studying the survival of cats following infection with the feline infectious peritonitis virus and at three years you find that the cumulative hazard is 4 [$H(t_3)$=4], then that would suggest that in three years after infection, we would expect to see four deaths. Obviously, only one death is possible, but it provides an indication that the probability of the cat surviving to three years post-infection is very low.

19.7.6 Relationship among survivor, failure and hazard functions

Some of the relationships between the survivor, failure and hazard functions have already been shown in previous sections. If one of the functions is known, the others can all be computed. For example, if the survivor function is known, the other functions are:

$$F(t) = 1 - S(t) \qquad f(t) = \frac{dF(t)}{dt} \qquad h(t) = \frac{f(t)}{S(t)} \qquad H(t) = -\ln S(t) \qquad \textit{Eq 19.7}$$

While survival experiences for groups of animals are usually shown by plotting the survivor function, the hazard function plays a key role in semi-parametric and parametric analyses.

19.7.7 Examples of hazard functions

A wide variety of hazard functions are possible, but constant and Weibull functions are the two most commonly encountered in survival analyses. Other forms used include the log-normal, log-logistic, gamma and Gompertz.

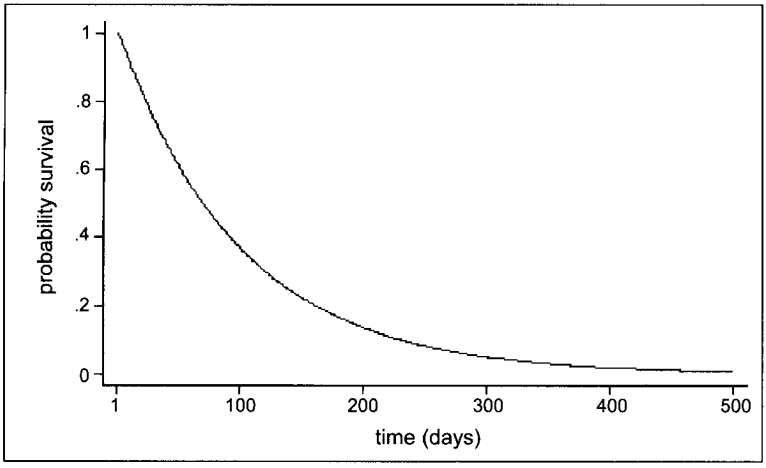
Constant hazard

A constant hazard is one which does not change over time. With a constant hazard (λ), the survivor function drops exponentially and survival times will have an exponential distribution. The hazard $h(t)$, density $f(t)$ and survivor $S(t)$ functions are:

$$h(t) = \lambda \qquad f(t) = \lambda e^{-\lambda t} \qquad S(t) = e^{-\lambda t} \qquad \textit{Eq 19.8}$$

The appropriateness of an exponential model can be assessed by plotting the cumulative hazard $H(t)$ (or equivalently $-\ln S(t)$) against t. If the exponential model is appropriate, the line will be straight. Fig. 19.7 shows a survivor function derived from a constant hazard of λ=0.01 per day.

Fig. 19.7 Survivor function from a constant hazard



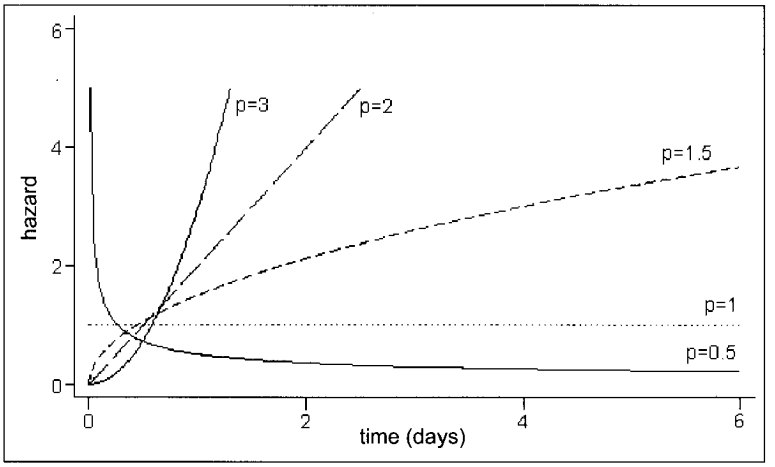
Weibull hazard

A Weibull hazard function depends on two non-negative parameters: a scale parameter (λ) and a shape parameter (p). If p=1 this function reduces to the exponential distribution. If p<1 then the hazard function decreases monotonically. If p>1 then the function is monotonically increasing with a value between 1 and 2 producing a curve that increases at a decreasing rate, p=2 produces a hazard function that increases linearly with time and p>2 produces a function that increases at an ever-increasing rate. The hazard and survivor functions are:

$$h(t) = \lambda p(t)^{p-1} \qquad S(t) = e^{-\lambda t p} \qquad \textit{Eq 19.9}$$

Fig. 19.8 shows Weibull hazard functions for several values of p. An increasing Weibull hazard function ($1<p<2$) might be appropriate for dairy cow conception data if the

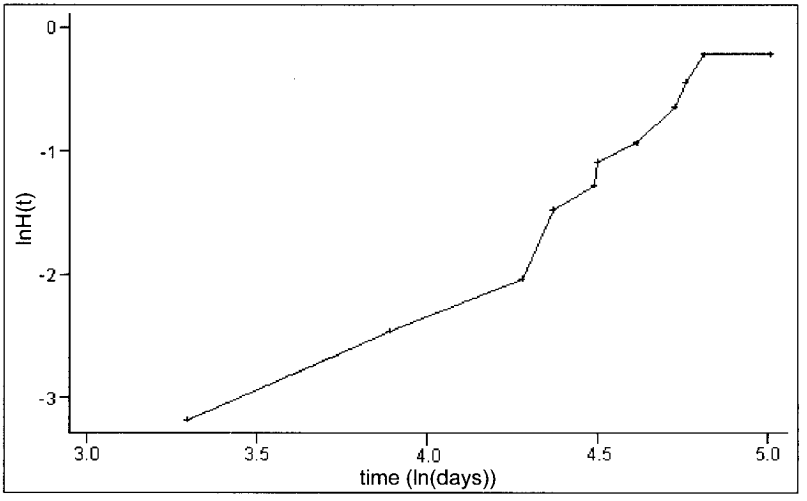
Fig. 19.8 Weibull hazard functions



fertility of the cow increases with time after parturition, but does so at a decreasing rate. A decreasing Weibull hazard function ($p<1$) might be appropriate for the survival of animals after surgery when the hazard is highest right after surgery and then decreases.

The suitability of the Weibull model can be assessed by evaluating the log-cumulative hazard plot [$\ln(H(t))$ versus $\ln t$]. If the data fit a Weibull distribution, the line on the graph should be an approximately straight line. The intercept and the slope of the line will be $\ln(\lambda)$ and p respectively. Fig. 19.9 shows the log cumulative hazard plot for the calf-pneumonia data. The line is approximately straight suggesting that a Weibull model might be appropriate for these data. The slope is approximately (0-(-3))/(5-

Fig. 19.9 Log cumulative hazard plot (calf-pneumonia data)



3.25)=1.75 suggesting that the hazard is increasing over time, but at a decreasing rate. Parametric survival models based on exponential and Weibull hazard functions are described in section 19.8.

Other distributions

One of the limitations of the Weibull model is that the hazard can only increase or decrease over time. **Gamma**, **log-normal** and **log-logistic hazards** can be used to deal with the situation in which the risk first increases and then decreases (or vice versa). Such a model would be appropriate in a situation where the risk of death was high early in an illness, drops to a lower level and then increases again over time. For example, a new intramammary infection with *Staph. aureus* in a dairy cow might produce a high risk of culling early (if acute clinical mastitis developed), followed by a sharp reduction in the risk and then a gradually increasing risk as the level of chronic udder damage increased over time. Detailed descriptions of these functions can be found in recent survival analysis texts (Collett, 1994, Hosmer and Lemeshow, 1999, Cleves et al, 2002).

19.8 Semi-parametric analyses

Non-parametric analyses are limited to evaluating the effect of one, or a small number, of qualitative variables on survival times. However, we often want to simultaneously evaluate the effects of multiple continuous or categorical explanatory variables. This requires that we model the survival data using a multivariable technique. The most commonly used form of multivariable analysis for survival data is the **proportional hazards model** (also known as the **Cox regression model**) (Cox, 1972). It is a semi-parametric model in that we do not have to assume any specific functional form for the hazard, but we do model the ratio of hazards as a linear function of the predictors.

19.8.1 Cox proportional hazards model

The proportional hazards model is based on the assumption that the hazard for an individual is a product of a baseline hazard (h_0) and an exponential function of a series of explanatory variables.

$$h(t) = h_0(t)e^{\beta X} \qquad \textit{Eq 19.10}$$

where $\beta X = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$. Equivalently, it can be expressed as:

$$HR = \frac{h(t)}{h_0(t)} = e^{\beta X} \qquad \textit{Eq 19.11}$$

where *HR* is the hazard ratio. The first formulation emphasises that the hazard for an individual is always a multiple ($e^{\beta X}$) of a baseline hazard, while the second formulation shows that it is the ratio of the hazards which is assumed to be constant over time. Two important features of this model are that no assumption is made about the shape of the baseline hazard (h_0) and that the model has no intercept. In fact, the intercept

(which in most regression models reflects the value of the outcome when all covariates (predictors) are zero) is subsumed into the baseline hazard which represents the hazard when all covariates are zero.

19.8.2 Hazard ratios

Based on Eq 19.11, the $\ln HR = \beta X$. Consequently, exponentiating the coefficient from a proportional hazards model produces a hazard ratio. Hazard ratios have interpretations similar to odds ratios and risk ratios. They represent the effect of a unit change in the predictor on the frequency of the outcome (which in this case is measured as a hazard).

Note You will sometimes encounter hazard ratios referred to as relative risks (or risk ratios), but this is not a correct use of the term and should be avoided. For example, if factor X_1 has an $HR=2$, then a unit change in X_1 will double the hazard of the outcome. If X_1 is a dichotomous variable and, because we are assuming that this HR is constant (over time), this means that at any point in time during the risk period 'failures' will be occurring at twice the rate in animals with $X_1=1$ than in animals with $X_1=0$. This is not equivalent to a doubling of the risk over the full study period.

Example 19.4 provides some examples of HRs derived from a dataset from a clinical trial of prostaglandin use in dairy cattle. A total of 319 cows in three herds were assigned randomly to receive prostaglandin (or not) at the time that the producer had indicated was the beginning of the breeding period (*ie* the number of days after calving that he would start breeding a cow that came into heat). The time from the onset of the breeding period to conception was the outcome of interest. The dataset (pgtrial) is described more fully in Chapter 27. The variables in Table 19.3 are those that we will use in examples in this chapter.

Table 19.3 Variables = the dataset pgtrial

herd	herd (cows from three herds were used in the study)
tx	treatment (1=yes, 0=no)
lact	age (lactation number - a continuous variable)
thin	body condition score at time of treatment (1=thin, 0=normal or fat)
dar	days at risk (number of days from end of voluntary wait period to either conception or censoring); this is the outcome of interest
preg	status of animal at end of 'dar' (1=pregnant, 0=censored)

As the main interest was in the effect of treatment, a pair of Kaplan-Meier survival curves (one for each treatment group) provides some initial insight into the possible effect of the treatment (Fig. 19.10). It appeared that treated cows conceived slightly more quickly, although the difference was most pronounced early in the breeding period.

19.8.3 Fitting the Cox proportional hazards model

Obtaining maximum likelihood estimates of the parameters in a Cox proportional hazards model requires an iterative estimation procedure (the Newton-Raphson

Example 19.4 Cox proportional hazards model
data=pgtrial

A Cox proportional hazards model was fit to the prostaglandin-trial data with herd, treatment lactation number, and body condition (-thin-) as predictors. The first table presents the model in terms of coefficients.

No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(5) = 9.50
Time at risk = 25018 Log likelihood = -1307.73
Prob > chi2 = 0.0908

Predictor	Coef	SE	Z	P	95% CI	
herd=2	-0.284	0.170	-1.68	0.094	0.617	0.048
herd=3	0.037	0.174	0.21	0.833	0.305	0.378
tx	0.184	0.125	1.46	0.143	-0.062	0.429
lact	-0.043	0.041	-1.04	0.297	-0.123	0.038
thin	-0.146	0.138	-1.06	0.291	-0.416	0.125

Although not statistically significant, treatment appears to increase the ln*HR* by 0.18 units. As we rarely think in terms of ln*HR*s, it is more common to present the results as *HR*s.

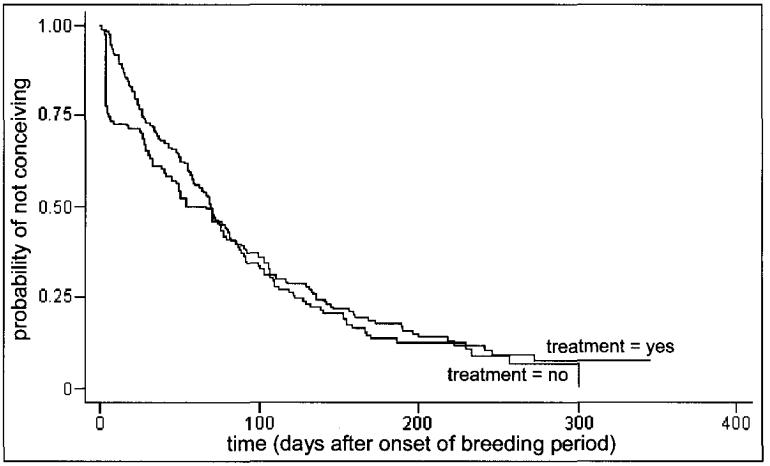
Predictor	HR	SE	Z	P	95% CI	
herd=2	-0.752	0.128	-1.68	0.094	0.539	1.050
herd=3	1.037	0.181	0.21	0.833	0.737	1.460
tx	1.202	0.151	1.46	0.143	0.940	1.536
lact	0.958	0.039	-1.04	0.297	0.884	1.038
thin	0.865	0.119	-1.06	0.291	0.660	1.133

Here, it appears that treatment increases the hazard of conception 1.2 times. If this effect is real (which appears questionable at this stage given the P-value of the *HR*), it means that at any point in time after the onset of the breeding period, conceptions were happening at a 20% higher rate in the treated cows than in the non-treated ones. Similarly, for each additional lactation the cows had experienced, the rate of conception dropped by approximately 4% (but this predictor had an even larger P-value).

procedure is most commonly used). As with a non-parametric Kaplan-Meier estimation procedure, a Cox model is only evaluated at the times at which failures occur. In fact, fitting a Cox model with no predictors produces exactly the same survival curve as a Kaplan-Meier estimation does. In both procedures, it is not the actual times at which failures occur which is important, it is only the order in which they occur that matters.

Because the order in which failures occur is critical for conducting the analysis, there must be a way of handling the problem of two (or more) failures being recorded at the same time. Details of various methods of dealing with ties can be found in texts on survival analysis but they fall into two general approaches. The first is called a **marginal calculation** or continuous-time calculation and is based on the assumption that the timing of the events was not really tied, but simply due to the fact that the

Fig. 19.10 Kaplan-Meier survival estimates by treatment (prostaglandin trial data)



timing of the failure was not recorded with sufficient precision to differentiate among 'tied' observations. The second is called the **partial calculation** and is based on the assumption that the events were actually tied and treats the problem as a multinomial problem. Exact calculation of the likelihood function under either assumption is very computationally demanding so a number of approximate methods have been developed. The most commonly used is the **Breslow approximation** of the marginal calculation and all examples in this chapter will use this approach.

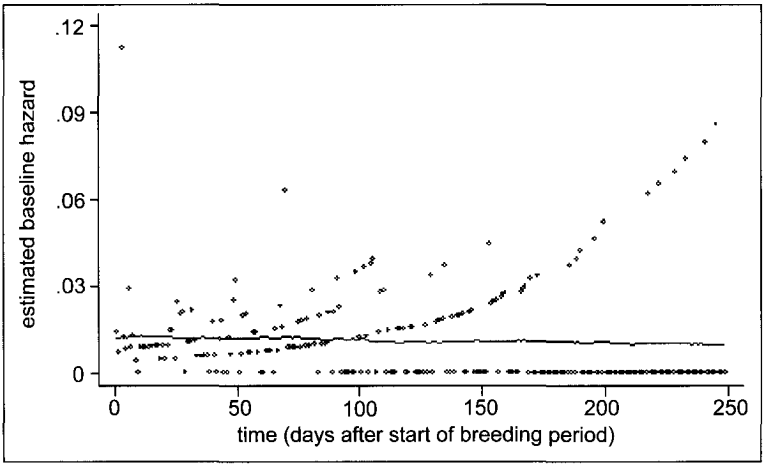
19.8.4 Baseline hazard

Although, as noted above, no assumption is made about the baseline hazard (h_0) and the Cox model does not estimate it directly, an estimate of it can be derived conditional on the set of coefficients estimated model. This baseline hazard represents the hazard in an individual for whom all predictors equal zero. For it to be meaningful, it is important that X=0 is reasonable for all predictors. If computed directly from the data in pgtrial using the model shown in Example 19.4, this would represent the hazard of conception in a non-treated, normal body condition cow in herd 1 in her 0th lactation. To avoid this nonsensical value for lactation, lactation should be modified so that a cow with a value of 0 is possible (*eg* rescale it by subtracting 1 so a cow in lactation 1 now has a value of 0). The contribution to the baseline hazard can only be computed at times (days) at which a failure occurred. If the hazard is assumed to be zero at all other days and a smoothed graph generated, you can get an estimate of the shape of the baseline hazard. Such a graph is shown in Fig. 19.11. The hazard of conception in non-treated, normal-weight, first-lactation cows in herd 1 starts at about 0.015 (1.5%) and then very gradually declines over time.

19.8.5 Model-building

In general, model-building procedures for Cox models are similar to those used for other regression-type models. Wald tests and likelihood ratio tests can be used to

Fig. 19.11 Baseline hazard estimate



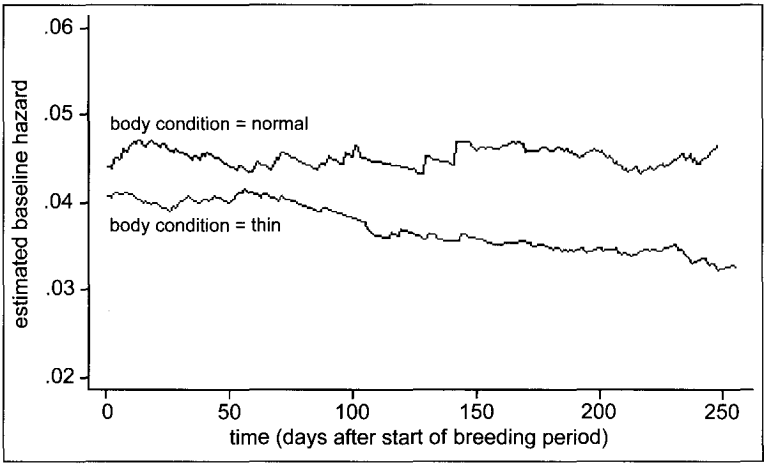
evaluate the significance of individual predictors or groups of predictors. Confounding and interaction can be assessed using methods presented for other regression-type models. Because the explanatory variables are related to the logarithm of the hazard ratio, it follows that interaction will be assessed on a multiplicative scale. There are, however, two issues that are specific to survival models: **stratified analysis** to allow for different baseline hazards in different groups of animals in the study, and the possibility of including **time varying covariates** (see section 19.8.6).

Although we made no assumption about the shape of the baseline hazard, we have assumed that it is appropriate for an animal with all X_k=0. Let's consider the effect of being 'thin' on the hazard of conception in the prostaglandin data. If we obtained a significant *HR* for -thin-, we would assume that it multiplies the h_0 by the *HR* and that this effect was constant over time. If we had reason to believe that the shape of the hazard was different in thin cows than in normal-weight cows, we could stratify the analysis on -thin- and obtain separate estimates of the baseline hazard in each group. Fig. 19.12 shows the results of this evaluation. It plots the kernel smoothed mean estimates of the baseline hazard in the two groups of cows. While non-treated, first-lactation, herd 1, thin cows conceive at a slower rate than comparable normal-weight cows, there is no strong evidence that the baseline hazards in the two groups have different shapes.

19.8.6 Time varying covariates

Up to now, we have focused on exposure factors that do not change their value over time. However, given the long-term nature of many survival studies, it is conceivable that the values of some of those predictors might change over time. For example, in the prostaglandin trial, if the body condition of the cows had been assessed periodically, rather than just once, some cows that were initially thin could have gained enough weight to be classified as normal or vice versa. If the value of a predictor changes during the study period, it is called a time varying covariate, and semi-parametric and parametric survival models are able to appropriately analyse these data.

Fig. 19.12 Baseline hazards (normal and thin cows)



Because there were no time varying covariates in the prostaglandin trial data, we will shift our attention to a study that evaluated the effects of a number of risk factors on the time to occurrence of infectious salmon anemia (ISA) outbreaks in salmon being reared in net-pens in ocean-based aquaculture operations. These data (isa_risk) are from 182 net-pens on 18 sites and are described more fully in Chapter 27. For this example, we will focus on a single predictor: whether or not there was (or had been) another outbreak at the site. At a site with no outbreaks, all records were censored at the end of the study period and there was a single record for each net-pen. For sites where an outbreak occurred, each net-pen would have two records. The first would describe the period up to the date of the first outbreak and would end in a censoring for all net-pens except for the one that had the first outbreak. The second record would span the period from the date of the first outbreak until the cage either had an outbreak or was censored. Example 19.5 shows how the data must be modified to account for a time varying covariate. (Day 0 represents the calendar day at the start of the study period.)

The use of time varying covariates is not limited to the situation in which a covariate changes at discrete time points. Continually varying covariates are used to evaluate interaction terms between predictors and time. In this case, the effect of the predictor changes continually with time. This is discussed further in section 19.8.8 because time varying covariates are one way of evaluating the proportional hazards assumption.

19.8.7 Validating the model

Validation of a Cox proportional hazards model will be covered in the following six sections. The components in the validation process include:

- evaluating the proportional hazards assumption (section 19.8.8)
- evaluating the assumption of independent censoring (section 19.8.9)
- evaluating the overall fit of the model (section 19.8.10)
- evaluating the functional form of predictors (section 19.8.11)
- checking for outliers (section 19.8.12)
- detecting influential points (section 19.8.13).

Example 19.5 Time varying covariate
data=isa_risk

Data were collected on a number of risk factors for outbreaks of ISA in net-pens of salmon at sea-cage sites. The period of risk was considered to start on 1 April 1997 (day=0) and carried on until the fish were harvested in the fall of 1997. Data from three net-pens at site 19 are:

		time		
site	net-pen	start	end	outcome
19	39	0	86	outbreak
19	46	0	211	censored
19	56	0	79	outbreak

Net-pen 46 did not have an outbreak and was censored on day 211. Net-pens 39 and 56 had outbreaks on days 86 and 79 respectively with the outbreak in net-pen 56 being the first outbreak at the site. In order to allow for a time varying covariate to indicate whether or not there had been another net-pen with an outbreak on the site, multiple records for each net-pen need to be created. The resulting data are as follows.

		time			site
site	net-pen	start	end	outcome	positive
19	39	0	79	0	0
19	39	79	86	1	1
19	46	0	79	0	0
19	46	79	211	0	1
19	56	0	79	1	0

Net-pen 39 now has one record for the period of days 0 to 79 during which the covariate (predictor) for the site being positive was 0 and which ended in a censoring. It has a second record for the period from day 79 to day 86 when the site was positive and which ended in an outbreak. Similarly, net-pen 46 has two records (representing the period before and after the site became positive), but both end in censorings because this net-pen did not have an outbreak. Net-pen 56 still has only one record because it was the first outbreak.

A Cox model fit to these data with the single predictor -pos- (*ie* site was positive) produces:

No of subjects = 182 Number of obs = 319
No of failures = 83 LR chi2(1) = 15.24
Time at risk = 28353 Log likelihood = -392.91
Prob > chi2 = 0.0001

Predictor	HR	SE	Z	P	95% CI	
pos	2.610	0.676	3.70	0.000	1.571	4.335

Although it appears that there were 312 observations, the number of subjects is correctly identified as 182. Once a site became positive, the rate of outbreaks in other cages at the site was 2.6 times higher than prior to the site becoming positive.

19.8.8 Evaluating the assumption of proportional hazards

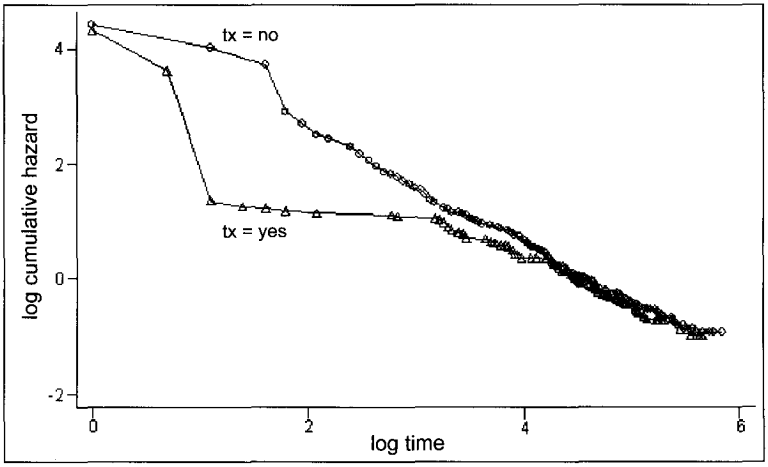
There are three general ways of testing the assumption of proportional hazards:

- graphical assessment
- the use of time varying covariates
- statistical assessment using Schoenfeld residuals.

Graphical assessment

For a categorical predictor, the assumption of proportional hazards can be tested by examining the **log-cumulative hazard plot** ($\ln H(t)$ vs $\ln t$) to check if the lines for the two (or more) study groups are parallel. If they are not parallel, then the assumption has been violated. Fig. 19.13 shows a log cumulative hazard plot for the prostaglandin data. It is clear that the lines are not parallel, at least up to ln(time) $\cong$3.5 (33 days), suggesting that the proportional hazards assumption has been violated. This seems reasonable because we would expect prostaglandin treatment to have a more pronounced effect shortly after administration than many weeks later.

Fig. 19.13 Log cumulative hazard plot by treatment



Time varying covariates

A term for the interaction between the treatment and time (or the log of the survival time) can be added to the model. This makes treatment into a time varying covariate because its effect will be allowed to vary over time. This is the same as saying that the *HR* for treatment changes over time. The effect of treatment can be allowed to vary with time in a linear fashion or to vary with ln(time) (or any other function of time for that matter). In Example 19.6 a Cox model has been fit in which the effect of treatment is allowed to vary with ln(time). The advantage of adding a predictor*time interaction term is that if the assumption of proportional hazards is violated, the addition of the interaction term can solve the problem (provided the change in effect over time can be appropriately modelled).

Example 19.6 Assessing proportional hazards assumption
data=pgtrial

A Cox model with a single predictor (treatment) was fit but the effect of treatment was allowed to interact with time on a log scale. ln(time) was chosen because it was assumed that the effect of treatment would drop off rapidly after administration and then more slowly as time went on (instead of a linear or straight-line decay in effect).

No of subjects = 319
No of failures = 264
Time at risk = 25018

Number of obs = 319
LR chi2(2) = 10.51
Log likelihood = -1307.22
Prob > chi2 = 0.0052

Predictor	HR	SE	Z	P	95% CI	
main effect						
tx	3.085	1.102	3.15	0.002	1.532	6.211
ln(time) interaction effect						
tx	0.759	0.072	-2.92	0.003	0.631	0.913

Treatment is now a significant predictor of time to conception. The treatment*ln(time) interaction term is also significant, confirming that the effect of treatment does vary with time (*ie* the proportional hazards assumption does not hold). In the presence of interaction, the effect of treatment can be better understood by computing the *HR* at a number of time points. The hazard ratio at time t is $3.08*0.759^{\ln(t)}$.

time (days)	ln(time)	HR
1.0	0	3.08
2.7	1	2.34
7.4	2	1.77
20.1	3	1.35
54.6	4	1.02
148.4	5	0.78

The effect of treatment drops off until by day 55, it has completely disappeared.

Schoenfeld residuals

Schoenfeld and **scaled Schoenfeld residuals** are based on the contribution that an observation makes to the partial derivative of the log partial likelihood (which is computed as part of fitting the Cox model). There is a separate set of residuals for each predictor in the model, each set corresponding to the partial derivative for that covariate. These residuals are only computed at observed survival times. Scaled Schoenfeld residuals are adjusted using an estimate of the variance of the residual.

A graph of the scaled Schoenfeld residuals for a given predictor, when plotted against time (or ln(time)) can provide a graphic assessment of the proportional hazards assumption.

This is particularly useful for continuous predictors because the log cumulative hazard plot is not useful for those variables. This graphical assessment can be enhanced by adding a smoothing line to indicate the overall trend. The residuals should hover around the 'zero' line, indicating no trend in the residuals over time. If the residuals trend up or down, it suggests that the effect of the predictor is varying over time. Fig. 19.14 shows a plot of the scaled Schoenfeld residuals for lactation against ln(time). The assumption of proportional hazards appears to be reasonable for this predictor.

Schoenfeld residuals also form the basis for a statistical test of the assumption of proportional hazards. The test checks for a non-zero slope of the scaled Schoenfeld residuals against time (or a function of time) using a generalised linear regression. It provides an overall assessment and a test for each predictor separately. Results of this test for the prostaglandin data are presented in Example 19.7. These suggest that a treatment*time interaction term does need to be added to the model.

19.8.9 Evaluating the assumption of independent censoring

One of the fundamental assumptions of survival models is that censoring is independent of the outcome of interest. This means that censored animals should have the same future survival expectation as non-censored animals (*ie* if the animals were not censored, they would have the same survival distribution as the non-censored animals). There are no specific tests to evaluate the independence of censoring and the event of interest. However, sensitivity analyses can be used to look at the extreme situations of complete positive or negative correlations between censoring and the event of interest.

Complete **positive correlation** would mean that every animal that was censored would have experienced the event of interest immediately if it had not been censored. This could be evaluated by refitting the model after recoding all of the censored observations so that they had the event of interest instead of being censored (at the time of censoring).

Complete **negative correlation** would mean that every animal that was censored would be guaranteed a long 'event-free' existence if it had not been censored. This could be evaluated by refitting the model after changing each censored animal's time at risk to a large, but plausible, value.

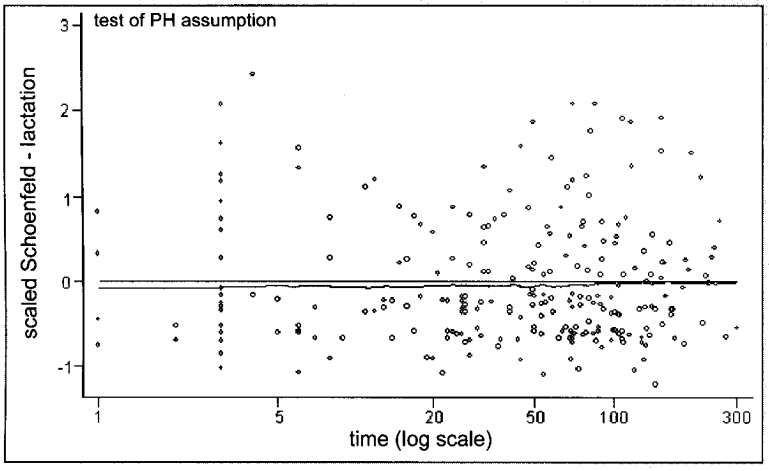
The above two analyses would provide the possible range of values that the coefficients of the factors of interest could possibly take if the assumption of independent censoring was badly violated. If gross violation of this assumption does not drastically alter the estimates of the parameters of interest, you can be confident that the actual bias in the parameter estimates will be small.

Example 19.8 presents the results of a sensitivity analysis designed to evaluate this assumption in the prostaglandin data.

Example 19.7 Testing the proportional hazards assumption
data=pgtrial

A Cox model with herd, treatment, lactation and body condition (-thin-) as predictors was fit to the prostaglandin trial data (without any time varying covariates). Schoenfeld and scaled Schoenfeld residuals were obtained. Fig.19.14 shows a smoothed plot of the scaled Schoenfeld residuals for lactation plotted against time on a log scale.

Fig. 19.14 Schoenfeld residuals for lactation



The statistical test for a non-zero slope for any of the predictors (against ln(time)) resulted in the following.

	rho	chi2	df	prob>chi2
herd=2	-0.03508	0.34	1	0.5591
herd=3	-0.01793	0.09	1	0.7600
tx	-0.16812	7.65	1	0.0057
lact	0.03007	0.28	1	0.5937
thin	-0.07995	1.81	1	0.1789
global test		10.43	5	0.0639

While the global test was borderline significant, it is clear that the assumption of proportional hazards was violated for treatment.

Example 19.8 Assumption of independence of censoring
data=pgtrial

A Cox model with herd, treatment, lactation and body condition (-thin-) as predictors was fit to the prostaglandin trial data (with treatment as a time varying covariate on the ln(time) scale). The model was then refit assuming complete positive and complete negative correlations between censoring and conception (see text for description of method). Negative correlation was based on assigning a -dar- of 400 to all censored cows. The results are summarised in the following table.

Variable	Original estimate	Assuming complete positive correlation	Assuming complete negative correlation
herd=2	-0.260	-0.199	-0.228
herd=3	0.023	-0.007	0.008
tx	1.089	0.983	0.927
lactation	-0.043	-0.006	-0.061
thin	-0.145	-0.141	-0.050
tx*ln(time)	-0.259	-0.209	-0.215

Both sensitivity analyses resulted in a small reduction in the coefficient for treatment, but the change was not large and the treatment effect remained highly significant (P-values not shown).

19.8.10 Evaluating the overall fit of the model

Two approaches to evaluating the overall fit of the model are: to evaluate graphically the distribution of the Cox-Snell residuals and to use a goodness-of-fit test similar to the Hosmer-Lemeshow test used for logistic regression.

Cox-Snell residuals are the estimated cumulative hazard for an individual at its failure (or censoring) time. If the model is appropriate, these residuals are a censored sample from a unit exponential distribution (*ie* an exponential distribution with a mean of zero and variance of 1). Consequently, the range of these residuals is zero to $+\infty$. Cox-Snell (CS) residuals can be used to assess the overall fit of a proportional hazards model by graphically assessing how close these residuals are to having a unit exponential distribution. To do this, you:

- compute the CS residual
- fit a new proportional hazards model with the CS residuals used as the 'time' variable (along with the original censoring variable)
- derive an estimate of the cumulative hazard function ($H(t)$) from this new model
- plot $H(t)$ against the CS residuals

If the residuals have a unit exponential distribution, the cumulative hazard should be a straight line with an intercept of 0 and a slope of 1.

For censored observations, the estimated cumulative hazard is an underestimate

of the true cumulative hazard for an individual (by virtue of the fact that we don't observe them for the full period until they have the outcome of interest). Consequently, Cox-Snell residuals are sometimes modified by the addition of a constant (either 1 or ln(2)=0.693) for censored observations. This is only important if a substantial proportion of the observations are censored. Because approximately 17% of the observations in the prostaglandin trial dataset were censored, this modification could be helpful. A comparison of plots of the cumulative hazard versus **modified Cox-Snell residuals** for models without and with treatment as a time varying covariate (Fig. 19.15) suggest that the latter might fit the data slightly better.

An overall **goodness-of-fit** test similar to a Hosmer-Lemeshow test for logistic regression models can be computed. The observed number of failures in groups defined by quantiles of risk from the fitted model are compared to the expected number of failures which are based on Martingale residuals (see section 19.8.11). However, this test has limited power to detect problems with model fit (see Example 19.9).

19.8.11 Evaluating the functional form of predictors

Martingale residuals can be used to evaluate the functional form of the relationship between a continuous predictor and the survival expectation for individuals. These residuals represent the difference between the observed final outcome for an individual and the cumulative hazard for that individual at the final point in time. (As such, they are more like typical residuals which represent a difference between an observed and a predicted value). Because they are based on the estimated cumulative hazard, these residuals are similar to Cox-Snell residuals except their range is from $-\infty$ to 1. The values of these Martingale residuals are:

- uncensored observations: 1- estimated cumulative hazard
- censored observation: 0 - estimated cumulative hazard

Consequently, residuals will be negative for all censored observations and for observations in which $H(t)>1$ (equivalent to $S(t)<0.37$).

To check for the functional form of continuous predictors, Martingale residuals should be computed from a null model (*ie* one with no predictors included). These residuals can then be plotted against each continuous predictor. A smoothing function (*eg* kernel smoothing) can be used to better visualise the relationship. If the relationship is linear, the smoothed Martingale residual line should be approximately straight. Fig. 19.16 shows a kernel smoothed graph of Martingale residuals against lactation number. It appears that a linear relationship is appropriate.

19.8.12 Checking for outliers

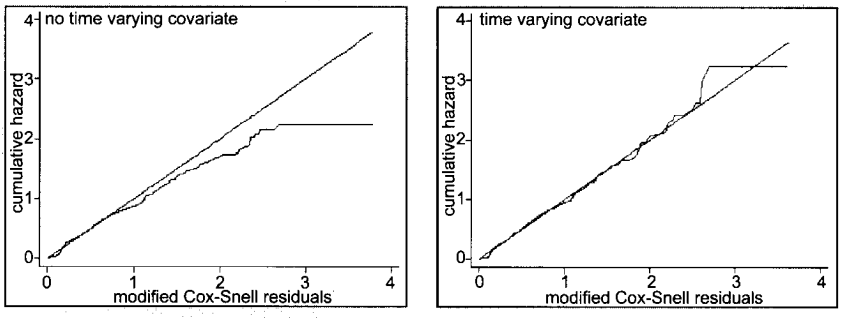
Two types of residual can be used to identify outliers (*ie* points that are not well fit by the model). These are **deviance residuals** and **score residuals** (also known as efficient score residuals).

Deviance residuals are Martingale residuals that have been rescaled so they are

Example 19.9 Evaluating overall fit of a model
data=pgtrial

Two Cox proportional hazards models were fit for the prostaglandin trial data with herd, treatment, lactation number, and body condition (-thin-) as predictors. In the first model it was assumed the effect of treatment was constant over time. In the second model, treatment was allowed to vary with ln(time). Fig. 19.15 shows cumulative hazard versus Cox-Snell residuals for the two models.

Fig. 19.15 a and b Goodness-of-fit plots



It appears that the model with treatment as a time varying covariate fits better.

An overall goodness-of-fit test was carried out with the data divided into five percentiles of risk. The results are shown in the tables below.

Model 1 (treatment no time-varying covariates)

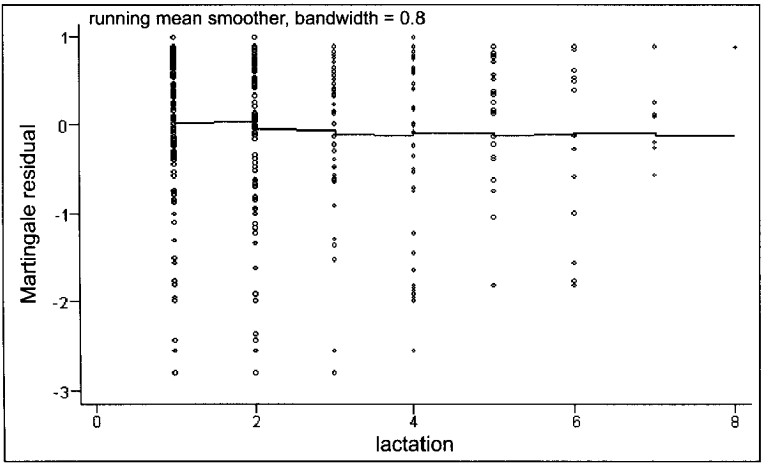
Quantile of risk	Observed	Expected	Z	p-Norm	Observations
1	56	54.287	0.233	0.816	76
2	50	63.217	-1.662	0.096	65
3	50	43.294	1.019	0.308	56
4	58	56.796	0.16	0.873	67
5	50	46.407	0.527	0.598	55
Total	264	264			319

Model 2 (treatment as a time-varying covariate)

Quantile of risk	Observed	Expected	Z	p-Norm	Observations
1	49	43.245	0.875	0.382	64
2	58	64.503	-0.81	0.418	70
3	49	53.491	-0.614	0.539	58
4	58	55.631	0.318	0.751	72
5	50	47.130	0.418	0.676	55
Total	264	264			319

Neither test identified the lack of fit inherent in the first model (without treatment as a time varying covariate). This highlights the limited power of overall goodness-of-fit tests to detect some inappropriate models.

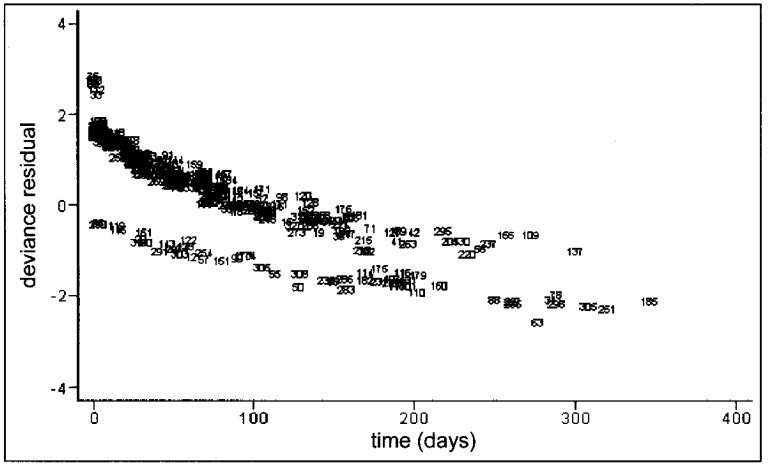
Fig. 19.16 Martingale residuals for lactation



symmetric around 0 (if the fitted model is appropriate). If plotted with an observation number as the plotting symbol, they can be used to identify outlying observations. Fig. 19.17 is a plot of deviance residuals from the model with -tx- as a time-varying covariate. The cluster of large positive residuals at the top left are residuals from 6 cows (33, 37, 68, 75, 103, 112) that conceived on day 1 or day 2 (before the large block of cows that conceived on day 3). The cumulative hazard was low on days 1 and 2 because relatively few cows conceived on those days (relative to the large pool of cows 'at risk' of conception). Hence, for any cow that did conceive, the Martingale and deviance residuals were 'large'.

Score residuals are a variation of Martingale residuals but are computed for each predictor (covariate) in the model. They have a 'leverage-like' property in that

Fig. 19.17 Deviance residuals

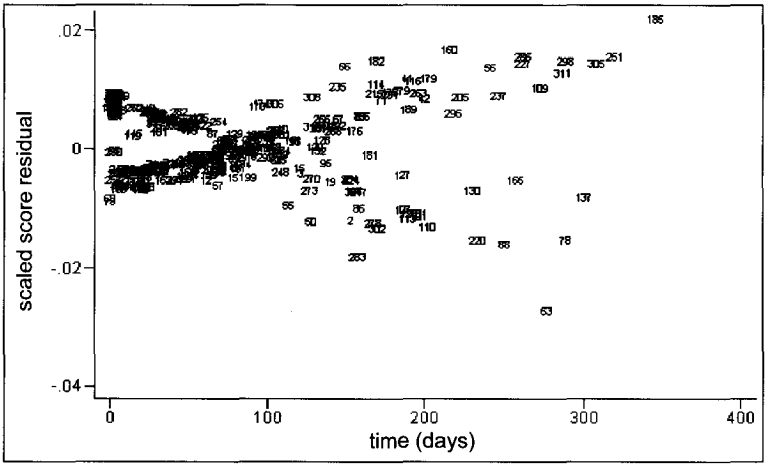


observations that are far from the mean of the predictor have larger (positive or negative) residuals. When plotted against time, they typically form a 'fan-shaped' pattern (with the centre of the fan at the mean of the predictor) and observations lying outside this 'fan' should be considered as outliers. They are most useful for identifying influential and poorly fit subjects.

19.8.13 Detecting influential points

Score residuals can be modified to compute a **delta-beta**-like parameter for coefficients in the model. This modification involves multiplying the score residual by the estimated variance of the coefficient (from the variance-covariance matrix of the coefficients) and produces what is called a **scaled score residual**. A plot of these residuals against time, with an observation identifier as the plotting symbol will identify observations which have a substantial influence on that coefficient. Fig. 19.18 shows a plot of the scaled score residuals for treatment against time. No cows stand out has having a huge influence, but cows 283 and 63 warrant some further investigation. These cows were both censored after long observation periods. They were also both treated. The main effect of these two cows is to reduce the estimated treatment effect.

Fig. 19.18 Scaled score residuals



19.9 PARAMETRIC MODELS

As noted previously, Cox proportional hazards models make no assumption about the shape of the baseline hazard, which can be a real advantage if you have no idea what that shape might be, or if it has a very irregular form. However, these models achieve this flexibility at a price. Because they only use information about the observations at times at which one or more fails, they do not efficiently use all of the information you have about the observations. For example, because the Cox model is based solely on the rank ordering of the observations, it makes no difference if two successive failures

are one day apart or one year apart. The length of the interval, which provides some valuable information in terms of survival times, is ignored. Consequently, if you can correctly specify the form of the baseline hazard, a parametric model will be more efficient (*ie* use more of the available information).

A parametric model could be written in the same way as a semi-parametric model:

$$h(t) = h_0(t)e^{\beta X}$$

but $h_0(t)$ is assumed to have a specified functional form. The major difference is that βX now includes an intercept term (β_0). (An alternative method of writing these models is described in section 19.10).

Two of the most commonly used shapes for $h_0(t)$ are the constant and Weibull forms. Each of these will be discussed briefly.

19.9.1 Exponential model

An exponential model is the simplest form of parametric model in that it assumes that $h_0(t)$ is constant over time (*ie* in the baseline group, the rate at which failures are occurring remains constant). Consequently

$$h(t) = c(e^{\beta X}) \qquad \textit{Eq 19.12}$$

where c is the constant baseline hazard. For any given set of predictor values, $e^{\beta X}$ will also have a unique value, so we will let $c' = c(e^{\beta X})$. Consequently:

$$H(t) = h(t)*t = c(e^{\beta X})*t = c't* \quad \text{and} \quad S(t) = e^{-c'*t} \qquad \textit{Eq 19.13}$$

The survival probability for any individual will have a decreasing exponential distribution.

Interpretation of coefficients

Coefficients for predictors in parametric models are interpreted the same way as coefficients from a Cox model. The exponentiated coefficient is the **hazard ratio** – a measure of the increase (or decrease) in the rate of the outcome that accompanies a unit change in the predictor. The intercept in the model is the estimate of the log of the (constant) baseline hazard. In Example 19.10 an exponential model is fit to the prostaglandin data. If this model was appropriate (which it isn't - but more on that later), the baseline hazard would be estimated to be $e^{-4.41}=0.012$. That is, cows in the baseline group are conceiving at a rate of 1.2% per day. **Note** Lactation was rescaled so that the baseline group was first lactation animals.

Evaluating the assumption of constant hazard

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The first is to generate an estimate of the baseline hazard from a Cox model and graph it to see if it approximately follows a straight, horizontal line. Fig. 19.11 showed that the baseline hazard fell gradually over time. A second approach is to fit a model with a piecewise-constant baseline hazard. In this case, the baseline hazard is allowed to vary across time intervals by including indicator variables for each of the

Example 19.10 Exponential regression
data=pgtrial

An exponential survival model was fit to the prostaglandin data after rescaling -lact- so that first lactation animals had a value of 0.

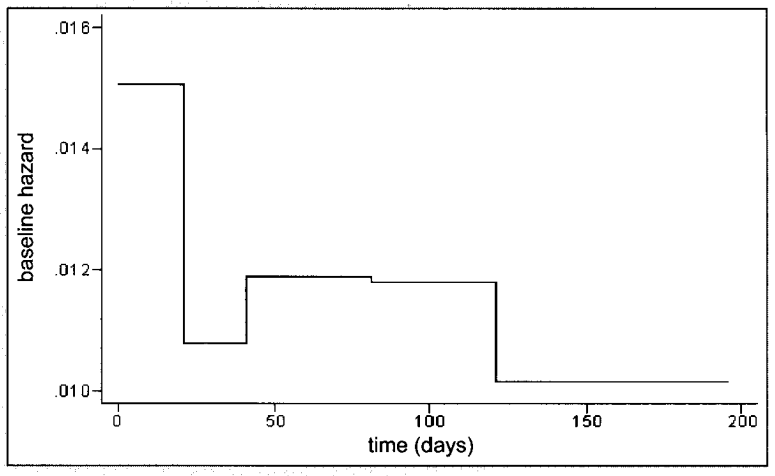
No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(5) = 11.42
Time at risk = 25018 Log likelihood = -528.356
Prob > chi2 = 0.0437

Predictor	Coef	SE	Z	P	95% CI	
herd=2	-0.315	0.169	-1.86	0.063	-0.647	0.017
herd=3	0.038	0.175	0.21	0.830	-0.306	0.381
tx	0.218	0.125	1.74	0.083	-0.028	0.464
lact=2+	-0.041	0.041	-1.01	0.314	-0.123	0.039
thin	-0.157	0.138	-1.14	0.256	-0.428	0.114
constant	-4.405	0.161	-27.28	0.000	-4.721	-4.089

The *HR* for treatment would be $e^{0.218}=1.24$ but it was only borderline significant. The baseline hazard would be estimated to be $e^{-4.40}=0.012$ or 1.2% per day.

The assumption that the baseline hazard was constant was evaluated by fitting a model with a piecewise-constant hazard (*ie* the baseline hazard was estimated separately for the time intervals 0-19, 20-39, 40-79, 80-119 and 120+ days). Fig. 19.19 shows a plot of the estimated baseline hazard up to day 200 (which captures most of the data).

Fig. 19.19 Baseline hazard (piecewise-constant)



From this graph, it is evident that early in the study period, the hazard falls and it might then rise slightly (at least up to day 120) making the assumption of a constant hazard inappropriate.

time intervals in the model. The baseline hazard is assumed to be constant within each time period, but can vary between time periods. This produces the step graph shown in Fig. 19.19. Here it appears that the hazard falls early on, followed by a rise (up to day 120). A third approach to evaluating the assumption of constant hazard is to evaluate the shape parameter from a Weibull model (see section 19.9.2).

19.9.2 Weibull model

In a Weibull model, it is assumed that the baseline hazard function has a shape which gives rise to a Weibull distribution of survival times. The baseline hazard is:

$$h_0(t) = \lambda p t^{p-1} \qquad \textit{Eq 19.14}$$

where λ is the scale parameter and p is the shape parameter. As noted previously these distributions are either monotonically increasing or decreasing (or flat) (see Fig. 19.8) If a vector of covariates (predictors) is added to a Weibull model, the formula for the hazard function becomes:

$$h(t) = \lambda p t^{p-1} e^{\beta X} \qquad \textit{Eq 19.15}$$

where βX includes an intercept term (β_0).

Evaluating the Weibull distribution

As was noted previously, the suitability of the assumption that the survival times follow a Weibull distribution can be assessed by generating a log-cumulative hazard plot. If the distribution is Weibull, this graph will show as a straight line. A rough evaluation can be obtained by generating a simple plot of $\ln H(t)$ vs $\ln(t)$ for all of the data. Fig. 19.13 shows a plot of $\ln H(t)$ vs $\ln(t)$ for each of the two treatment groups in the prostaglandin data. The baseline hazard will be included in the non-treated group and that line was approximately straight suggesting that the Weibull model might be appropriate. The step graph of the baseline hazard (Fig. 19.19) however, suggests that the Weibull model, although preferable to the exponential, might not be ideal because the hazard initially falls and then rises. A Weibull model would assume that it continued to fall with time. The estimate of the shape parameter from the Weibull model gives an indication of whether the hazard is falling ($p<1$), constant ($p=1$) or increasing ($p>1$) with time. If p equals, or is close to, 1, it suggests that the exponential model might be adequate. For the prostaglandin data (Example 19.11), $p=0.867$, indicating that overall, the hazard is falling with time.

19.10 Accelerated failure time models

Parametric models can be written in one of two ways: as a proportional hazards model (which is what has been presented thus far) or as an accelerated failure time model (AFT). Some models (*eg* exponential and Weibull) can be written in either form, while others can only be written as AFT models.

Example 19.11 Weibull model
data=pgtrial

A Weibull model was fit to the prostaglandin data.
No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(5) = 9.96
Time at risk = 25018 Log likelihood = -524.174
Prob > chi2 = 0.0764

_t	Coef	SE	Z	P	95% CI	
_Iherd_2	-0.289	0.169	-1.71	0.088	-0.621	0.043
_Iherd_3	0.038	0.175	0.22	0.825	-0.304	0.381
tx	0.205	0.125	1.63	0.102	-0.041	0.450
lact2	-0.041	0.041	-1.01	0.315	-0.122	0.039
thin	-0.136	0.138	-0.99	0.324	-0.406	0.134
constant	-3.790	0.259	-14.64	0.000	-4.2	-3.282
ln_p	-0.143	0.051	-2.80	0.005	-0.243	-0.043
p	0.867	0.044			0.784	0.958
1/p	1.154	0.059			1.044	1.275

The treatment effect is similar to that seen in the exponential and Cox models and intermediate to those two models in terms of statistical significance. The shape parameter (p) from the Weibull distribution indicates that the hazard is falling with time (*ie* $p<1$).

The general form of an AFT model is:

$$\ln t = \beta X + \ln\tau \quad \text{or} \quad t = e^{\beta X}\tau \qquad \textit{Eq 19.16}$$

where $\ln t$ is the natural log of the time to the failure event, βX is a linear combination of explanatory variables and $\ln\tau$ is an error term with an appropriate distribution. **Note** The values of the βs in this representation will not be the same as the βs in a proportional hazards representation.

From Eq 19.16 it can be seen that τ is the distribution of survival times when $\beta X=0$ (*ie* $e^{\beta X}=1$). τ is assumed to have a specific distribution (*eg* Weibull, log-normal). If τ has a log-normal distribution, then the log of survival times will have a normal distribution which is equivalent to fitting a linear model to ln(survival times) (assuming you can ignore the problem of dealing with censored observations).

Eq 19.16 can be rearranged as follows:

$$\tau = e^{-\beta X}t \quad \text{or} \quad \ln(\tau) = -\beta X + \ln(t) \qquad \textit{Eq 19.17}$$

The linear combination of predictors in the model (βX) act additively on log(time) or multiplicatively on time (*ie* they accelerate or decelerate the passage of time by a multiplicative factor) where $e^{-\beta X}$ is called the **acceleration parameter** because if:

- $e^{-\beta X}>1$, then $t<\tau$ so time passes more quickly (*ie* failures expected sooner)

- $e^{-\beta X}=1$, then $t=\tau$ so time passes at a 'normal' rate (*ie* no effect of predictors)
- $e^{-\beta X}<1$, then $t>\tau$ so time passes more slowly (*ie* failures expected later)

As indicated above, the exponential and Weibull models can be written either as proportional hazards models or as AFT models. Other parametric models (*eg* log-normal, log-logistic, gamma) can only be written as AFT models (the predictors in these models do not necessarily multiply the baseline hazard by a constant amount). The relationship between the coefficients from a proportional hazards expression (β_{ph}) of a Weibull model and an AFT expression (β_{aft}) is:

$$\beta_{aft} = \frac{-\beta_{ph}}{p} \qquad \textit{Eq 19.18}$$

where p is the shape parameter from the Weibull model.

19.10.1 Coefficients in AFT models

A coefficient in an AFT model represents the expected change in the ln(survival time) for a 1 unit change in the predictor. For example, assume you have a dichotomous predictor (X with a coefficient of 1.6. If, in the absence of X, a study subject is expected to fail at t=5 days (ln(t)=1.61), the presence of X would increase the expected ln(survival time) to 1.61+1.6=3.21 or the survival time to 24.8 days. The presence of X in a subject which was expected to survive 30 days would result in an increase expected survival time from 30 to 149 days. As you can see, in absolute time, factors have a greater impact at longer expected survival times.

An alternative interpretation is to exponentiate the coefficient to compute a **time ratio** (TR). A coefficient of 1.6 produces a TR=4.95 which means that the presence of X increases the expected survival time by a factor of almost 5 times.

An example of a log-normal survival model expressed in AFT terms is shown in Example 19.12. Because it appeared that the baseline hazard from the prostaglandin data first declined and then rose, either a log-normal or log-logistic model might be more appropriate than a Weibull model which requires the model to either increase or decrease continually.

19.11 Multiple outcome event data

In all of the material presented so far in this chapter, we have assumed that there was only one possible occurrence of the outcome of interest (*eg* onset of pneumonia in calves, conception in dairy cows). However, in some instances multiple outcome events are possible, and these fall into three general classes.

- Multiple different failure events - These arise in situations where you want to evaluate the effect of a predictor on multiple possible outcomes such as an evaluation of the use of a nutritional supplement in dairy cows after calving on the time to first service, the time to achieving positive energy balance and the

Example 19.12 Accelerated failure time (log-normal) model

A log-normal model was fit to the prostaglandin data and the results are shown here.

No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(5) = 13.85
Time at risk = 25018 Log likelihood = -533.499
Prob > chi2 = 0.0167

Predictor	Coef	SE	Z	P	95% CI	
herd=2	0.293	0.238	1.23	0.217	-0.172	0.759
herd=3	-0.046	0.244	-0.19	0.852	-0.524	0.433
tx	0.061	0.054	-3.08	0.002	-0.045	0.168
lact=2+	0.061	0.054	1.13	0.259	-0.045	0.168
thin	0.064	0.193	0.33	0.740	-0.313	0.441
constant	3.948	0.228	17.28	0.000	3.501	4.40
ln σ	0.423	0.045	9.49	0.000	0.336	0.510
σ	1.526	0.068			1.399	1.666

Expressing the results at time ratios (TR) produces:

Predictor	Time ratio	SE	95% CI	
herd=2	1.341	0.319	0.842	2.136
herd=3	0.955	0.233	0.592	1.542
tx	0.581	0.102	0.411	0.821
lact=2+	1.063	0.058	0.956	1.182
thin	1.066	0.205	0.731	1.555

In this model, treatment is a highly significant predictor of survival time and the TR for treatment suggests that it reduces the time to conception by a factor of 0.58 (*ie* a 42% reduction). σ is the SD of the (assumed) normal distribution of the log survival times. **Note** Although -tx- is now highly significant, this model does not treat -tx- as a time varying covariate. Allowing the effect of -tx- to vary with ln(time) substantially reduced the log likelihood (*ie* the model with -tx- as a time varying covariate fits the data much better). Data from the latter model are not presented, but the program for fitting the model is available in Chapter 28.

time to peak milk production. These are sometimes referred to as competing risks data.

- Multiple 'same' endpoints (not ordered) - These arise in situations where multiple possible outcomes of the same event are possible, but there is not necessarily any ordering to them (*eg* time to onset of clinical mastitis in each of the quarters of a cow). One way of dealing with these is to change the unit of observation to the quarter, but in many cases, most of the risk factors will be at the cow level.
- Multiple 'same' endpoints (ordered) - These are also called **recurrence data**. They arise when it is possible for the outcome event to occur multiple times

in the same animal (*eg* breedings, cases of clinical mastitis). They key feature to these is that there is a natural ordering to them (*ie* the second case cannot happen before the first case). These types of data are the focus of this section.

19.11.1 Models for recurrence data

Three approaches to modelling recurrence data have been reviewed (Wei and Glidden, 1997). These will be summarised here, but details of structuring data appropriately for these analyses is presented in Cleves (1999) and an example is shown here in Example 19.13.

Anderson-Gill (AG) model
This model is a generalised proportional hazard model and is the simplest approach to analysing recurrence data. The risk of recurrence is assumed to be independent of previous events, although the assumption of independence can be relaxed by including a time-dependent covariate for the number of previous occurrences. The model is fit by assuming each subjects 'at-risk' time starts over again after each outcome is observed. If an animal is not considered to be at risk for a defined period after the occurrence of a case, then the time not at risk can be excluded (interval censored or gap). For example, it is common when defining cases of clinical mastitis that 7-14 days elapse between cases for the second occurrence to be considered a new case. An Anderson-Gill model fit to some hypothetical data is shown in Example 19.13.

Prentice-William-Peterson (PWP) model - conditional risk sets model
This model is a proportional hazard model that is conditional on previous occurrences. It is equivalent to carrying out a stratified analysis with the strata defined by the number of previous outcome events. All first occurrences would be in the first stratum, the second stratum would consist of second cases, but only animals that had experienced a first case would be at risk *etc.* Time at risk for each outcome can be measured either from the start of the study period or from the time of the previous event. An example of the latter approach is shown Example 19.13.

Wei-Lin-Weissfeld model - marginal risk sets model
This model is based on an approach similar to that which could be used for survival data with multiple different failure events. Strata are set up for each possible failure event (up to the maximum number observed in the data) and a proportional hazards model fit for each of the strata and then pooled to derive a single estimate of the coefficients in the model. Separate estimates of the coefficients for each stratum could be obtained, but the overall results shown in Example 19.13 provide a single pooled estimate of the effect of the predictor.

In each of the above three models, the multiple observations within an animal are not independent. In Example 19.13, this lack of independence was dealt with by using robust standard error estimates (as proposed by Lin and Wei (1989)). Robust standard errors are described in more detail in Chapter 24. An alternative approach to dealing with the lack of independence is to use a frailty model (section 19.12).

Example 19.13 Multiple failure event models
data=hypothetical data

Some hypothetical data were constructed according to the following specifications

- 2,000 animals
- 1,000 animals with a risk factor (X) and 1,000 without
- X increased the risk of failure (X reduced ln(survival times) by 20%)
- the number of events observed followed a Poisson distribution with a mean of 1.5
- animals with 0 events were censored, all other animals had a censoring time on or after their last event

The data structures for animals 3 and 4 for each of the models fit is shown below. Animal #3 had 3 events (days 31, 48, 61) and was censored on day 66, while animal 4 had a single event on day 54 and was censored on day 94.

			Anderson-Gill		Prentice-William-Peterson		Wei-Lin-Weissfeld	
ID	Event	Outcome	Start	End	Start	End	Start	End
3	1	1	0	31	0	31	0	31
3	2	1	31	48	0	17	0	48
3	3	1	48	61	0	13	0	61
3	4	0	61	66	0	5	0	66
3	5						0	66
4	1	1	0	54	0	54	0	54
4	2	0	54	94	0	40	0	94
4	3						0	94
4	4						0	94
4	5						0	94

The results (Cox proportional hazards model coefficients) from each of the models fit are shown below. There was a single predictor -X- in each model.

Model	Coef	SE	Lower CI	Upper CI
Anderson-Gill	0.481	0.052	0.380	0.582
Prentice-William-Peterson	0.551	0.045	0.463	0.638
Wei-Lin-Weissfeld	0.587	0.060	0.470	0.705

The effect estimates (coefficients) are roughly similar, although the Anderson-Gill model produces a somewhat lower estimate.

19.12 FRAILTY MODELS

As noted in previous sections, predictors in survival models (semi-parametric and parametric) act multiplicatively on the baseline hazard (*ie* the hazard for an individual is a multiple of the baseline function). In a frailty model, an additional latent (unobserved) effect (*ie* the frailty) acts multiplicatively on the hazard. The frailty is not measured directly, but is assumed to have a specified distribution and the variance of the distribution is estimated from the data. The addition of a frailty to the model is useful if the standard model does not adequately account for all of the variation in the data.

There are two general types of frailty model: individual frailty and shared frailty models (Gutierrez, 2002). In an individual frailty model, the additional variance is unique to individuals and serves to account for additional variability in the hazard among individuals in much the same way that the negative binomial model accounts for more variability than a Poisson model. In a shared frailty model, groups of animals are assumed to have a common frailty so this model is analogous to a random effects model (see Chapters 20-24). Each of these will be discussed and examples presented based on hypothetical data and/or those from the prostaglandin trial.

19.12.1 Individual frailty models

An individual frailty model can be written as follows:

$$h(t|\alpha) = \alpha h(t) \qquad \textit{Eq 19.19}$$

Conditional on the frailty, the hazard at any point in time is multiplied by a factor α, which is assumed to have a distribution with a mean of 1 and a variance of θ. Two commonly assumed distributions of α are the gamma and the inverse Gaussian. The frailty can be thought of as representing the effects of unmeasured predictors (which, if they had been measured, would act multiplicatively on the hazard).

A frailty effect can account for apparent changes in the hazard in a population over time. This can best be seen with some simulated data (Example 19.14). In this example, the addition of a frailty term (random effect) to a Weibull model helps explain some of the variability in the unconditional (population) hazard function.

Hazard ratios need to be interpreted with caution in individual frailty models. The *HR* at any time t represents the shift in the hazard due to a unit change in the predictor, conditional on the frailty α (*ie* assuming a comparable frailty). In general, the population hazards might not be proportional over time and the hazard ratio only represents the effect of the predictor at time 0. In general, the effect of the predictor on the population hazard will diminish over time in favour of the frailty effect. With gamma frailties, the population hazard ratio tends to 1 as time approaches infinity, while for an inverse Gaussian frailty the *HR* tends toward the square root of the *HR*.

Example 19.15 shows the addition of a gamma frailty to the Weibull model of the prostaglandin data (with no time varying predictors).

Example 19.14 The effect of frailty on the population hazard
data=hypothetical

A hypothetical dataset of 100 observations was created so that the hazard was assigned the following values in each time period:

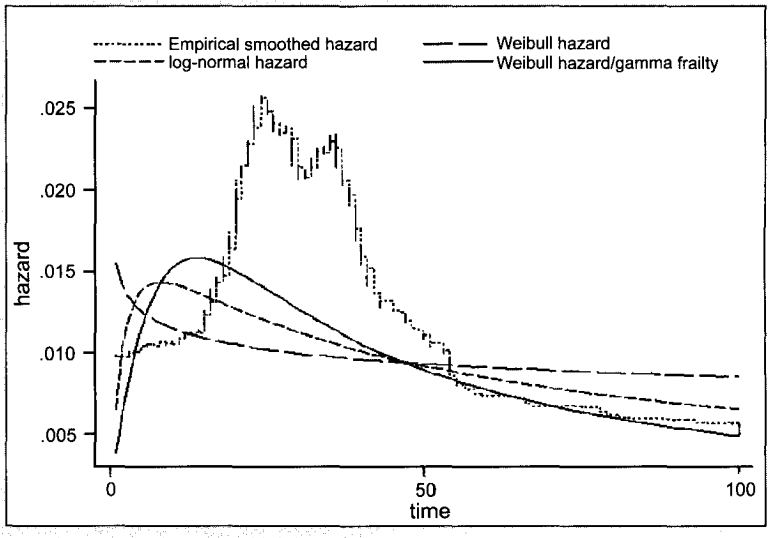
- 1-19 days: hazard = 0.01/day
- 20-39 days: hazard = 0.02/day
- 40-59 days: hazard = 0.005/day
- 60-100 days: hazard = 0.0025/day

Computer-generated random numbers (uniform distribution) were used to determine survival or failure on any given day.

Fig. 19.20 shows the empirical hazard (Kaplan-Meier estimate) from the data along with the predicted hazard from a:

- Weibull model (log likelihood = -1464.6),
- log-normal model (log likelihood = -1426.3)
- Weibull model with a gamma frailty (log likelihood = -1405.5)

Fig. 19.20 Estimated hazard functions



Clearly the Weibull model is inappropriate because the hazard initially increases and then decreases and a Weibull model is restricted to monotonically increasing or decreasing hazards. While none of parametric models fit the data particularly well, the log-normal model would be preferable to the Weibull model. However, the addition of a gamma frailty parameter to the Weibull model accounts for some of the variation in the hazard that is not accounted for by the Weibull hazard function. Based on the graph, and the log-likelihood estimates from each of the models, this model appears to fit the data the best. (**Note** The shape parameter for the Weibull/gamma model is 1.75 indicating a monotonically increasing hazard. The frailty parameter accounts for the decrease in the estimated hazard after day 20).

Example 19.15 Individual frailty model - prostaglandin trial data
data=pgtrial

A Weibull model with a gamma frailty was fit to the prostaglandin trial data.
No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(5) = 9.96
Time at risk = 25018 Log likelihood = -524.174
Prob > chi2 = 0.0764

Predictor	Coef	SE	Z	P	95% CI	
herd=2	-0.289	0.169	-1.71	0.088	-0.621	0.043
herd=3	0.039	0.175	0.22	0.825	-0.304	0.381
tx	0.205	0.125	1.63	0.102	-0.041	0.450
lact=2+	-0.041	0.041	-1.01	0.315	-0.122	0.039
thin	-0.136	0.138	-0.99	0.324	-0.406	0.134
constant	-3.790	0.259	-14.64	0.000	-4.297	-3.282
/ln_p	-0.143	0.051	-2.80	0.005	-0.243	-0.043
/ln_θ	-14.871	756.182	-0.02	0.984	-1496.961	1467.219
p	0.867	0.044			0.784	0.958
θ	0.000	0.000			0	

The variance of the gamma frailty was estimated to be zero (*ie* no frailty effect at all), suggesting that the Weibull hazard might be appropriate for these data.

Shared frailty models

In a shared frailty model, it is assumed that a number of individuals share a common frailty as opposed to the frailty being distinct for each individual. Consequently, the shared frailty can be thought of as representing the effects of unmeasured predictors which those individuals have in common. These can represent the random effect of a grouping variable such as herd. (See Chapters 20-24 for more discussion of random effects). A shared frailty would be an appropriate way of dealing with the lack of independence observed when we have multiple failure times in an individual. (The frailty would represent the common characteristics of the individual that affect time to each event occurrence.)

A shared frailty model can be written as follows:

$$h_i(t|\alpha_i) = \alpha_i h_i(t) \qquad \textit{Eq 19.20}$$

where α_i represents the frailty for the i^{th} group. Example 19.16 shows a shared frailty model fit to the prostaglandin trial data, with the frailty common to the herd. Normally, we would not fit a random effect (shared frailty) when there were only two herds, so this has been done for example purposes only.

Example 19.16 Shared frailty
data=pgtrial

A shared frailty model (Weibull distribution with gamma distributed frailty common to all cows in a herd) was fit to the prostaglandin trial data.

No of subjects = 319 Number of obs = 319
No of failures = 264 LR chi2(3) = 4.79
Time at risk = 25018 Log likelihood = -526.591
Prob > chi2 = 0.1881

Predictor	Coef	SE	Z	P	95% CI	
tx	0.186	0.125	1.48	0.139	-0.060	0.431
lact=2+	-0.044	0.041	-1.09	0.274	-0.124	0.035
thin	-0.115	0.131	-0.87	0.383	-0.372	0.143
constant	-3.853	0.245	-15.73	0.000	-4.333	-3.373
ln_p	-0.147	0.051	-2.88	0.004	-0.247	-0.047
ln_θ	-4.566	1.684	-2.71	0.007	-7.866	-1.266
p	0.863	0.044			0.781	0.954
θ	0.010	0.018			0.000	0.282

The variance of the gamma distribution is significantly different from 1 (P-value for H_0:lnθ=0 (or θ=1) is 0.007). Comparing the likelihoods from models with the frailty term included (lnL=-526.6) and without (lnL=-527.0) (testing H_0:θ=0) suggests that the variance is not significantly different from zero. However, including the frailty term did result in a slight increase in the coefficient for treatment.

19.13 Sample size considerations

Computation of sample sizes for studies with survival time as the outcome can be a complex process. For studies where the primary focus is the comparison of survival times across two (or more) groups, as it often is in controlled trials, one approach is to compute the sample size required to have a desired power in an analysis based on an unweighted log-rank test. If an assumption of proportional hazards is likely not valid, basing the sample size on that required for a weighted version of the test (*eg* Tarone-Ware or Harrington-Flemming tests) might be more appropriate.

However, there are many factors which will influence the required sample size. Some of the following have been discussed under sample size estimation in Chapter 2 and some are unique to studies of survival time.

1. Sample size might need to be increased to account for multiple predictors in the analysis, and/or to adjust for clustering of the data (*ie* non-independence among observations) (see Chapter 2).
2. As pointed out in Chapter 11, multiple comparisons (often arising from interim analyses), losses in the follow-up process and subgroup analyses are common features of controlled trials which require adjustment to the sample size.

3. The shape of the baseline hazard function might not be known in advance of the study so a sample size estimate based on a non-parametric test (*eg* log-rank) would be appropriate.
4. The possibility of non-proportional hazards needs to be considered.
5. In controlled trials, crossover might occur in which animals could move from one treatment group to another (*eg* treated to not-treated if the owner fails to comply with treatment instructions).
6. Recruitment of animals into the study could take place over time which might affect the length of follow-up period for animals recruited.

A general discussion of sample size issues can be found in Friedman et al (1998). A brief discussion of some of the issues identified above and a description of a software program for computing samples sizes for survival analysis studies has recently been published (Royston and Babiker, 2002).

SELECTED REFERENCES/SUGGESTED READING

1. Cleves M. Analysis of multiple failure-time data with Stata. Stata Tech Bull 1999; 49: 30-39.
2. Cleves MA, Gould WW, Gutierrez RG. An introduction to survival analysis using Stata. College Station, TX: Stata Press, 2002.
3. Collett D. Modelling survival data in medical research: texts in statistical science. New York: Chapman and Hall, 1994.
4. Cox DR. Regression models and life-tables (with discussion). J R Stat Soc B 1972; 34: 187-220.
5. Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. Prev Vet Med 2001; 50: 127-144.
6. Friedman LM, Furberg CD, Demets DL. Monitoring response variables. Fundamentals of clinical trials. New York: Springer-Verlag, 1998.
7. Gutierrez RG. Parametric frailty and shared frailty survival models. The Stata Journal 2002; 2: 22-44.
8. Hosmer DW, Lemeshow S. Applied survival analysis. Regression modelling of time to event data. New York: John Wiley & Sons, 1999.
9. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; 53: 457-481.
10. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. J Am Statis Assoc 1989; 84: 1074-1078.
11. Royston P, Babiker A. A menu driven facility for complex sample size calculation in randomized controlled trials with a survival or binary outcome. The Stata Journal 2002; 2: 151-163.
12. Thysen I. Application of event time analysis to replacement, health and reproduction data in dairy cattle research. Prev Vet Med 1988; 5: 239-250.
13. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. Statist Med 1997; 16: 833-839.

SAMPLE PROBLEMS

1. This problem is based on some artificial data ..... when you have completed these analyses ... DO NOT send them off for publication in a peer-reviewed journal.

   You are interested in the efficacy of radiation and chemotherapy as adjuncts to surgery for treating dogs with malignant lymphosarcoma. By working collaboratively with seven veterinary schools in major European cities, you are able to carry out a prospective study in which dogs are randomly assigned to receive chemotherapy, radiation therapy, both or neither. Randomisation is done using a random numbers program at each centre so you do not have exactly equal numbers in all groups.

   You enroll 300 dogs in the study and follow them (as best you can) for the rest of their lives (you have to be patient - this is a long-term study). At the end of the study you have a database (called lympho) with the following variables in it.

Variable	Description
clinic	the clinic identification number (1 .. 7)
dog	the dog's identification number (1 .. 300)
age	the age of the dog in years when it was diagnosed with lymphosarcoma
rad	whether or not the dog received radiation therapy (0=no, 1=yes)
chemo	whether or not the dog received chemotherapy (0=no, 1=yes)
died	whether the dog died (1) or was lost to follow-up (0)
months	the number of months after the start of therapy before the dog died or was lost to follow-up

   a. Use Kaplan-Meier life tables to evaluate the effects of radiation and chemotherapy on the survival expectations of these dogs. Generate a graph of survival times for each of the four possible treatment combinations.
   b. Build a Cox proportional hazards model which includes -age-, -rad- and -chemo-. (Ignore the clinic identifier; we will come back to that later). Is there any evidence of interaction or confounding among the three variables? Would confounding between age and radiation therapy or age and chemotherapy be possible?
   c. Interpret the results from this final model in terms that a lay audience would understand (*ie* pretend you are presenting them to a local kennel club).
   d. Is the assumption of proportional hazards valid for all predictors in the model. Use both graphical and statistical techniques to assess this assumption.
   e. Is the assumption of independent censoring valid?
   f. Evaluate the overall fit of the model. Does it seem reasonable?
   g. Is it reasonable to assume that the effect of age on the log of the hazard ratio is linear?
   h. Are there any dogs whose survival time was not well predicted by the model? What effect does removing these observations have on the model? Should you

remove these dogs?

i. Are there any dogs which seem to have an undue influence on the model? What effect does removing these observations have on the model? Should you remove these dogs?
j. Compute the baseline hazard and survival values. Before doing this, create a new age variable so that a dog of age 24 months has a value of 0. (Otherwise, the baseline hazard that you compute is going to be the hazard for a dog of 0 months of age). Evaluate the baseline hazard and survivor functions by graphing them against time. What shape does the hazard appear to have?
k. Fit a Weibull model to the data. What is the shape of the baseline hazard? Try fitting other baseline hazard distributions (log-normal and log-logistic). Do they fit the data better?
l. Does adding an individual frailty term to the model make much difference to the model?
m. Up to now, we have ignored any potential effect of -clinic-. It is conceivable that dogs at some clinics have better (or worse) survival expectations due to a variety of factors (*eg* breed distributions at the clinics, expertise and skill of the clinicians). Fit a shared frailty model to investigate this possible difference among clinics. What is the estimated variance of this shared frailty term? Is it significantly different from zero (*ie* is there evidence of shared frailty)?
n. Because the baseline hazard is not particularly well fit by any of the distributions explored above, fit a piecewise exponential distribution (*ie* assume a constant hazard over periods of time) to see if you can better fit the observed baseline hazard.

20

INTRODUCTION TO CLUSTERED DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Determine if clustering is likely to be present in your data.
2. Understand why clustering might be a problem, particularly as related to estimating standard errors of coefficients.
3. Understand what impact clustering might have on your analysis of either continuous or discrete data.
4. Use fixed effects models as one approach to dealing with clustering.

20.1 INTRODUCTION

In common usage, a cluster denotes a set of objects (*eg* individuals) in a small group. In statistics, cluster analysis aims to identify clusters among the observations, based on the similarity of their outcomes and possibly their physical distance. Our usage of **clustered data** is similar but does not pertain to cluster analysis. We think of clusters as observations that share some common features (that are not explicitly taken into account by explanatory variables in a model). This type of clustering is always derived from the data structure, of which the most common example is a hierarchical data structure. It is usually expected to lead to dependence between the responses of observations in a group (or cluster) because the shared feature makes the outcomes 'more similar' than otherwise. Thus, two alternative and occasionally encountered terms for these data are **hierarchical data** and **correlated data**.

Before proceeding, recall that statistical dependence between observations (for example, Y_1 and Y_2) is measured by covariance or correlation (which equals the covariance divided by the respective standard deviations):

$$\rho = \text{corr}(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\text{SD}(Y_1)\,\text{SD}(Y_2)}, \qquad \text{where} \quad -1 \le \rho \le 1 \qquad \textit{Eq 20.1}$$

Similarity between observations corresponds to positive values of ρ and the dependence increases the further the value is from zero.

20.2 CLUSTERING ARISING FROM THE DATA STRUCTURE

In this section, we discuss here clustering which arises from animals sharing a common environment, clustering in space (*eg* geographical proximity) and repeated measurements within the same individual.

Common environment

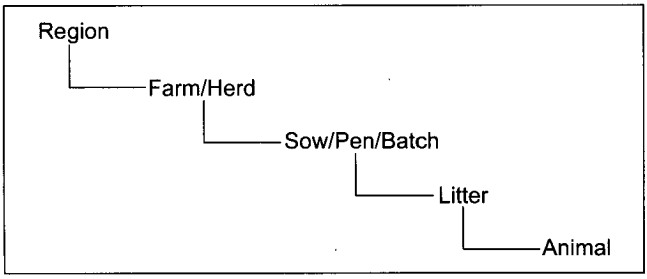
Cows within a herd, puppies within a litter, and quarters within a cow are all examples of clustering in an environment. We usually assume that the degree of similarity among all pairs of observations within such a cluster are equal. Clustering is not necessarily restricted to a single level. For example, pigs might be clustered within a litter which might be clustered within a pen of pigs, which might be clustered in a farm which might be clustered in a region, as shown in the Fig. 20.1. Such data are called hierarchical or multilevel data. The structure shown in Fig. 20.1 is a 5-level structure. In practice, we deal more often with data that have a 2-level or a 3-level structure.

Spatial clustering

The hierarchy in Fig. 20.1 suggests that farms in the same region are similar. It sometimes seems natural to replace or extend this relationship by one where the dependence between farms is directly related to (inversely proportional to) the distance between them. Spatial models incorporate the actual locations of study subjects (in this

example the subjects are farms but they could also be the actual locations of cows in a tie-stall barn). If such detailed information is not available or detailed spatial modelling is not desirable (*eg* due to sparse data), spatial clustering might be accounted for by hierarchical level(s).

Fig. 20.1 A typical hierarchical data structure in veterinary epidemiology



Repeated measurements

Repeated measures arise when several measurements of a variable are taken on the same animal (or other unit of observation) over a period of time. Daily milk weights in a cow are highly correlated because the level of milk production on one day, is likely to be quite close to the production on the day before and the day after. Multiple measurements of lactation total milk production across lactations within a cow are also repeated measurements, but would not be so highly correlated. We might think of repeated measures as a special type of hierarchical clustering (*eg* in Fig. 20.1 an additional level could be added at the bottom of the hierarchy for repeated measurements on the animal). However, just as with spatial clustering, several special considerations apply. Observations close together in time are likely to be more highly correlated than measurements with a longer time span between them. Also, repeated measurements might occur at any level in the hierarchy, not just a the lowest level. For example, if a study on pig production involved several batches within a farm, the batch level would then correspond to repeated measures over time on the farm.

Diagrams such as Fig. 20.1 are highly recommended to determine and present data structures, as long as their defaults with regard to spatial and repeated structures are kept in mind. Note that the data structure pertains not only to the outcome but also to the predictor variables and it is very useful to know whether predictors vary or were applied at particular levels. We elaborate on this idea in the context of the simplest two-level experimental design: the split-plot design. Section 20.2.3 briefly discusses how the effects of predictors vary in their interpretation at the different levels of a hierarchy.

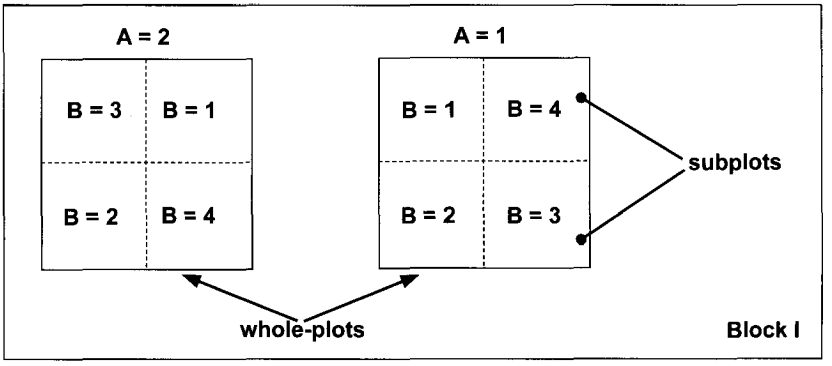
20.2.1 Split-plot design

The split-plot concept and terminology dates back to the early 20th century when statistical methods were developed in the context of agricultural field trials. Consider the planning of an experiment involving two factors A and B with a and b levels, respectively. The special feature of the design is that factor B is practically applicable

to smaller units of land (plots) than factor A. In the field trial context, we might think of A as a large-scale management factor such as pesticide spraying by plane and B as a small-scale factor such as plant variety. The experimental units for factor A are called **whole-plots**. The design needs some replication, and we assume we have c blocks of size a at our disposal, giving a total of ac whole-plots. The blocks would typically be separate pieces of land or experimental sites. A minor modification of the design occurs if the ac whole-plots are not laid out in blocks but are just replicates; the same principle applies, but for simplicity we describe the design with blocks. Within each block, the design would now be laid out in a two-step procedure, as illustrated in Fig. 20.2.

1. randomly distribute the levels of factor A onto the a whole-plots,
2. divide each whole-plot into b **subplots**, and randomly distribute the levels of factor B onto the subplots.

Fig. 20.2 Split-plot layout within one block, with a = 2 whole-plots and b = 4 subplots



As an animal-production example, we might have a herd-management factor A (*eg* tie-stall versus free-stall barns) and a treatment B applicable to individual animals (*eg* vaccination with one of four vaccines). Thus, the whole-plots would be the herds, and the subplots the animals. The blocks could be groups of similar herds, *eg* in the same region, or there could be replication instead of blocks (*eg* pairs of herds – one tie stall and one free stall). A split-plot design corresponds to a 2-level hierarchy with whole-plots as the upper level and subplots as the bottom level.

In the analysis of a split-plot experiment, the two factors A and B cannot be expected to be treated equally because they are applied to different experimental units. In particular, effects of the whole-plot factor A should be compared to the variation between whole-plots (corresponding to the first step of the design construction), and effects of the subplot factor B to the variation between subplots. It follows that it is necessary (and possible!) to split the total variation into variations between and within whole-plots. These variations are estimated independently from each other and with different accuracy (degrees of freedom). Usually the whole-plot variation will be considerably larger than the subplot variation, and factor A is estimated with less precision than factor B. The interaction between A and B 'belongs to' the subplot variation because differences between B-levels within any A-level can be determined within the whole-plots. This makes the split-plot design particularly attractive in situations where the

principal interest is in the main effect of factor B and its interaction with factor A. In the example above, this would correspond to estimating the effects of the vaccines and determining if the vaccines worked differently in tie-stall compared with free-stall barns.

20.2.2 Variation at different levels

The split-plot design with its 2-level structure (*eg* cows within herds) illustrated how variation in the outcome of interest resides at the different levels of the hierarchy and how predictor variables explain variation at these different levels. One important implication is that the amount of unexplained variation at the different levels indicates what can be achieved by a detailed study of the units at the different levels. For example, a large unexplained variation between herds might indicate a substantial room for improvement in the outcome of interest, if we were able to understand why some herds do better than others. Generally, interventions targeted at the level where the greatest variation resides would seem to have the greatest chance of success. Explorative studies prior to interventions are one example of when the clustering of the data within the hierarchical structure is of primary interest (Dohoo et al, 2001).

20.2.3 Clustering of predictor variables

While the focus of our discussion to this point has been in the variation in the outcome of interest, we have also noted that predictor variables occur at various levels and might also be clustered. There is a wealth of potential relationships that can be examined when the hierarchical structure of the data is taken into consideration. For example, if data are recorded at the cow level, but clustered at the herd level we can examine:

- cow-level factors (*eg* lactation number) that affect a cow-level outcome (lactation total milk production),
- herd-level factors (*eg* barn type) that affect a cow-level outcome,
- herd-level factors (*eg* barn type) that affect a herd-level outcome (*eg* average lactation total milk production for the herd),
- cow-level factors (*eg* lactation number) that affect a herd-level outcome (*eg* average lactation total milk production for the herd), where the cow-level factors could either be recorded individually or aggregated to the herd-level (*eg* average lactation number for the herd),
- herd-level factors (*eg* barn type) that might alter a cow-level relationship (*eg* is effect of lactation number on milk production different in tie-stall and free-stall barns?) or vice versa.

Correctly evaluating the range of effects outlined above requires correct identification of the hierarchical structure of the data.

20.3 Effects of clustering

Aside from any interest we might have in questions pertaining to the data structure, the reason for our interest in clustering is that it must be taken into account to obtain valid

estimates of the effects of interest. This is because the assumption of independence inherent in most of the statistical models reviewed up till now in the book will be invalidated by the clustering.

To start with, let us address two common questions: 1. what happens if clustering is ignored?, and 2. if the data show no dependence, can clustering be ignored? If the presumption of these questions is whether one can escape the nuisance of accounting for clustering if it is not 'influential', we must raise a sign of warning. Today's standard statistical software offers a variety of easily accessible options to account for clustering, and we find it hard to justify scientifically the use of a flawed method (even if only slightly) when better methods are readily available. If 'no dependence' means that a significance test of correlation turned out non-significant, it might be worthwhile to recall that the data showing no (significant) evidence against independence is by no means a proof of independence (by the distinction between Type I and Type II errors of statistical tests). Remember, "absence of evidence is not evidence of absence" (Carl Sagan).

Having said that, it might be fruitful for the understanding of the concept of clustering to examine the consequences of ignoring it. Perhaps not too surprisingly, the answer to the question depends on the statistical model used. Linear and logistic regression are discussed in more detail in the sections below. However, one general effect of ignoring clustering is that the standard errors (SEs) of parameter estimates will be wrong and often too small. This is particularly true if the factor of interest is a group-level factor (*eg* a herd-level factor such as barn type), or if it is an individual-level factor that is also highly clustered within groups (*eg* occurrence of milk fever – some cows are affected and others not, but some herds have a lot of cases while others have few). For a two-level structure and a group-level predictor, it is possible to compute a **variance inflation factor** (*VIF*) (section 20.3.3) for a cluster-adjusted analysis relative to an unadjusted analysis.

Unfortunately, the simple *VIF* calculation lead to a widespread, but incorrect, belief that clustering always and only causes variance inflation. The discussion of the split-plot design illustrated the separation of the total variation into variation between and within whole-plots, with different values and degrees of freedom for each level. Therefore, if the data show these variations to be respectively large and small, the cluster-adjusted (split-plot) analysis will actually give smaller standard errors for subplot predictors – and larger standard errors for whole-plot predictors. It also follows that in a dataset with only a few herds (even if there is little clustering within herds), ignoring the hierarchical structure will lead you to grossly overestimate the power for evaluation of herd-level factors because it is the number of herds that determines the appropriate degrees of freedom, not the number of animals within herds. However, accounting for the data structure in the analysis might lead to smaller SEs for an animal-level factor. A final, less clear-cut effect of ignoring clustering is in the weighting of observations from different clusters. If the number of cows in different herds is highly variable, an unadjusted analysis gives unreasonably large weight to large herds. In summary, ignoring clustering can lead to other deficiencies than variance inflation, and in answer to question 2. above, even when 'no dependence' is seen, one would usually want to

use cluster-adjusted methods to properly take into account the data structure.

20.3.1 Clustering for continuous data

Least squares estimation for linear (regression) models yields unbiased estimates of the regression coefficients, even if clustering is present and ignored (Diggle et al 1994). This, perhaps intuitively surprising, fact is however of limited practical use because the corresponding SEs might be strongly affected by ignoring clustering. Thus, without reliable standard errors and test statistics to assess the precision and significance of the estimates, the statistical analysis does not go very far. Also, even if the estimates are unbiased, they might be very inefficient. By means of two simulated datasets, Example 20.1 illustrates how clustering might affect the standard errors. In this example, we use a linear mixed model (Chapter 21) to account for clustering, but other approaches are presented in Chapter 23.

20.3.2 Clustering for discrete data

Estimation procedures in regression models for discrete data (*eg* logistic and Poisson regression models) are asymptotically unbiased which means that with infinitely large samples, they produce correct parameter estimates. However, with limited sample sizes, some bias in the estimates will be present. If the data are clustered and the clustering is ignored in the analysis, the variance of the SEs of the estimates will (in most cases) be underestimated as was seen in models for continuous data. The larger 'true' variance in the parameter estimate means that the parameter estimate might be far from the true value, but this will not be readily apparent, resulting in (apparently) more biased estimates.

The link between the mean and the variance in discrete models makes them quite sensitive to clustering, but also it provides a tool to detect clustering – by comparing the dispersion in the data with the dispersion predicted by the model. As a very simple example, assume that the overall prevalence of disease in a population was 50% and there were no differences among herds (*ie* no clustering within herds). If data were collected from 10 herds of 20 cows each, the herd prevalence values might look like this due to random variation:

.3 .4 .45 .5 .5 .5 .5 .55 .6 .7

The variance of these herd prevalences is 0.012 (σ=0.108). However, if there was 100% clustering within herds and five herds had all cows affected and five had none, the herd prevalence values would be:

0 0 0 0 0 1.0 1.0 1.0 1.0 1.0

Now the variance of the herd prevalence is 0.278 (σ=0.527). Because the logistic model estimates the variance from the overall prevalence, it would estimate the variance to be the same as in the first situation and this underestimates the actual variance because it ignores the between herd variability. Example 20.2 illustrates the practical implication of ignoring clustering for two simulated datasets. In this example, we use a logistic (generalised linear) mixed model to account for the clustering (Chapter 22), but other approaches are described in Chapter 23.

Example 20.1 Clustered continuous data
data=simulated

Two simulated datasets, each consisting of cows in 100 herds, were created. Herd sizes ranged from 20 to 311 cows (μ=116). Herd mean daily milk production varied randomly between herds (μ=30 kg/day, σ_h=7 kg/day). with larger herds tending to have higher production. Individual cow daily milk production values were normally distributed around the herd average (with σ=8 kg/day) unless the factor -X- was present, in which case the milk production was 5 kg higher. The single predictor -X- was added to each dataset with the herd prevalence of -X- varying between datasets. In dataset 1, -X- was a herd-level factor so all cows in 50 herds had X=1 and all cows in 50 herds had X=0. In dataset 2, -X- was a cow-level factor, present in half of the cows in each herd.

For each dataset, two or three models were fit. In the first, an ordinary linear model (a simple two-sample comparison) ignoring herd was fit. In the second, a linear mixed model was used to account for the clustering within herds. In the third, herd average values of milk production were computed and analysed with respect to -X- (also a two-sample comparison); this was only appropriate for dataset 1 in which -X- was a herd-level variable.

Regression coefficients and SEs for analyses of two simulated datasets

		Logistic model		Logistic mixed model		Herd average linear model	
Dataset	Parameter	Estimate	SE	Estimate	SE	Estimate	SE
1: -X- at	-X-	3.557	0.200	3.796	1.495	3.779	1.497
herd level	constant	30.021	0.146	31.137	1.058	31.166	1.059
2: -X- at	-X-	4.982	0.199	4.968	0.149		
cow level	constant	29.257	0.141	30.646	0.728		

In dataset 1, all the estimates for -X- are a long way from the true value (5) but this is due to random variation in the generation of the data. Most importantly, ignoring clustering produces SEs that are much lower than they should be. Controlling for clustering by computing herd average values for milk production and analysing those with respect to presence/absence of -X- produces almost exactly the same values as those observed from the linear mixed model.

In dataset 2, both estimates for -X- are close to the true value because estimation of a cow-level effect is much more precise than a herd-level effect. The linear mixed model gives a **reduced** SE for -X-, because the SE is derived from the within-herd variation which is smaller than both the between-herd variation and the total variation. For the constant (average milk production for cows with X=0 across herds), the correct SE involves the between-herd variation, and when clustering is ignored, the SE is again far too small.

Example 20.2 Clustered binary data
data=simulated

To the same (fictitious) 100 herds as in Example 20.1, a binary outcome -dis- (disease) was added. In both datasets, the effect of -X- corresponded to an *OR* of 2, or a regression coefficient of ln2=0.693 on the logistic scale. The disease level of non-exposed cows was set at P=0.2, corresponding to a value of ln(0.2/0.8)=-1.4 on logistic scale. Herd effects varied on the logistic scale with a standard deviation of 1. As before, dataset 1 had -X- as a herd-level factor (with -X- present in 50 herds), and dataset 2 had -X- as a cow-level factor (with -X- present in 50% of the cows in each herd).

For each dataset, two models were fit: an ordinary logistic regression ignoring herd clustering (a 2X2-table analysis), and a logistic mixed model to account for herd clustering.

Regression coefficients and SEs for analyses of two simulated binary datasets

		Logistic model		Logistic mixed model	
Dataset	Parameter	estimate	SE	estimate	SE
1: -X- at herd level	-X-	0.529	0.042	0.649	0.204
	constant	-1.242	0.033	-1.311	0.146
2: -X- at cow level	-X-	0.586	0.042	0.697	0.046
	constant	-1.250	0.032	-1.361	0.111

In both datasets, the most conspicuous difference between the two analyses is that the simple logistic model underestimates the standard errors. The parameter estimates of the mixed logistic model are somewhat closer to the true value in this case, but the SEs show that it could easily have been the other way around. Note that the SEs for the logistic mixed model in dataset 2 are less than in dataset 1 (because a within-herd design is more powerful than a between-herd design), but still larger than the incorrect SE from the logistic model. This is an effect of the variance in a discrete model being estimated from the mean, and not reflecting the actual variation in the dataset.

20.3.3 Variance inflation as a result of clustering

The effect of clustering on variance estimates can most easily be seen in the situation in which a group (*eg* herd) level factor is being evaluated, but the outcome (*eg* milk production) is measured at the individual (*eg* cow) level. In this case, it is the variance of the herd mean milk production which is important for statistical testing. The magnitude of the effect of clustering on this variance (estimate) depends on both the **intra-class correlation** (*ICC*), and the size of the clusters. The *ICC* is the correlation between two observations within a cluster. If we assume that this correlation (ρ) is the same in all herds, then the variance of a herd mean milk production ($\text{var}(\bar{y})$) for a herd of size m is:

$$\text{var}(\bar{y}) = \frac{\sigma^2}{m}\left[1+(m-1)\rho\right] \qquad \textit{Eq 20.2}$$

where σ^2 is the variance among individual cow milk production values (and the variance is assumed constant within the cluster). **Note** If there is no clustering (*ie* ρ=0), then this formula is the usual one for the variance of a group mean). The quantity $[1+(m-1)\rho]$ is sometimes referred to as the **variance inflation factor** (Wears, 2002). In section 2.10.6 you saw how this quantity can be used to adjust sample size estimates for clustering when computing sample sizes. Table 20.1 shows how both the group size and the magnitude of the *ICC* affect how much the variance needs to be inflated to adequately account for clustering. *ICC*s have been computed for herd-level clustering of a number of infectious diseases and were found to range from 0.04 (*Anaplasma marginale* in cattle) to 0.42 (bovine viral diarrhea in cattle), but most were less than 0.2 (Otte, 1997).

Table 20.1 The effect of group size (m) and the ICC (ρ) on the variance of group means when dealing with clustered data (from Eq 20.2)

ρ	m	VIF	Comment
0	any	1	No within-group clustering = no variance inflation
1	m	m	Complete within-group clustering effectively makes the sample size equal to the number of groups
0.1	6	1.5	A low ICC with a moderate group size can have as much impact as a
0.5	2	1.5	high ICC with a very small group size
0.1	101	11	Large group sizes, even with a low ICC, result in a very high variance inflation factor

Finally, a few notes on the use of *VIF*s. First, they apply to cluster means and therefore more generally to between-cluster effects, but not to within-cluster effects. Second, because the *VIF*s depend only on the *ICC* and the cluster size, they are equally applicable to discrete as continuous outcomes. However, the underlying assumption of equal variances within each cluster will not hold for discrete data with within-cluster predictors, because the variance varies with the prevalence of the outcome (which will change depending on the distribution of within-cluster predictors).

20.4 INTRODUCTION TO METHODS OF DEALING WITH CLUSTERING

The next three chapters of the book deal with statistical models involving clustering. Our primary focus is on **mixed**, or **random effects models** which, with recent advances in computer software and power, have become widely accessible. These models are reviewed for continuous and discrete data in Chapters 21 and 22, respectively. Many more methods exist, and a few of these are briefly reviewed in Chapter 23. In addition, frailty models for dealing with clustering in survival data are introduced (briefly) in section 21.5. We also revisit the repeated measures and spatial data structures that only partly fall within the hierarchical framework, in section 22.6. Among the many special approaches for discrete data, we cover estimation using **generalised estimation equations** (GEE) in section 23.3. The present section contains some introductory remarks on detection of clustering, and a discussion of simpler, traditional approaches to dealing with clustering using fixed effects and stratification.

20.4.1 Detection of clustering

The primary resource for detection of clustering is the researcher's awareness. Whenever data are collected from individuals that are managed in a group, we should suspect that the data might be clustered. More generally, this is the case whenever animals share common features of potential impact that are not accounted for by explanatory variables. Any hierarchical structure of the origin or management of individuals might introduce clustering, as shown in Fig. 20.1. Also, repeated measures and spatial data structures should always be noticed and examined.

One might expect some general statistical test for clustering to be 'standard' (in common use), but this is not so. We offer two explanations. One is that clustering is dealt with differently in discrete and continuous data, and in different statistical models. One general approach is to extend a statistical model with an additional parameter (or effect) for clustering, estimate that parameter and test whether it differs significantly from zero (no clustering). This approach has been introduced in Chapter 18 where addition of an extra variance parameter to the Poisson model, produced a negative binomial model. In discrete models such as logistic and Poisson regression, one might also compare the actual variation in the data with the expected variation (from the binomial or Poisson distributions) by a goodness-of-fit statistic, which, if significant, indicates **overdispersion**, potentially a result of clustering. A second reason why testing for clustering is less common than one might expect, is that even small amounts of clustering might have substantial impact on variance estimates, as illustrated in section 20.3.3. Therefore, one is often inclined to keep a clustering effect in the statistical model even if it is not statistically significant, particularly if it shows 'some effect' and is strongly suggested by the data structure.

20.4.2 Fixed effects and stratified models

As indicated above, methods for dealing with clustering will be covered in more detail in the next three chapters. However, we will first discuss one simple and previously common approach to dealing with clustering which has been used in previous chapters of this book – that is to include the group identifier as a **fixed effect** in the regression model. Let us for the sake of the discussion, refer to the groups as herds and the within-groups subjects as cows. In fixed effects models, dummy (indicator) variables representing the 'group' (*eg* herd) are included in the model. The fixed effects analysis then effectively estimates a separate parameter for each herd. This has the effect of separating the variation between herds from the residual variation and results in more appropriate tests of significance for within-herd factors.

There are several major drawbacks to this approach. The first is that one cannot include any herd-level predictors (*eg* barn type) in the model because they will be included in the herd effects. The second drawback is that the model does not contain any dependence between cows in the same herd (*ie* the model contains only the within-herd variance as the between-herd variance is removed by the fixed effects), and therefore does not properly inflate the variance on means across herds (*eg* the calving to conception interval for heifers treated with a certain vaccine). Another way

of saying this is that any inferences made are specific to the actual herds, where very often one would want conclusions to refer to a more general population of herds. A third drawback is that with many herds it requires the fitting of a large number of parameters (one for each herd), and the parameter estimates in the model might become unstable if there are relatively few observations per group. Because we are not usually interested in the actual effects of each herd these fixed effects are often considered 'nuisance' parameters. A fixed effect model based on the data presented in Example 20.1 is shown in Example 20.3.

Example 20.3 Fixed effects model for continuous data
data=simulated

For dataset 2, from Example 20.1, with -X- as a cow-level factor, a linear model was fit with effects of -X- and herds (essentially, a two-way ANOVA). The 99 coefficients for the herds 2-100 are not shown below.

Regression coefficients and SEs for fixed effects of a simulated dataset

	Linear mixed model		Fixed effects linear model	
Parameter	Estimate	SE	Estimate	SE
-X-	4.968	0.149	4.968	0.149
constant	30.646	0.728	24.324	1.800

The estimates and SEs for -X- are identical to those from the linear mixed model in Example 20.1. In the fixed effects model, the constant corresponds to the mean of herd 1, and therefore differs from the overall mean (across all herds) from the linear mixed model.

The fixed effects approach applies equally to discrete models, and has the same drawbacks. As most discrete models use an iterative estimation procedure, the consequences of having a large number of 'nuisance' parameters in the model might be more serious than for normal distribution models. The fixed effects approach is illustrated in Example 20.4 for the previously used binary dataset.

Another simple approach to dealing with clustered binary data and a dichotomous within-herd factor is to carry out a **stratified analysis** using the Mantel-Haenszel procedure described in Chapter 13, with strata defined by the clustering variable. Such an analysis, based on the data from Example 20.2, is also shown in Example 20.4.

Despite the above-mentioned drawbacks of fixed effects modelling, it might still be a useful approach to account for herd-level clustering, particularly when:

i. there are no herd-level predictors,
ii the number of herds is reasonably small, and
iii there is more interest in the specific herds than assuming they represent a population.

Example 20.4 Stratified analysis and fixed effect model for binary data
data=simulated

For dataset 2 from Example 20.2 with -X- as a cow-level factor, the crude *OR* for -X- was $e^{0.586}$=1.797 from Example 20.2, and Mantel-Haenszel combined *OR* was $e^{0.698}$=2.009. Furthermore, a logistic model was fit with fixed effects of herds and -X-. The 99 coefficients for the herds 2-100 are not shown below.

Regression coefficients and SEs for fixed effects and stratified analyses of a simulated binary dataset

	Logistic mixed model		Logistic fixed effects model		Stratified MH analysis	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
-X-	0.697	0.046	0.704	0.046	0.698	0.046
constant	-1.361	0.111	-2.130	0.632		

Both estimates and SE for -X- from the MH procedure and the fixed effects model are very close to the results of the logistic mixed model.

SELECTED REFERENCES/SUGGESTED READING

1. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. Oxford: Oxford University Press, 1994.
2. Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. Prev Vet Med 2001; 50: 127-144.
3. Otte MJ, Gumm ID. Intracluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys. Prev Vet Med 1997; 31: 147-150.
4. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. Academic Emergency Medicine 2002; 9: 330-341.

SAMPLE PROBLEMS

1. Open the pig dataset and evaluate the relationship between worm count and average daily gain.
   a. First fit a simple linear regression. What is the effect of-worms- on -adg-.?
   b. Next, control for 'farm' by including them in the model as fixed effects.
      i. What happens to the coefficient for worms? Why?
      ii. What happens to the SE of the coefficient for worms? Why?
   c. What is the relationship between -worms- and -adg- at the farm level? (Hint - collapse the data so you have one record per farm with the farm average worm counts and daily gains?
      i. Does this help explain any of the things you observed in 1.b.?
2. Using the same dataset, explore the relationship between atrophic rhinitis (-ar-) and pneumonia (pn=lu>0).
   a. First fit a simple logistic regression. What is the effect of -ar- on -pn-?
   b. Next, control for 'farm' by including them in the model as fixed effects.
      i. What happens to the coefficient for -ar-? Why?
      ii. What happens to the SE of the coefficient for -ar- ? Why?
   c. What is the relationship between -ar- and -pn- at the farm level?
      i. Does this help explain any of the things you observed in 2.b.?

21

MIXED MODELS FOR CONTINUOUS DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Write an equation for a model that contains both fixed and random components.
2. Compute the variance for each level of a multilevel model.
3. Determine how highly correlated observations are within a cluster.
4. Determine if predictors have the same (fixed), or different (random slopes) effects across clusters.
5. Compute the variance of the outcome (a complex function) in models containing random slopes.
6. Evaluate the statistical significance of fixed and random effects in a model.
7. Evaluate residuals from a multilevel model.
8. Determine the optimum Box-Cox transformation for the outcome in order to normalise the residuals from a model.
9. Choose among a variety of correlation structures that might be appropriate for repeated measures or spatial data.

21.1 INTRODUCTION

Mixed models (for continuous data) contain parameters or effects of two types:

- **fixed**, or mean effect, such as ordinary regression coefficients in a linear regression model (Chapter 14),
- **random**, or variability around the mean effect, explaining some of the error term.

Mixed models can be used to take into account that the data have a hierarchical or multilevel or nested structure, and are sometimes referred to by these terms. Although other methods exist for analysing hierarchically structured data, the use of mixed models has become a popular choice during the last decade, due to advances in computing power. Mixed models also apply to many other data structures, but our focus is on hierarchical data and we only briefly discuss repeated measures and spatial data (section 21.5). Mixed models are also known as **variance component models**. Variance components are the technical/mathematical constructs used to decompose the variance (variation, variability) in a dataset into (a sum of) several components that can each be given a useful interpretation.

The dataset scc_40 (described in more detail in Chapter 27) is used to illustrate the methods numerically. It is comprised of data from 40 herds selected from a much larger dataset that was collected to study problems related to mastitis and milk yield. We will take the (log) somatic cell counts (SCC) as the outcome. The data structure is 3-level hierarchical: 14,357 tests within 2,178 cows within 40 herds. The tests were performed on each cow approximately monthly throughout one lactation, and thus constitute **repeated measures** per cow. In this section we include only a single test per cow, the first test recorded in the cow's lactation. This gives a 2-level structure of the 2,178 cows in 40 herds; herd sizes ranging from 12 to 105. The 2-level dataset is denoted scc40_2level. Obviously, any inferences to real associations of predictors with the outcome should not be based on results from such subdatasets. The variables used in the examples in this chapter are listed below. For clarity, we use the term season for the quarters of the year without claiming to infer any seasonal effects from two years of data.

Table 21.1 Selected variables from the dataset scc_40

Variable	Level of measurement	Description
herdid	3:herd	herd identification
cowid	2:cow	cow identification
test	1:test	approximate month of lactation for test: 0,1,2,...,10
h_size	3:herd	herd size (averaged over study period)
c_heifer	2:cow	cow parity with values 1 (heifer) and 0 (older cow)
t_season	1:test	season of test with values 1 (Jan, Feb, Mar), 2, 3 and 4
t_dim	1:test	days 'in milk' (since calving) on test day
t_lnscc	1:test	(natural) log of somatic cell count

21.2 LINEAR MIXED MODEL

Linear mixed models extend the usual linear regression models (Chapter 14) of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_i, \quad i = 1,...,n \qquad \textit{Eq 21.1}$$

We will take as our outcome Y the log somatic cell counts and as our regressors $X_1,...,X_k$ the continuous and dummy variables necessary to represent the chosen predictors. Further, the errors $\varepsilon_1,...,\varepsilon_n$ are assumed independent and ~ $N(0,\sigma^2)$. This equation (and its assumptions) would be meaningful if we considered one test per cow and there was no clustering in herds (*eg* we might have data from only one herd). It is worth noting that, in this model, the observations $Y_1...$, Y_n are independent and all have the same variance:

$$\text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2$$

So far, there is no trace of variance components. However, in reality we have recordings in several (40) herds, and we would like the herds to enter our model as well, because we know that there might be some variation of cell counts across herds. Previously, we have included herds in the model by including a set of (40-1) indicator variables and estimating a separate β for each of them. A **mixed model** with a **random herd effect** is written:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_{\text{herd}(i)} + \varepsilon_i \qquad \textit{Eq 21.2}$$

Note For the sake of simplicity, a single index notation will be used for all multilevel data. The subscript i denotes the individual (lowest level) observation. In the example above, $u_{\text{herd}(i)}$ refers to the herd containing the i^{th} individual. If there are 40 herds, u could have one of 40 values. An alternative notation uses multiple indices such as $u_j + \varepsilon_{ij}$ where j refers to the herd and i to the i^{th} individual in the j^{th} herd.

The explanatory variables and the β-parameters in Eqs 21.1 and 21.2 are unchanged. These are usually termed the **fixed effects**, in contrast to the last two terms which are **random effects**. The only new term in Eq 21.2 is $u_{\text{herd}(i)}$, a random herd effect for the herd of the i^{th} cow (*eg* u_7 for cows in herd 7). Thus, in total we have 40 random effects: u_j, j=1,...,40 in the model. Random simply means that it is modelled as a random variable, in contrast to a fixed parameter (according to a 'frequentist' or non-Bayesian view; see section 23.4 for further discussion and the alternative Bayesian approach). Let's defer the question as to why we model herd as a random term for now, and first look at the assumptions for u and ε:

$$u_j \sim N(0, \sigma_h^2), \quad \varepsilon_i \sim N(0, \sigma^2)$$

where all u_j and ε_i are independent.

Thus, we assume the impact of each herd to be a random fluctuation with mean zero (and consequently centred at the mean determined by the fixed effects) and standard deviation σ_h. Therefore, the parameter σ_h^2 can be interpreted as the overall random variation in log cell counts between herds. Furthermore, we could calculate:

$$\text{var}(Y_i) = \text{var}(u_{\text{herd}(i)}) + \text{var}(\varepsilon_i) = \sigma_h^2 + \sigma^2 \qquad \textit{Eq 21.3}$$

In effect, we have decomposed the total variance to a sum of the variance between herds and the error variance (or the variance within herds). The σ^2s are the variance components; Example 21.1 shows how they might be interpreted.

Example 21.1 Variance components and random effects
data=scc40_2level

This dataset contains one observation from each of 2,178 cows from 40 herds. In a 2-level random effects model for -t_lnscc- with no fixed effects (a 'null' or 'empty' model'), the variance components were estimated at:

$$\sigma_h^2 = 0.148 \quad \text{and} \quad \sigma^2 = 1.730$$

Thus, the total (unexplained) variance was 0.148+1.730=1.878. It is often useful to compute the fractions at the different levels; here we have 0.148/1.878=7.9% of the variance between herds and 84.3% within herds. We can also give a direct interpretation of σ_h^2: 95% of the herd effects should be within an interval of $\pm 1.96\sigma_h = \pm 0.754$. As the overall mean ($\beta_0$) was 4.747, this means that most herd mean -lnscc- values lie between 3.993 and 5.501.

Random effects modelling of herds can be motivated in different ways. Strictly speaking it corresponds to effects (herds) in the model being randomly selected from a population. Sometimes, in a study, this could be the case, but it might be reasonable to assume that the herds are generally representative of the population even if they were not randomly selected. In our example, the 40 herds were randomly selected from the full set of study herds, which constituted all milk-producing herds in a certain geographical area of Denmark. Consequently, these 40 herds were representative of this region. With random effects, the focus shifts from the individual herd to the variability in the population σ_h^2. In a study with only a few herds of particular interest (possibly because they were individually selected for the study), one might prefer to model herds by fixed effects (*ie* β-parameters) instead (see sections 20.4.2 and 23.2.1).

Mixed models can be used to take into account more general hierarchical data structures by **inserting random effects for all levels** above the bottom level (which is already present in the model as the error term ε). For example, a 3-level structure with animals in herds in regions would lead to random effects for both herds and regions and we then split the variation into three terms: $\text{var}(Y_i) = \sigma_r^2 + \sigma_h^2 + \sigma^2$. Note, however, that modelling the somatic cell count dataset with multiple tests per cow in the lactation in a similar model with random effects for both cows and herds causes problems due to the long series of repeated measures on each cow (section 21.5). In mixed models, the predictors might reside at any level of the hierarchy. As a particular example, the split-plot design (section 20.2.1) could be analysed by a mixed model with random effects for the whole-plots. In epidemiology, we often work with datasets in which predictors explain

variation at several levels (section 20.2.2); the mixed model analysis takes this into account. Finally, the one exception to the 'random effects for every level' rule is that the top level could be modelled by fixed effects, if (and only if!) there are no predictors at that level. As discussed above, this situation often occurs when the top level (*eg* herd or region) is not a random sample of a larger population and does not have a large number of elements. Example 21.2 shows some of the possible changes to a linear mixed model when fixed effects are included.

Example 21.2 Mixed model estimates for 2-level somatic cell count data
data=scc40_2level

A linear mixed model with herd size, heifer, season and days in milk was fit to the 40-herd, 2-level scc data (see Table 21.1 for codes).

	Coef	SE	t	P	95% CI	
h_size (*100)	0.408	0.377	1.08	0.286	-0.355	1.172
c_heifer	-0.737	0.055	-13.3	0.000	-0.845	-0.628
t_season = spring	0.161	0.091	1.78	0.076	-0.017	0.339
t_season = summer	0.002	0.086	0.02	0.986	-0.168	0.171
t_season = fall	0.001	0.092	0.02	0.987	-0.179	0.182
t_dim (*100)	0.277	0.050	5.56	0.000	0.179	0.375
constant	4.641	0.197	23.5	0.000	4.245	5.038

Note that, because of the random herd effects, the constant refers to the log somatic cell count in an average herd, not to the value of an average cow across the population of cows. As herds differ in size, these means are not necessarily the same. For example, if the highest cell counts were obtained in the largest herds (even if the -h_size- estimate hardly indicates this to be the case), then the cow average would typically be higher than the herd average. The cow and herd averages are analogous to weighted and unweighted averages in multistage sampling (section 2.8). The other regression coefficients are interpreted in the usual way.

In addition, the estimated variance components (also with standard errors (SEs)) were:

$$\sigma_h^2 = 0.149\,(0.044) \quad \text{and} \quad \sigma^2 = 1.557\,(0.048)$$

In a linear regression model, adding predictors always reduces the unexplained variation. Intuitively, one would expect a similar effect in a mixed model at the levels affected by added predictors. But, by comparison, in Example 21.1, we note a reduced value for σ^2 and a slightly increased value for σ_h^2. It is not unusual that adding fixed effects to hierarchical models redistributes the variation across the levels and thus increases some of the variance components and, sometimes, even the total variation (the sum of all variance components). No simple intuitive explanation can be offered; see Chapter 7 in Snijders and Bosker (1999) for details and ways of defining measures of the variance explained by fixed effects.

21.2.1 Intra-class correlation coefficient

The model assumptions allow us to examine the dependence or correlation between observations from the same herd. In a linear model, all observations are independent, but in mixed models this is no longer so. The correlation between observations within the same group (in our example, herd) is described by the intra-class correlation coefficient (*ICC* or ρ). For a 2-level model (21.2), the *ICC* equals the proportion of variance at the upper level; from Example 21.1:

$$\rho = \frac{\sigma_h^2}{\sigma_h^2 + \sigma^2} = \frac{0.148}{0.148 + 1.730} = 0.079 \qquad \textit{Eq 21.4}$$

Thus, a low *ICC* means that most of the variation is within the groups (*ie* there is very little clustering), while a high *ICC* means that the variation within a group is small relative to that between groups.

Generally in mixed models with homogeneous variances and independent random effects, correlations are assumed to be the same between any two observations in a group and can be computed by a simple rule. Recall (Eq 20.1) that the correlation is the ratio between the covariance of the two observations in question and the product of their standard deviations. As all observations have the same variance, the denominator of this ratio is always the total variance, *ie* the sum of all variance components. The numerator is obtained by noting which random effects are at the same level for the two observations in question, and summing the respective variance components. For the 2-level model, this rule gives Eq 21.4 for observations in the same group and zero correlation for observations in different groups. If region was added as a third level to the model, the correlation between cows in the same herd (and hence within a region) would be:

$$\rho \text{ (cows in same herd)} = \frac{\sigma_r^2 + \sigma_h^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2} \qquad \textit{Eq 21.5}$$

Similarly, the correlation between cows in different herds in the same region would be:

$$\rho \text{ (cows in same region, but different herds)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2} \qquad \textit{Eq 21.6}$$

Example 21.3 shows similar computations for a 4-level model. The correlation in Eq 21.6 referred to cows in different herds but an intuitively more appealing value might be the correlation **between herds** – more precisely, **between herd means**. The correlation between means of two herds of size m is

$$\rho \text{ (herds of size } m \text{ in same region)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2 / m} \qquad \textit{Eq 21.7}$$

When m is large, the contribution of σ^2/m to the formula is small and might be ignored (see Example 4.7 of Snijders and Bosker, 1999 for further discussion).

Example 21.3 Intra-class correlations in a 4-level mixed model
data=reu_cfs

Dohoo et al (2001) used 4-level mixed models to analyse the (log) calving to first service intervals for cattle in Reunion Island. Their model had several fixed effects which we denote $X_1,\ldots,X_k$, so that the model could be written:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{cow}(i)} + v_{\text{herd}(i)} + w_{\text{region}(i)} + \varepsilon_i$$

The variance components for the unexplained variation were:

$$\text{region}: \sigma_r^2 = 0.001, \quad \text{herd}: \sigma_h^2 = 0.015, \quad \text{cow}: \sigma_c^2 = 0.020, \quad \text{lactation}: \sigma^2 = 0.132$$

The fact that the first three variance components were small once again points out that there is little similarity (in terms of calving to first service interval) between lactations within a cow, between cows within a herd or between herds within a region. In the original study, the authors suggested that management of reproductive performance should focus on individual lactations within individual cows, because this is where most of the unexplained variation resided.

From the estimates we could compute a total variance of 0.168 and the following correlations between observations (lactations):

lactations of the same cow:	$\rho = (0.001 + 0.015 + 0.020) / 0.168 = 0.214$
lactations of different cows in the same herd:	$\rho = (0.001 + 0.015) / 0.168 = 0.095$
lactations of cows in different herds in the same region:	$\rho = 0.001 / 0.168 = 0.006$

21.2.2 Vector-matrix notation

Notation involving vectors and matrices allows us to write the linear and linear mixed models in a compact and clear form. The linear regression model (Eq 21.1) can be written

$$Y = X\beta + \varepsilon$$

where Y, β and ε are (column) vectors and X is the so-called design matrix, comprised of a column of 1s followed by the k columns containing the values of the k predictors of the model. (**Note** Our usage of X_{ji} for the element in the i^{th} row and j^{th} column of X contrasts usual matrix notation but is of no serious consequence because we do not pursue any computations with matrix notation.) Similarly, linear mixed models such as Eq 21.2 can generally be written as:

$$Y = X\beta + Zu + \varepsilon \qquad \textit{Eq 21.8}$$

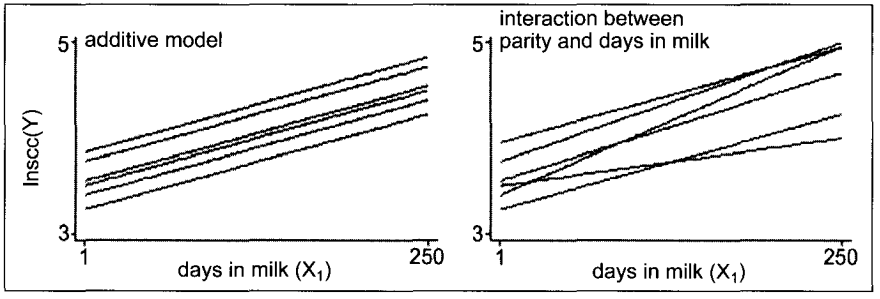
where u is a vector of all random effects (except for ε) and Z is the design matrix for the random part of the model. Our assumptions for this model (up to section 21.5) are that all random variables are normally distributed with mean zero, and that all the errors are independent, have the same variance and are independent of the random effects.

21.3 RANDOM SLOPES

21.3.1 Additive and non-additive modelling

Before turning to the extension of the mixed model (Eq 21.2) with a random slope, we consider in more detail one implication of the model assumptions. Let's focus on a quantitative explanatory variable, for instance, days in milk. Assume these values to be in X_1, and assume the model has a linear term for X_1 with a positive regression coefficient (β_1), and no interaction terms with X_2 (parity of the cow). Then the predicted log somatic cell counts from the model for different cows in different parities, as a function of X_1 will be parallel lines, as outlined on the left in Fig. 21.1. Each line represents the predicted value for cows of a specific parity. If an interaction term between parity and days in milk was added, this would produce non-parallel lines (for different parities), as outlined on the right.

Fig. 21.1 Schematic graphs of additive and non-additive modelling of a continuous predictor (days in milk) for a continuous outcome (lnscc)



Exactly the same interpretation is valid for cows in different herds: in an additive model (Eq 21.2) the regression lines corresponding to different herds are parallel, and the random herd effects can be read as the vertical distances between the lines. Thus, Eq 21.2 assumes the impact on the logarithmic cell counts of a change in days in milk (*eg* 10-day increase) to be the same for all cows in all herds (parallel lines).

21.3.2 Random slopes as non-additive herd effects

An assumption of additive herd effects (parallel lines) might not be biologically obvious because other factors such as breed or herd management factors (inherent in the herd effects) could influence the relationship. Adding an interaction between parity and X_1 to the model means that separate slope (β-) parameters for the regression of Y on X_1 are estimated. Adding an interaction between herds and X_1 means that slopes vary randomly between herds according to some distribution, in addition to the intercepts varying between herds. A model with random slopes for a single fixed effect (X_1) is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_{\text{herd}(i)} + b_{\text{herd}(i)} X_{1i} + \varepsilon_i \qquad \textit{Eq 21.9}$$

where in addition to the previous assumptions, we assume for the random slopes that the b_{herd}s $\sim N(0, \sigma_1^2)$. The parameter σ_1^2 is interpreted as the variation in slopes between

herds. The regression parameter β_1 is now the overall or average slope for X_1, which is then subject to random fluctuations between herds. As a rough rule, with probability 95%, the slope in a given herd would lie in the interval $\beta_1 \pm 2\sigma_1$. The choice of whether the slopes should be modelled as random or fixed effects usually follows the choice for the random effects themselves. That is, if herds are modelled as random, any slopes varying between herds should also be random. In social science applications, such models (with random slopes) are often referred to as hierarchical models, but we will call them random slope models.

We have not yet specified the assumptions about the relationship between b_{herd}s and the other random variables, and it is undesirable to assume random effects at the same level to be independent. In our example, the two random effects at the herd level (u_{herd} and b_{herd}) correspond to intercept and slope for the regression on X_1 at the herd level. Recall that slope and intercept are usually strongly negatively correlated (although centring the variable might remove this correlation). Consequently, we usually estimate a correlation or covariance between the herd intercept and slope. It is useful to display the three parameters: σ_h^2 , σ_1^2 and the covariance σ_{h1}, in a 2X2 matrix as follows:

$$\begin{pmatrix} \sigma_h^2 & \sigma_{h1} \\ \sigma_{h1} & \sigma_1^2 \end{pmatrix}$$

and the correlation between the herd intercepts and slopes is computed as $\sigma_{h1}/(\sigma_h\sigma_1)$. Example 21.4 shows the effect of adding a random slope to the SCC data.

Example 21.4 Random slopes of -t_dim- for somatic cell count data
data=scc40_2level

Adding a random slope of -t_dim- to the model of Example 21.2 gave almost the same regression coefficient (0.0027) but with a somewhat increased SE (0.0006), and the random effect parameters (with SEs) were:

$$\begin{pmatrix} \sigma_h^2 & \sigma_{h1} \\ \sigma_{h1} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.210\,(0.068) & -0.0011\,(0.0011) \\ -0.00059\,(0.00037) & 0.0000043\,(0.0000026) \end{pmatrix} \text{ and } \sigma^2 = 1.541(0.048)$$

The value of σ_1^2 is very small because the regression coefficient for -t_dim- is already small. It suggests that 95% of the slopes for -t_dim- lie roughly within 0.0027 ± 0.0040. The correlation between intercepts and slopes is negative and strong ($-0.00059/\sqrt{0.210 * 0.0000043} = -0.62$), as expected. However, the values of σ_1^2 and σ_{h1} are only moderately larger than their respective SEs, so it is not obvious whether the random slopes add much to the model. We will later see how to compute a statistical test for the random slopes (it is far from significant). Note finally that a model with random slopes for -h_size- would not be meaningful; random slopes are possible **only for variables at a lower level than the random effects** themselves in order to be interpreted in the way we have done.

21.3.3 Caveats of random slopes modelling

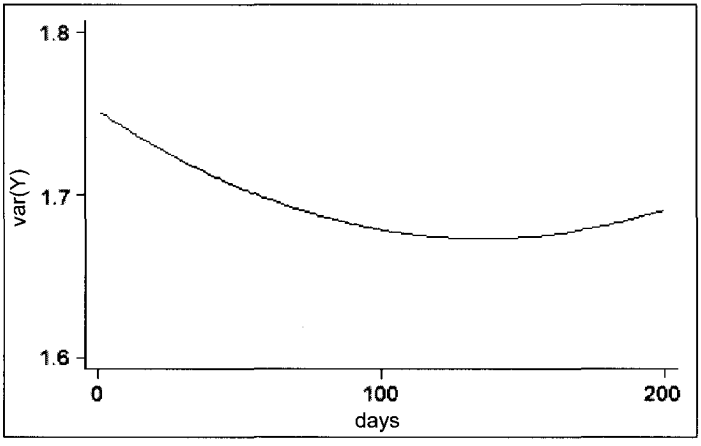
As intuitively appealing as the random slopes might appear, we must raise a few warning signs in their use. Usually, in epidemiological studies our main interest is in the fixed effects, and it is then wise policy not to build models with too many variance parameters. In our experience, it is rarely useful to have more than one or two random slopes in a model, and random slopes should usually only be included for statistically significant and clearly interpretable predictors.

One reason why random slopes should be used cautiously is that the **variance of the model is no longer constant**. To illustrate, we compute the variance components for the random slopes model of Eq 21.9:

$$\begin{aligned}\mathrm{var}(Y_i) &= \mathrm{var}(u_{\mathrm{herd}(i)}) + \mathrm{var}(b_{\mathrm{herd}(i)}X_{1i}) + 2\,\mathrm{cov}(u_{\mathrm{herd}(i)}, b_{\mathrm{herd}(i)}X_{1i}) + \mathrm{var}(\varepsilon_i) \\ &= \sigma_h^2 + X_{1i}^2\sigma_1^2 + 2X_{1i}\sigma_{h1} + \sigma^2 \end{aligned} \qquad \textit{Eq 21.10}$$

This equation involves the values of the explanatory variable X_1. In consequence, the variance is no longer the same for all observations but a function of X_1. Also, there is no longer a unique decomposition of variance in the model. For moderate magnitudes of σ_1^2 and σ_{h1} one might arrive at approximately the same decomposition of variance within the most relevant range of X_1. It is always recommended to plot the resulting variance function from a random slopes model, and if possible, convince yourself that it makes biological sense. Fig. 21.2 shows the variance function of the random slopes model for the somatic cell count data. The dependence of the total variance on X_1 is rather weak because the major portion of the variance is at the cow/test level; nevertheless, the dependence on X_1 is biologically reasonable. Mastitis in cows is more dynamic early in lactation (so we might expect more variance in -t_lnscc- early in lactation) and rises again late in lactation.

Fig. 21.2 Variance function of random slopes model for somatic cell count data



Random slope models have been introduced for continuous predictors (where the relationship between Y and X is a regression). However, interactions between categorical variables and random effects are possible as well, although not interpretable as random slopes. As before, an additive model assumes the impact of each categorical predictor to be the same in all herds, and sometimes one might want to allow it to vary between herds. It's simplest to specify such models for a dichotomous predictor: treat its 0-1 representation as if it was a continuous variable. If the variable takes several (j) categorical values, one might create (j-1) indicator variables and proceed in the same way. Be aware that such models quickly grow to contain a lot of covariance terms, and that they could produce very different variances for the different categories.

Example 21.5 shows the effect of adding a random slope for a dichotomous predictor in the SCC data.

Example 21.5 Random slopes of -c_heifer- for somatic cell count data
data=scc40_2level

Adding a random slope (of heifer) to the model from Example 21.2 produces a regression coefficient of -0.431(0.202) and the variance parameters (with SEs):

$$\begin{pmatrix} \sigma_h^2 & \sigma_{h1} \\ \sigma_{h1} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.202\,(0.062) & -0.076\,(0.042) \\ -0.076\,(0.042) & 0.051\,(0.039) \end{pmatrix} \text{ and } \sigma^2 = 1.546\,(0.048)$$

The two herd level variance contributions of this model are 0.202 for non-heifers and 0.202+0.051+2*(-0.076)=0.101 for heifers. We see how the covariance is part of the variance calculation, so it should not be assumed to be zero when dealing with random slopes for categorical predictors. The data thus seem to indicate both smaller mean and less variation of somatic cell counts for heifers than older cows. This makes biological sense based on our knowledge of mastitis.

21.4 STATISTICAL ANALYSIS OF LINEAR MIXED MODELS

In mixed models there are several methods of analysis, and the principal estimation procedure, which is based on the likelihood function (section 21.4.1), does not have closed-form expressions for the estimates but involves running several steps of an estimation algorithm. This requires some extra attention by the researcher to the statistical software to ensure that it employs the desired estimation procedure and to ensure that it is capable of analysing the data at hand. Statistical software differ in the range of models that can be analysed, in their ability to handle large data structures (many units at any level beyond the lowest one) and in their user interface. Specialised hierarchical or multilevel software has been developed to deal with huge data structures; a good source of information is the website of the Multilevel Models Project at the Institute of Education, University of London, UK (http://multilevel.ioe.ac.uk). Unlike the rest of this text, in which Stata was used for all analyses, the analyses

presented in this chapter and a few analyses in Chapters 22 and 23 have been carried out using the SAS system or the MLwiN program. (Without going into details regarding different software, just a cautionary note that some packages require the data to be sorted according to the hierarchical levels before analysis; failure to do so might lead to meaningless results.)

In most ways the mechanics of the analysis of linear mixed models is similar to the analysis of linear models, because the actual estimation procedure is taken care of by the software program, which also outputs many of the same quantities (*eg* estimates and SEs, tests of individual parameters and confidence intervals, as already shown in Example 21.2).

21.4.1 Likelihood-based analysis

Parameter estimation in normal linear mixed models is based on the likelihood function derived from the normal distribution assumptions. Roughly speaking, the likelihood function for any set of parameters gives the 'probability' of the observed data under that set of parameters (see section 16.4). Then it is intuitively reasonable to seek the set of parameters that maximises this probability – the maximum likelihood estimates. Because of the complicated form of the likelihood function, closed-form formulae for the maximum likelihood estimates generally do not exist. Therefore, parameter estimation employs an **iterative procedure** in which tentative estimates are gradually improved from their starting values to final convergence. As with all iterative procedures, caution must be exercised so that convergence is achieved. The estimation software should take care of this, but any messages that the iterative procedure has not converged are true causes for alarm. If the iterative procedure fails to converge, it sometimes helps to provide sensible starting values of the variance parameters; however, most commonly it signals a misspecified model.

Two variants of maximum likelihood estimation are available for mixed linear models: genuine **maximum likelihood** (ML) (also known as **full information maximum likelihood** or FIML) and **restricted maximum likelihood** (REML) estimation. From a theoretical point of view, REML estimates are unbiased, whereas ML estimates often have less variance; the weighting of these properties is not straightforward, but in practice the difference is usually negligible compared with the standard errors of the estimates. Both variants give 'asymptotically correct' values (*ie* when the number of observations at all levels of the hierarchy grows very large) and enable a full mixed model statistical inference. Therefore the choice between the two is essentially a technicality and a matter of taste; in the authors' experience, REML is the more commonly used. All results shown in this chapter are based on REML estimation.

Before proceeding with the statistical inference based on the likelihood function, it is worth mentioning an estimation approach based on the ANOVA table. It is simpler to implement and offered by more software packages. By and large, this approach is obsolete by today's standard, but in **balanced datasets** it will give the same estimates for the variance components and similar statistical tests for fixed and random parameters as the REML analysis. A dataset is balanced when every combination of

predictor values ('treatments') occurs the same number of times in the data. While this is frequently the case in experimental, factorial designs, it is rarely so in observational studies (in particular, if the data contain continuous predictors). The idea of the method is to compute variance components as linear functions of the mean squares of the ANOVA table, suitably chosen to make the variance component estimates unbiased. Therefore, closed-form expressions are available and they require little calculation beyond the ANOVA-table. Thus, the method is an add-on to a fixed effects analysis rather than a 'real' mixed models analysis, and herein lies its drawback: not all aspects of the statistical inference are managed correctly, *eg* standard errors are not readily available.

One particular example of an ANOVA-based method is still in quite common use – estimation of the *ICC* for a 2-level structure from a one-way ANOVA using the formula:

$$\rho \approx \frac{\text{MSM} - \text{MSE}}{\text{MSM} + (m-1)\text{MSE}} \qquad \textit{Eq 21.11}$$

where m is the (average) number of observations per group. If the groups are all of the same size (balanced data), this gives the same value as computing the *ICC* from likelihood-based variance components using Eq 21.4. When the data are unbalanced, the likelihood-based estimate is preferred. For the 2-level somatic cell count data, the above formula yields ρ=0.076; Eq 21.4 gives a value of 0.079.

21.4.2 Inference for fixed part of model

Although not evident from the material presented in Example 21.2, the reference distribution for fixed parameters is not the same as in linear models. Generally, the statistical inference is no longer exact but approximate, and the approximations are only 'asymptotically exact'. When the number of observations grows very large (at all hierarchical levels), the reference distribution approaches a standard normal distribution – thus one option for the reference distribution. However, with small or moderate numbers of observations at some of the hierarchical levels, a standard normal distribution might be too liberal as the reference, because it overestimates the degrees of freedom. Some software programs offer a finite sample approximation (Satterthwaite approximation) based on a t-distribution with a degree of freedom reflecting the design and the parameter under consideration. With a reference distribution in place, tests and confidence intervals are computed in the usual manner, *eg* a 95% confidence interval of $\beta_1 \pm t(0.975,\text{df})\text{SE}(\beta_1)$.

Approximate tests computed from the estimate and its SE are usually termed **Wald tests** (see section 6.4.2), and a multiple version exists for tests involving several parameters, *eg* for several indicator variables of a categorical variable. Tests based on comparing the attained value of the likelihood function (not the restricted likelihood from REML) in models with and without the parameter(s) of interest are possible as well but usually offer little advantage over Wald tests, and we leave them to the next section. Example 21.6 illustrates the inference of fixed effects in the SCC data.

Example 21.6 Fixed effects for 2-level somatic cell count data
data=scc40_2level

The parameters of -c_heifer-, -t_dim- and -t_season- all have finite sample reference t-distributions with about 2,150 degrees of freedom, which corresponds roughly to the residual degrees of freedom at the cow/test level. The Wald tests indicate both the effects of heifer and days in milk are clearly significant. A multiple Wald test for -t_season- gives F=2.07 which is non-significant in $F(3, 2167)$ with a P-value of 0.10.

The finite sample reference distribution for -h_size- is $t(38.1)$ or $\approx t(38)$ (based on the Satterthwaite approximation), reflecting that it is a herd-level predictor, and that the 40 herds would leave only 38 degrees of freedom for the herd-level residual. Therefore, the effect of -h_size- is estimated with considerably less precision than the other predictors, and not surprisingly, it shows up clearly non-significant. The finite sample reference distribution for the constant has similar degrees of freedom, reflecting the previously noted fact that it refers to a herd mean value.

21.4.3 Inference for random part of model

Even though the software usually outputs both variance parameters and their SEs, the latter should not be used to construct Wald-type confidence intervals or tests, because the distribution of the estimate can be highly skewed.

Variance parameters can be tested using likelihood-based (**likelihood ratio**) tests, although we usually retain random effects corresponding to hierarchical levels despite their non-significance (unless the variance is estimated to be zero). To illustrate, a likelihood ratio test in Eq 21.9 for the hypothesis H_0:σ_h=0 is calculated as G^2=-2($\ln L_{full}$-$\ln L_{red}$) where the full and reduced models refer to the models with and without the herd random effects, and L refers to values of the likelihood function. Either ML or REML likelihood functions might be used. Generally, the value of G^2 is compared with an approximate χ^2-distribution with the degrees of freedom equal to the reduction in number of parameters between the two models. Snijders and Bosker (1999) note that reference χ^2-distributions are conservative when testing a variance parameter being equal to zero, and recommend halving the P-value obtained from the χ^2-distribution to take into account that the alternative hypothesis is one-sided (H_a:σ_h>0).

For random effect parameters, symmetric confidence intervals are usually inappropriate. Two alternative methods are suggested in the literature: bootstrapping (Goldstein, 1995, section 3.5) and profile-likelihood intervals (Longford, 1999). Bootstrapping is a general statistical technique primarily aimed at estimating standard errors and calculation of confidence intervals in situations too complex for analytical methods to be manageable; however, bootstrap confidence intervals require specialised software (*eg* MLwiN). In brief, a profile-likelihood confidence interval (with approximate 95% coverage) includes the values (σ^*) of the parameter, for which twice the log-likelihood with the parameter under consideration fixed at the particular value (*ie* σ=σ^*), drops

less than 3.84 (the 95% percentile in $\chi^2(1)$) from twice the log-likelihood value of the model. If your software allows you to fix a variance in the model, a crude search for such parameter values is simple to carry out. Example 21.7 illustrates the inference for random parameters in the SCC data.

Example 21.7 Herd random effect for 2-level somatic cell count data
data=scc40_2level

The table below gives values for twice the log likelihood function (based on REML) for various somatic cell count models in this chapter and likelihood-ratio test statistics for model comparisons (comparing all models with the one presented in Example 21.2).

Model	2lnL	G^2	df	P-value
no herd random effect	-7346.93	97.01	1	0.000
model from Example 21.2	-7249.92	-	-	-
random slope of -t_dim-	-7243.90	6.02	2	0.049
random slope of -c_heifer-	-7244.13	5.79	2	0.055

The table shows strong evidence against the hypothesis of no (random) variation between herds, and it also shows that extensions of the model with random slopes for -t_dim- and -c_heifer- are both borderline significant.

A 95% confidence interval for σ_I^2 was obtained by the profile-likelihood method and a crude search of parameter values around the estimate of 0.149. For example, fixing σ_I^2 at a value of 0.25 gave a 2lnL value of -7252.90, which is still within 3.84 of the model's value (-7249.92); therefore, the value 0.25 belongs to the 95% confidence interval. The resulting interval was (0.085,0.269), which is asymmetric and more appropriate than the symmetric interval: 0.149±1.96*0.044=(0.063,0.235).

21.4.4 Residuals and diagnostics

Residuals and diagnostics play a similar, crucial role for model-checking in mixed models as they do in ordinary linear models. The mechanics and interpretations are analogous (see sections 14.8 and 14.9). Moreover, the additional model assumptions for the random effects should be evaluated critically together with the other assumptions. Mixed models contain additional 'residuals' – one set per random effect in the model. These residuals are, in reality, predicted values of the random variables in the model (sometimes called **best linear unbiased predictors** (BLUPs)). They include not only the effects for the hierarchical levels but also the random slopes, *ie* in a model with random intercepts and slopes, there are two sets of residuals at the corresponding level. Langford and Lewis (1998) recommend inspecting first the residuals at the highest hierarchical level, and then gradually work downwards. Thus, before looking at individual cows being influential or not fitted well by the model, we examine the same questions for the herds. This is because several of the cows being flagged could stem from the same herd, so the 'problem' might be with the herd rather than with the individual cow.

Example 21.8 Residuals and diagnostics for somatic cell count data
data=scc40_2level

We present here the residuals and diagnostics for the 10 most extreme herd random effects (the analysis of cow-level residuals and diagnostics follows similar lines as in Chapter 14). The computations were done using the MLwiN software. The normal plot of the standardised residuals did not indicate any serious deviations from the normal distribution (not shown).

herd number	raw residual	standardised residual	deletion residual	leverage	DFITS
40	-0.831	-2.426	-2.599	0.113	0.405
7	-0.787	-2.310	-2.454	0.117	0.389
8	-0.445	-1.299	-1.311	0.114	0.204
15	-0.403	-1.137	-1.141	0.083	0.151
39	-0.370	-1.067	-1.069	0.103	0.158
...	...	...	...	...	...
35	0.516	1.489	1.513	0.103	0.224
34	0.523	1.696	1.740	0.202	0.365
32	0.600	1.733	1.780	0.103	0.264
6	0.666	1.983	2.064	0.130	0.344
18	0.688	2.546	2.753	0.300	0.712

Herd 18 stands out somewhat with the highest values of residuals, leverage and DFITS. The magnitude of the residuals is hardly anything to worry about, but the influence seems appreciable. When analysing the data without this herd, the effect of -h_size- increases by more than 50% and approaches significance. Herd 18 turns out to have the smallest value of -h_size-, but the largest value of -t_lnscc-.

Unfortunately, we need to dampen the reader's enthusiasm (faced with the prospect of a multilevel residual analysis); currently, residuals and diagnostics are available to a varying degree with different software for mixed models. Also, they are not straightforward to compute directly from the parameters of the model. Although unsatisfactory from a scientific point of view, this tends to imply, in practice, that you confine yourself to the model-checking available in the software being used. Example 21.8 presents herd-level residuals and diagnostics for the SCC data.

21.4.5 Box-Cox transformation for linear mixed models

In section 14.9.5, we discussed the Box-Cox method of choosing the 'best' power (λ) transformation of our data to match the assumptions of a linear model. We assumed the method to be implemented in available software and did not go into details with how the optimal λ was calculated. A Box-Cox analysis is however, to our knowledge, not readily available elsewhere for mixed models, so we give the necessary details to enable the analysis for transformation of the outcome.

Recall that we confine the analysis to a set of 'nice' λ-values, *eg* for a right-skewed distribution, we might search for the best value among λ=1, 1/2, 1/3, 1/4, 0, -1/4, -1/3, -1/2, -1, -2. Among these λ=1 corresponds to no transformation, λ=0 to natural log transformation, and λ=-1 to reciprocal transformation. Finding the approximate optimal λ-value involves the following steps:

1. compute the mean of the lnY-values and denote this value by $\overline{\ln Y}$; also denote the total number of observations as n,
2. for each candidate λ-value, compute for each observation i the transformed value

$$Y_i^{(\lambda)} = \begin{cases} (Y_i^{\lambda} - 1)/\lambda & \text{for } \lambda \neq 0 \\ \ln Y_i & \text{for } \lambda = 0 \end{cases}$$

   and analyse these $Y^{(\lambda)}$-values by the same mixed model as the untransformed values, and record the model's attained log-likelihood value, $\ln L^{(\lambda)}$,
3. compute the value of the profile log-likelihood function as

$$\text{pl}(\lambda) = \ln L^{(\lambda)} + n(\lambda - 1)\overline{\ln Y} \qquad \textit{Eq 21.12}$$

   and plot the function to identify approximately the λ where pl(λ) is maximal. This is the optimal power transformation of the outcome. An approximate 95% confidence interval for λ consists of those λ-values with a value of pl(λ) within 3.84 of the optimal pl-value.

We demonstrate the procedure in Example 21.9 using the SCC data.

Recall (from Chapter 14) that the optimal Box-Cox value does not guarantee 'well -behaved' residuals (at all hierarchical levels), and that transformation could shift problems from one model assumption to another (*eg* from skewed residuals to heteroscedasticity). Therefore, even after transformation, all the residuals should be examined. If well-behaved residuals at some hierarchical level cannot be achieved by transformation, one might turn instead to models with non-normal random effects; such models are currently only available within the Bayesian framework for hierarchical models (section 23.4).

21.5 Repeated measures and spatial data

We have already touched upon why repeated measures and spatial data structures are not completely satisfactorily modelled by the hierarchical mixed models presented so far (section 20.2). For example, considering repeated measures on animals over time as a 2-level hierarchical structure (tests within animals) does not take time ordering of the tests into account in the random part of the model. Where animals within a herd can be interchanged without altering the meaning of the data, observations over time cannot. Generally, one would expect two adjacent measures to be more highly correlated than two very distant ones. The hierarchical mixed models we have discussed so far assume correlations are the same between any pairs of measures (section 21.2.1). This pattern or structure of the correlations is called **compound symmetry** or **exchangeable**, and our first step in extending the mixed model class towards more realistic repeated

Example 21.9 Box-Cox analysis for somatic cell count data
data=scc40_2level

The data contain n=2,178 observations and the mean (natural) logarithmic cell count is 4.7569865. The following table and graph give a Box-Cox analysis:

λ	1	0.5	0.33	0.25	0
lnL for $Y^{(\lambda)}$	-17224.41	-9807.93	-7551.10	-6543.36	-3624.96
pl(λ) from (21.11)	-17224.41	-14988.29	-14492.78	-14313.90	-13985.68
λ	-0.10	-0.25	-0.33	-0.5	-1
lnL for $Y^{(\lambda)}$	-2553.74	-1043.18	-281.48	1246.43	5188.00
pl(λ) from (21.11)	-13950.53	-13994.08	-14061.23	-14294.64	-15533.43

Fig. 21.3 Profile-likelihood function for Box-Cox analysis of SCC data

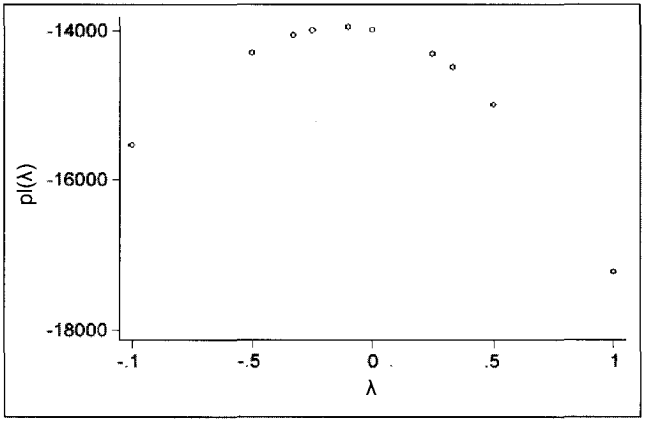


Table and figure indicate the optimal value of λ to be close to, but slightly less than, zero, but a 95% CI for λ does not include zero; the large number of lowest-level observations causes the CI to be very narrow. With the optimal transformation so close to the log-transformation, the Box-Cox analysis supports our choice of analysing the log somatic cell counts.

measures and spatial models is to consider alternative, more appropriate correlation structures.

21.5.1 Correlation structure

To conveniently display the dependence between measurements $(Y_1,...,Y_m)$ on the same animal (in the repeated measures context), or more generally among a set of correlated measurements, we introduce the covariance matrix cov(Y) and the correlation matrix corr(Y) – (mxm)-matrices holding all the covariances, or correlations, between pairs of measurements:

$$\mathrm{cov}(Y)=\begin{pmatrix} \mathrm{var}(Y_1) & & & & \\ \mathrm{cov}(Y_1,Y_2) & \mathrm{var}(Y_2) & & & \\ \mathrm{cov}(Y_1,Y_3) & \mathrm{cov}(Y_2,Y_3) & \mathrm{var}(Y_3) & & \\ \vdots & \vdots & \vdots & \vdots & \\ \mathrm{cov}(Y_1,Y_m) & \mathrm{cov}(Y_2,Y_m) & \mathrm{cov}(Y_3,Y_m) & \cdots & \mathrm{var}(Y_m) \end{pmatrix}$$

$$\mathrm{corr}(Y)=\begin{pmatrix} 1 & & & & \\ \mathrm{corr}(Y_1,Y_2) & 1 & & & \\ \mathrm{corr}(Y_1,Y_3) & \mathrm{corr}(Y_2,Y_3) & 1 & & \\ \vdots & \vdots & \vdots & \vdots & \\ \mathrm{corr}(Y_1,Y_m) & \mathrm{corr}(Y_2,Y_m) & \mathrm{corr}(Y_3,Y_m) & \cdots & 1 \end{pmatrix}$$

The matrices are symmetric, so for clarity, the values above the diagonal have been left blank.

Table 21.2 lists some of the more common correlation structures for repeated measures in the case of m=4 repeated measures on the same animal. For simplicity, we show only the correlation matrix in all cases except the last one but, if variances are assumed to be equal (σ^2), the covariances are simply the correlations multiplied by σ^2.

The first two correlation structures are well known and included mainly to familiarise the reader with the display. Recall that the correlation ρ in the compound symmetry structure can be expressed in terms of the variance components σ_h^2 and σ^2 as $\rho=\sigma_h^2/(\sigma_h^2+\sigma^2)$ (Eq 21.4).

The simplest structure showing the desired decay in correlation with increasing distance between observation is **ar(1)**. However, in practice the decline of correlations is often too strong (fast), and the two slightly more complicated correlation structures with additional parameters (**arma(1,1)** and **Toeplitz**) are often useful as well. The unstructured correlation structures are included here mainly for completeness because, with long series of repeated measures, the number of parameters involved grows so large that they become difficult to estimate and interpret.

In our discussion so far, we have paid little attention to the actual recording or measurement times of the measurements. First, you need to ensure that these are properly entered into the model. For example, it makes a difference with most correlation structures whether recordings were taken at times (1,2,3,4), at times (3,4,5,6) or at times (1,2,5,6). Second, the autoregressive and Toeplitz structures are most meaningful if the measures are equidistantly spaced in time (*eg* monthly recordings). This is by far the simplest situation for building repeated measures models (something worth considering when designing a study involving repeated measures!). The data could have non-equidistant recordings either within each animal or between animals so that measures are taken at different times and with different intervals for

different animals. Such data structures raise the need to incorporate into the matrices the actual recording times.

Table 21.2 Repeated measures correlation structures for four repeated measures per animal

Name	Correlation structure	Interpretation
uncorrelated or independent	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\0&1&&\\0&0&1&\\0&0&0&1\end{pmatrix}$	uncorrelated (for normal data: independent) observations
compound symmetry, or exchangeable	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\\rho&1&&\\\rho^2&\rho&1&\\\rho^3&\rho^2&\rho&1\end{pmatrix}$	hierarchical, mixed model (same correlation between all pairs of observations)
ar(1), or first order autoregressive	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\\rho&1&&\\\rho^2&\rho&1&\\\rho^3&\rho^2&\rho&1\end{pmatrix}$	repeated measures or time-series model with power decay of correlations
arma(1,1), or first order autoregressive moving average	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\\gamma&1&&\\\gamma\rho&\gamma&1&\\\gamma\rho^2&\gamma\rho&\gamma&1\end{pmatrix}$	extended repeated measures or time series model with power decay
Toeplitz, or stationary	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\\rho_1&1&&\\\rho_2&\rho_1&1&\\\rho_3&\rho_2&\rho_1&1\end{pmatrix}$	repeated measures with unconstrained correlations at different spacings
unstructured	$\mathrm{corr}(Y)=\begin{pmatrix}1&&&\\\rho_{12}&1&&\\\rho_{13}&\rho_{23}&1&\\\rho_{14}&\rho_{24}&\rho_{34}&1\end{pmatrix}$	repeated measures with entirely unconstrained correlations
unstructured with inhomogeneous variances, or non-stationary	$\mathrm{cov}(Y)=\begin{pmatrix}\sigma_1^2&&&\\\sigma_{12}&\sigma_2^2&&\\\sigma_{13}&\sigma_{23}&\sigma_3^2&\\\sigma_{14}&\sigma_{24}&\sigma_{34}&\sigma_4^2\end{pmatrix}$	repeated measures, unconstrained variances and correlations

For non-equidistant repeated measures or spatial data, denoted by $d_{ii'}$ the distance between observations i and i'. With repeated measures, the $d_{ii'}$ would be the difference between the recording times of observation i and i', and for spatial data the distances would be actual physical distances (*eg* between herds). Table 21.3 lists some examples of correlation structures defined from such distances. The models are **isotropic** because only the distances, not the actual locations i and i' are used.

Table 21.3 Spatial (or non-equidistant repeated measures) correlation structures

Name	Correlation structure	Interpretation
power, or exponential	$\text{corr}(Y_i, Y_{i'}) = \rho^{d_{ii'}} = e^{-d_{ii'}/\theta}$	power decay with distance; note the relationship: $\rho = e^{-1/\theta}$)
power, or exponential, with nugget effect	$\text{corr}(Y_i, Y_{i'}) = \frac{\sigma^2}{\sigma^2 + \sigma_0^2} \rho^{d_{ii'}}$	power decay with distance, close observations not fully correlated
Gaussian	$\text{corr}(Y_i, Y_{i'}) = e^{-d_{ii'}^2/\theta})$	exponential-quadratic decay with distance
linear	$\text{corr}(Y_i, Y_{i'}) = \begin{cases} 1 - \rho d_{ii'} & \text{if } \rho d_{ii'} < 1 \\ 0 & \text{if } \rho d_{ii'} \geq 1 \end{cases}$	linear decay with distance

Note The power structure is the extension of ar(1) to non-equidistant data.

21.5.2 Mixed models with complex correlation structures

Recall that, in the linear mixed model (Eq 21.8):

$$Y = X\beta + Zu + \varepsilon$$

we assumed the components of ε to be independent, and modelled the hierarchical structure using the Zu part of the model. Now we will allow dependence corresponding to a particular correlation structure within some sets of ε-values. In the repeated measures context, each set contains all the repeated measures for an animal, and in the spatial context, each set contains a particular group of observations for which we want to model a spatial correlation (*eg* herds within a certain region).

The statistical analysis of such 'extended' mixed models evolves along the same lines as previously discussed, only with additional variance parameters to be estimated. Therefore parsimonious models for cov(Y) are recommended, to avoid overspecification of the model and unexpected impacts of the covariance structure on the other parameters. The choice of correlation structure can be formalised by using likelihood ratio statistics to test nested correlation-structure models against each other. For example, the compound symmetry and ar(1) models can be tested against both arma(1,1) or Toeplitz models, but they cannot be tested against each other. Models with the same number of parameters can be compared in models fit by their log-likelihood values (the higher log-likelihood model is generally preferred). Model selection criteria

such as the AIC or BIC (Chapter 15) could also prove useful here. Example 21.10 examines different correlation structures for the full SCC data.

Example 21.10 Repeated measures analysis for somatic cell count data
data=scc_40

Several correlation structures were examined for the full 40-herd somatic cell count data, using the same fixed effects as in the previous examples. The model is now a 3-level model with tests within cows within herds.

Model/correlation structure	Correlation parameters	Estimated ρ 1 month	Estimated ρ 2 months	-2ln likelihood
compound symmetry	1	0.541	0.541	39004.73
ar(1)	1	0.661	0.437	38563.57
non-equidistant power	1	0.673	0.453	38574.30
arma(1,1)	2	0.657	0.578	37802.21
Toeplitz	10	0.657	0.578	37795.72

The table illustrates how the different structures adapt to the data. In terms of statistical significance, the Toeplitz model is no better than the arma(1,1) model, which in turn is clearly preferable to the structures with only one parameter. The estimated correlations for tests one and two time steps (for the non-equidistant structure considered equivalent to 30 days) apart shows the deficiency of the one-parameter models. The compound symmetry structure does not allow for a smaller correlation for two time steps, and the autoregressive-type structures produce too rapidly decaying correlations.

For comparison with the results of the 2-level data, we also present a table of estimates for the fixed effects and the two variance components from the arma(1,1) model:

	Coef	SE	t	P	95% CI	
h_size (*100)	0.627	0.306	2.05	0.047	0.009	1.245
c_heifer	-0.777	0.040	-19.22	0.000	-0.857	-0.698
t_season = spring	0.034	0.022	1.54	0.125	-0.009	0.078
t_season = summer	0.039	0.027	1.57	0.117	-0.010	0.087
t_season = fall	-0.007	0.023	-0.32	0.752	-0.052	0.037
t_dim (* 100)	0.328	0.014	24.08	0.000	0.301	0.354
constant	4.516	0.154	29.25	0.000	4.205	4.827

In addition, the estimated correlation parameters (also with SEs) were:

$\gamma = 0.657$ (0.008) and $\rho = 0.880$ (0.006)

and the variance components were:

$= 0.104$ (0.028) and $\sigma^2 = 1.378$ (0.027)

We conclude with some additional remarks about the consequences of using a hierarchical mixed model for repeated measures data. First, the general mixed model approach allows us to test the adequacy of compound symmetry relative to some of the 'genuine' repeated measures structures. In our example, the compound symmetry structure fitted the worst of all structures examined. However, with the large dataset and an average of 6.8 observations per cow, we were not surprised to find clear evidence against the equal correlations assumption. For a smaller series of repeated measures, for example, m=2 to 4, there might be little evidence of decreasing correlation with distance in time, and therefore a compound symmetry structure might work quite well. Note also that the correlation structure applies to the unexplained variation after adjusting for fixed effects, and if there are strong time-level predictors, the errors might show little autocorrelation. It is also interesting that the compound symmetry analysis will tend to give a too liberal inference for predictors at the time level (including interactions with time). The theory of repeated measures models in factorial designs also tells us that for predictors at the animal level, the choice of correlation structure is less critical.

SELECTED REFERENCES/SUGGESTED READING

1. Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in cattle. Prev Vet Med 2001; 50: 127-144.
2. Goldstein H. Multilevel statistical models. 2d ed. London: Arnold. 1995.
3. Langford IH, Lewis T. Outliers in multilevel models (with Discussion). J R Stat Soc A 1998; 161: 121-160.
4. Longford N. Standard errors in multilevel analysis. Multilevel Newsletter 1999; 11: 10-13.
5. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modelling. London: Sage Publications, 1999.

SAMPLE PROBLEMS

1. Using the reu_cc data, explore multilevel models for the log calving to conception intervals.
   a. As a start, draw a diagram of the hierarchical structure. Include in the diagram the predictors -heifer- and -ai- by indicating at which level they reside.
   b. Build a 2-level model with fixed effects of the two predictors as well as cow random effects.
      i. Which fixed effects are significantly different from zero?
   c. Turn your attention to the variance parameters.
      i. Compute the proportion of variance residing at the two levels (lactation and cow) of the model.
      ii. Give also the intra-class correlation coefficient (*ICC*) for two observations on the same cow.
      iii. If a cow has a long calving-to-conception interval in one lactation, is it likely that she will have a long one in the subsequent lactation?
   d. Add herd random effects to the model.
      i. How do the estimates for -heifer- and -ai- change?
      ii. Does the residual variance change much?
      iii. Recompute the *ICC* for lactations within a cow.
      iv. Compute also the *ICC* of two observations on different cows in the same herd
   e. Test the significance of each of the two random effects using likelihood ratio test statistics.
2. Again using the reu_cc data, examine the validity of model assumptions.
   a. Refit the 3-level model from Problem 1 with -heifer- and -ai- as fixed effects.
      i. Examine graphically the normality of residuals at all three levels (lactation, cow, herd), preferably (depending on your software) using the standardised residuals.
      ii. Plot (standardised) residuals against predicted values, preferably (depending on your software) at all three levels, to detect any signs of heteroscedasticity.
   b. Repeat the steps from a. with -cc- as the outcome instead of -lncc-, and comment on any differences noted.
   c. Turn next your attention to transformation of the outcome (-cc-).
      i. Carry out a Box-Cox analysis to determine approximately the optimal power transformation.
      ii. Analyse the data using the selected optimal transformation, and compare the results and residuals with the previous analysis of -lncc-.
   d. Add the log calving to first service interval (-lncfs-) to the model as a fixed effect, using again -lncc- as the outcome.
      i. Notice the changes in model estimates, and compute a statistical test for the effect of -lncfs-.
      ii. Examine the residuals from the model, in particular at the lowest level, and note any problems with the model's assumptions.
      iii. Plot -lncc- against -lncfs-, and try to understand the model's behaviour when adding -lncfs- as a predictor.
3. Explore random slopes models for the pig data.

a. Build a 2-level model for the average daily gain (-adg-) with random farm effects and fixed effects of the predictors -sex-, -pn-, -worms- and -ar2-.
   i. Examine which fixed effects are necessary in the model (explore also interaction terms).
   ii. Assess the significance of the farm random effects, and compute the *ICC* for pigs in the same farm.

b. Add farm random slopes for -pn- (and include also the covariance term).
   i. Do the random slopes seem to improve the model? - compute a likelihood ratio test to quantify your impression of the estimates.

c. Add farm random slopes for -ar2- (again including the covariance term).
   i. Do the random slopes seem to improve the model? - compute a likelihood ratio test to quantify your impression of the estimates.
   ii. Compute the total variance of observations with ar2=0 and ar2=1.
   iii. Confirm by another method (*eg* simple descriptive statistics) that the data show different variations at the two levels of ar2.
   iv. What do you conclude about the effect of ar2 on the mean value and variation of average daily weight gains?

4. Explore repeated measures models for milk yields in the somatic cell count data.

a. Draw a diagram of the hierarchical structure of the scc_40 dataset, including the predictors -t_dim-, -t_season-, -t_lnscc-, -c_heifer-, and -h_size-.

b. Build an initial hierarchical mixed model for -ecm- with random effects of herds and cows (and thus a compound symmetry correlation structure), as well as fixed effects of the above predictors.
   i. Examine whether the relationship between -ecm- and the predictors -t_lnscc- and -t_dim- can be modelled as linear to a reasonable approximation (our main focus in this exercise is not on the fixed effects). If not, extend the fixed part of the model suitably.
   ii. Examine whether the fixed part of the model should contain interaction terms.
   iii. Compute the + for two observations on the same cow, and explain why the correlation structure of the model might be inadequate for these data.

c. Fit a model with an autoregressive correlation structure for the repeated measures on the same cow.
   i. Compare the fit of this model with the previous (compound symmetry) model; note that you cannot test the two models against each other but you might use the value of the log-likelihood or the AIC.
   ii. Compute the estimated correlation between two tests one month apart. Repeat for tests two months apart.

d. Fit other correlation structures to the repeated measures on the same cow (depending on the capabilities of your software).
   i. Compare the models using likelihood ratio tests (for nested models) or the log-likelihood and AIC values (for non-nested models).
   ii. Compare the estimated correlations for tests one and two months apart.
   iii. Which correlation structure seems to be preferable (balancing the data fit and model parsimony)?
   iv. What impact did the choice of correlation structure have on the fixed effects?

22

MIXED MODELS FOR DISCRETE DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the differences between linear mixed models (continuous data) and generalised linear mixed models (GLMMs) (discrete and continuous data) and the role of the link function in the latter.

2. Fit random effects logistic and Poisson models.

3. Use a latent variable approach to compute the intra-cluster correlations for binary outcomes.

4. Use either quasi-likelihood or maximum likelihood methods for fitting GLMMs.

5. Assess the statistical significance of both fixed and random effects in GLMMs.

6. Evaluate residuals (except at the lowest level) to assess the adequacy of a GLMM that you have fit.

7. Compute a dispersion parameter to adjust for over- or underdispersion in GLMMs that you have fit using quasi-likelihood methods.

22.1 INTRODUCTION

In both theory and practice, it has proven more difficult than one might have anticipated to generalise the mixed models approach from continuous to discrete data . One effect of these difficulties is the existence of a wide variety of generalisations of mixed models to discrete data, some of them only for a particular type of discrete data (usually binary) and some of them within wider frameworks. In this chapter, we review the model class most analogous to linear mixed models: the **generalised linear mixed models.** In order to fully appreciate this analogy, the reader is encouraged to review mixed models first (Chapter 21).

Our main focus here will be on binary data (logistic regression with random effects, section 22.2) but the random effects extension applies to a flexible class of discrete models which include multinomial and Poisson regressions. A Poisson regression with random effects is presented in section 22.3. As in Chapter 21, our mixed models will reflect a hierarchical structure but it is also possible to build models for other data structures. However, (and this goes generally for mixed models for discrete data) the statistical analysis is more difficult than for continuous data, requires more care and choices by the researcher (of which the choice of software is an important one). This field is still growing and advancing but we attempt to give the applied researcher a snapshot of its present state.

We will use two binary data examples to illustrate the methods: one on pneumonia in pigs and another on first service conception risks in cows. The first dataset, pig_adg, stems from a 2-level hierarchy (pigs within farms) and we will consider only a single pig-level, dichotomous predictor. These data will also be used in Chapter 23 to illustrate some of the alternative methods to deal with clustering for discrete data. The second dataset, reu_cfs, contains a 3-level hierarchy (lactations within cows within herds) and two lactation-level predictors.

22.2 LOGISTIC REGRESSION WITH RANDOM EFFECTS

We consider again the example of animal disease observed in several herds (*eg* the pig-pneumonia data). The logistic regression analogue of Eq 21.2 for the probability p_i of the i^{th} animal being diseased is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{herd}(i)} \qquad \textit{Eq 22.1}$$

where $u_{\text{herd}(i)}$ is the random effect of the herd (which contains animal i), assumed to be $u_{\text{herd}(i)} \sim N(0, \sigma^2_h)$, the X_is are the predictor values for the i^{th} animal, and the relationship between the probability p_i and the binary outcome Y_i is unchanged: $p(Y_i=1)=p_i$. The only change from the ordinary logistic regression model is the herd random effects term. Example 22.1 shows that adding random effects can have an appreciable impact on the model.

Example 22.1 Random effects logistic model for pig-pneumonia data
data=pig_adg

Data on both atrophic rhinitis and enzootic pneumonia were recorded on 341 pigs at slaughter. Causally, it was assumed that atrophic rhinitis might increase the risk of pneumonia, through destruction of the pig's air-filtering mechanism (nasal turbinates). The atrophic rhinitis score was converted to a dichotomous variable (-ar_g1-) indicating the presence/absence of an atrophic rhinitis score greater than 1. Similarly, the lung score was converted to a dichotomous variable (-pn-) representing the presence/absence of pneumonia.

The unconditional association between -pn- and -ar_g1- was:

		ar_g1		
		1	0	Total
pn	1	109	77	186
	0	66	89	155
Total		175	166	341

Odds ratio = 1.909
95% CI = (1.212,3.010)
Chi-sq = 8.687
P-value = 0.003

These statistics indicate a moderate but clearly significant association between -pn- and -ar_g1-. However, we have ignored the fact that the pigs came from 15 farms, and the prevalence of -pn- actually varied from 17% to 95% across farms. Consequently, it appears that we should be concerned about farm effects. The logistic regression with random effects (Eq 22.1) gave the estimates:

	Coef	SE	Z	P	95% CI	
ar_g1	0.437	0.258	1.69	0.091	-0.069	0.943
constant	0.020	0.301	0.07	0.948	-0.571	0.610

In addition, the estimated variance of herd random effects (with SE) was:

$$\sigma_h^2 = 0.879\ (0.434)$$

We shall later see how to compute the significance of the random effect (it is highly significant). The regression coefficient for -ar_g1- should be compared with the log of the simple odds ratio (ln1.909=0.647). Accounting for the herd effects reduces the association considerably, and it is no longer significant. In other words, the farms had both a clustering and confounding effect.

22.2.1 Analogies and differences to a linear mixed model

We have seen that a mixed logistic regression model adds the random effects to the fixed effects, both on a logistic scale. So, bearing the logistic scale in mind, we build the models in a similar way to linear mixed models and they might include multiple random effects and possibly random slopes as well. The statistical analysis also has strong similarities in the way confidence intervals and tests are computed.

The 2-level model (Eq 22.1) has a correlation structure similar to its linear mixed model analogue, with equal correlations between animals within the same herd and independence between herds. However, we have to be careful here: the correlations within a herd are the same only for animals with the same fixed effects. In our example, all -ar_g1- positive animals within a herd are equally correlated, and the same for all -ar_g1- negative animals. This difference between animals with different predictor values may seem strange and is usually small in practice (unless the predictor has a very strong effect). It is one of the many consequences of modelling the fixed and random effects on the logit scale. Nevertheless, the model is perfectly valid as a method to account for correlation (or clustering) between animals in the same herd.

The interpretation of fixed effects in a linear mixed model was essentially unaffected by the added random effects. Again, the modelling on the logit scale complicates the interpretation of models such as Eq 22.1. In a strict sense, the model has a 2-step interpretation which is perhaps best understood by imagining how data would be generated by the model. For an animal i in the j^{th} herd, we would first select the herd random effect (uj) according to its $N(0, \sigma_h^2)$ distribution and compute p_i from the fixed effects and the selected u_j-value. We would then select the outcome Y_i as positive with probability p_i or negative with probability $1\text{-}p_i$. A common shorthand for this 2-step interpretation is that Eq 22.1 is 'conditional on' the random effects. This interpretation of the model means that when exponentiating a regression coefficient (for -ar_g1- in the example) to obtain the odds-ratio (*ie* exp(0.437)=1.55), the odds-ratio refers to comparing pigs with and without atrophic rhinitis **in a particular herd** (corresponding to a selected herd random effect, no matter the actual u_j-value). Frequently this is called a **subject-specific** (in our example, a herd-specific) estimate, as opposed to a **population-averaged** estimate, which would refer to the odds-ratio for comparing pigs with and without atrophic rhinitis **from any herd** in the population of herds. Therefore, if we think of the odds ratio as the answer to questions such as 'how much is the risk increased?' (in our example, the risk of pneumonia for an 'ar'-pig versus a healthy pig), the subject-specific estimate answers the farmer's question and the population-averaged estimate answers the slaughterhouse's question (where pigs are submitted from many different herds). That these two questions have different answers challenges our intuition, but is an incontestable fact. However, the answers are usually closely related, though (see section 23.3 for further discussion of subject-specific and population-averaged estimates).

22.2.2 Interpretation of variance parameter(s)

In Eq 22.1, the herd random effect variance σ_h^2 has no direct interpretation in terms of the probabilities of disease. The equation shows that it refers to the variation between herds of the disease probabilities on a logit scale. We can still interpret σ_h^2 qualitatively: a value of zero means no variation between herds (and therefore no clustering) and a large positive value means a high degree of clustering. However, the (correct) statement that the logits of probabilities vary within $\pm 1.96\sigma_h$ across herds with a probability of 95%, is not very intuitive.

In linear mixed models, the variance parameters could be interpreted as **variance components**, but in models of discrete data, we have problems with this interpretation. If we compare Eq 22.1 with the linear mixed model (Eq 21.2), the error term or ε_i is missing in the logistic equation. This is because the distribution assumption is on the original scale – in our example $Y_i \sim \text{bin}(1,p_i)$, so that the errors in the model stem from the binomial (binary) distribution instead of a normal distribution. Recall that in this binary distribution the variance equals $p_i(1-p_i)$. Now the total variance in the data, $\text{var}(Y_i)$, is no longer just the sum of the error variance and the random effects variance, as they refer to different scales. Even worse, the total variance is not constant because the binomial variance varies with p, so a single decomposition of the variance does not exist. Several recent papers have reviewed computation of variance components and intra-class correlation coefficients (*ICC*s, sometimes also denoted variance partition coefficients) in mixed logistic regression, and a number of different methods have been suggested (Goldstein et al, 2002a; Browne et al, 2003; Rodriguez and Elo, 2003; Vigre et al, 2003). We confine ourselves to explaining a simple approximation method based on **latent variables** (latent variables were introduced in Chapter 17).

The simplest approach to getting both the individual and herd variances on the same (logistic) scale is to associate with every animal i a latent continuous measure, Z_i, of the 'degree' of sickness. The observed binary outcome Y_i is then obtained simply as whether the degree of sickness exceeds a certain threshold. In formulae, if we denote the threshold by t, then $Y_i=1$ if $Z_i>t$, and $Y_i=0$ when $Z_i \leq t$. Sometimes this may seem a plausible theoretical construct, and sometimes less so. Mathematically speaking, any model for Z_i is then translated into a model for the binary outcomes. In particular, Eq 22.1 is obtained exactly when $t=0$ and

$$Z_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{herd}(i)} + \varepsilon_i \qquad \textit{Eq 22.2}$$

where the fixed effects and the herd effects are exactly as before, and where the error terms ε_i are assumed to follow a so-called logistic distribution with mean zero and variance $\pi^2/3$. (The logistic distribution is similar in shape to the normal distribution, and for most practical purposes, it is equivalent to assume either of these distributions.) Eq 22.2 is a linear mixed model for Z_i! Therefore, computation of variance components and *ICC*s for Z_i-variables follows the rules of Chapter 21 (see Example 22.2):

$$\text{var}(Z_i) = \text{var}(u_{\text{herd}(i)}) + \text{var}(\varepsilon_i) = \sigma_h^2 + \pi^2/3$$

$$\rho = \sigma_h^2/(\sigma_h^2 + \pi^2/3)$$

To summarise, the latent variable approach allows interpretation in terms of variance components and *ICC*s by fixing the error variance at $\pi^2/3$ (Example 22.2). We should keep in mind that the strict interpretation is for the latent variables, and the values are only approximate for the binary outcomes. In particular, as noted, the variances and correlations are not constant for the binary outcomes but depend on the predictors; this dependence has disappeared for the latent variables. Experience with different methods for computing *ICC*s indicates that the latent variable *ICC* tends to be somewhat larger than the true *ICC* for the binary outcome (see the above-cited papers).

Example 22.2 Variance components and *ICC* for the pig-pneumonia data
data=pig_adg

Based on the model presented in Example 22.1, we calculate a total variation of 0.879+3.290=4.167, and an *ICC* (and proportion of variance at the herd level) of

$$\rho = 0.879/4.167 = 0.21.$$

Intra-herd correlations ranging from 0.04 to 0.42 (with most values <0.2) have been observed for a number of infectious diseases of animals (Otte and Gumm, 1997).

22.3 Poisson regression with random effects

A Poisson regression model with exposure n and herd random effect u can be written:

$$\ln(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{herd}(i)},$$

$$Y_i \sim \text{Poisson}\,(n_i \lambda_i) \qquad \textit{Eq 22.3}$$

The random effect is added to the fixed effects in a similar way as for logistic regression (see Example 22.3).

22.4 Generalised linear mixed model

The examples of mixed models in the first two sections extend to a larger class of models called generalised linear mixed models (GLMMs). These models are constructed similarly to those already shown by adding random effects to generalised linear models. We, therefore, set out to present the class of generalised linear models.

22.4.1 Generalised linear model

Generalised linear models (GLMs) were developed in the 1970s to provide a common framework for a wide range of statistical models, including both continuous and discrete distributions, with similar model-building and analysis to linear models based on the normal distribution (McCullagh and Nelder, 1989). The cornerstone of GLMs is the **link function**: the idea that linear modelling of predictors should be allowed to take place on a different scale from the scale of the observations. The link function makes that transition between the observation's mean and the linear modelling. This idea may have been triggered by realising the problems of linear modelling of the observation's mean for bounded distributions. For example, modelling disease probabilities as a linear function of predictors may easily lead to predicted values outside the allowed range of probabilities (*ie* between 0 and 1). In logistic regression, we model instead the $\text{logit}(p)=\ln[p/(1-p)]$ as a linear function of predictors. In GLM terminology, the logit function is the link function. The logit function, which maps the unit interval (0,1) onto the entire real axis is shown on the left of Fig 22.1. Intuitively, this is like 'stretching' the interval. The graph on the right shows its inverse function, $\text{logit}^{-1}(s)=e^s/(1+e^s)$.

Example 22.3 Random effects Poisson model
data=tb_real

In Examples 18.1 and 18.3, Poisson and negative-binomial models were fit to data on the incidence of new TB cases in cattle and cervid herds in Canada. The simple Poisson model (Example 18.1) was clearly inappropriate due to overdispersion. Below are the results from a random effects Poisson model with random herd effects which were assumed to have a normal distribution.

log likelihood = -143.56

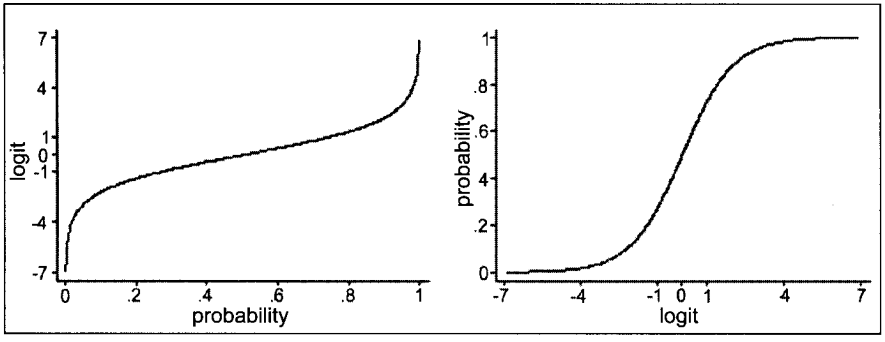
	Coef	SE	Z	P	95% CI	
type_2	-0.395	0.333	-1.19	0.236	-1.047	0.257
type_3	-0.239	0.487	-0.49	0.623	-1.193	0.715
type_5	-0.110	0.801	-0.14	0.891	-1.680	1.460
sex	-0.339	0.208	-1.63	0.103	-0.747	0.069
age_1	2.716	0.747	3.64	0.000	1.252	4.180
age_2	2.466	0.725	3.40	0.001	1.044	3.888
constant	-10.716	0.872	-12.29	0.000	-12.425	-9.007
logpar	(offset)					

In addition, the estimated variance of herd random effects was

$$\sigma_h^2 = 1.685\ (0.587).$$

Compared with the negative binomial model, the type of animal remains completely insignificant while the coefficients for sex and age groups have generally moved slightly away from the null and their P-values have gone down. The random effects Poisson model fits the data substantially better than the negative binomial model because the log-likelihood is -143.6 compared with -157.7; the models are not nested (so a likelihood ratio test does not apply) but they have the same number of parameters (so log-likelihood values can be compared directly).

Fig. 22.1 Logit function (left) and inverse logit function



In theory, the link function can be arbitrary, but in practice, it is restricted to a few common choices for each distribution of Y. For binary/binomial data, two occasionally encountered alternatives to the logit function are the so-called **probit** function (inverse cumulative probability for the standard normal) and the **complementary log-log** function. The statistical inference using logit and probit links is usually similar, but parameter estimates are scaled roughly by the factor $\pi/\sqrt{3}$ (*ie* logistic regression estimates are numerically larger than those from a probit regression). For count data (and a Poisson or negative binomial distribution), the (natural) log is the most common link function but one might also encounter the identity function (*ie* no transformation). Also for ordinal data (and a multinomial distribution), the logit is the most common link. No formal statistical procedure exists to select the 'best' link function. We would usually use the most common one for the data type at hand, or perhaps (especially if the model showed lack of fit) try some of the alternatives and choose the one that gives the best fit to the data.

For the sake of completeness, we summarise the discussion by listing all the components of a generalised linear model:

1. a link function,
2. a distribution of the outcome Y,
3. a set of explanatory variables (in a design matrix X), linked to the mean of the i^{th} observation, μ_i=E(Y_i), by the equation:

$$\text{link}(\mu_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} \qquad \textit{Eq 22.4}$$

4. an assumption of independence between the outcomes.

Unless you are already familiar with GLMs, we recommend that you take some time to revisit the chapters on binary data (Chapter 16) and count data (Chapter 18), and assure yourself how the models fit into the GLM framework.

In previous chapters you have gone through the statistical analysis of several GLMs, so we will not repeat all the details here. It might be worthwhile, though, to note some major differences from linear models. The most obvious difference is that in a GLM with a non-identity link, all parameters are obtained on a transformed scale, and in order to give meaningful interpretations, we need to backtransform our results to the original scale, using the inverse link function. Model-specific rather than general methods are used, *eg* the odds-ratio for logistic regression.

Moreover, comparing Eqs 22.4 and 21.1, we note again that the error term ε_i is missing. This is because the distribution assumption is on the original scale and not on the transformed scale. In many GLMs (including logistic and Poisson regression), the variance is not a separate parameter but a function of the mean. This implies that estimation of the mean and variance in GLMs are mutually dependent. Sometimes the actual data show more or less variability than expected by the variance formula; this is called over- or underdispersion. It is possible to incorporate an additional over/underdispersion parameter into the models (see section 22.5.5).

Estimation in GLMs is generally based on the likelihood function (*ie* estimates are maximum likelihood estimates), unless the additional dispersion parameter is included,

in which case there is no longer a genuine likelihood, and we talk instead about quasi-likelihood estimates. The difference between the two likelihoods and estimation principles is of little practical importance for GLMs but becomes more of an issue with GLMMs (section 22.5). Statistical inference is approximate (unless special procedures for exact inference are used) and to a large extent, based on the (quasi)-likelihood function or equivalently the **deviance** (which is essentially just another name for twice the log-likelihood function).

22.4.2 Generalised linear mixed model

Turning a GLM into a GLMM is straightforward: add the desired random effects on the transformed scale specified by the link function in the same way as we did in the logistic and Poisson regressions. Using the vector-matrix notation introduced in section 21.2.2, the resulting equation for the **linear predictor** specifying the modelling of fixed and random effects becomes:

$$\text{link}(\mu) = X\beta + Zu \qquad \textit{Eq 22.5}$$

where u is the set of random effects and Z is the design matrix for the random part of the model. The random effects are assumed normally distributed with mean zero but possibly involving some non-zero correlations. Among the four assumptions listed in section 22.4.1, Eq 22.5 replaces Eq 22.4 but the other three assumptions are still valid, although 2. and 4. are now conditional on the values of the random effects (section 22.2). Also the discussion of correlation structure and interpretation of fixed effects and variance parameters from section 22.2 carries over to GLMMs, except that the latent variable approach for computing variance components and *ICC*s only works for binary/binomial data. Example 22.4 introduces the data and model we use to illustrate statistical methods for GLMMs.

22.4.3 Other random effects models

Mixed models with the random effects on an original scale (instead of the transformed scale as in a GLMM) do exist, and we briefly mention two of them here.

The **beta-binomial** model has been used extensively in veterinary epidemiology (*eg* Donald et al, 1994). As indicated by the name, it is a model for binomial data incorporating beta-distributed random effects for probabilities. One major advantage of this model is that the likelihood function is given by a relatively simple and explicit formula (which is not the case for GLMMs), and therefore the model is numerically simpler to compute than GLMMs. As one of its drawbacks, it is a 2-level model and has no natural extension to several hierarchical levels. Furthermore, it does not, in a natural way, allow for predictors at the lowest level; it is essentially a model for grouped or replicated binary data.

The **negative binomial** distribution was introduced in Chapter 18 as an extension of a **Poisson** distribution with overdispersion. Overdispersion meant that, in addition to $Y \sim \text{Poisson}(\lambda)$, there is random variation in the mean (λ) according to some distribution (and hence, the var(Y) is greater than the mean). Such variation may be attributed to

Example 22.4 Generalised linear mixed modes (random effects logistic regression) for first service conception data
data=reu_cfs

In a study of reproductive measures, the success or failure of first-service conception (-fscr-) was one of the outcomes evaluated. The study comprised 3,027 lactations distributed on 1575 cows in 50 herds on Reunion Island. The data were analysed in a 3-level random effects logistic regression model (*ie* with random effects of cows and herds). Strictly speaking, the different lactations of each cow were repeated measures over time, and one might question whether the compound symmetry correlation structure in a hierarchical mixed model is adequate (sections 21.5 and 22.6). However, the very short series of repeated measures per cow (1.9 observations on the average) does not realistically allow any more complex modelling. The model contained two dichtomous, lactation-level predictors: -heifer- (primiparous vs multiparous) and -ai- (artificial insemination vs natural breeding).

	Coef	SE	Z	P	95% CI	
ai	-1.017	0.130	-7.80	0.000	-1.273	-0.762
heifer	-0.064	0.097	-0.66	0.510	-0.254	0.126
constant	0.577	0.129	4.47	0.000	0.324	0.829

In addition, the estimated variances of the cow and herd random effects were, respectively:

$$\sigma_c^2 = 0.262\ (0.120) \text{ and } \sigma_h^2 = 0.089\ (0.039)$$

Our first impression of these estimates is that there is no effect of parity and a clear, negative effect of artificial insemination on the conception rates. Both the random effects seem small but their significance is difficult to assess. The *ICC*s between two observations from the same cow (ρ_c) and between two observations on different cows in the same herd (ρ_h) can be estimated using the latent variable approach (sections 21.2.1 and 22.2.2):

$$\rho(\text{lactations of same cow}) = \frac{0.262 + 0.089}{0.262 + 0.089 + \pi^2/3} = 0.096$$

and

$$\rho(\text{lactations of different cows in same herd}) = \frac{0.089}{0.262 + 0.089 + \pi^2/3} = 0.024$$

'inter-subject variability' – a heterogeneity between subjects not accounted for by the Poisson model. If λ has a gamma distribution with shape parameter $1/\alpha$ and scale parameter $\alpha\mu$ (equivalently: mean μ and variance $\alpha\mu^2$), then Y is a negative binomial distributed with mean μ and variance $\mu+\alpha\mu^2$, as shown in Eq 18.9. This distribution may also be called a compound or mixture Poisson model. Note that these random effects cannot be used for modelling of a hierarchical structure, because they are already incorporated into the negative binomial distribution and because they are at the lowest (subject) level. However, if the clustering in the data that made a Poisson distribution inadequate in reality might have derived from a known hierarchical structure, a Poisson regression with corresponding random effects might be preferable to a negative

binomial regression because it directly models the structure. In Example 22.3 it was noted that the random effects Poisson model fit the data much better than the negative binomial model.

22.5 Statistical analysis of GLMMs

Despite the apparent simplicity of models such as Eq 22.1 and Eq 22.3, analysis of GLMMs is not straightforward, even in the logistic and Poisson regression settings. In contrast to most other models in the book, even the estimation of parameters is not clear-cut. A number of different methods exist, and in some situations, they give appreciably different results. No definitive answer exists at this point as to which method is preferable. We outline briefly the methods available and indicate if, and where, they are discussed in this text.

1. Maximum likelihood estimation (section 22.5.2): the likelihood function involves an integral over each random effect, which must be approximated by a summation and therefore makes ML estimation computationally demanding for large models.
2. Quasi-likelihood or iterative weighted least squares estimation (section 22.5.1): existing algorithms for linear mixed models and GLMs are combined to produce slightly different variants of an algorithm, which is fast and computationally simpler than ML estimation.
3. Bootstrap methods (not discussed further): algorithms added to the quasi-likelihood methods intended to remove bias, which lends these algorithms extra computational demands and complexity, and requires specialised software (MLwiN).
4. Bayesian MCMC (Markov chain Monte Carlo) estimation (section 23.4): based on an entirely different statistical approach (Bayesian statistics) and simulation-based estimation. This is a field which has seen rapid development during the last decade.

All results shown so far in this chapter have been from maximum likelihood estimation. But how does one determine which method is best, in general, for one's own data? The standard answer is to use simulation, *ie* generate artificial data from a model with known values of all parameters and then compare the results of different methods with those known values. Such simulation studies are regularly published in statistical journals (*eg* Browne and Draper 2003), and you could also carry out your own simulation study for the data structure at hand (Stryhn et al, 2000).

22.5.1 Quasi-likelihood estimation

A quasi-likelihood function could be thought of as a substitute for a (real) likelihood function whenever the latter does not exist or is too difficult to compute. In the early 1990s, when computers were much less powerful, several algorithms employing an iterative weighted least squares scheme were developed to maximise quasi-likelihood functions for GLMMs. These algorithms are referred to by many different acronyms, typically containing the letters QL (for quasi-likelihood) or ILS (for iterative and least

squares), and usually in combination with a G for generalised or a W for weighted or an R for reweighted or restricted. The main idea of the iterative weighted least squares methods is to compute an 'adjusted' variate on the scale given by the link function (*eg* logistic scale) in each step of the iteration. Technically, the adjusted variate is obtained by a Taylor expansion of Y around the current estimated mean, but one may think of it as a continuous version of the discrete outcome. Estimation for this adjusted variate is carried out using estimation procedures for linear mixed models (weighted REML or ML estimation). The procedure continues until convergence of the parameter estimates. Again, for the technically interested reader, some common options in the procedure are mentioned below:

- first or second order Taylor expansion, the latter being considered more accurate whenever the procedure converges,
- ML or REML estimation for the adjusted variate, the latter being the more commonly used,
- MQL or PQL form of the adjusted variate (M=marginal, P=predictive or penalised), the former being computationally more robust by omitting estimates of random effects in the linear predictor, but the latter being considered more appropriate for subject-specific inference (section 22.2.1).

The three options can be combined arbitrarily (depending on the facilities of the software package used). Algorithms of this type are available in many general statistical packages plus the specialised hierarchical or multilevel software (see section 21.4). Example 22.5 shows results from one of these algorithms.

Example 22.5 Quasi-likelihood estimation of a GLMM
data=reu_cfs

A quasi-likelihood estimation (second order, PQL, REML) of the first-service conception data performed using MLwiN software gave the estimates:

	Coef	SE	Z	P	95% CI	
ai	-0.995	0.123	-8.12	0.000	-1.235	-0.755
heifer	-0.064	0.093	-0.69	0.490	-0.247	0.119
constant	0.567	0.123	4.47	0.000	0.326	0.809

In addition, the estimated variances of the cow and herd random effects were, respectively:

$$\sigma_c^2 = 0.153\ (0.080) \text{ and } \sigma_h^2 = 0.088\ (0.034)$$

All the estimates and SEs are in close agreement with those from Example 22.4, except the cow level variance which is only about 60% of the previous value. We interpreted the disagreement about this value as a bias of the quasi-likelihood estimation procedure; a simulation study confirmed that with these data the quasi-likelihood procedure would consistently give too low cow-level variance estimates (Stryhn et al, 2000).

A long list of (statistical) papers from the 1990s discuss the different versions of algorithms and their implementation in software packages (*eg* Zhou et al, 1999). For well-behaved data, the different variants of the algorithms give very similar results (taking into account the standard errors of the estimates). One should whenever possible use the 'best' possible of the above options (second order, REML, PQL). More importantly, any 'strange-looking' estimates or standard errors should cause the model to be examined carefully and the results to be confirmed with other models or estimation methods.

Early simulation studies showed that estimates from some of the iterative least squares algorithms for GLMMs could be markedly biased towards the null. The bias might affect both fixed and random effect parameters, but the latter are particularly sensitive. The general consensus seems to be that particular caution should be exercised if:

- the number of replications at a hierarchical level is 'small' (*eg* less than 5),
- the corresponding random effect is 'large' (*eg* the variance exceeds 0.5).

In our example, the number of cow-level replications was indeed small, with only an average of 1.9 observations per cow.

22.5.2 Maximum likelihood estimation

Maximum likelihood (ML) estimation in GLMMs would, at first sight, seem to be our first choice, because of the overall strengths of the method (good statistical properties of the ML estimates) and the access to likelihood-based inference (*eg* likelihood ratio tests). However, ML estimation has for many years had the reputation of being unfeasible for any GLMM beyond the simplest 2-level models, due to the massive and difficult computations required. Recent advances in computer power and software have changed this judgement; the ML analyses of this chapter used the powerful Generalised Linear Latent And Mixed Models (gllamm) macro for multilevel modelling implemented in Stata (Rabe-Hasketh et al, 2002). It is likely that, within a few years, ML estimation will become the standard estimation approach for all but huge GLMMs. Even if the method's numerical side now looks promising, we outline why computation of the likelihood function is so difficult and give some cautions (complex procedures always have pitfalls, even if the complexities are hidden in the software).

For simplicity, consider the 2-level logistic regression model (Eq 22.1) and let us begin by focusing on a single herd – herd 1. Given the value of u_1 (the random effect of herd 1), the conditional likelihood of the observations from that herd is binomial,

$$L_1(\beta|u_1) = \prod_{i:\,\text{herd}(i)=1} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

and the full (sometimes denoted marginal) likelihood for those animals is obtained by integration over the distribution of the random effect u_1:

$$L_1(\beta) = \int L_1(\beta|u_1)(2\pi\sigma_h)^{-1/2}\exp\left(-\tfrac{1}{2}u_1^2/\sigma_h^2\right)\mathrm{d}u_1 \qquad \textit{Eq 22.6}$$

The integration weights the possible values of u_1 according to their likelihood in a normal distribution with mean zero and standard deviation σ_h. Integrals such as Eq 22.6 cannot be solved analytically, and therefore a numerical integration or **quadrature** becomes necessary, approximating the integral by a weighted sum of values of the integrand (*ie* the function being integrated) at a number of selected quadrature points. In such a scheme, you need to decide on the number of quadrature points and the way they are selected. Generally, increasing their number improves both accuracy and calculation time. Also, it is generally recommended that an **adaptive** approximation method be used, where the quadrature points (and their weights) are successively adapted to the integrand.

So far we have dealt only with observations from one herd. Observations from different herds are independent, so the full likelihood function for the entire dataset is obtained as a product of terms such as Eq 22.6 over the total set of herds. We trust it is not necessary to write out the equation to make the point that, not only computing, but also maximising, a quadrature approximation to such a multiple integral with respect to the fixed and random effects parameters of the model is a formidable task. Extension to multiple levels and/or multiple random effects at the same level rapidly increases the complexity of the problem.

To summarise, a few recommendations and cautions for the use of ML estimation for GLMMs:

- ML estimation might be computationally unstable or the approximation of the likelihood function may be insufficient; it is highly recommended, therefore, that the stability of results be checked by trying different starting values of the algorithm and/or different variants of the numerical integration procedure, such as a different number of quadrature points as well as adaptive procedures,
- ML estimation should always be compared with other approaches (either quasi-likelihood estimation or other approaches for clustered data), and caution should be exercised if major differences appear,
- ML estimation in GLMMs may be impractical for model selection (because of computational demands); it is then considered legitimate to use computationally simpler methods for (part of) the model selection and confirm the results by running selected models also by ML estimation.

In Example 22.6, we examine the stability of the quadrature behind the ML estimates.

22.5.3 Statistical inference

Statistical inference in GLMMs is generally only approximative (asymptotically correct when the number of observations at all hierarchical levels is large). Fixed effects parameters are usually assessed by Wald-type confidence intervals and tests, however likelihood-based inference (profile likelihood CIs and likelihood ratio tests, see section 21.4) may be preferable, in particular when the parameters are highly correlated or not well determined. (As in GLMs, Wald-type statistics are useless for parameters that are 'out of bounds', *eg* in logistic regression when one category of a predictor has no cases.) However, likelihood-based inference is only feasible when

Example 22.6 Checking maximum likelihood estimation of a GLMM
data=reu_cfs

In Example 22.4, ML estimation was used to fit a random effects logistic regression model to the Reunion Island first-service conception risk data. This model was refit using a range of number of quadrature points in the estimation procedure. The fixed and random effects estimates from each estimation were:

	Number of quadrature points			
Estimate	6	8	10	12
ai	-1.017	-1.017	-1.017	-1.017
heifer	-0.064	-0.064	-0.064	-0.064
constant	0.576	0.577	0.577	0.577
cow variance	0.262	0.262	0.262	0.262
herd variance	0.090	0.089	0.089	0.089

This model seems very stable as the estimates changed very little as the number of quadrature points was increased.

ML estimation is used. Reference distributions are most commonly 'asymptotic', *ie* the standard normal or χ^2-distributions. The resulting inference may be too liberal if replication is sparse at the level of the parameter of interest (Stryhn et al, 2000), and some software packages give the option of using similar approximations as for linear mixed models using t- and F-distributions. In general, no clear guidelines can be given about the accuracy of approximative inference in GLMMs. As in linear mixed models, Wald-type statistics are inappropriate for variance parameters, which should therefore be assessed by likelihood-based inference (Example 22.7) or alternative procedures such as bootstrapping.

Example 22.7 Statistical inference in a GLMM
data=reu_cfs

The tests and confidence intervals given for the fixed effects in Example 22.4 are 'asymptotic'; for example, the 95% CIs are computed as $\beta \pm 1.96 * SE(\beta)$. Both predictors vary (potentially) at the lowest level and even if some variation may reside at higher levels (in particular, the cow level) there would seem ample replication to justify inference based on the standard normal distribution.

To compute tests for the random effects of the model, we note the log-likelihood value of the fitted model (-2010.85) and refit the model without the random effect of interest. The models without cow random effects and herd random effects had log-likelihood values of (-2014.11) and (-2017.93), respectively, so that the corresponding χ^2-statistics with 1 df were 6.52 and 14.2, and thus both significant. Recall from section 21.4.3 that P-values should be computed as half the tail probability from the $\chi^2(1)$-distribution to account for the one-sided alternative hypothesis. It is perhaps interesting to note that the herd random effect was clearly the most significant of the two; this is not at all obvious from the estimates and standard errors.

22.5.4 Model-checking

The standard tools for model-checking, residuals and diagnostics, are even less developed and accessible for GLMMs than for linear mixed models (section 21.4.4). The main new point for GLMMs (compared with linear mixed models) is that, because the model has no normally distributed error terms at the lowest level, the corresponding residuals and diagnostics at that level are difficult to assess. As an extreme example, in a binary model all the lowest-level residuals are dichotomous and cannot be expected to conform to a normal distribution. In this case, the residuals at the lowest level are not very informative. Unfortunately, the problems with the lowest-level residuals could penetrate to the higher levels if there is little replication. Reference distributions and points for residuals and diagnostics are therefore difficult to use rigorously, and one is advised instead to look for data points that are extreme in some way relative to the rest of the data. Example 22.8 discusses the residuals from our 3-level Reunion Island data.

GLMM analogues of some of the special statistics for discrete data, such as the Hosmer-Lemeshow test for goodness of fit in a logistic regression model, are not available. However, quasi-likelihood estimation procedures allow for estimation of an over/underdispersion parameter as in a GLM. This parameter gives an indication of how the distribution specified as part of the GLMM fits to the data (section 22.5.5).

22.5.5 Over- and underdispersion in GLMs and GLMMs

Overdispersion has been mentioned in a number of places in this book; this section summarises both estimation and interpretation of over- and underdispersion. A dataset is said to contain **overdispersion** (underdispersion) if the variance in the data is larger (smaller) than expected from its mean and the assumed probability distribution (with an inherent relation between the mean and variance), see sections 18.5.3, 18.5.5 and 20.3.2 for examples involving the Poisson and binomial distributions. To begin with, it follows that over- and underdispersion are always **relative to** an assumed probability distribution/model. Changing (improving) the model may cause the anomalous dispersion to vanish. It also follows that over- and underdispersion are only meaningful for distributions where the variance is determined by the mean – and therefore, not for the normal distribution.

Interpretation of over- and underdispersion

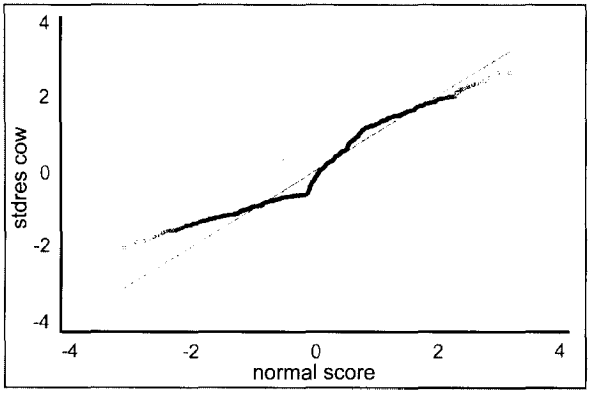
Strictly speaking, overdispersion means lack of fit – that the distribution just does not fit the data. It may be caused by omission of an important factor affecting the data such as a clustering in the data (associated with the levels of that factor). Such clustering may be attributable to unobservable quantities, for example herd effects caused by management factors or litter effects caused by genetic and environmental factors. Another common cause of overdispersion is positive correlation between observations, *eg* in a series of measurements over time. Before proceeding with any statistical methods for overdispersion, make sure that the overdispersion is not caused by a simple oversight and cannot be remedied by improving the model.

Underdispersion is possible as well but less common in practice. It can result from a negative correlation between observations, the standard example being competition in a group of animals for a limited resource (*eg* feed). Because underdispersion means a better fit than expected to the data of our model, we often tend not to worry much about it, maybe it was just 'good luck'. By ignoring an appreciable underdispersion and pretending the dispersion to be as predicted by our model (when it is in reality smaller), our statistical inference becomes conservative – which may be considered the appropriate approach for 'a case of good luck'. Underdispersion (as well as very small values of one-sided test statistics) may however also indicate something strange to be going on in the data. There are some (in)famous examples of data manipulation that were revealed by very strong underdispersion (*ie* the data fit too well to the hypothesis examined!). In general, if strong over- or underdispersion is present in the data, one should always explore the data (residuals) and search for possible explanations.

Example 22.8 Residuals from a 3-level GLMM
data=reu_cfs

The 3-level logistic regression of Example 22.4 (Reunion Island first-service conception data) has residuals at all three hierarchical levels but the lowest-level residuals are of little use in this case so we disregard them completely. A normal (Q-Q) plot for the 1,575 cow-level standardised residuals is given in Fig 22.2. The plot shows a curious pattern, far from a straight line but instead with three separate, almost straight, lines. One must realise that with typically only 1-3 observations per cow and only four different sets of predictor values, the cow-level residuals cannot realistically be expected to look like a normal distribution sample. With closer scrutiny, each approximately linear part of the graph correspond to cows with the same response pattern. For example, the lower part of the plot corresponds to cows without any first-service conceptions in the dataset and the upper part of the plot to cows that conceived at first service in all lactations. It seems almost impossible to assess from the plot whether there are problems with the model assumptions at the cow level.

Fig. 22.2 Normal plot for cow-level residuals of 3-level model for Reunion Island data

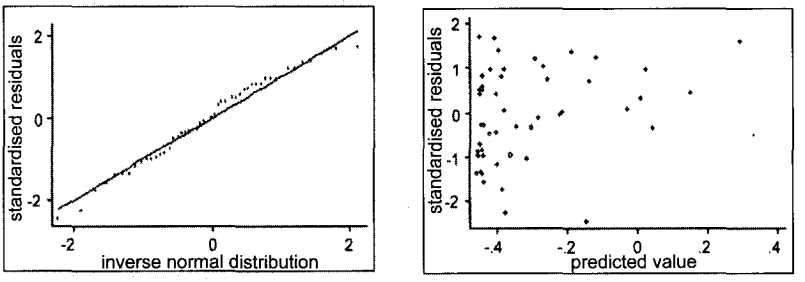


(*continued on next page*)

Example 22.8 (*continued*)

Fig. 22.3 shows the herd-level residuals depicted in a normal plot and plotted against the herd-level predicted values (including cow-level predictors). The normal plot is somewhat skewed due to lack of herds with strongly positive residuals; however, when comparing with the lower tail of the distribution, you can see that only two negative residuals are more extreme than in the upper tail. The plot against the fitted values shows a grouping of predicted values at the lower end of the scale but no particular patterns in the residuals.

Fig. 22.3 Normal plot (left) and plot against predicted values for herd level residuals of 3-level model for Reunion Island data



The residuals shown in this example were computed using MLwiN.

Over- and underdispersion in a GLM

Assuming that the anomalous dispersion is not caused by model deficiencies (as discussed above) and therefore corresponds to a 'natural' dispersion, a dispersion (or scale) parameter ϕ can be introduced into a GLM to incorporate the dispersion into the model. For our two main discrete distributions, this amounts to assuming:

$$\begin{aligned} Y \sim \text{bin}(n, p): &\quad \text{E}(Y) = np \quad \text{and} \quad \text{var}(Y) = \phi\, np(1\text{-}p) \\ Y \sim \text{Poisson}(\lambda): &\quad \text{E}(Y) = \lambda \quad \text{and} \quad \text{var}(Y) = \phi\, \lambda \end{aligned} \qquad \textit{Eq 22.7}$$

Thus, $\phi > 1$ corresponds to overdispersion, $\phi < 1$ to underdispersion and $\phi = 1$ to the variance following exactly the relationship of the distribution. When a dispersion parameter is assumed, the distributions are no longer truly binomial or Poisson distributions; therefore, no likelihood function is available for the model, and one must resort to quasi-likelihood estimation.

Overdispersion is detected by calculating goodness-of-fit statistics for the model, either the deviance G^2 or the Pearson χ^2-statistic. Under the true model ($\phi = 1$) both are expected to approximately follow a χ^2-distribution with degrees of freedom (df) equal to the number of observations minus the number of estimated parameters (which in Eq 22.4 would be $k+1$). **Note** For binomial data, the approximation requires replication – that the binomial denominators $n_i > 1$, see section 16.11.1 for a discussion of binomial versus binary data. If the data are binary and no grouping corresponding to a common linear predictor for several observations is possible, the values of χ^2 and G^2 are meaningless, and overdispersion cannot be modelled.)

If these statistics are significantly larger than expected, overdispersion exists and ϕ must be estimated. The simplest estimate applies to a model where ϕ is constant across the entire dataset (thus, in our example independent of the binomial denominators n_i), where one simply calculates $\phi = \chi^2/df$. Estimation of ϕ based on G^2 is less useful, in particular in binomial models with small denominators. All other parameter estimates are unchanged compared with a model with $\phi = 1$, but the estimated standard errors are multiplied by $\sqrt{\phi}$, and Z- and χ^2-statistics are divided by $\sqrt{\phi}$ and ϕ, respectively. This procedure is valid regardless of the value of ϕ but caution should be exercised with values below 1. As discussed above, values <1 may be ignored and distributional dispersion assumed instead. In Example 22.9, we revisit overdispersion in a binomial dataset.

Example 22.9 Overdispersion for artificial binomial data

To illustrate, we return briefly to the artificial data of section 20.3.2 with 10 herds of 20 cows each and 100% clustering of a binary outcome in herds, which gave the proportions:

0 0 0 0 0 1.0 1.0 1.0 1.0 1.0

The Pearson statistic is χ^2=200, so that ϕ=200/(10-1)=22.22. Note that the calculation is based on the 10 groups corresponding to the herds. The variance of these proportions is σ^2=0.278 and the expected variance from a binomial distribution (20,0.5) is 0.5*0.5/20=0.0125. It is seen that ϕ approximates the ratio between the actual and expected variances very closely (0.278/0.0125=22.24). For this particular setting, where the overdispersion can be attributed to clustering in herds, it would seem more appropriate to extend the model with herd effects instead of modelling the overdispersion. See also section 23.2.2 about using overdispersion factors to account for clustering.

Over- and underdispersion in GLMMs

GLMMs also allow for an additional dispersion parameter (ϕ), which, as noted before, can only be estimated using quasi-likelihood methods. With random effects in the model to take into account any hierarchical clustering, the interpretation of overdispersion is more difficult, although additional (unrecognised) clustering may be a possibility.

It is commonly experienced in GLMMs that the dispersion parameter is estimated at a value below 1 (underdispersion); see Example 22.10 for one such case. A simulation study has indicated values of ϕ in the range 0.8-1 to be a possible artifact generated by the quasi-likelihood estimation procedure (Jacob, 2000). Facing a value in this range, one would usually be content to fix ϕ at 1 (most software will allow you to do this). Generally, the presence of a dispersion parameter in the model may affect the variance parameters considerably; in our experience it is preferable to fix ϕ at 1 whenever that seems sensible, in order to facilitate interpretation of the model. Values of ϕ clearly below 0.8 are definitely suspect and the model should be seriously mistrusted, unless verified by other analyses. One model deficiency that has been shown to generate serious underdispersion is autocorrelated errors in a repeated measures model (Goldstein et al, 2002b).

Example 22.10 Dispersion parameter in a GLMM
data=reu_cfs

Quasi-likelihood estimation (Example 22.5) produced a value of ϕ=0.932 (0.032) for the 3-level logistic model. Despite the fact that it is about three standard errors away from 1, we do not take this value to indicate problems with assuming a binary/binomial distribution for the variance.

22.6 Repeated measures and spatial data

The linear mixed model approach of incorporating correlation structures into the model's error component (ε) runs into the serious problem in GLMMs that the linear predictor (Eq 22.5) does not contain an error component! The reason is that a GLM(M) models the mean and variance on different scales: the mean on the scale given by the link function but the variance on the observation scale. This makes it more difficult to incorporate the correlation structures discussed for linear mixed models into a GLMM and is one of the reasons why no general GLMM-type class of models exists for repeated measures and spatial structures. Instead, models are, to a large extent, developed specifically for the most interesting data types: binary and count data. The literature in this field is large, rapidly developing, and beyond the scope of this book. We give a few notes and references to what has been done within the GLMM framework. Note also that the generalised estimating equations approach (section 23.3) can be used for repeated measures and spatial data structures with discrete outcomes.

The standard hierarchical mixed model with random effects corresponding to the unit of the repeated measures is valid, provided one is willing to assume a compound symmetry correlation structure. This was exactly the type of model used for the conception data from Reunion Island. To detect violations of compound symmetry may require much more data than in the continuous case because the information content is lower in discrete data – something that certainly is true for binary observations.

A GLMM with a correlation modelled at the original scale (Barbosa and Goldstein 2000) and a multivariate multilevel logistic model (Yang et al, 2000) have been developed, but both require specialised software. Quasi-likelihood estimation software may allow specification of a repeated measures or spatial model for the adjusted variate computed in each step of the iteration (section 22.5.1). This will lead to repeated measures or spatial correlation structures (Gotway and Wolfinger 2003), although covariance parameters specified in this way may have no direct interpretation in the discrete model.

SELECTED REFERENCES/SUGGESTED READING

1. Barbosa MF, Goldstein H. Discrete response multilevel models. Quality and quantity 2000; 34: 323-330.
2. Browne WJ and Draper, D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. http://www.maths.nottingham.ac.uk/personal/pmzwjb/index.html 2003 Submitted.
3. Browne WJ, Subramanian SV, Jones, K and Goldstein, H. Variance partitioning in multilevel logistic models that exhibit over-dispersion. http://www.maths.nottingham.ac.uk/personal/pmzwjb/index.html 2003 Submitted.
4. Donald A, Gardner IA, Wiggins AD. Cut-off points for aggregate testing in the presence of disease clustering and correlation of test errors. Prev Vet Med 1994; 19: 167-187.
5. Goldstein, H, Browne, WJ and Rasbash, J. Partitioning variation in multilevel models. Understanding Statistics 2002; 1: 223-232.
6. Goldstein H, Browne WJ, Rasbash J. Multilevel modelling of medical data. Stat Med 2002; 21: 3291-3315.
7. Gotway CA, Wolfinger RD. Spatial prediction of counts and rates. Stat Med 2003; 22: 1415-1432.
8. Jacob M. Extra-binomial variation in logistic multilevel models - a simulation. Multilevel Modelling Newsletter 2000; 12: 8-14.
9. McCullagh P, Nelder JA. Generalized linear models. 2d ed. London: Chapman and Hall, 1989.
10. Otte MJ, Gumm ID. Intra-cluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys. Prev Vet Med 1997; 31:147-150.
11. Rabe-Hesketh S, Pickles A, Taylor C. Generalized linear latent and mixed models. Stata Technical Bulletin 2000; 53: 47-57.
12. Rodriguez G, Elo I. Intra-class correlation in random-effects models for binary data. The Stata Journal 2003; 3: 32-46.
13. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modelling. London: Sage Publishers, 1999.
14. Stryhn H, Dohoo IR, Tillard E, Hagedorn-Olsen T. Simulation as a tool of validation in hierarchical generalized linear models. Proc of Intl Symp on Vet Epidem and Econom. Breckenridge, Colorado, 2000.
15. Vigre H, Dohoo IR, Stryhn H, Busch ME. Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumoniae* and *Mycoplasma hyopneumoniae* at different levels in Danish multisite pig production facilities. Prev Vet Med 2003; accepted.
16. Williams, D. Modelling binary data. 2d ed. Boca Raton: Chapman and Hall/ CRC, 2002.
17. Yang M, Heath A, Goldstein H. Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. J R Stat Soc A 2000; 163: 49-62.
18. Zhou XH, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. American Stat 1999; 53: 282-290.

SAMPLE PROBLEMS

1. For a 2-level logistic regression problem, see Sample Problem 1 in Chapter 23.
2. Using a dataset on preweaning mortality in piglets (prewmort), explore 3-level random effects logistic regression models.
   a. Consider the prewmort data, with -lmort- (a binary indicator for preweaning mortality in a litter) as the outcome and -quarter- (of the year), -sow_tx-, -sow_parity and -herdtype- as predictors.
      i. Draw a diagram of the hierarchical structure, including the predictors.
   b. Fit a 3-level logistic model with fixed effects of the predictors as well as sow and farm random effects.
      i. Which fixed effects are significantly different than zero?
      ii. Modify the fixed part of the model to include only significant terms and, if necessary, relevant interaction terms.
      iii. Interpret the fixed effects in terms of odds ratios.
      iv. Give the subject-specific interpretation of the odds ratios, and explain the difference to a population-averaged interpretation.
   c. Turn your attention to the variance parameters.
      i. Use the latent variable approach to compute the proportion of variance residing at the three levels (litter, sow and farm) of the model.
      ii. Still using the latent variable approach, compute also the intra-class correlation coefficients (*ICC*s) for two observations from the same sow, and for two observations from the same herd.
   d. Maximum likelihood estimation is required to carry out tests for variance parameters; skip this point if your software does not allow for a 3-level model.
      i. Assess the significance of each of the two variance parameters using likelihood ratio tests (recall, that since variances can only be positive the P-value is computed as half of the tail probability obtained from the chi-square distribution).
   e. Quasi-likelihood estimation is required to estimate an additional dispersion parameter; skip this point if your software does not allow it.
      i. Compute an additional over- or underdispersion parameter to assess the data's compliance with the binomial variance assumption. Does it seem reasonable to assume a binomial variation?
      ii. If you did item d as well, compare the parameter estimates from maximum likelihood and quasi-likelihood estimation, in particular assess whether the latter estimates would seem to be biased towards the null.
   f. Your software should allow you to compute 'residuals' (estimated random effects) at the sow and farm levels, and preferably also corresponding standardised residuals.
      i. Inspect the herd level (standardised) residuals for the presence of extreme herds and use a normal plot to assess the residual's agreement with a normal distribution. Compute also any further model diagnostics at the herd level that your software may offer, and draw conclusions about the model's fit at the herd level.
      ii. Same question for the sow level (standardised) residuals. Recall that a close agreement with a normal distribution cannot be expected (why?).

23

ALTERNATIVE APPROACHES TO DEALING WITH CLUSTERED DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the uses, advantages and limitations of simpler methods to deal with clustering, such as fixed effects and stratified modelling, correction factors, robust variance estimation and aggregation of hierarchical levels.

2. Understand the differences between population-averaged and subject-specific modelling.

3. Use generalised estimating equations for analysing clustered data, in particular repeated measures data.

4. Understand the basic differences between Bayesian and classical (likelihood-based or frequentist) statistical approaches.

5. Use a Bayesian hierarchical model with non-informative priors and Markov chain Monte Carlo estimation for analysing clustered data.

23.1 Introduction

The previous two chapters have presented mixed models (*ie* models containing both fixed and random effects) as an approach for dealing with the problem of clustering (lack of independence among observations) in a dataset. As noted, these mixed models are very flexible and can handle any number of levels of hierarchical clustering as well as more complex data structures. However, some unresolved issues remain. As discussed in section 22.6, the mixed models approach is not as successful with repeated and spatial structures for discrete data as it is for continuous data. Also, mixed models are limited by their common assumption of normally distributed random effects; in practice, you will encounter data that clearly do not meet that assumption. From a more philosophical point of view, one might argue that, in our analyses, we should only make the absolutely necessary distributional assumptions and for 'nuisance effects', rely on robust procedures that are less affected by the peculiarities of the data. This would follow the trend in modern statistics toward non- and semi-parametric procedures, as seen for example in survival analysis. Finally, complex mixed models might be difficult to fit due to the size of the data or to numerical difficulties. Consequently, simpler alternatives are valuable, if for no other reason than to confirm the results of the mixed model analysis.

This chapter first reviews a number of methods to deal with clustering which are easier to implement than mixed models. In some situations, these simpler methods might be adequate and more easily carried out by an investigator. Following these simple approaches, we give a short introduction to two other large model classes or statistical approaches that can deal with clustering: generalised estimating equations (GEE) and Bayesian estimation using Markov chain Monte Carlo (MCMC) methods. More precisely, the GEE approach was devised specifically to deal with clustered or other complex data structures, whereas the Bayesian and MCMC approach gives an alternative view on all aspects of statistics, of which models for clustered data are only one example. To illustrate the methods, we will use two previously encountered datasets throughout: the somatic cell count data (scc_40) from Chapter 21 and the pig-pneumonia data (pig_adg) from Chapter 22.

23.2 Simpler methods

23.2.1 Fixed effects and stratified modelling

In a number of examples throughout this book, group (*eg* herd) identifiers have been included as fixed effects in regression-type models, primarily to account for confounding due to factors at the group level (section 13.8 also discussed stratified (Mantel-Haenszel) analysis to control confounding). We have already shown in section 20.4.2 how these same approaches can also be used to deal with the issue of clustering within groups and have discussed their advantages and disadvantages, and the choice between taking group effects as fixed or random (section 21.2). Only a few summary remarks are given here.

Recall that herd random effects have an interpretation in terms of a population (the variance of which we estimate) and that **fixed effects** are specific to the actual herds in the study – and might, therefore, be more appropriate if there is little interest (or sense) in the population interpretation. In other cases, variance estimates might be one of the primary outcomes of interest in an investigation. Perhaps the most severe restriction of fixed effects modelling is that it does not allow for group (herd) level predictors. Technically, fixed effects analyses are usually much easier to carry out. In particular, facing limited model-checking facilities for mixed models in some software, you might be tempted to base part of the model-checking on the fixed effects versions of a mixed model, although, strictly speaking, this is incorrect. Also from a technical aspect, it should be kept in mind when using fixed effects models for discrete data (Example 23.1), that estimating a large number of group (herd) parameters might adversely affect maximum likelihood estimation procedures. **Mantel-Haenszel-type stratified** analyses are limited to binary outcomes and a single within-group predictor; for multifactorial problems.

Example 23.1 Fixed effects and stratified models for pig-pneumonia data
data=pig_adg

Recall that the binary outcome is the presence of pneumonia and our sole predictor is -ar_gl-, a dichotomous variable indicating the presence of atrophic rhinitis. An ordinary logistic regression model gave a value for its regression coefficient (with SE) of 0.647 (0.220), but adding herd random effects reduced the value to 0.437 (0.258). The fixed effects model estimate is 0.365 (0.268) and appears to adjust reasonably well for both herd confounding and clustering; no adverse effect of estimating 15 herd parameters is apparent. The same can be said of the Mantel-Haenszel estimate (of the log odds ratio) of 0.346 (0.261).

23.2.2 Factors to correct for clustering

This section summarises two ways of correcting an analysis in which clustering has not been taken into account in the model. These involve an estimate of the **intra-class correlation coefficient** (**ICC**) (sections 20.3.3 and 21.2.1) or an estimate of the **overdispersion** (section 22.5.5), and using one of these to adjust the standard error (SE) of regression coefficients. Note that these methods rely on the simplistic premise that clustering affects only the SEs of estimates (and, generally, when not taken into account, leads to SEs that are too small). Our previous examples, including those of simulated data in Chapter 20, have shown that this is not always the case. Therefore, not all uncorrected analyses might be 'repaired' by increasing the SEs, and the researcher must pay particular attention to the requirements for these correction factors to be meaningful.

Using a correction factor assumes a 2-level structure (*eg* animals within herds). Example 20.1 shows how the effect of clustering on the variance of herd means depends on both the *ICC* and the herd size. If both of these are the same in all herds,

an analysis involving only herd-level factors, but ignoring clustering in herds, might be corrected by inflating the standard errors of regression coefficients by the square root of the variance inflation factor, assuming you have a value for the *ICC*. In practice, herds are rarely of the same size and the *ICC* is only constant between herds in special cases. One such case is normally distributed data, but then, you might as well use a 2-level linear mixed model (which with today's software, should not pose any numerical problems). The method might be acceptable as an approximation even without these conditions (Donner, 1993), but the validity of the approximation is difficult to assess.

Generalised linear models (GLMs) allow for an additional **dispersion** (or scale) parameter (ϕ) to take into account if the 'natural dispersion' in the data does not match the distribution used (*eg* binomial or Poisson, see Eq 22.7). As before, for a 2-level model with only herd-level predictors, this might be used to adjust for the inflation in variance at the herd level. In principle, this correction is valid for unequal herd sizes as well. Let's look at the binomial example to see how that would work. Denote by Y_i and n_i the number of positive outcomes and the total number of animals in herd i, respectively. Then the model's assumptions are:

$$\mathrm{E}(Y_i) = n_i p_i \qquad \text{and} \qquad \mathrm{var}(Y_i) = \phi\, n_i p_i (1-p_i) \qquad \textit{Eq 23.1}$$

Here ϕ is assumed to be independent of the group sizes n_i, and clearly that is not necessarily true! It makes us realise that using an overdispersion parameter assumes a particular form of the variance inflation across groups. For grouped binomial data, other possibilities exist, *eg* the Williams method (Collett, 2002) which also affects the parameter estimates. For moderately varying herd sizes, the two methods do not differ much (they are identical for equal herd sizes). Also, the two models for grouped binary data that we have discussed in this book – a random effects logistic regression and a beta-binomial distribution – have varying variation inflations across groups. Both the beta-binomial distribution and the Williams procedure assume

$$\mathrm{var}(Y_i) = [1+(n_i-1)\phi]\, n_i p_i (1-p_i)$$

and in the random effects model the inflation depends also on the probability p_i. We illustrate the methods by the pig-pneumonia data (Example 23.2) although it does not directly fit into the framework (because the predictor -ar_g1- is not at the herd level).

The advantages of the simple overdispersion approach are its numerical simplicity and that it does not assume a parametric form of the random effects. You can also use ordinary regression model diagnostics after fitting the model. The size of the overdispersion parameter provides an estimate of the severity of the clustering problem. The disadvantages are a potential problem in estimating ϕ (when there is sparse replication), assuming the overdispersion to be constant when group sizes (n_i) differ strongly, the lack of likelihood-based inference and its limitation to grouped (binomial) data. As noted in the introduction, using overdispersion more generally to compensate for non-modelled hierarchical clustering is not recommended., particularly because there is little reason to believe that the only effect of clustering is to increase the standard errors.

Example 23.2 Overdispersion for pig-pneumonia data
data=pig_adg

In order to have the predictor (-ar_g1-) at the group level, we must redefine the groups – instead of herds they should be the groups of pigs within herds with the same predictor value. That is, there are two groups (ar-positive and ar-negative) in each herd for a total of 2*15=30 groups. All ar-positive groups have the same predicted value by the model, and overdispersion means that these 15 counts of pn-positive pigs disperse more than we would expect from their predicted value; the same principle applies to the ar-negative animals. In the presence of clustering in herds, these counts would indeed show overdispersion. Therefore, the overdispersion parameter does take into account extra variation arising from herd clustering. What the model and approach does not take into account, and this is clearly unsatisfactory, is that we would expect strong correlation between the two counts from the same herd.

Estimation in this grouped dataset gives the Pearson χ^2=78.24 with 28 df, and thus an overdispersion value of ϕ=χ^2/df=2.794. Further, the regression coefficient is unchanged at 0.647 and the SE has been increased to 0.368 (computed as 0.220*$\sqrt{2.794}$). The method fails to adjust for the confounding and/or clustering in the data. It is interesting to note that the Williams method applied to the same data gives the value 0.424 (0.362) and thus, seems to perform better here; however, we do not go into details with this method.

23.2.3 Robust variance estimation

In a 'usual' regression model (linear, logistic *etc*), the SEs of the coefficients in the model are based upon the assumption that the model is true in all respects and that the errors are independent and follow the appropriate distribution (Gaussian for a linear model) or binomially distributed (for a logistic model). If these assumptions are met and you had an infinite sample, the estimated β would be correct and you would have an SE of zero.

There is an alternative approach to computing the variance (and hence the SE) of β that is referred to as robust variance estimation, or Huber-White variance estimation, or 'sandwich' variance estimation (so called because, in matrix notation, the formula for the variance matrix of the βs looks like a sandwich). These estimates are less sensitive to the assumptions on which a model-based estimation is built but they also have a slightly different interpretation. The SEs simply estimate the expected variability in the βs if repeated samples of the same size as the dataset were drawn from the original population, and thus, are somewhat analogous to bootstrap SEs (Guan 2003). As such, they are more robust to violations of any of the assumptions on which the model is based and usually (but not necessarily) produce larger SEs (and hence, wider CIs) for parameters than the usual variance estimates. While robust SEs might also be computed for discrete data, the 'robustness' is less obvious with discrete data because model misspecifications might affect not only the variances but also the estimates themselves. The robust variance estimate can also be allowed to vary across clusters, which is important when dealing with clustered data, because it relaxes the assumption of independence to require only independence of observations across clusters, not within clusters. A more complete discussion of alternative variance estimation procedures

(including sandwich estimators and others) can be found in Hardin and Hilbe (2001).

The advantages of robust variance estimation are that it is simple to use (if implemented by your software) and does not require specific assumptions about the nature of the clustering. For linear models, it provides SEs that are robust to different violations of the model assumptions (*eg* distribution of errors and heteroscedasticity). One disadvantage of this approach is that it provides no information about the magnitude or origin of clustering. Further, it has no impact on the point estimates of the parameters, which might be considered particularly critical for non-normal data, and the SEs differ in their interpretation from usual SEs. Finally, it might also be said that robust variance estimation is part of the generalised estimating equations (GEE) approach to clustered data (section 23.3) which offers more control over the modelling without requiring additional assumptions. We illustrate the robust variance method by two examples: discrete data (Example 23.3) and continuous data (Example 23.4).

Example 23.3 Robust variance estimation for pig-pneumonia data
data=pig_adg

The regression coefficient for -ar_g1- is not affected by employing robust variance procedures, only its SE. Cluster-specific robust variance estimation, using farms as clusters, increased the SE to 0.276. We previously noted the inadequacy, for this example, of methods that do not affect the regression coefficient.

23.2.4 Aggregation of levels

The hierarchical structure in a dataset might contain many levels, as shown in the 5-level structure of Fig. 20.1. However, sometimes we decide to exclude some levels from our analysis, and in this section, we give a few comments related to two common scenarios. In order to estimate the variation and the random effects at the different levels, a certain minimal amount of replication is necessary at all levels. This is intuitively obvious because, if, for example, all batches contained only a single litter, then there would be no way of distinguishing between batch and litter effects. Another potential problem for the estimation procedure is a strongly variable replication at one of the hierarchical levels (*eg* if some batches contain only one litter while others contain up to 10 litters). To detect such problems, it is worthwhile to compute the range and average of the number of replications at each hierarchical level. There is no definitive rule as to the minimal replication but, whenever the average number of replicates is less than 2 and/or more than half of the units are unreplicated, problems can be anticipated. To illustrate the arbitrariness of such a rule, the Reunion Island dataset, analysed extensively in Chapter 22, had on average only 1.9 lactations per cow but no numerical problems were encountered. If some levels need to be omitted in the hierarchy, it is useful to keep those at which principal predictors reside and those showing a lot of variation in a null model or based on descriptive statistics.

Example 23.4 Robust variance estimation for somatic cell count data
data=scc_40

We show the set of estimates and SEs for the model with four predictors used in Chapter 21, both for simple linear regression (ignoring the 3-level hierarchy) and with robust variance estimates clustered on cows. For convenience, the values from Chapter 21 (mixed model, repeated measures) are restated as well.

Model	Linear regression			Mixed; arma(1,1)	
Variable	β	SE	Robust SE	β	SE
h_size (*100)	0.898	0.057	0.127	0.627	0.306
c_heifer	-0.743	0.021	0.044	-0.777	0.040
t_season=spring	0.105	0.028	0.030	0.034	0.022
t_season=summer	0.121	0.029	0.035	0.039	0.027
t_season=fall	0.016	0.030	0.031	-0.007	0.023
t_dim (*100)	0.314	0.013	0.015	0.328	0.014
constant	4.318	0.041	0.079	4.516	0.154

Considerable increases in robust SEs are seen, mostly for the cow and herd level variables. (When clustering on herds, the SEs for -h_size- and the constant further increase to 0.365 and 0.195, respectively.) The parameter estimates for -t_dim-, -c_heifer- and even -h_size- are in reasonable agreement between linear regression and the mixed model. The conspicuous difference in -t_season- estimates of the linear and mixed models is essentially due to cow effects – the sampling periods of the cows in the dataset are not equally distributed over the year.

For discrete data, in particular binary data, it is not uncommon to encounter problems with high correlation and strong underdispersion at the lowest level. As discussed in section 22.5.5, the proper statistical procedure in such cases is not clear; we discuss this by way of Example 23.5.

23.3 GENERALISED ESTIMATING EQUATIONS

Generalised estimating equations (GEE) were introduced in two papers by Liang and Zeger in 1986, as a set of estimating equations to obtain parameter estimates for discrete and continuous repeated measures data. The idea has proven not only durable but also extendable to include other data structures (*eg* spatial data), statistical inference accompanying the estimates, as well as many variants of estimating equations. A recent (statistical) monograph (Hardin and Hilbe, 2003) is devoted entirely to GEE methods, which today are one of the most popular approaches in the health and biological sciences.

Example 23.5 Aggregation of the lowest level for pig-seroconversion data
data=ap2

Vigre et al (2003) observed that seroconversion to *Actinobacillus pleuropneumoniae* was strongly clustered in batches of pigs in multisite production systems. On average, each of the 36 batches of pigs consisted of about 30 pigs, and in 17 batches more than 90% of the pigs seroconverted, in four batches between 50% and 85% of the pigs seroconverted, and in the remaining 15 batches no pigs seroconverted. A 3-level logistic regression model (with predictors at all levels) had an dispersion parameter of ϕ=0.2, indicative of serious underdispersion and which could be interpreted as a very poor fit of the binary model. Also, the 3-level analysis showed some numerical instability when fitted using quasi-likelihood estimation. It was therefore decided to aggregate the data to the batch level by defining a batch as positive if at least one pig seroconverted, and as negative if otherwise. The mean pig's age at slaughter was computed for each batch as well, and the data were analysed by a 2-level model using this batch-level predictor. From a biological perspective, it was considered perfectly acceptable to designate batches as seroconverted or not, given the strong clustering in batches, so the 2-level model was preferred to a 3-level model with its obvious estimation problems and difficult interpretations of the variance components.

The scope of this textbook means we cannot do justice to this concept; however, we will describe the original (and probably still most popular) GEE method to obtain population-averaged estimates for clustered data. This method is based on correlations, in a working correlation matrix. We mention that recently, an alternative variant for binary data based on alternating logistic regression (Carey et al, 1993), has received renewed attention and been favoured for binary data (Hardin and Hilbe, 2003; we draw repeatedly from this reference throughout the section without specific mention).

23.3.1 Population-averaged versus subject-specific modelling

The distinction between **population-averaged** (PA) and **subject-specific** (SS) modelling for clustered data was introduced in section 22.2.1 where generalised linear mixed models (GLMMs) were referred to as subject-specific. Here we give more details and examples (largely following Diggle et al, 2002), in particular for the PA approach. Our first remark, however, is that the PA and SS interpretations of regression coefficients are equivalent for linear models. This is not due to the usual normal distribution assumption but to the fact that the linear predictor is modelled on the original scale; in the terminology of GLMs, the link function is the identity function and there is no shift of scale. Therefore, the proper reference for our discussion is a GLM, and we also assume a 2-level structure. In the original context of repeated measures, the data consisted of several observations (*eg* over time) on different subjects. In the context of our usual hierarchical clustering, we might instead have our subjects clustered in groups (*eg* animals in herds). To avoid any confusion of this double use of 'subjects', we shall refer to the upper level of the structure simply as clusters or groups.

The difference between the PA and SS approaches is in the way the clustering or grouping of the data is dealt with. As previously seen, subject-specific (or cluster-

specific) models include a random effect for each cluster in the linear predictor of the model. The assumptions for the random effects (*ie* their distribution and correlation) imply a particular form of the distribution of the set of observations within a cluster, including their correlation structure. Population-averaged or marginal models involve only the **marginal** means, *ie* the expected values for a particular set of predictors **averaged** across the **population of clusters**, and do not include specific effects for each cluster. PA models might therefore dispense with the assumptions for the distribution of the set of values within a cluster. To show the difference between the two types of model in simple formulae, denote by Y our observations and by u the random effects for the clusters (in an SS model). Then GLMs of SS and PA types are based on the equations:

$$\begin{aligned} &\text{subject-specific:} && \text{link}\,[\mathrm{E}(Y|u)] &&= X\beta^{\mathrm{SS}} + u \\ &\text{population-averaged:} && \text{link}\,[\mathrm{E}(Y)] &&= X\beta^{\mathrm{PA}} \end{aligned} \qquad \textit{Eq 23.2}$$

where, as before, $X\beta$ is our shorthand for the fixed part of the model, and $\mathrm{E}(Y|u)$ is the mean of Y conditional on the value of u (as discussed in section 22.2.1). As indicated by the notation in Eq 23.2, the SS and PA regression parameters are not identical (unless there is no clustering or link). Generally the PA parameters are closer to the null ('attenuated') than their SS counterparts. The difference depends on the amount of clustering and is often small relative to estimation error. For a logistic regression model, we have the approximation:

$$\beta^{\mathrm{PA}} \approx \beta^{\mathrm{SS}} \Big/ \sqrt{1 + 0.346\sigma_h^2} \qquad \textit{Eq 23.3}$$

where σ_h^2 is the (herd) cluster variance. For example, for PA parameters to be more than 10% lower than SS parameters, we need $\sigma_h^2 \geq 0.68$.

The selection of the most appropriate model type (SS or PA) depends on the predictor(s) being examined. Consider, as an example, a clinical trial of the effect of a treatment (compared to a placebo) on the risk of a cow developing milk fever. The study is carried out in multiple herds and the breed of the cow (Holstein versus Jersey) is also recorded. The final model includes terms for the two dichotomous variables: treatment and breed. The β^{SS} for treatment in a subject-specific model estimates the effect of the treatment in a **particular herd** on the risk of milk fever (compared with the risk in the same cow if she was not treated). This makes biological sense for cows staying in the same herd. On the other hand, the β^{PA} gives the effect (assumed decrease in prevalence of milk fever) of introducing a programme for treatment against milk fever **across all herds**. Thus, interest has been shifted from the individual herd to effects across herds. The parameters for breed are interpreted similarly, but the SS interpretation for breed would seem of less interest for herds with cows of a single breed (it refers to the altered risk of milk fever if all cows were replaced by cows of another breed), and the PA estimate seems more appropriate by comparing breeds across herds. Note also that an SS interpretation would become almost meaningless for herd-level predictors that are inherent in the herd (*eg* related to its location). This problem with an SS interpretation of a predictor that is unchangeable for the subject is more common in repeated measures data where subjects are individuals, for example, with predictors such as sex or race/breed.

Our main example of a statistical method based on PA estimation is GEE (next section); another example is the beta-binomial model briefly discussed in section 22.4.3.

23.3.2 Generalised estimating equations for population-averaged model

Let us initially explain the meaning of 'estimating equation'. When using maximum likelihood estimation, the parameters are chosen to maximise the log-likelihood function. In practice, maximising a function involves computing the (partial) derivatives of the function with respect to its parameters and equating these to zero. These would be the estimating equations for ML estimation (and the derivatives of the log-likelihood function is called the score function). Except for very simple cases, the equations do not have an explicit solution and must be solved iteratively. The approach we are going to take here involves GLMs and a partially specified model, so that no likelihood function is available. Nevertheless, estimation is based on iterative solution of similar **generalised estimating equations**.

The population-averaged model in Eq 23.2 can be estimated by using the GEE method, using only assumptions about the marginal mean and variance (and information about the grouping of the data). Even if no assumptions about the form of the correlation of the data within the groups are made, the estimating equations involve a **working correlation matrix** containing the estimated correlations among individuals within a cluster, in each cycle of the iterations. This matrix can be given different forms (independent, compound symmetry, autoregressive, unstructured *etc* as in section 21.5.1) to tailor the estimating algorithm toward one's perception of the data structure. Because the matrix is not part of the model, its form is not as crucial as in a fully parametric model. Theoretically, the GEE method gives asymptotically unbiased estimates even if the working correlation matrix is misspecified (it might, however, lead to loss in efficiency). Several options are available for variance estimation but generally it is recommended that the robust (or empirical) variance estimate (section 23.2.3), also asymptotically unbiased, be used.

As to the choice of working correlation structure, you should first and foremost be guided by your understanding of the data. For hierarchically clustered data (*eg* pigs in farms), anything but a compound symmetry (or exchangeable) correlation structure would seem unreasonable. Particular caution should be exercised with negatively correlated binary data. In this case, an ordinary logistic model with robust standard errors, section 23.2.3) has been recommended (Hanley et al, 2000). For repeated measures data, one would usually choose a structure that allows for decreasing correlations with distance in time. It might also be tempting to try an unstructured correlation to see what patterns the data show when not constrained by a particular structure. However, large correlation structures imply estimation of a large number of 'working parameters' and numerical problems might be encountered especially in unbalanced datasets. Recently a criterion (QIC) similar to Akaike's information criteria has been developed to guide the choice of correlation matrix (Pan, 2001). We illustrate the GEE method in Example 23.6 (discrete data) and Example 23.7 (continuous data).

Example 23.6 Generalised estimating equations for pig-pneumonia data
data=pig_adg

A GEE analysis with a compound symmetry structure for the working correlation matrix and robust standard errors gave a regression coefficient of 0.354 (0.216). For comparison with the previous random effects estimate (0.437), we might compute its PA counterpart using Eq 23.3: $\beta^{PA} \approx 0.437/\sqrt{1+0.346*0.879} = 0.383$. Thus, the difference between the two estimates is not entirely due to their different interpretations; however, relative to the SEs, the difference is small. The working correlation matrix had a correlation of 0.18 (between pigs in the same farm), which is quite similar to ρ=0.21 computed in Example 22.2.

The advantage of the GEE method (and many of its generalisations) is that it has robust theoretical properties with few model assumptions. It is also computationally feasible for large datasets and can be fit with a wide range of working correlation structures. It is one of the few general methods for use with discrete repeated measures and spatial data.

One drawback of the GEE method is its basic limitation to a single level of clustering (however, the alternating logistic regression approach allows for two levels). In situations in which you have multiple levels of clustering (*eg* multiple observations within cows, within herds), it might be possible to obtain reasonable estimates for predictors below the herd level by allowing for clustering at the cow level and assuming that the herd effects will be incorporated into the within-cow correlations. A second drawback, because the model is not fully specified, is that a standard likelihood-based inference is not available; however, alternative methods for model selection and model-checking have been developed in recent years. Finally, we should remember that PA models provide estimates of coefficients with slightly different interpretations than mixed (SS) models.

23.4 Bayesian analysis

Little known outside statistical science, there exist two different approaches to statistical inference, which have different concepts and philosophical bases and will, in general, lead to different results. The rivalry between the two schools has persisted over decades, with neither of them emerging as the clear winner. Many statisticians cling to the middle ground believing that each of the two approaches has its weaknesses and strengths which make each of them attractive in particular situations. However, most (introductory) statistics courses are taught within the non-Bayesian (classical, likelihood-based, frequentist) framework with no reference to the Bayesian view.

Bayesian analysis has gained in popularity in recent years, and has for example been applied to complex problems in veterinary epidemiology such as risk assessment or comparison of diagnostic tests without a gold standard (*eg* Hanson et al, 2003) and to the analysis of multilevel data (Dohoo et al, 2001). The scope of practical Bayesian

Example 23.7 Generalised estimating equations for somatic cell count data
data=scc_40

We analysed these data using linear mixed models for repeated measures in Example 21.10. Because of the lack of link function, the SS and PA parameters coincide. The difference of the GEE approach lies therefore, entirely in the estimation method. The table shows parameter estimates from GEE analyses clustered at the cow level with compound symmetry, autoregressive (ar(1)) and unstructured working correlation matrices. The table also gives values of the working correlations one and two time steps apart; the values for the unstructured correlation were obtained by averaging the corresponding values in the matrix. Some software implementations of GEE (*eg* in SAS) will fit stationary (Toeplitz) structures without excluding incomplete sets of repeated measures; the results were close to those shown for the unstructured correlations.

	Working correlation matrix structure					
	Compound symmetry		Autoregressive		Unstructured	
Variable	β	SE	β	SE	β	SE
h_size (*100)	0.826	0.123	0.799	0.124	0.755	0.121
c_heifer	-0.777	0.042	-0.755	0.042	-0.772	0.041
t_season=spring	0.015	0.023	0.054	0.024	0.031	0.022
t_season=summer	0.026	0.026	0.060	0.026	0.033	0.024
t_season=fall	-0.022	0.025	0.003	0.025	-0.010	0.023
t_dim (*100)	0.336	0.014	0.315	0.014	0.327	0.013
constant	4.415	0.074	4.424	0.074	4.454	0.072
ρ (1 month)	0.555		0.671		0.647	
ρ (2 months)	0.555		0.451		0.592	

These values should be compared to those of Example 23.4. The parameter estimates for -c_heifer- and -t_dim- are in reasonable agreement between all models, including the uncorrected analysis. This may be said also for -h_size- when considering its large SE (from the linear mixed model). For -t_season-, the GEE estimates adds further to the disagreement already seen in Example 23.4. The best agreement with the mixed model is achieved by the unstructured correlations, but in particular the estimates obtained by the autoregressive correlation structure differ markedly. These data demonstrate that choice of working correlation structure is not always of minor importance for the fixed effects, even in a large dataset. The correlations show good agreement with the mixed model estimates, and still indicate both the compound symmetry and autoregressive structures to be inadequate. The comparison of SEs between the GEE, mixed model and uncorrected analysis follow the hierarchical levels. The herd level SEs are inflated relative to the uncorrected analysis but not to the level of the mixed model; this is no surprise because GEE does not take into account herd-level clustering. The cow-level variable (-c_heifer-) has similar SEs of GEE and mixed model analysis, and both larger than the uncorrected analysis, and the test-level SEs are similar in all analyses.

In summary, there is good agreement between the GEE and linear mixed models analysis, but the former is limited by its lack of likelihood-based inference (*eg* for choosing a correlation structure) and standard errors for correlation parameters.

inference has been increased widely by the invention and recent advances of a simulation-based tool for statistical inference: **Markov chain Monte Carlo (MCMC)** estimation. Mixed models analysis by the Bayesian approach is entirely based on MCMC methods.

We hope the reader will bear with us for the inevitable inadequacy of a few pages' introduction to a full, new statistical approach. Our aim can only be to give little more than a superficial impression of the ideas and steps involved in a Bayesian analysis of clustered data. Recent textbooks on applied Bayesian analysis in the health and biological sciences (*eg* Congdon, 2001, 2003) would be the proper starting point. Most Bayesian analyses require specialised software, and the standard choice is the (free) BUGS programme developed by the Medical Biostatistics Unit in Cambridge (http://www.mrc-bsu.cam.ac.uk/bugs/). BUGS is short for Bayesian analysis using Gibbs sampling, which is a particular type of MCMC analysis. The analyses of this section were carried out using the MLwiN software (version 1.2).

23.4.1 Bayesian paradigm

Bayesian methodology owes its name to the fundamental role that **Bayes' theorem** (see Eq 23.4) plays in it. In Bayesian reasoning, uncertainty is attributed not only to data but also to the parameters. Therefore, all parameters are modelled by distributions. Before any data are obtained, the knowledge about the parameters of a problem are expressed in the **prior distribution** of the parameters. Given actual data, the prior distribution and the data are combined into the **posterior distribution** of the parameters. The posterior distribution summarises our knowledge about the parameters after observing the data. The major differences between classical and Bayesian inference are outlined in Table 23.1, and will be detailed in the sections that follow.

Table 23.1 Bayesian versus classical approaches to statistics

Concept	Classical approach	Bayesian approach
parameter	constant	distribution
prior information on parameters	none	prior distribution
base of inference	likelihood function	posterior distribution
parameter value	(ML) estimate	statistic of posterior distribution, eg median, mode, mean
parameter range	confidence interval	probability range of posterior distribution
hypothesis statement	test	Bayesian factors

Let us briefly indicate the way the prior and the data are merged, and denote by Y the data, by θ the parameter (vector), and

- $L(Y|\theta)$ – the likelihood function,
- $f(\theta)$ – the prior distribution for θ,
- $f(\theta|Y)$ – the posterior distribution for θ

where the $f(\cdot)$s are either probability functions (discrete data) or probability densities (continuous data). With these definitions, Bayes' theorem says:

$$f(\theta|Y) = const(Y)\, L(Y|\theta)\, f(\theta) \qquad \textit{Eq 23.4}$$

where $const(Y)$ is a constant depending on Y but not on θ. Thus, the posterior distribution for θ is essentially constructed by multiplying the likelihood and the prior, and is a sort of compromise between the two. In complex models, the constant depending on Y in Eq 23.4 is virtually impossible to calculate; therefore, simulation methods such as MCMC have had a great impact on Bayesian analysis.

23.4.2 Statistical analysis using the posterior distribution

Even if it might seem awkward to discuss the posterior before the prior distribution, let us see a simple example of Bayesian analysis (Example 23.8) before turning to the discussion of how to choose the prior distribution. The net result of a Bayesian analysis is a **distribution**, and the analysis might, therefore, be conveniently summarised by a graph (Fig. 23.1). Point estimates and confidence intervals are not truly Bayesian in spirit, but values such as the mean, median or mode, and intervals comprising a certain probability mass of the posterior (sometimes called **credibility intervals**) might be calculated from the posterior distribution. The two most commonly used point values are median and mode, the latter also called a maximum *a posteriori* (MAP) estimate.

23.4.3 Choice of prior distributions

Generally, it can be said that the strength and weakness of Bayesian methods lie in the prior distributions. In highly multidimensional and complex problems, it is possible to incorporate model structure by means of prior distributions; such an approach has been fruitful, for example, in image analysis. The posterior of one analysis can also be taken as the prior for a subsequent study, thereby enabling successive updates of the collected and available information (empirical Bayes method). On the other hand, the choice of prior distributions might seem open to a certain arbitrariness, even if subjectivity in the prior does not contradict the Bayesian paradigm. In the past, priors have often been chosen in a particular form allowing for explicit calculation of the posterior (**conjugate priors**) but, with access to MCMC methods, these have decreased in importance.

A common choice of prior (in particular among less-devoted Bayesian researchers) is a **non-informative** (flat or diffuse) prior, which gives minimal preference to any particular values for θ. As an extreme case, if we take $p(\theta) \equiv 1$ in Eq 23.4, the posterior distribution is just the likelihood function. So, for example, maximising the posterior (MAP estimate) yields exactly the maximum likelihood estimate. Therefore, we would by and large expect Bayesian inference with non-informative priors to be similar to likelihood-based inference. To take $p(\theta)$ constant is not always possible, but an alternative is a normal distribution with mean zero and a very large variance, effectively making values in a large interval around zero equally probable. For a variance parameter, where values below zero are impossible, a standard non-informative distribution is a gamma distribution for the inverse of the variance.

Example 23.8 Bayesian analysis of proportions

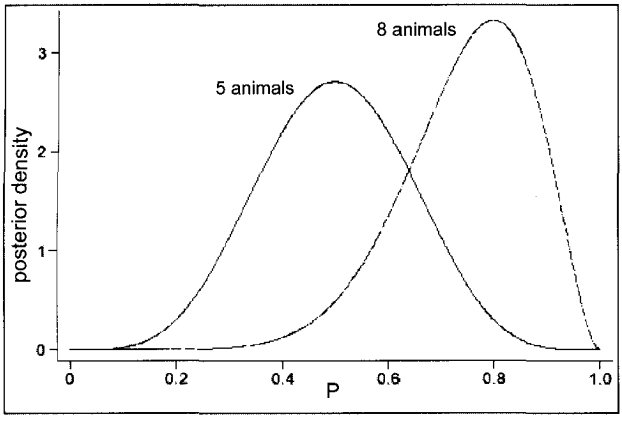
Assume that we test 10 animals for a disease with highly variable prevalence. In one scenario five of the animals tested positive; in another, eight animals tested positive. What information have we obtained about the disease prevalence in these two scenarios?

Recall that all Bayesian analyses involve *a priori* distribution, in this case for the disease prevalence P. Assume (somewhat unrealistically) that we had no particular prior information (due to the high variability of the disease) so that *a priori* all values of p would seem equally likely. Then we could choose a uniform distribution on (0,1) as our prior; this is an example of an non-informative prior (section 23.4.3). The probability density of the uniform distribution is constant (1). The likelihood function for observing the number of positive animals out of 10 are the probabilities of the binomial (10, p). Therefore, if we observe Y positive animals, the posterior distribution has density:

$$f(p|Y) = const(Y) * p^Y (1-p)^{10-Y} * 1 = const(Y)\, p^Y (1-p)^{10-Y}$$

This probability density corresponds to a beta-distribution with parameters (Y+1,10-Y+1). Fig. 23.1 shows the beta-distributions with parameters (6,6) and (9,3) corresponding to observed values of Y=5 and Y=8, respectively.

Fig. 23.1 Posterior distributions after 5 and 8 out of 10 animals tested positive



If we wanted to summarise our knowledge about p after the testing into a single value, we could use the mean, median or mode of the distribution; for the two beta-distributions, they equal (0.5, 0.5, 0.5) and (0.75, 0.764, 0.8), respectively. These values can be compared with the usual estimates P=0.5 and P=0.8; the agreement of the mode and maximum likelihood estimate is no coincidence! If we wanted to summarise our knowledge about P into a 95% interval, we could choose the interval with endpoints equal to the 2.5% and 97.5% percentiles of the distribution; for the two beta-distributions they are (0.234,0.736) and (0.482,0.940). These intervals might be compared with the (exact) binomial confidence intervals of (0.187,0.813) and (0.444,0.975). The confidence intervals are wider than the credibility intervals.

23.4.4 Markov chain Monte Carlo (MCMC) estimation

Note This section uses a notation somewhat inconsistent with the rest of the book in order to stay reasonably in line with the usual notation in the field.

Markov chains

A Markov chain is a process (or sequence) $(X_0,X_1,X_2,\ldots)$ of random variables which satisfies the Markov property (below). The variables take values in a state space which can be either finite (*eg* {0,1}), discrete (*eg* {0,1,2,3…}) or continuous (*eg* an interval, possibly infinite). The value of X_0 is the initial state of the chain, and the steps of the chain often correspond to evolution over time. The **Markov property** is a strong assumption about the probability distribution of the process (X_t):

distribution of $(X_t,X_{t+1},X_{t+2},\ldots)$ given $(X_0,X_1,\ldots X_t)$ =
distribution of $(X_t,X_{t+1},X_{t+2},\ldots)$ given only X_t ***Eq 23.5***

In words, the future (of the process) depends on the past only through its present state. Thus, the chain has a 'short memory'. Some examples of Markov chains are processes describing games, population sizes and queues. Examples of non-Markov processes are periodic phenomena and growth curves. Our interest here is in **homogeneous** chains in which development does not change over time. For such chains the Markov condition (Eq 23.5) implies that whenever the chain has reached state x, it evolves from there as if it was restarted with $x_0=x$. The importance of homogeneous chains is that under some further, technical conditions they converge to limiting distributions as time runs. That is, distr$(x_t) \rightarrow \pi$ as time runs, where π is the limiting (or **stationary**) distribution (and in this case not the number 3.1415926…). This implies for example that $p(x_t=x) \rightarrow \pi(x)$.

The simplest example of a homogeneous Markov chain has state space {0,1}. The states 0 and 1 could, for example, correspond to disease states (healthy/sick) or system states (busy/idle). The transitions from one state to the next are governed by a transition matrix

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

where $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$

For example, from state 0 the process continues to state 1 with probability p_{01} (and stays in state 0 with probability p_{00}). This chain has a stationary distribution whenever all probabilities are non-zero, and $\pi(1)= p_{01}/(p_{01}+p_{10})$.

Markov chain Monte Carlo estimation

The idea of Markov chain Monte Carlo estimation is simple, yet surprising. Suppose we were interested in a particular distribution π, but that quantities from this distribution were difficult to calculate because its analytical form is unknown (the distribution we have in mind is a posterior distribution from a complicated model). Suppose furthermore, that we were able to devise a Markov chain (X_t) such that distr $(X_t) \rightarrow \pi$. Then, in order to calculate statistics from π , we could run our Markov chain for a long

time, for example, up to time step T (where T is large), to make the distribution of all X_t for $t \geq T$ a good approximation to π. Then in order to calculate, for example, the mean of the distribution π we could simply average over a sample of observations from the chain after time T. In a formula this would appear as:

$$\mathrm{E}(\pi) \approx \frac{1}{n} \sum_{t=T+1}^{s=T+n} X_t \qquad \textit{Eq 23.6}$$

Note that our sample from (X_t) is nothing like an independent sample (it is n successive values from a Markov chain which will be correlated). Despite the correlation, we can still use the formula to estimate $\mathrm{E}(\pi)$; however, our precision will be less than if we had an independent sample, and very much so if there is strong correlation in the chain. Other statistics than the mean might be computed from the limiting distribution as well. The initial part of the chain, $X_0,\ldots,X_T$ is called the **burn-in** period.

Apparently the flaw of this idea is the necessity to construct a Markov chain with π as the limiting distribution, when we haven't even got an analytical form for π! But that turns out to be possible for many multidimensional statistical models where π is known only up to a proportionality constant (such as *const*(Y) in Eq 23.4). To construct a Markov chain one needs to specify its transition mechanism (in the example above, the transition matrix P), whereas the starting value is of minor importance. There are two major, general techniques for doing this: **Gibbs sampling** and **Metropolis-Hastings sampling** (technically, Gibbs sampling is a special case of Metropolis-Hastings sampling but usually is considered to be a separate method). One major practical complication involved in MCMC estimation is the length of the burn-in period, in order to make estimation from Eq 23.6 valid. Constructed Markov chains might converge rapidly or very slowly to their limiting distribution, sometimes so slowly that the chain is useless for estimation purposes. Therefore it is crucial to have tools for monitoring the convergence and the required length of burn-in periods. The MCMC software will provide some diagnostics tools for monitoring. We will not go into details with these or with the construction of the Markov chains, only mention that Gibbs sampling is feasible for linear mixed models with conjugate priors, whereas Metropolis-Hastings sampling can be applied generally but might result in highly correlated and very slowly converging chains.

We illustrate the MCMC techniques on the two recurring datasets of (Examples 23.9 and 23.10). All prior distributions were taken as non-informative using the default values of the software.

The two examples demonstrated that good agreement between likelihood-based and Bayesian estimation with non-informative priors can be achieved (without asserting this to always be the case). One additional advantage of the Bayesian approach is that the models can quite easily be extended to include, for example, non-normal random effects (provided that Markov chains can be constructed with good convergence).

This section included only examples with non-informative priors. As previously mentioned, the real strength of the Bayesian approach lies in its ability to combine informative priors and data; however, such models are beyond our present scope.

Example 23.9 Bayesian MCMC analysis of pig-pneumonia data
data=pig_adg

A Metropolis-Hastings chain with a burn-in of 10,000 samples and the estimation based on 100,000 subsequent samples gave the following results of the posterior distribution for the regression coefficient of -ar_g1-:
median = 0.436, mode = 0.434, standard deviation = 0.260, 95% interval = (-0.070,0.948)

The median (or mode) and standard deviation are almost identical to the estimate and SE of the previous (likelihood-based) GLMM analysis. The MCMC analysis also gave a value (median) of the farm variance σ_h^2 of 0.990 (0.658). The estimate is somewhat higher than the ML estimate, and the standard deviation is about 1.5 times the SE.

Example 23.10 Bayesian MCMC analysis of somatic cell count data
data=scc40_2level

Two MCMC analyses were carried out using the 2-level somatic cell count dataset (the full dataset was not used to avoid the complications of repeated measures correlation structures). One analysis used Gibbs sampling (the recommended method for linear mixed models), the other used Metropolis-Hastings sampling (for fixed parameters). In theory, both procedures are valid provided convergence of the chains. In the table below, we restate for convenience also the linear mixed model estimates from Example 21.2.

Method	Mixed model		Bayesian and MCMC			
option	REML estimation		Gibbs sampling		Metropolis-Hastings	
Variable	β	SE	β*	SE#	β*	SE#
h_size	0.408	0.377	0.405	0.387	0.384	0.394
c_heifer	-0.737	0.055	-0.737	0.056	-0.738	0.055
t_season=spring	0.161	0.091	0.161	0.091	0.159	0.090
t_season=summer	0.002	0.086	0.001	0.086	0.000	0.086
t_season=fall	0.001	0.092	0.002	0.092	0.000	0.091
t_dim	0.277	0.050	0.278	0.050	0.277	0.050
constant	4.641	0.197	4.641	0.202	4.647	0.206
herd variance	0.149	0.044	0.150	0.048	0.151	0.049
error variance	1.557	0.048	1.558	0.048	1.558	0.048

*median of posterior distribution
#standard deviation of posterior distribution

The Gibbs sampled chain converged more rapidly and showed less correlation, so only 20,000 samples were used for estimation after a burn-in of 10,000 samples. The Metropolis-Hastings chain showed high correlation for some of the fixed parameters and therefore, estimation was extended to 100,000 samples. Overall, the agreement between the three sets of estimates is very good. The only noteworthy disagreements are in the herd-level parameters. The Metropolis-Hastings estimate for -h_size- is somewhat off the other two estimates, but the chain for this parameter was extremely highly correlated and thus, the posterior distribution not estimated well. Also the posterior distributions for -h_size- and the constant show slightly higher standard deviations than the SEs from REML estimation.

23.5 SUMMARY OF CLUSTERED DATA ANALYSIS

A variety of approaches for dealing with clustered data has been presented in this and previous chapters. We conclude with a comparative table of estimates for the pig-pneumonia data (Example 23.11), and a summary table for the methods covered. Table 23.2 gives only a very brief summary; consult the respective sections for details.

Example 23.11 Summary of analyses for pig-pneumonia data
data=pig_adg

Variable	Model	β	SE
ar_g1	logistic	0.647	0.220
	overdispersion	0.647	0.368
	robust variance	0.647	0.276
	fixed effects	0.365	0.268
	stratification	0.346	0.261
	GLMM	0.437	0.258
	GEE	0.354	0.216
	Bayesian	0.438	0.260

The estimates and SE show reasonable agreement between the stratified, fixed and random effects models as well as GEE and Bayesian estimation. These five approaches would seem acceptable choices for analysis.

Table 23.2 Summary of approaches for clustered data

	Properties/Features				
Method to account for clustering	adjusted SE	adjusted β	>1 level of clustering	estimate of ρs	Comments on scope or use of method
linear mixed model	yes	yes	yes	yes	
GLMM	yes	yes	yes	yes	subject-specific model
fixed effects	yes	yes	no	no	restrictions in predictors
stratification	yes	yes	no	no	specific designs
overdispersion factor	yes	no	no	no	restricted range of GLMs
robust SE	yes	no	no	no	mainly continuous data
GEE	yes	yes	(no)	(yes)	PA (marginal model)
Bayesian mixed models	yes	yes	yes	yes	different statistical approach

Note The GEE method yields correlations as part of the working correlation matrix, and the alternating logistic regression version of GEE for binary data allows for two levels of clustering.

SELECTED REFERENCES/SUGGESTED READING

1. Browne WJ. MCMC Estimation in MLwiN. Notes 2002, available from http://multivel.ioe.ac.uk/dev/devleop.html
2. Carey V, Zeger SL, Diggle PJ. Modelling multivariate binary data with alternating logistic regressions. Biometrika 1993; 80: 517-526.
3. Collett D. Modelling binary data, 2d ed. Boca Raton: Chapman &Hall/CRC, 2002.
4. Congdon P. Bayesian statistical modelling. Wiley, 2001.
5. Congdon P. Applied Bayesian modelling. Wiley, 2003.
6. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of longitudinal data, 2d ed. Oxford: Oxford University Press, 2002.
7. Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in cattle. Prev Vet Med 2001; 50: 127-144.
8. Donner A. The comparison of proportions in the presence of litter effects. Prev Vet Med 1993; 18: 17-26.
9. Guan W. From the help desk: Bootstrapping standard errors. The Stata Journal 2003; 3: 71-80.
10. Hanley JA, Negassa A, Edwardes MD. GEE: Analysis of negatively correlated binary responses: a caution. Stat Med 2000; 19: 715-722.
11. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 2003; 157: 364-375.
12. Hanson TE, Johnson WO, Gardner IA. Hierarchical models for the estimation of disease prevalence and the sensitivity and specificity of dependent tests in the absence of a gold-standard. J Agr Biol Env Sci 2003 (in press).
13. Hardin J, Hilbe J. Generalized linear models and extensions. College Station: Stata Press, 2001
14. Hardin JW, Hilbe JM. Generalized estimating equations. Boca Raton: Chapman & Hall/CRC, 2003.
15. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001; 57: 120-125.
16. Vigre H, Dohoo IR, Stryhn H, Busch ME. Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumoniae* and *Mycoplasma hyopneumoniae* at different levels in Danish multisite pig production facilities. Accepted Prev Vet Med 2003.

SAMPLE PROBLEMS

1. Compare simple and more complex methods for taking a 2-level structure into account for binary data on first service conceptions on Reunion Island.
   a. In order to reduce the data to a 2-level structure, use only the first lactation of each cow in the reu_cfs data.
   b. Fit logistic regression models with -heifer- and -ai- as fixed effects, whereby:
      i. ignoring herd
      ii. including herd as a fixed effect
      iii. using robust variance estimates
      iv. using GEE estimation with compound symmetry working correlation matrix
      v. using herd random effects and maximum likelihood estimation
      vi. using herd random effects and quasi-likelihood estimation.
   c. Compare the estimates of fixed effects obtained by the different procedures as well as the values related to the clustering.
      i. Which estimates have subject-specific and population-averaged interpretations, respectively?
      ii. Compare also the estimates to those obtained in the 3-level analyses in Chapter 22.
   d. Summarise your experience with these data by classifying each of the above approaches i.-vi. as either unacceptable, acceptable for descriptive purposes, or acceptable for statistical inference.
2. Explore GEE estimation for continuous repeated measures data, using the milk yields in the somatic cell count data.
   a. Set up the 'standard' GEE procedure (based on the working correlation matrix) for-ecm-withcowsassubjectsandsuitablychosen(seeSampleProblem21.4)fixed effects of -t_lnscc-, -t_dim-, -t_season- and -c_heifer-.
      i. Which are the proper distribution and link function for the model when viewed within the generalised linear model framework?
      ii. Does GEE estimation for this model give estimates with subject-specific or marginal interpretation? (caution: this is a trick question!)
   b. Run the GEE procedure with different working correlation structures, as specified below, and for each of these record the fixed parameters and the 'working parameters' in the correlation matrix:
      i. compound symmetry structure
      ii. autoregressive (first order) structure
      iii. stationary or Toeplitz structure
      iv. unstructured correlations
      v. any other structure of interest to these data that your software offers.
   c. Compare the values obtained by the different GEE estimations, and compare them also with the values obtained in Sample Problem 21.4.
      i. Do the regression coefficients agree reasonably well?
      ii. Do the standard errors of regression coefficients agree reasonably well?
      iii. Do the correlation parameters agree reasonably well, and do both approaches (linear mixed model and GEE estimation) suggest a similar correlation structure for the repeated measures within a cow?
      iv. How would you expect the GEE estimation with cows as subjects to

perform for a herd level predictor (*eg* -h_size-)? Confirm your conjecture by running a model that includes -h_size-.

3. Using the previously analysed data and model (Sample Problem 22.2) for preweaning mortality, explore Bayesian estimation by MCMC procedures.
   a. Set up of a Bayesian hierarchical logistic regression model.
      i. Include fixed effects of -sow_tx-, -sow_parity- and -quarter- (three dummy variables to represent the four categories), with non-informative prior distributions for the regression coefficients (*eg* the default values of your software).
      ii. Include random effects of sows and herds, with non-informative prior distributions for the variances (*eg* the default gamma distributions for the inverse variances).
      iii. If you are familiar with the model diagrams used in the WinBugs software, draw a diagram of the model.
   b. Fit the model using the default MCMC method (typically Metropolis-Hastings sampling) and reasonably low values for the burn-in period (*eg* 1,000 samples) and the monitoring (estimation) chain length (*eg* 5,000 samples).
      i. Evaluate the trajectories for convergence and autocorrelation in each component of the chain. Do any of the parameters cause concern?
      ii. If needed, try to improve the estimation by increasing the burn-in period and/or the monitoring chain length.
      iii. Compare the posterior distribution obtained for each parameter with the previously obtained estimates and confidence intervals, using the median (or mode) of the posterior as a point value and a central 95% interval of the posterior distribution.
      iv. Explore the dependence of the estimates on the prior distributions by changing these to become more or less informative than the default settings.
   c. As an alternative to logistic regression, probit regression was briefly mentioned in section 22.4.1 and in this last ('advanced') part of the problem we give the opportunity of exploring this model in practice. Recall that probit regression differs from logistic regression solely by the link function. Most software for GLM(M)s allows you to use either the logit or probit links. Furthermore, some software for Bayesian analysis using MCMC (*eg* MLwiN) offers a potentially more efficient MCMC method (Gibbs sampling) for a probit regression model, using the model's connection to the normal distribution (Browne, 2002).
      i. Fit a usual (non-Bayesian) 3-level probit regression model with the same effects as before.
      ii. Compare the parameter estimates to those of the logistic regression model, using the rough rule that probit regression estimates (regression coefficients and random effect standard deviations) are scaled down by a factor of $\pi/\sqrt{3}=1.81$ relative to logistic regression.
      iii. Set up the Bayesian 3-level probit regression model with non-informative priors for all parameters, and run the MCMC estimation.
      iv. Evaluate as previously the convergence of the chain, and compare with the chain for the logistic regression model.
      v. Compare the posterior distributions with the non-Bayesian estimates.

24

META-ANALYSIS

Objectives

After reading this chapter, you should be able to:

1. Choose among various options (including meta-analysis) for summarising data from a collection of published studies.
2. Carry out a literature review and data-extraction process to provide data suitable for a meta-analysis.
3. Calculate summary estimates of effect, evaluate the level of heterogeneity among study results and choose between using fixed- and random-effects models in your analysis.
4. Present the results of your meta-analyses graphically.
5. Evaluate potential causes of heterogeneity in effect estimates across studies.
6. Evaluate the potential impact of publication bias on your study results.
7. Determine if your results have been heavily influenced by an individual study.

24.1 INTRODUCTION

When making decisions about animal health interventions, we would like to use all of the information available in order to make the most informed decision. Unfortunately, the information in the literature is often inconclusive and could be conflicting. For example, the introduction of the use of recombinant bovine somatotropin (rBST) in the United States in 1994 initiated a substantial discussion of the potential effects of the use of the drug on the risk of clinical mastitis in dairy cows. If, in 1998, you carried out a review of all randomised clinical trials of rBST which reported risk ratios (or the data required to calculate a risk ratio) for clinical mastitis, you would have found 20 studies (representing 29 groups of cows) (Dohoo et al, 2003b). The point estimates of the risk ratio (*RR*) in those studies ranged from 0.67 (*ie* a reduction in risk) to 4.87 (a substantial increase in risk) (see Example 24.1). However, the effect was not statistically significant in 28 of the 29 groups studied. This might have led you to conclude that there was no effect of rBST on the risk of mastitis. Nonetheless, you might be left wondering if the variation in results was more than would be expected due to chance variation and what the power of each study to detect an effect was.

Similarly, if you carried out an evaluation of the effects of rBST on milk production (measured as 3.5% fat-corrected milk), you would have found data on 28 groups of cows in 19 different studies (Dohoo et al, 2003a). The point estimates ranged from a loss of 0.7 kg/day to a gain of 10.6 kg/day. Although there was a wide range of point estimates, the vast majority were over 3 kg/day and 23 of the 28 groups had statistically

Example 24.1 Individual point estimates of risk ratio for effect of rBST on clinical mastitis

data=bst_mast

Twenty studies, containing data from 29 separate groups of cows had sufficient data to be able to calculate the risk ratio of the effect of rBST on clinical mastitis. The individual point estimates from each of the 29 groups were:

study	group	RR	study	group	RR	study	group	RR
1	1	0.83	6	11	1.00	15	21	1.19
1	2	0.91	7	12	0.96	15	22	1.26
2	3	1.08	8	13	0.95	16	23	1.40
3	4	1.30	8	14	1.31	16	24	0.67
3	5	0.90	9	15	1.45	16	25	1.11
4	6	1.75	10	16	1.02	17	26	4.87
4	7	1.45	11	17	1.40	18	27	2.60
4	8	0.83	12	18	1.80	19	28	4.00
4	9	1.35	13	19	1.73	20	29	1.37
5	10	2.50	14	20	1.91			

significant increases in production. Consequently, while it was clear that there was an effect, you might be interested in what the average effect was and why it varied from study to study.

If you wanted to carry out a more formal review of the available data on the effect of rBST on mastitis risk, there are several approaches that you could take, and we discuss each of these in turn.

Study-by-study assessment

The first possible approach would be to consider each study individually and to subjectively take into account the unique circumstances of each study. However, you would soon find that each of the individual studies had very limited power to detect a moderate effect of rBST and the precision of each estimate of the RR was low. It is also likely that, with so much data available, you would like some form of summary estimate of the effect derived from all of the studies.

Narrative review

The second approach would be to carry out a traditional narrative review in which you qualitatively assess each of the studies individually and then subjectively combine the conclusions from each study into an overall conclusion. If there are a limited number of studies to be reviewed, this might be the best approach, although it has several limitations. First, it is subjective in nature and thus prone to reviewer bias. In deriving an overall estimate of effect, there is also a tendency to weight all studies equally, and as will be seen later, they should not all receive equal weight. Finally, this type of review might fail to detect meaningful effects which were not statistically significant in any individual study due to the lack of power of those studies.

Pooled analysis

A third approach would be to contact the authors of all of the individual studies and request the original data from each study. These could then be pooled into a single dataset and reanalysed, taking into account the clustered nature of the observations (within study and within herds in each study). This would provide an excellent overall estimate of effect but is very time consuming and expensive.

Meta-analysis

The fourth option would be to carry out a meta-analysis based on the results from each of the individual studies. A meta-analysis has been defined as: "The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976). It is a formal process for combining results from a number of studies that is being used increasingly in human medicine and, to a more limited extent, in veterinary medicine. Meta-analyses have been used most commonly to combine results from a series of controlled trials and this chapter will focus on that application. However, they can also be used to combine results from a series of observational studies as was done in a recently published meta-analysis of the effects of disease on reproductive performance in dairy cows (Fourichon et al, 2000). For a discussion of some of the specific issues related to this latter application, see Chapter 32 of Rothman and Greenland (1998) or Egger et al (2001). A more complete description of

meta-analyses can be found in journal articles (Dickersin and Berlin, 1992; Normand, 1999), textbooks (Egger et al, 2001; Sutton et al, 2000; Petitti, 1999) or at the website of the Cochrane Collaboration (Cochrane Collaboration, 2001). The last reference relates to an international organisation set up to help health-care professionals make informed decisions through the use of systematic reviews of health research.

24.2 Objectives of meta-analysis

The objectives of a meta-analysis are to provide an overall estimate of an association or effect based on data from a number of scientific studies and to explore reasons for variation in the observed effect across studies. It accomplishes this by imposing a systematic methodology on the review process. Because it combines data from multiple studies, there is a gain in statistical power for detecting effects. When computing an overall estimate of effect, it takes into account both the individual study estimates and the precision of those estimates (standard errors) so that the results from each study are weighted appropriately.

Meta-analyses can be used to review existing evidence prior to making clinical or animal-health policy decisions, or as a precursor to further research by better quantifying what is already known, and identifying gaps in the scientific literature. A meta-analysis might be combined with a traditional narrative review and hence, should be thought of as complementary to that review process.

24.3 Meta-analysis process

The steps involved in carrying out a meta-analysis are:

1. specify the question to be answered
2. define inclusion/exclusion criteria for studies to be included in the review
3. find all of the relevant studies
4. evaluate study quality and select relevant studies
5. extract the relevant data from each study
6. conduct the analysis
7. interpret the results.

24.3.1 Specifying the question

When specifying the question to be answered, you need to keep in mind what is most important from a clinical or animal-health policy objective, rather than letting data availability drive the study objective. It is often more desirable to address a more general question, which will broaden the eligibility characteristics for studies to be included in the review, rather than to address a very specific, but restrictive, question.

> ...far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
>
> Tukey (1962) quoted in Dickersin and Berlin (1992)

For example, a review of the ability of β blockers to reduce the short-term risk of myocardial infarction was based on studies in which 12 different drugs had been used (Freemantle et al, 2002) rather than focusing on a single specific drug. This enhanced the generalisability of the results.

24.3.2 Define the inclusion/exclusion criteria

The first step in determining what studies are to be included in the meta-analysis is the decision about what outcome(s) is to be evaluated. Obviously, only studies reporting the outcome of interest can be included. In general, issues related to the internal and external validity of each study will also need to be considered. Some specific issues which you might also want to consider when deciding what studies should be included in the meta-analysis include (but are not limited to):

- study design issues
  - should only randomised controlled trials be considered or should other forms of studies (historical control trials, observational studies) be included?
  - should the review be restricted to double-blind studies?
- should only studies which control for potential confounders (*eg* animal age) be included?
- other study characteristics
  - should a minimum study size criterion be employed?
  - should studies be limited to those carried out in a specific geographic region?
- logistic concerns
  - should the review be limited to studies published in a specific language(s)?
  - should only studies for which complete reports are available be used?

24.3.3 Find the studies

The literature review on which a meta-analysis is based must be both complete and well-documented. The most commonly used approach to ensuring that all published studies are found is to carry out computer-based literature searches of the major electronic databases (*eg* Medline, Agricola, Index Veterinarius and the Veterinary Bulletin) and to follow this with a review of the reference lists in all of the papers identified through the computer-based search. The search process, including the names and date ranges of all databases searched along with the search strategy (*eg* keywords used) must be documented. When selecting studies for inclusion, any exclusions not already specified in the inclusion/exclusion criteria need to be documented.

One of the difficult issues to address is whether or not the review should include unpublished studies. The potential effects of publication bias are discussed in section 24.4.5, but identifying and obtaining results from unpublished studies is a difficult task. In some cases, databases of funded research projects could be used to identify studies that have been conducted, but not published. Alternatively, personal contact with investigators working in the field might identify unpublished studies.

24.3.4 Evaluation of study quality

There are two general approaches to the evaluation of study quality. The first is to establish a checklist of criteria that must be met for a study to be included. For example, while a meta-analysis might be based only on randomised controlled trials (*ie* the eligibility criterion), other issues of study design (*eg* double blinding, specific formal method of randomising treatment allocation, clear criteria for eligibility of subjects in the trial) could also be evaluated. If a study does not meet all of these additional criteria, you might decide to exclude it from the meta-analysis. However, if very stringent criteria are set, you might end up excluding most studies.

The second approach is to evaluate study design issues and assign a quality score to the study. This quality score can be used to weight the studies in the meta-analysis (*ie* poor quality studies receive less weight when estimating the summary effect). This introduces a degree of subjectivity to the meta-analysis so should be carried out with caution and the method of assigning the quality score clearly defined. Alternatively, you can evaluate the quality score as a potential source of variation in study results (do poor-quality studies have more dramatic results than high quality studies?).

24.3.5 Extraction of the relevant data

The layout and presentation of results in epidemiologic studies is highly variable. This is particularly true for observational studies, but it is even an issue when reviewing randomised controlled trials. The two fundamental pieces of information that you need from each study are the point estimate of the outcome(s) of interest and a measure of the precision of that estimate (SE or CI). In some cases, these are not presented directly, but sufficient data are available to allow you to compute the required information. For example, in the rBST studies referred to above, the primary outcome for most studies was a measure of productivity, but the number of cows in each study group which had one or more clinical cases of mastitis was also reported. From these data, the risk ratio for mastitis and its CI could be computed and used in the meta-analysis.

For outcomes measured on a binary scale (*eg* occurrence of clinical mastitis), you need to decide if you will extract and record a relative measure of effect (*eg* risk ratio - *RR*) or an absolute measure (*eg* risk difference - *RD*). It is generally more meaningful to use relative measures for summarising effects. The summary estimate can then be applied to specific populations in which the overall risk of disease is known (or can be estimated) to compute an absolute effect of the intervention. Regardless of which measure of effect is used, you should record the frequency of disease (*eg* risk) in the control group as this might be a source of heterogeneity of study results (see section 24.4.4).

Before starting the data-extraction process, you need to develop a template on which to record all of the fundamental information about the study, including any information required in the evaluation of the quality of the study or to evaluate as a possible cause of heterogeneity among study results. If resources permit, it is desirable to carry out duplicate data extraction (*ie* data extracted independently by two investigators) followed by a comparison of the two datasets to identify and resolve any differences.

When carrying out the data extraction, it is also important to watch for duplicate reporting of results. In some cases, data from an individual study might be published in multiple locations (*eg* a company report and a peer-reviewed journal publication) but must only be included in the meta-analysis once. Example 24.2 describes the literature review and data-extraction process for the meta-analysis of rBST. These data are used for all subsequent examples in this chapter.

Example 24.2 Literature review and data extraction for meta-analysis

The meta-analysis of the effects of rBST on dairy cattle productivity and health was carried out by an expert panel of the Canadian Veterinary Medical Association at the request of Health Canada. The data for the meta-analyses were obtained through the following process. A literature review of three electronic databases covering the period 1984 to 1998 identified a total of 1,777 references related to rBST. A review of the titles identified 242 manuscripts that potentially contained results from randomised clinical trials or were relevant reviews of the subject. These were all reviewed by the panel members and 60 identified as useful for the review. These were combined with 26 unpublished study reports provided as part of the company's submission in support of the request for registration of the drug. From all of these reports (n=86), 53 studies (representing 94 distinct groups of cows) were found to contain original data from randomised controlled trials. Estimates of effect (n=546) on the various outcomes of interest were obtained and used in the meta-analyses.

Only data relating to milk production (3.5% fat-corrected milk) and the risk of clinical mastitis are presented in this chapter. A more detailed description of the methods used and estimates of effects on other parameters can be found in the panel's report (Dohoo et al, 1999) or subsequent manuscripts (Dohoo et al, 2003a,b).

24.4 Analytical procedures in meta-analyses

There are a number of important issues to be addressed when carrying out the analyses for a meta-analysis review. These include:

- whether to base the analysis on a fixed- or random-effects model
- how to compute the summary estimate
- how to present the data
- an evaluation of possible reasons for heterogeneity of study results (*ie* why different studies produce different estimates)
- a search for evidence of publication bias
- an evaluation of the influence that individual studies have on the outcome.

24.4.1 Fixed versus random effects models

A fixed-effects model is based on the assumption that the effect of the factor being investigated is common across studies and that any variation among studies is due only to random variation and can be written as:

$$Y_i \sim N(\mu, \sigma_i^2) \qquad \textit{Eq 24.1}$$

where Y_i is the effect estimate from each study, μ is the overall (average) effect and σ_i^2 is the known variance (SE^2) from each study.

A random-effects models assumes that a distribution of effects exists, resulting in heterogeneity among study results. It is most common to assume that the study effects have a normal distribution and they can be written as:

$$Y_i \sim N(\mu_i, \sigma_i^2)$$
$$\mu_i \sim N(\mu, \sigma^2) \qquad \textit{Eq 24.2}$$

where Y_i, μ and σ_i^2 are as in Eq 24.1, μ_i are the random effects for each study and σ^2 is the between-study variance. Random-effects models generally produce a point estimate of the summary effect that is similar to that obtained from fixed-effects model, but which has a wider confidence interval than a fixed-effects model.

A statistical test of heterogeneity (referred to as a Q statistic) can be used to determine if the variability in the individual study estimates of effect is greater than would be expected due to chance (Egger et al, 2001). The formula for this statistic depends on the weighting procedure selected for producing the summary estimate (section 24.4.2), but in all cases it is expected to have a χ^2 distribution with k-1 df (where k is the number of studies). This test has relatively low power for detecting heterogeneity when the number of studies is small, so the possibility of heterogeneity of effects should not be ruled out simply because the test yields a non-significant P-value. You might want to relax the P-value required for assessing heterogeneity (*eg* 0.1 instead of 0.05). If there is any evidence of heterogeneity, potential causes of that variability should be investigated (section 24.4.3). An estimate of the variance of the (assumed) normally distributed study results can also be obtained.

Results from fitting both fixed- and random-effects models of the effect of rBST on milk production are shown in Example 24.3.

24.4.2 Summary estimate

Regardless of whether a fixed- or random-effects model is used, a system of weighting study results based on the precision of their estimates must be employed to compute the weighted average summary effect. The most commonly used procedure is to weight the estimates by the inverse of their variance (see Egger et al, 2001 for relevant formulae). This procedure is applicable for pooling results from models of continuous (linear regression, ANOVA) and discrete (logistic, Poisson regression) data. All examples used in this chapter are based on this approach.

Alternative approaches based on the Mantel-Haenszel procedure or an approach attributable to Peto are available (Egger et al, 2001). The former might be better than the inverse variance approach if data are sparse (such as in studies where the outcome is

Example 24.3 Fixed- vs random-effects models
data=bst_milk, bst_mast

Both fixed- and random-effects models were fit to both the milk production data and mastitis data from the meta-analysis of the effects of rBST on dairy cow productivity and health. In all models, the inverse variance approach (section 24.4.2) was used to assign weights to the study results.

Milk production (28 studies):

Method	Pooled estimate (kg/day)	SE	Z	P	95% CI Lower	95% CI Upper
Fixed	4.465	0.159	28.078	0.000	4.153	4.777
Random	4.434	0.297	14.911	0.000	3.851	5.016

The Q statistic for heterogeneity was 79.9 with 27 degrees of freedom (P=0.000) indicating there was strong evidence of heterogeneity among study results. Potential reasons for this heterogeneity will be explored in Examples 24.5 and 24.6. As expected, the point estimates for the summary effect were quite similar, but the random-effects model produced wider confidence intervals.

Based on the random-effects model, the estimate of the between-study variance was 1.42 (SD=1.2) suggesting that 95% of the effects of rBST should lie between 4.4-2*1.2=2.0 kg/day and 4.4+2*1.2=6.8 kg/day.

Mastitis (29 studies):

Method	Pooled estimate (RR)	Z	P	95% CI Lower	95% CI Upper
Fixed	1.271	4.016	0.000	1.131	1.429
Random	1.271	4.016	0.000	1.131	1.429

SEs have not been computed because the analysis is carried out on the lnRR scale. The Q statistic for heterogeneity was 16.4 with 28 degrees of freedom (P=0.96) suggesting there was no evidence of heterogeneity among study results. **Note** Because Q <df, the between-study variance is assumed to be zero and the results from the fixed- and random-effects models are identical.

a relatively rare event). The latter is applicable to studies in which odds ratios or time-to-event outcomes are being pooled.

The most commonly used random-effects model is the DerSimonian and Laird model (DerSimonian and Laird, 1986) in which the study effects are assumed to follow a normal distribution, with the variance of that distribution being estimated from the data. It is most commonly used with inverse variance weighting, but could alternatively be based on Mantel-Haenszel weights. If the heterogeneity statistic (Q) is less than its df, then the variance of the distribution is assumed to be zero and this model produces results identical to a fixed-effects model.

24.4.3 Presentation of results

One of the most important outputs from a meta-analysis is a graphic presentation of the results with the most commonly used format referred to as a **forest plot** which displays the point estimate and confidence interval of the effect observed in each study along with the summary estimate and its confidence interval. Fig. 24.1 shows a forest plot for the effects of rBST on the risk of clinical mastitis and the elements of the plot are described in Example 24.4.

In some cases, it might be desirable to order the individual studies according to some criteria such as year of completion (to see if there is a trend over time) or quality score (to see if study quality affects the observed effects).

24.4.4 Evaluating heterogeneity

There are a variety of possible causes of heterogeneity of study results. Heterogeneity might be due to real differences among studied populations in their response to treatment or due to differences in study protocols. If there is evidence of heterogeneity, the summary effect must be interpreted with caution because it represents an average effect, rather than a specific effect which is applicable to any given population. When heterogeneity is observed, it is important to try to determine its cause.

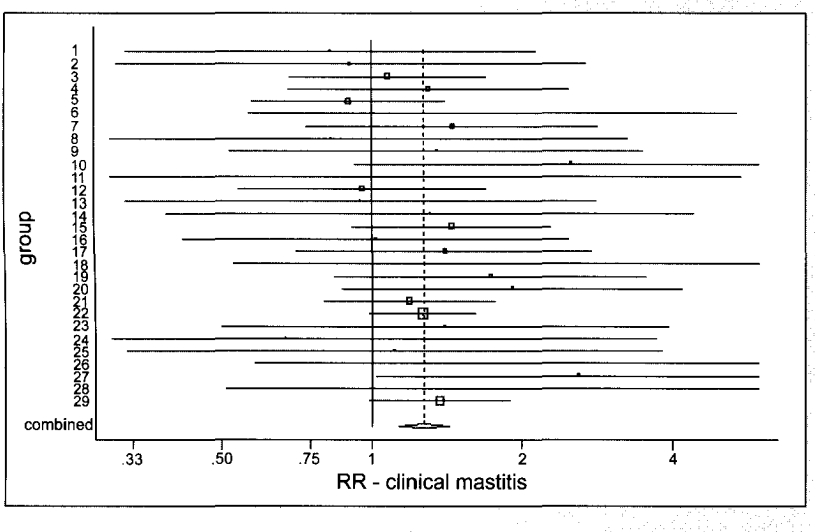
Two approaches to investigating causes of heterogeneity are to carry out stratified analyses or use meta-regression techniques. In a stratified analysis, the data are stratified according to a factor thought to influence the treatment effect, and separate meta-analyses carried out in each of the strata. The disadvantage to this approach is that individual strata might contain relatively few studies. Example 24.5 presents a stratified (by parity group) meta-analysis of the effects of rBST on milk production.

The second approach is to carry out a **meta-regression** with one or more factors that might influence study results included as predictors. A meta-regression is simply a weighted regression of the study results on the factors of interest (weights equal to the inverse variance of each study's results are most commonly used). If the number of studies is limited, factors might be investigated one at a time, or if there are sufficient data, a multivariable regression model could be built. Example 24.6 shows a meta-regression of the effects of rBST on milk production on parity group, daily drug dosage and duration of treatment (with each factor investigated individually).

Example 24.4 Forest plot
data=bst_mast

Fig. 24.1 shows a forest plot of the risk ratios for the effect of rBST on the risk of clinical *mastitis*.

Fig. 24.1 Forest plot



In these plots, each horizontal line represents the results from a single study (or distinct group of cows within a study). Each line is labelled with a unique label (the group number). The length of the line represents the 95% confidence interval for the parameter estimate from the study. **Note** Some lines have been truncated at 6 or 0.3. The centre of the shaded box on each line marks the point estimate of the parameter from that study, and the area of the box is proportional to the weight assigned to the study in the meta-analysis. Studies with large boxes have had a strong influence on the overall estimate. The dashed vertical line marks the overall estimate of the effect. The <> at the bottom of the dashed line shows the confidence interval for the estimate of the overall effect. The solid vertical line marks the value where rBST would have no effect (*ie* RR=1).

As you can see, there was considerable variability among the individual study point estimates of the RR and only one of them was statistically significant (CI excludes 1). However, as seen in Example 24.2, this variability was not greater than what would be expected due to chance (given the generally small size of most of the studies). Group 22 had the largest influence on the summary result (*ie* largest weighting).

Example 24.5 Stratified meta-analysis
data=bst_milk

Separate meta-analyses of the effect of rBST on milk production (kg milk/day) were carried out for each of the three parity groups: (primiparous, multiparous and no separation by parity) (*ie* studies which did not stratify on the basis of age).

Parity group	Number of groups	Estimate	Heterogeneity P
Primiparous	6	3.303	0.01
Multiparous	7	4.360	0.68
No separation by parity	15	5.060	<0.01

Parity seems to account for some of the heterogeneity among studies, but the results are not clear cut. Within the groups of multiparous cows, there was no longer any evidence of heterogeneity. However, there was still heterogeneity among the studies based solely on primiparous cows. You might have expected groups in which data from all parities were combined to have an effect intermediate to the other two groups, but this was not the case. However, the number of studies within each group was quite small, so the summary effects must be interpreted with caution.

24.4.5 Publication bias

When carrying out a meta-analysis, you need to consider whether or not it is likely that there are studies that have been completed, but for which the results have not been published. Study results that are not statistically significant or which are unfavourable to the sponsor of the study might be less likely to be published than significant, favourable results. Unfortunately, it is often very difficult to obtain unpublished study results. However, if you have any indication that unpublished results constitute a substantial portion of data available, then you should make an effort to obtain them. On the other hand, one argument against including unpublished results in a meta-analysis is that those results have not been peer reviewed and thus, do not have one of the key components in assuring data quality.

There are three general approaches to dealing with the problem of publication bias. The first, as described above is to contact investigators directly to obtain unpublished results, or to at least determine how many unpublished results there are. A second approach is to estimate how many studies with 'null' results (*ie* no observed effect) would have to exist before a summary effect from your meta-analysis would become non-significant.

The third approach is based on an evaluation of the relationship between study results and their precision. A **funnel plot** displays each study's estimated effect plotted against its SE. If publication bias is a problem, there will likely be a number of studies with large effects and large SEs but an absence or shortage of studies with large standard errors and small or no effects. These latter studies are the ones not published due to publication bias. Fig. 24.2 shows a funnel plot for the rBST–mastitis meta-analysis in Example 24.7.

Example 24.6 Meta-regression for evaluating causes of heterogeneity
data=bst_milk

Separate meta-regressions were carried out to evaluate the effects of parity group, duration of treatment and daily dosage on the effects of rBST on milk production.

Parity	Coef	SE	Z	P	No of studies = 28 σ^2 estimate = 1.186 95% CI	
parity=2+	1.966	0.718	2.74	0.006	0.558	3.37
parity=combined	1.210	0.833	1.45	0.146	-0.422	2.843
constant	3.068	0.609	5.04	0.000	1.874	4.262
Duration					σ^2 estimate = 1.072	
duration	-0.008	0.004	-2.11	0.035	-0.015	-0.001
constant	6.081	0.827	7.35	0.000	4.461	7.702
Dosage					σ^2 estimate = 1.471	
daily dosage	0.047	0.033	1.43	0.152	-0.017	0.111
constant	3.013	1.037	2.91	0.004	0.981	5.045

Both parity group and study duration were significant predictors of the observed study effects. The parity effects had an overall significance of P=0.02 and the coefficients mirrored the effects seen in the stratified analyses (Example 24.5). For each additional day in study duration, the effect of rBST decreased by 0.008 kg/day. There was a trend towards greater treatment effects with increasing daily dosage, but this was not statistically significant. The σ^2 are the estimates of the between-study variance after adjustment for the predictor in the meta-regression. In each of the analyses, the estimate of the between-study variation (σ^2) lay between 1.0 and 1.5.

There are a number of statistical tests based on the principle of the funnel plot. These evaluate the relationship between study results and their SEs using a rank correlation (Begg's test: Begg and Mazumdar, 1994) or a linear regression approach (Egger's test: Egger et al, 1997, or meta-regression). If an association exists, you conclude that publication bias might be influencing your results. However, the tests are not appropriate where you are looking for either positive or negative effects because they would not be able to detect a shortage of study results in the 'middle' of the funnel.

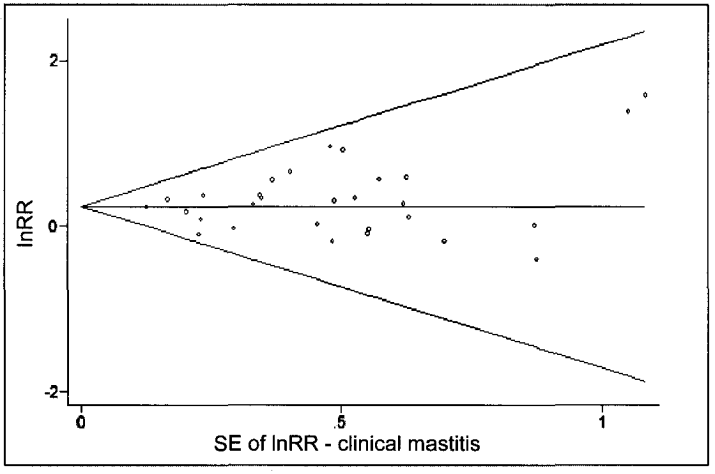
24.4.6 Influential studies

As in most regression-based models, it is important to determine if individual studies are having a profound influence on the summary estimate derived from a meta-analysis. If they are, you need to determine whether or not this is warranted. It might well be that one study was much larger than the others and consequently provides a much more precise estimate of the effect. In this situation, you need to evaluate that study to determine if it was of sufficiently high quality that you can accept the results.

Example 24.7 Publication bias
data=bst_mast

A funnel plot (ln*RR* vs SE of ln*RR*) was generated from the rBST–mastitis data.

Fig. 24.2 Funnel plot



If publication bias was a serious concern, we would expect a substantial number of studies with large SE and large effects (either positive or negative) to appear on the graph and relatively few studies with small effects (*ie* around the null value of 0) but large SE. There is little evidence of publication bias in these data.

One way to evaluate the effects of individual studies is to sequentially delete the studies from the meta-analysis and determine how the estimate of the summary effect changes (Example 24.8). The revised point estimates can all be plotted in an **influence plot** (see Fig. 24.3).

24.5 Use of meta-analysis

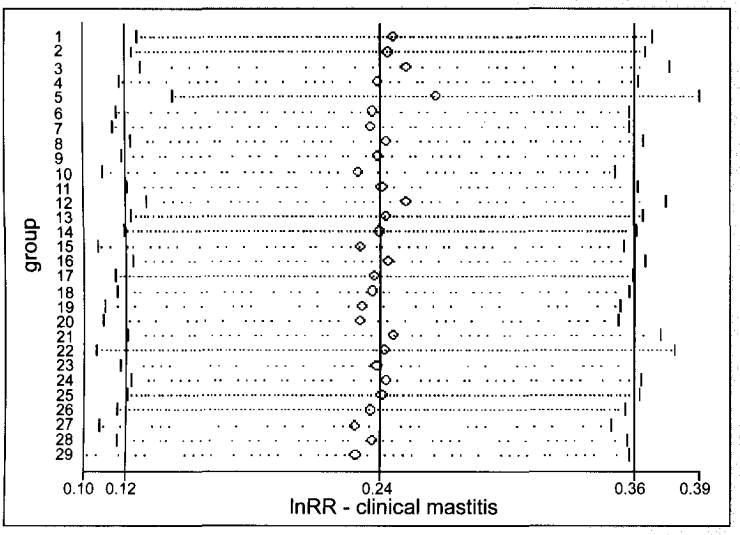
As indicated, the most common use of meta-analysis is for summarising data from a series of controlled trials. They have been used less in veterinary medicine than in human medicine because we seldom have multiple trials of a single product (or closely related group of products) on which to base a meta-analysis. However, with the increasing desire of the profession to have reliable field-based evidence of the efficacy of products used, the availability of clinical-trial data will increase.

Meta-analysis can also be used in research programmes. They might either serve as a 'definitive study' by combining the results from many previous studies or they can be used to help design future studies by providing the best estimate of effect for use in

Example 24.8 Influential studies
data=bst_mast

An influence plot was generated to determine the effect of removing individual studies from the meta-analysis of rBST on the risk of clinical mastitis.

Fig. 24.3 Influence plot



No individual study (group of cows) had an undue influence on the summary effect estimate. Omitting Group 5 had the largest effect and in this case the ln(*RR*) rose from 0.24 to 0.27 (equivalent to a rise in *RR* from 1.27 to 1.31). This is a relatively small change, indicating that no individual study had a particularly large influence on the summary *RR* estimate.

sample-size calculations. If a series of studies is being conducted, the results of a meta-analysis can also provide a 'stopping rule' by identifying when sufficient evidence of the efficacy of a product exists to warrant halting research on it. A meta-analysis might also identify factors that strongly influence study results (*ie* contribute to heterogeneity) and guide future research into those effects.

Meta-analysis can also be used to help guide policy decisions. For example, the meta-analysis of the effects of rBST on dairy cattle health and production was one of the pieces of information used by Health Canada when making a decision regarding the registration of the drug for use in Canada (in this case the decision was to not register the drug).

SELECTED REFERENCES/SUGGESTED READING

1. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994; 50: 1088-1101.
2. Cochrane Collaboration. The reviewer's handbook. http://www.cochrane.org/cochrane/hbook.htm, 2001.
3. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7: 177-188.
4. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science. Epidemiol Rev 1992; 14: 154-176.
5. Dohoo IR, DesCôteaux L, Dowling P, Fredeen A, Leslie KE, Preston A et al. Report of the Canadian Veterinary Medical Association Expert Panel on rBST. Health Canada, 1-420, 1999. Ottawa, Canada, Health Canada. http://www.hc-sc.gc.ca/english/archives/rbst/animals/
6. Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Dowling P, Preston A et al. A meta-analysis review of the effects of rBST. 1. Methodology and effects on production and nutrition related parameters. Can J Vet Res 2003a; in press.
7. Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Shewfelt W, Preston A et al. A meta-analysis review of the effects of rBST. 2. Effects on animal health, reproductive performance and culling. Can J Vet Res 2003b; in press.
8. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 1997; 315: 629-634.
9. Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London: BMJ Books, 2001.
10 Fourichon CSH, Seegers H, Malher X. Effect of disease on reproduction in the dairy cow: a meta-analysis. Theriogenology 2000; 53: 1729-1759.
11. Freemantle N, Cleland J, Young P, Mason J, Harrison J. β Blockade after myocardial infarction: systematic review and meta regression analysis. British Medical Journal 2002; 318: 1730-1737.
12. Glass GV. Primary, secondary and meta-analysis of research. Educ Res 1976; 5: 3-8.
13. Normand SLT. Meta-analysis: formulating, evaluating, combining and reporting. Stat Med 1999; 18: 321-359
14. Petitti DB. Meta-analysis, decision analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine, 2d ed. London: Oxford University Press, 1999.
15. Rothman KJ, Greenland S. Modern epidemiology, 2d ed. Philadelphia: Lippincott -Raven, 1998.
16. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. New York: John Wiley & Sons, 2000.
17. Tukey J. The future of data analysis. Ann Math Stat 1962; 33: 1-67.

SAMPLE PROBLEMS

1. A new treatment (treatment X) has been developed to aid in the treatment of *Staph. aureus* mastitis in dairy cows. The manufacturer claims that when used in conjunction with conventional antibiotic dry-cow therapy, the product increases the treatment success rate. There have been seven clinical trials carried out and they are summarised below. Your assignment is to:
   a. Come up with the best possible estimate of the effect of this treatment.
   b. Determine if the estimate of effect depends on whether the investigators took the age of the cow into consideration in the analyses.
   c. Determine if the estimate of effect depends on whether it was used in conjunction with cloxacillin or a cephalosporin-based dry-cow product.

Clinical trial

1. This was a trial carried out in 162 *Staph. aureus* infected cows with 51 of 82 cows that received treatment X in addition to cloxacillin being *Staph aureus* negative at calving while 27 of 80 cows receiving only cloxacillin were negative.

2. In this 116-cow study in which the antibiotic used was cloxacillin, the logistic regression output was as follows (outcome was 'cure' of infection):

Factor	Coef	SE
Tx X	-0.27	0.37
age	-0.03	0.001

3. The logistic regression output from this 158-cow study in which the antibiotic used was cephalosporin was (outcome was 'cure' of infection):

Factor	OR	CI
Tx X	1.6	0.85-2.95
age	0.9	0.85-0.94

4. In this 54-cow study, cephalosporins were used along with treatment X and 19 of 28 cows that got Tx X eliminated *Staph. aureus* while only 8 of 26 control cows did.

5. The logistic regression output from this 145-cow study in which the antibiotic cloxacillin was used (outcome was 'cure' of infection):

Factor	Coef	CI
Tx X	0.87	0.21-1.53
age	-0.06	-0.09 - -0.03

6. The logistic regression output from this 71-cow study in which the antibiotic used was cephalosporin was (outcome was 'cure' of infection):

Factor	OR	CI
Tx X	3.07	1.17-8.03
age	0.8	0.7-0.9

7. This was a trial carried out in 44 *Staph. aureus* infected cows with 17 of 23 cows that received treatment X in addition to cloxacillin being *Staph. aureus* negative at calving while 5 of 21 cows receiving only cloxacillin were negative.

25

ECOLOGIC AND GROUP-LEVEL STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. List the three major categories of variable used in ecologic models. Describe their attributes and apply these to a specific research question.

2. Describe the constructs of a linear model at the individual and group level and the constraints on estimating incidence rate ratios at the group level.

3. Describe how within-group misclassification, group-level confounding and group-level interaction can effect causal inferences.

4. Describe the basis of the ecologic and atomistic fallacies.

5. Identify scenarios where ecologic studies are less likely to produce cross-level inferential errors.

6. Describe the rationale for using non-ecologic group-level studies in epidemiologic research.

25.1 INTRODUCTION

The initial part of this chapter deals with studies in which groups of subjects are sampled, and analyses are conducted at the group level when the researcher wishes to make inferences to individuals. These are called **ecologic studies**. The primary analytic feature of an ecologic study is that we do not know the joint distribution of the risk factor(s) and the disease within each group. In other words, although we know the proportion exposed and the risk or rate of cases within each group, we do not know the proportion of exposed cases, typically because we lack individual level data on the risk factor, the disease, or both (Rothman and Greenland, 1998).

For example, in an ecologic study of the role of selected micro-organisms as potential causes of respiratory disease (BRD) in pens of feedlot cattle, we would know the pen-level incidence of BRD and the pen-level frequency of infection with each organism; however, we would not know the joint distribution of BRD and each organism. The lack of this piece of information can lead to inferential problems. Thus, given a positive association between infection with a micro-organism and higher rates of BRD, it is possible that the animals developing BRD are those that are not infected with the organism in question.

Ecologic studies might be called **exploratory** if there is no direct measure of the exposure of interest or if there is no specific exposure variable being studied. For example, if a study portrayed the rate of disease (*eg E coli* 0157 in humans) by administrative area on a map, we might use previous knowledge of local features (*eg* cattle density) to explain the observed spatial variation in rates of disease, even though there was no direct measurement of this factor in the study. Ecologic studies might be called **analytic** if the exposure factor is measured and included in the analysis.

In general, ecologic studies can be conducted using the same approaches as used for studying individuals; namely by:

1. comparing the frequencies of exposure and disease among **a number of** groups at a given point in (or during a limited period of) time, similar to cross-sectional studies, or
2. estimating the changes in both exposure and disease frequencies during a given period in one or more groups (often in just one group) as in cohort or case-control studies, or
3. a combination of the two types.

If the groups are small, the analysis should account for the different precision of disease rates by group. Spatial analysis might require adjustment for spatial correlation. Temporal studies might need to adjust for a lag period and inferences might need to take account of changes in diagnostic standards. Studies that include an extended period of time might have to account for and try to separate the age, period, and cohort effects on the outcome. This leads to an identifiability problem as these three components are interlinked and cannot be assessed independently (Osmond and Gardner, 1989; Robertson et al, 1999 for a discussion). Studies that combine both among-group and temporal approaches might provide a more thorough test of the hypothesis than either

approach alone. We begin our discussion by asking ourselves why we might study groups, especially if we want to make inferences to individuals?

25.2 Rationale for studying groups

Particularly in veterinary medicine, the group (*eg* the herd) is often the sampling unit as well as the unit of interest; these are not ecologic studies (Carver et al, 2002). The aggregate level, for example, litters of animals, hives of bees, sea-pens of fish, flocks/barns of poultry, mobs of sheep *etc* is often of more interest than the elements or components (*ie* individual piglets, bees, chickens, fish, sheep *etc*) of the group. The recent increase in the use of spatial statistics often focuses on even larger aggregates such as cities, districts, watersheds, and so forth. **Providing the variables are measured at the group level and any inferences are directed towards this level this poses no particular problems**. See section 25.7 for further discussion of non-ecologic group-level studies. It is often the intent, however, to make inferences about individuals based on the results from the group-level analysis, and in doing this, one must be very careful (reasons for this are discussed subsequently). Nonetheless, the major advantages of studying groups are:

Measurement constraints at the individual level Often, it is difficult to measure exposure at the individual level (*eg* level of pollutants, dietary intake) so an average for the group might suffice. In other circumstances, the variation in diet within an individual might be large, whereas the group average might adequately reflect exposure to specific nutrients for the purposes of the study.

Exposure homogeneity If there is little variation in exposure among individuals within a group, it might be difficult to assess the exposure's impact on them. For example, if all animals within a group are managed the same, one might need to study groups to observe the apparent effect of different management schemes. Hence, using groups with a wider variation in level or type of exposure than exists within groups would be helpful.

Interest in group-level effects These arise naturally if one is studying the impact of area-wide programmes, or area-wide exposures. For example, in many circumstances vaccines, different rations, types of housing, and treatments (*eg* water or feed-based antimicrobials) can only be delivered, or implemented practically, at the group level. Hence, farms or groups are of interest.

Simplicity of analysis Often it appears to be easier to display and present group-level rather than individual-level data. However, group-level analyses might hide serious methodological problems if we are attempting to make inferences to individuals (see section 25.4).

25.3 Types of ecologic variable

The categorisation of variable types within ecologic studies is still dynamic (see Diez-Roux, 1998a,b and McMichael, 1999, for a discussion). For our purposes, we will use three categories: aggregate, environmental and global variables.

25.3.1 Aggregate variables

Aggregate variables are summaries of measurements made on individuals within the group such as: the proportion exposed, the average age, average nutrient intakes *etc*. They can relate to the predictor variables, the outcome variable, or both. When a disease is the outcome, it is usually measured using rates because most groups are open; if closed, then a risk-based approach can be used. This type of variable is also called a **derived variable**. The type of derived variable used in ecologic studies is that which is formed, at least in part, by aggregating individual observations to form a summary variable (usually the mean) for the group (*eg* proportion exposed, feed conversion ratio, average daily gain, average somatic cell count, disease rate, mortality rate *etc*).

25.3.2 Environmental, or contextual, variables

Usually these are physical characteristics of the group such as local weather, level of pollutants in the area, or herd characteristics such as bulk-tank somatic cell count, characteristics of water supply (*eg* deep well versus surface water), and management strategy (*eg* teat-dipping strategy or colostrum-feeding protocol). The key feature of these variables is that they have an analogue at the individual level (*eg* the colostrum-feeding protocol might state that every calf gets a litre of colostrum within four hours of birth; whereas the individual level factor would indicate whether this particular calf received that amount of colostrum within that time period). Often we do not actually measure these variables at the individual-level because of practical constraints and for analysis, we assign the same value of the variable to every individual within the group. This approach becomes especially tenuous as the within-group variance in that factor increases. For example, a farmer might say that all calves get adequate colostrum, but in fact, only a small proportion actually receives it in the appropriate time or manner so serious misclassification results. In addition, it might well be that there is an interaction between the factor at the individual level (*eg* titre to agent *X*) and the contextual variable for the same factor (*eg* percentage of animals with a protective titre), as in herd immunity and these need to be identified for proper inference.

25.3.3 Group, or global, variables

These variables reflect a characteristic of groups, organisations or places for which there is no analogue at the individual level (*eg* population density). Global variables include farmer characteristics, and herd characteristics or management strategies such as herd size, open versus closed herd policy, density of housing, reproductive strategies, and some disease prevention programmes.

25.4 ISSUES RELATED TO MODELLING APPROACHES IN ECOLOGIC STUDIES

We begin by noting that, at the group level, both predictor and outcome ecologic variables often are measured on a continuous scale, even though factors might be dichotomous at the individual level; this is particularly true when aggregate variables are used. As mentioned, if the outcome at the group level is classified as dichotomous (*eg* disease present or absent) and the inferences are at the group level, the study is not an ecologic study and can be pursued with the same features and constraints as ordinary observational studies (Chapters 7-10). With aggregate variables, because the outcome reflects the average rate or risk for the group, a natural scale for modelling group level variables is the linear regression model (as outlined in Chapter 14) in which we regress the grouped outcome variable on the grouped exposure variables. Some prefer to use a Poisson model (see Rothman and Greenland, 1998, pp 464-465 for other examples of analytic approaches. Ducrot et al, 1996, also discuss these in the context of veterinary medicine).

As an example of the linear model approach, we can imagine the continuous outcome Y representing the risk or rate of disease (*eg* 0.15 per animal-year for the first group) modelled as a linear function of the exposure (*eg* 0.3 of the calves in the first group j do not receive early adequate colostrum) and perhaps adjusting for the effects of one or more confounders (*eg* the average age of calves in each group). The model could be specified as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where X_1 is the proportion receiving adequate colostrum and X_2 is the average age. Environmental or global variables might be entered and analysed as either dichotomous, ordinal or continuous variables. The linear model would provide an incidence rate difference (ID_G) from the exposure which is estimated as β_1, conditional on the other variables in the model. In many analyses the outcome might need to be transformed to better meet the assumptions of the linear model, and a weighted regression might be needed to account for the different levels of precision by group (because of differences in the number of study subjects). The outcome often should be weighted by the group size, the reciprocal of the within-group variance, or some function relating to the within-group homogeneity of exposure.

A 'nice' feature of a linear model is that if the rate (or risk) difference is constant across groups at the individual level, assuming no other biases, the rate difference at the group level will be of the same magnitude. In contrast, if the rate ratio is constant at the individual level, a logit model of the outcome will not produce unbiased estimates at the group level (Rothman and Greenland, 1998, p 468).

Associations between predictors and dichotomous outcomes at the individual level are usually based on ratio measures. However, a problem with using ratio measures at the group level in linear models is that, for aggregate variables, these estimates often force us to extrapolate our inferences to groups with no exposure and to groups with 100% exposure; rarely do we have these groups in our data. For example, from a simple linear model β_0 is the rate in non-exposed (X=0) groups and $\beta_1+\beta_0$ is the rate in exposed

groups (X=1). Hence, the incidence rate ratio at the group level is:

$$IR_G = \frac{\beta_0 + \beta_1}{\beta_0} = 1 + \frac{\beta_1}{\beta_0} \qquad \textit{Eq 25.1}$$

Hence, valid inferences about ratio measures requires totally exposed and non-exposed groups.

As in linear models (Chapter 14) issues of confounding and interaction are dealt with by including these variables in the model. Control of individual level confounders in an ecologic analysis, however, is less successful than it is in an individual analysis because control is performed by using average or proxy data, hence attenuating associations. Also, risk factors in ecologic analysis tend to be more highly correlated with each other than they are at the individual level making it difficult to isolate the effect of individual risk factors. When other variables are included in the model, the previous estimation method for IR_G must be extended to account for their effect. In order to accomplish this, we usually set the value of these variables (that is the Xjs) to their mean as shown in Eq 25.2.

$$IR_G = \frac{\left(\beta_0 + \beta_1 + \sum \beta \overline{X}\right)}{\beta_0 + \sum \beta \overline{X}} \qquad \textit{Eq 25.2}$$

where $\sum \beta \overline{X}$ is the sum of the products of the other coefficients in the model and the mean values of the other X variables.

Some researchers prefer to use standardised outcomes, such as (standardised morbidity/mortality ratios (*SMR*s) to control confounding and they regress these standardised outcomes on the group level explanatory variables. Typically age, sex, and breed are included in the *SMR*. However, this approach does not prevent confounding unless the explanatory variables are also standardised in the same manner, and usually sufficient data to achieve this are not available.

Interaction is usually modelled in the same manner as with individual analyses using a product term (*eg* $X_1 * X_2$). However creating this term based on group means is not equivalent to taking the average of the terms created at the individual level. Thus, this approach has a different (often lower) level of ability to detect an interaction. One particular type of interaction that is important to identify is a contextual effect where the group-level factor modifies the same factor's effect at the individual level. To identify this contextual effect, we create a cross-product term between the factor at the group and the individual level and test its significance.

25.5 Issues related to inferences

The major inferential problems that arise are because of heterogeneity of exposure and of confounders within the group. Thus a finding, at the group level, that exposure

increases (or decreases) the risk of disease by, for example, three times, does not mean that this is true at the individual level. Indeed, it might not mean that the exposed subjects are the ones having the highest individual risk of becoming cases. This error in inference is termed the **ecologic fallacy** (see section 25.7.2 for atomistic fallacy). In addition, even without the ecologic fallacy, the group-level bias almost always exaggerates the magnitude of the true association away from the null, but occasionally it reverses the direction of the association.

We now examine the three major causes of ecologic bias – within-group bias, group-level confounding and group-level interaction – in more detail.

25.6 Sources of ecologic bias

25.6.1 Within-group bias

Within-group bias can be caused by confounding, selection bias or misclassification. Here we discuss only misclassification of individual-level exposure and its effects on observations at the group level.

If aggregated exposure variables are used, the exposure level of groups is defined by combining individual exposure observations. Imperfect exposure classification of individuals in turn leads to errors in the estimates of both the individual level association and the group-level association. As noted in Chapter 12, non-differential exposure misclassification at the individual level biases the observed association toward the null, but, in ecologic studies, it biases the association **away** from the null. The effect of this bias on the rate ratio derived from an ecologic linear regression model can be predicted if the necessary data are known as indicated in Eq 25.3:

$$IR_G = 1 + \frac{IR - 1}{Se + Sp * IR - IR} \qquad \textit{Eq 25.3}$$

where *Se* is the individual-level sensitivity, *Sp* is the individual level specificity, and *IR* the true individual-level incidence rate ratio. The ID_G is also biased by the factor (*Se*+*Sp*-1). This bias can be quite large as shown in Example 25.1. Also, when exposure (or disease) prevalence of groups is based on a small sample of individuals within each group, measurement error at the individual level is compounded by sampling error (hence the earlier referral to extreme values of outcomes with small group sizes). For more details on this bias, see Brenner et al (1992).

Example 25.1 The effect of individual-level exposure misclassification on group-level results

We begin with the correctly classified study population in two farms (j=1,2) in Table 25.1.

Table 25.1 Correctly classified population structures

	Farm 1			Farm 2		
Correctly classified	Exposed	Non-exposed	Totals	Exposed	Non-exposed	Totals
Number of cases	50	40	90	100	30	130
Animal-time (t_j)	200	800	1000	400	600	1000
Rate (I_j)	0.250	0.050	**0.090**	0.250	0.050	**0.130**
Group proportion exposed		**0.20**			**0.40**	

The data in **bold typeface** are the numbers one would use for the analysis at the group level if there was no misclassification. Note that in Farm 1, 20% of the animal-time is exposed (200/1000), while in Farm 2, this is 40% (400/1000). At the individual level, the IR=5 and the ID=0.20. The regression coefficients for the group level analysis are obtained by solving the two equations for the two unknowns: $0.09=\beta_0+\beta_1*0.2$ and $0.13=\beta_0+\beta_1*0.4$ which gives the following model $Y=0.050+0.2X$. The ID_G=0.20 and

$$IR_G = 1+\frac{0.2}{0.05} = 1+4 = 5$$

Based on an exposure sensitivity of 0.8 and an exposure specificity of 0.9, and using the general approach shown in section 12.6, we would observe the data shown in Table 25.2.

Table 25.2 Misclassified population structure

	Farm 1			Farm 2		
Incorrectly classified	Exposed	Non-exposed	Overall rate	Exposed	Non-exposed	Overall rate
Number of cases	44	46	90	83	47	130
Animal-time (t_j)	240	760	1000	380	620	1000
Rate (I_j)	0.183	0.061	**0.090**	0.218	0.076	**0.130**
Group proportion exposed		**0.24**			**0.38**	

At the individual level, (based on the misclassified data pooled over the farms) the IR=3.04 and the ID=0.137. Here, the exposure misclassification leads to biased estimates of the proportion of animal-time exposed on each farm; the difference between these becomes smaller and hence, the apparent effect of exposure becomes larger. Using the same approach to obtain the regression coefficients, the model is $Y=0.0214+0.286X$. At the group level, the misclassified IR_G is 14.3 and the ID_G is 0.29. Thus, a non-differential misclassification at the individual level has biased the group IR_G and ID_G away from the null at the group level.

25.6.2 Confounding by group

If the background rate of disease in the unexposed individuals varies across groups, this sets up a group-level correlation of exposure and outcome. Such confounding can arise from the differential distribution of extraneous individual-level risk factors across groups (note that these risk factors need not (although they can) be confounders at the individual level (*ie* within groups)), or from the occurrence of group-level confounders (*ie* here the covariates are associated with both exposure and disease at the group level). Example 25.2 explains this phenomenon.

Example 25.2 The effects of confounding on group-level results

In this example, E_1 is the exposure of interest at the individual level and E_2 is the potential individual-level confounder (both binary). At the group level, these are represented by the variables X_1 and X_2, respectively (for simplicity, we omit subscripts for farms), both measured on the continuous scale (**bold typeface** in table). Consider these data from three farms:

Farm A	E_2+		E_2-		E_2 pooled	
	E_1+	E_1-	E_1+	E_1-	E_1+	E_1-
Cases	52	74	5	7	57	81
t_a	260	740	260	740	520	1480
I_a	0.20	0.10	0.02	0.01	0.11	0.055
IR_a		2		2		2
	X_1=p(E_1+)=0.26		**X_2=p(E_2+)=0.50**		**Y=p(D+)=0.068**	
Farm B	E_2+		E_2-		E_2 pooled	
	E_1+	E_1-	E_1+	E_1-	E_1+	E_1-
Cases	56	52	8	8	64	60
t_b	280	520	420	780	700	1300
I_b	0.20	0.10	0.02	0.01	0.09	0.046
IR_b		2		2		2
	X_1=p(E_1+)=0.35		**X_2=p(E_2+)=0.40**		**Y=p(D+)=0.062**	
Farm C	E_2+		E_2-		E_2 pooled	
	E_1+	E_1-	E_1+	E_1-	E_1+	E_1-
Cases	60	30	14	7	74	37
t_c	300	300	700	700	1000	1000
I_c	0.20	0.10	0.02	0.01	0.74	0.037
IR_c		2		2		2
	X_1=p(E_1+)=0.50		**X_2=p(E_2+)=0.30**		**Y=p(D+)=0.056**	

(continued on next page)

Example 25.2 (*continued*)

Examining these data from the individual's perspective, we observe that the true (individual) *IR*s for E_1 and E_2 are 2 and 10, respectively. Both ratios are constant across farms so there is no interaction at the individual level. Also, there is no confounding by E_1 or E_2 within farms (as E_1 and E_2 are independent). However, because the prevalence of E_2 varies by farm, this results in an association of farm with Y that is independent of E_1. Consequently, the group-level estimate of the effect of E_1 (*ie* using X_1) may be biased. At the farm level a simple linear regression of Y of X_1 yields Y=0.080-0.049X_1 and the ecological estimate of IR_G is (0.031/0.080)=0.39 suggesting that exposure is sparing. Controlling for exposure 2 in the analysis does not prevent the bias with the equation being Y=0.038+0.000X_1+0.060X_2. The ID_G is zero, and using the mean prevalence of exposure for X_2 of 0.40, when X_1 changes from 0 to 1 we have

$$IR_G = 1 + \frac{(.038 + .000 + .4 * .06)}{(.38 + .4 * .06)} = 2.00$$

This adjustment brings the IR_G for exposure 1 to the null value suggesting 'no effect.' Unfortunately, because we rarely have sufficient information to know whether or not the group and individual level results agree, relating group findings to individuals is fraught with difficulties.

25.6.3 Effect modification (interaction) by group

In a linear model, bias will occur at the group level if the rate difference at the individual level varies across groups. We should recall that although we use a logit scale (usually) at the individual level, we often use a linear model at the group level. This introduces a non-linearity into the comparison of the results which might evidence itself as interaction in the linear scale. Such variation can arise from the differential distribution of individual level effect modifiers across groups, or due to effect modification by a group-level factor (Example 25.3).

25.6.4 Summary of confounding and interaction at the group level

To summarise the previous discussion, cross-level (*ie* ecologic) bias will not occur if :

- the incidence rate difference, within groups, is uniform across groups, and
- if there is no correlation between the group-level exposure and the rate of the outcome in the unexposed.

The only (but huge) drawback to these criteria is that individual-level data are required to evaluate them and these data rarely are available.

On the other hand, if individual-level effect modifiers are differentially (*ie* unequally) distributed across groups, ecologic bias will occur as a result of the consequent group-level effect modification. If extraneous risk factors are differentially distributed across groups, ecologic bias will occur as a result of group-level confounding, **regardless** of whether the extraneous risk factor is a confounder at the individual level or not. Controlling for the extraneous risk factor in the ecologic analysis will generally remove only part of the bias.

Example 25.3 Effect modification by group

Consider the following data from three farms:

	Farm A		Farm B		Farm C		Total	
	E+	E-	E+	E-	E+	E-	E+	E-
Cases	120	30	120	36	120	42	360	108
Animal-time (t)	1000	1000	800	1200	600	1400	2400	3600
I	0.12	**0.03**	0.15	**0.03**	0.20	**0.03**	0.15	**0.03**
IR	4.0		5.0		6.7		5.0	
ID	0.09		0.12		0.17		0.12	
X_1 = p(E+)	**0.5**		**0.4**		**0.3**			
Y = p(D+)	**0.075**		**0.078**		**0.081**			

First let's examine the data from the perspective of the individual. We observe that the effect of the exposure *E* (as denoted by *IR*, or the *ID*) varies by farm. Thus some farm-level factor is interacting with the exposure *E*, and with a large enough sample, this might be declared as significant interaction on either the additive or the multiplicative scale (see Chapter 13). Note, that there is no confounding by any group (*ie* farm-level) factor at the individual level because p(*D*+|*E*-)=0.03 in all three farms. Thus, farm *per se* is not a cause of disease at the individual level. Also, because there is no confounding, the crude *IR* of 5.0 provides an unbiased estimate of the effect at the individual level. There is, however, interaction because some factor at the farm level is making the impact of exposure (whether measured by *IR* or *ID*) to vary, across farms, and this effect increases as the prevalence of *E*+ decreases.

An ecologic analysis at the farm level would only use the aggregated summary data (**bold typeface**) from the table. The ecologic linear regression of *Y* on *X* yields:

$$Y = 0.09 - 0.03X$$

and the ecologic estimate of IR_G would be:

$$1 + (-0.03/0.09) = 0.67$$

Clearly this is not anywhere near the individual-level *IR* of 5. Thus, the effect modification by group has led to an ecologic bias that actually reversed the direction of the association at the individual level.

It is clear we need to be careful when making inferences about individuals based on group-level analyses; yet, group-level analyses will continue to be used. So, how can we help avoid some of these problems? Well, the misclassification issue is best resolved by reducing the level of errors, but the bias away from the null is still a reality and needs to be considered in all group-level studies. With respect to confounding and interaction, again these are real problems. But, both the confounding and effect modification examples used here are taken from scenarios where group-level analyses are unlikely to be rewarding because most of the variation is at the individual level. Because the outcome varies little across groups, research should focus on the individual level.

In general, ecologic bias will be less of a problem when:

1. The observed range of exposure level across groups is large. Linear regression analysis of ecologic data is especially sensitive to problems of limited among-group exposure variation. If this is the situation you are faced with, consider using other model forms, such as exponential and log-additive models;
2. The within-group variance of exposure is small; therefore in selecting study populations minimise the within-group and maximise the among-group exposure variation (sometimes using smaller, more homogeneous, groupings helps accomplish this);
3. Exposure is a strong risk factor and varies in prevalence across groups (hence the group-to-group variation in incidence is large), and
4. The distribution of extraneous risk factors is similar among groups (*ie* little group-level confounding).

Despite the pitfalls, a recent editorial reminds us that we should continue our struggle to gain valid knowledge from group level studies (Webster, 2002). While the biases discussed very likely occur frequently, the effects might be small and need not prevent us making valid inferences to individuals. In this regard, we should treat these potential biases in the same manner we do in individual-level studies; try to understand, quantify and minimise them.

25.7 NON-ECOLOGIC GROUP-LEVEL STUDIES

A number of epidemiologists have noted that our discipline initially focused on groups as the unit of interest and only recently has it shifted that emphasis to individuals. In general, it is their view that we should strive to refocus on groups. If the individual is really the level of interest then multilevel models (Chapters 21-23) allow us to include core information from higher levels of organisation, and investigate any contextual effects. However, there is also a need to focus inferences on groups *per se* (McMichael, 1995, 1999; Diez-Roux, 1998a,b).

In thinking about studying groups and whether we should be making inferences to groups or individuals, Rose (1985) stated that it is helpful to distinguish between two questions.

1. What is the etiology of a case?
2. What is the etiology of incidence?

Both questions emphasise that there is more than one cause of a given disease or condition. The first question about causes of cases requires that we conduct our study at the individual level. With individual animals as our principal or only level of interest we identify causes of disease in individuals. In this context, within a defined population (group), the use of the ratio measures of association to identify potential causes, and measure their strength, assumes a heterogeneity of exposure within the study population. In the extreme, if every subject is exposed to a necessary cause, then the distribution of cases (in individuals) would be wholly determined by individual susceptibility determined by the other components of the sufficient causes (for example, a genetic

component, not the widespread (albeit essential) exposure). In general, Rose notes that the more widespread or prevalent a risk factor is, the less it explains the distribution of cases within that population. Hence we might even conclude that a prevalent necessary cause was of little causal importance – it might even be considered normal background exposure.

In addition to this inferential problem, when we focus on individuals, we often treat any group-level factors that are present as nuisance variables, whether through using a fixed effect or a random effect modelling approach. In this context, we have not tried to explain the group-to-group variation, just deal with it. As was discussed in Chapter 23, in choosing the appropriate aggregation level to study, it is useful to examine the proportion of variance that can be attributed to the individual and to the group because this is a useful guide for focusing future investigations. Even if our focus is on individuals, it is also useful to investigate if the effect of an exposure factor on individuals depends on that, or other factors, at the group level (the contextual effects). Herd immunity is one example where we know this to be a real biological phenomenon; the prevalence of disease in a group might have a similar important effect on the nature of the disease (*eg* timing and/or dosage of first exposure) in individuals.

To address the question about causes of incidence in populations, we must investigate the determinants of group or population means (*eg* why is the disease more common in group 'A' than in group 'B'?). To do so, we need to study the characteristics of groups to identify factors that act causally by shifting the distribution of disease of the entire group. For their success, group-level studies require either a large variance of exposure levels across groups, a large study size (*ie* number of groups), or a combination of the two. Obtaining a sufficient number of groups (*eg* herds) to give a study reasonable power has often been a practical limitation of group-level studies. Nonetheless, in both herd-health management, and veterinary public-health activities, we have a particular need to know the determinants of incidence, be they groups, herds or geographic areas, in order to help prevent disease in the population.

25.7.1 The group as the aggregate-scale of interest

Virtually all epidemiologists are aware of the hierarchical organisation of the populations we study. These levels of organisation range from subcellular units, to cells, organs, body systems, individuals, aggregates of individuals (households of people, families, litter mates, pens and herds of non-human animals), neighbourhoods, states, nations *etc*. The key point is that each higher level of organisation subsumes all the properties of lower levels, but has additional unique properties of its own (Susser, 1973; Krieger, 1994; Diez-Roux, 1998a; Ducrot et al, 1996). From this, it would seem crucial that risk-factor identification is conducted in the light of the appropriate population level context, but with an awareness of risk factors at other levels of organisation. Moving beyond the primarily biologic individual-based explanations of disease causation does not imply denying biology, but rather involves viewing biologic phenomenon within their global and environmental contexts.

A natural level of aggregation as the unit of interest for veterinarians is the farm (or

kennel) as veterinary clinicians are often required to be responsible for the health care of all animals within that farm. The reason(s) we emphasise aggregates of animals as the unit of concern could in large part reflect the relative economic value of the individual; the single fish in a sea pen, the broiler chicken in a poultry house, or a single sheep in a mob is of little economic importance to the group, and therefore to its owner. The same is true to a decreasing extent of individual pigs and beef cattle. Individual dairy cattle are of more relative economic value and perhaps because of this, the majority of epidemiologic studies in dairy cattle have tended to focus on individuals. Studies of health problems in horses and companion animals are usually focused at the individual level, and a logical level on which to aggregate them for population approaches is not easily apparent. However, an obvious need when considering population control in pets is to move beyond the simple individual-animal-oriented approach of spaying the pet or constraining contact, to examining the social and biological contexts of domestic and feral pets. Similarly in vaccination programmes, if we are principally vaccinating (or prophylactically medicating) the low-risk group, we will have little impact on the disease in the population, even when a significant proportion of the population is vaccinated.

The previous ideas relating to focusing on levels beyond the individual would suggest that when researching, for example, food safety issues, while it might be necessary to include features of individual micro-organisms such as *E. coli* 0157, and/or factors which influence its survival at the individual/farm/flock level, one must also understand the operation of modern farms and modern meat-processing plants, as well as the impact of the industry structure, and the centralisation of food processing that has been under way recently in the food industry. The same comments apply to researching large-scale disease outbreaks in the food-animal industries such as BSE in cattle; regardless of its origin, one cannot deny that the spread of this disease was aided and abetted by the structure of the animal feed-stuff industries. Wing, 1998, as an example, has commented on the need to work at the large scale in resolving many of our current important problems, especially those relating to farming and the environment.

In addition to the need to conduct research at the population level to help resolve endemic diseases, collective experience has been that disease control programmes for contagious or exotic diseases need to be directed more at the population than at the individual level. Despite our most advanced tests for identifying infected individuals, at the end stages of many national-level infectious disease control programmes, the optimal strategy for disease control is almost always to focus control on groups not individuals.

25.7.2 The group as the level of inference

The desired level of inference links to the level of analysis. In some studies the intent is to identify causal factors of cases by investigating individual-level risk factors, whereas in others it might be to make inferences about causal factors of incidence by focusing on the group level. However, as noted in earlier sections, if one is trying to make inferences about one level (a lower level) from data collected at a higher level, then such cross-level inferences are open to considerable bias. If we are interested in

the interaction between animal-level and group-level variables, then that aspect can be studied using analyses aimed at individuals but with an appropriate group-level variable (*eg* prevalence of disease) included to allow the interaction to be identified.

Previously, we examined some of the features that can help us avoid the ecologic fallacy when making inferences about the effect of an exposure on individuals when we use group level or ecologic studies. In that context, correct meant the group-level findings were consistent with the findings at the individual level. However, despite our discussion on this point, given the pervasiveness of reductionism in biomedical science, it is likely that the **atomistic fallacy** (using data from lower levels to make inferences about higher levels) is undoubtedly the more common of the two errors. We certainly risk making this error if our explanations of disease in populations are based primarily on what we know about disease in individuals. However, little is written about this fallacy. The difference in our assessments of these errors likely reflects the prevailing scientific view about what constitutes valid causal inferences. It seems that ecologic fallacies are viewed as serious problems because the associations, while true at the aggregate level, are not true at the individual level; whereas in the atomistic fallacy, the facts at the cellular or individual level are deemed to be correct, regardless of how correct, or useful (or useless) that knowledge is for efficient and effective disease prevention in populations.

In addition to the atomistic fallacy, a long-held axiom is that if one is interested in populations one must study populations (McMichael, 1995). This axiom arises in part because the physical, chemical, biological and sociological/managerial properties at the higher level likely differ from those at the lower level, and in part because there are a host of sociological/managerial factors and some biological factors which operate principally at the group level. A simple physical-chemical example is that the properties of oxygen and hydrogen tell us very little about the properties of water. Also as Schwartz (1994) observes, we should not confuse characteristics of a group with that of its individuals, "a hung jury might be indecisive but its members might be anything but indecisive."

In our research endeavours, we should not look at group-level studies as only crude attempts to uncover individual-level relationships. Many criticisms of ecologic studies are based on the questionable assumption that the individual level of analysis is the most appropriate (Schwartz, 1994). In fact, the health status of an individual, is itself an aggregated measure, because it is body cells/systems, not individuals that become diseased. The threshold for disease being present in an individual usually is based on a set of criteria, some quantitative, some qualitative. Most often, as epidemiologists, we define the cutpoint(s) for 'having the disease' and then ignore the tremendous variance in severity and effects of that disease in most of our studies (because these are not our primary interest). In a similar vein, we need to study disease at the group level, where a herd might be categorised as diseased or not and we might ignore the proportion of animals with disease (*eg* if one is attempting to establish disease-free groups, then this approach is workable). However, in other studies the dichotomisation of disease presence or absence (or presence beyond a specified cutpoint) might be too crude an approach because one is forced to discard valuable information about the extent or

severity of disease at the herd level. In this situation, it might be preferable to retain the level of disease (or outcome) as a quantitative statement about disease frequency, even though there is no intent on making inferences below the group level.

In order to optimally interpret some of our group-level studies, a major issue is to differentiate the causal inferences we make about associations at the group level from inferences we might make relative to the effect of that same (or apparently similar) variable at the individual level (Schwartz, 1994; Diez-Roux, 1998a). For example, if variable X_1 at the individual level indicates seroconversion to a specific agent, then $X_2=(\Sigma X_1/n)$ at the group level inherently carries more information than just the proportion that seroconverted; by its nature a group with a low level of X_2 likely has different dynamics of infection than one with a high level of X_2. For example, as noted, it could influence the timing of initial exposure to an agent, and this is often an important factor in the type of syndrome that might result.

In conclusion, it is clear that there are numerous problems in using aggregated data to make inferences about events in individuals. Multilevel analyses allow us to include important factors from higher levels of organisation when studying individuals, including contextual effects. However, appropriately designed studies that focus on groups are needed to identify factors of importance in the distribution of health and disease in populations.

SELECTED REFERENCES/SUGGESTED READING

1. Brenner H, Greenland S, Savitz DA. The effects of non-differential confounder misclassification in ecologic studies. Epidemiology 1992; 3: 456-459.
2. Carver DK, Fetrow J, Gerig T, Krueger T, Barnes HJ. Hatchery and transportation factors associated with early poult mortality in commercial turkey flocks. Poult Sci 2002; 81: 1818-1825.
3. Diez-Roux AV. Bringing context back into epidemiology: Variables and fallacies in multilevel analyses. Am J Pub Hlth 1998a; 88: 216-222.
4. Diez-Roux AV. On genes, individuals, society and epidemiology. Am J Epidemiol 1998b; 148: 1027-1032.
5. Ducrot C, Legay J, Grohn Y, Envoldsen C, Calavas D. Approach to complexity in veterinary epidemiology; example of cattle reproduction. Natures-Sciences-Societies, 1996; 4: 23-33.
6. Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med 1992; 11: 1209-1223.
7. Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. Int J Epidemiol 1989; 18: 269-274.
8. Greenland S, Robins J. Ecologic studies: Biases, misconceptions, and counter examples. Am J Epidemiol 1994; 139: 747-760
9. Krieger N. Epidemiology and the causal web: Has anyone seen the spider? Soc Sci Med 1994; 39: 887-903.
10. McMichael AJ. The health of persons, populations, and planets: epidemiology

comes full circle. Epidemiology 1995; 6: 633-636.
11. McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 1999; 149: 887-897.
12. Morgenstern, H. Ecologic studies in Rothman KJ and Greenland S. Modern epidemiology, 2 ed. Philadelphia: Lippincott-Raven, 1998.
13. Osmond C, Gardner MJ. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. Am J Epidemiol 1989; 129: 31-35.
14. Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. J Clin Epidemiol 1999; 52: 569-583.
15. Rose G. Sick individuals and sick populations. P.A.H.O. Epidemiological Bulletin 6: 1-8, 1985.
16. Rothman KJ, Greenland S. Modern epidemiology, 2d ed. Philadelphia: Lippincott-Raven, 1998.
17. Schwartz, S. The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. Am Jour Pub Hlth 1994; 84: 819-824.
18. Susser M. Causal Thinking in the health sciences: concepts and strategies of epidemiology. Oxford University Press, Toronto, 1973.
19. Webster T. Commentary: Does the spectre of ecologic bias haunt epidemiology? Int J Epidemiol 2002; 31: 161-162.
20. Wing S. Whose epidemiology, whose health. Int Jour Hlth Serv 1998; 28: 241-252.

Sample Problems

1. Using dataset feedlot, ascertain if there are any significant associations between the serological data and the occurrence of BRD or the weight gain of calves in the first 28 days in the feedlot. The variables in this dataset are shown below:

Variable	Description
group	group identification
tag	eartag number
province	province of feedlot
brd	bovine respiratory disease (Y/N)?
brsvpos	arrival titre to brsv positive?
brsvsc	seroconversion to brsv ?
bvdpos	arrival titre to bvd virus positive?
bvdsc	seroconversion to bvd?
ibrpos	arrival titre to ibr virus positive?
ibrsc	seroconversion to ibr virus?
pipos	arrival titre to pi3 virus positive?
pisc	seroconversion to pi3 virus?
phcypos	arrival titre to Ph (now Mh)?
phcysc	seroconversion to Ph cytotoxin?
phaggpos	arrival titre to Ph (now Mh) agglutinins positive?
phaggsc	seroconversion to Ph (now Mh) agglutinins?
hspos	arrival titre to Hs positive?
hssc	seroconversion to Hs ?
wt0	arrival weight (kg)
wt28	28 day weight (kg)

   a. You might investigate whether or not you need to control for group effects, province of origin, or arrival weight in both of these models. (For this exercise we will ignore intervening variables, although you might examine the associations with just the viral agent titre data, and then add the bacterial agent titre data. For the weight gain model you might run the model with and without -brd-. We will also ignore the assessment of model fit!)
   b. What do you conclude about the role of specific agents as potential causes of BRD?
   c. What do you conclude about the role of specific agents as causes of change in weight gain ?
2. Now, create a summary dataset based on the mean values of -brd- , the province of origin, the arrival and 28-day weight, and the mean proportion positive on arrival and the mean proportion seroconverting for each of the organisms listed in the dataset. Before creating the group-level file create a new variable for weight gain

for each calf. Then obtain the average of this for the group level file.

a. With the proportion developing BRD as the outcome, ascertain if arrival weight, province of origin and any of the serologic variables are associated with BRD. If you enter -brd- into the final model, does it alter the coefficients of the other variables? Is this what you might expect?
b. Is the average of the weight gain by group similar to the difference between the average arrival weight and the average weight at day 28? Why do you think they are/(are not) similar?
c. Regress the average of the weight gain on province and the serological variables to ascertain if there are any significant associations with weight gain. Is arrival weight a confounder?

3. Return to the questions at the end of Chapter 20 assessing the role of atrophic rhinitis in pigs and their lung scores and weight gain. Compare the results you obtained from the individual-level and the group-level analyses.
   a. Is there a rational explanation/interpretation of the group level results?

26

A STRUCTURED APPROACH TO DATA ANALYSIS

OBJECTIVES

After reading this chapter, you should be able to:

1. Conduct a detailed analysis of a complex dataset arising from an epidemiologic study with a minimum of wasted time and a maximum probability of avoiding serious errors in the analysis.

2. Congratulate yourself on getting through all of the material in this text, provided you didn't skip directly to this final 'substantive' chapter.

26.1 Introduction

When starting into the analysis of a complex dataset, it is very helpful to have a structured approach in mind. In this chapter, we provide one template which, we trust, will be applicable in most situations. Others, with experience in data analysis, might have different approaches and we would not suggest that what is presented below is either the 'only' approach, or necessarily the 'best' one – as with models, every approach is imperfect but some are useful. However, for individuals getting started in veterinary epidemiology, the following will serve as a template which can be used to guide their initial efforts in data analysis.

For most, there is a strong tendency to want to jump straight into the sophisticated analysis which will provide the ultimate answer for your study. This rarely works out, but, in order to satisfy your curiosity ... go ahead and try it anyway. Just don't waste more than an hour on it and ignore whatever results you get, as they will inevitably be wrong. Having thus satisfied that primal urge to take the short cut to the end, you can proceed with a structured approach to the analysis.

We will work through the process in a logical sequence, starting with the handling of data-collection sheets and ending with keeping track of results. However, bear in mind that data analysis is an iterative process which often requires that you back up several steps as you gain more insight into your data.

However, before you start any work with your data, it is essential that you construct a plausible **causal diagram** of the problem you are about to investigate. This will help identify which variables are important outcomes and predictors, which ones are potential confounders and which might be intervening variables between your main predictors and outcomes. Keep this causal diagram in mind throughout the entire data-analysis process. **Note** With large datasets, it will not be possible to include all predictors as separate entities. This can be handled by including blocks of variables (*eg* farm-management practices) in the diagram instead of listing each variable.

26.2 Data collection sheets

It is important to establish a permanent storage system for all original data collection sheets (survey forms, data-collection forms *etc*) that makes it easy to retrieve individual sheets if they are needed during the analysis. If animals (or groups of animals) in the study have identification numbers, this makes a convenient way to store (and later retrieve) individual files. Some things to consider when dealing with the file are as follows.

- Do not remove originals from this file. If you need to take a specific sheet for use at another location, make a photocopy of the sheet.
- Never ship the original to another location without first making copies of all forms. (You don't want to lose your whole study because the post office or courier loses your package).

- Set up a system for recording the insertion of data collection sheets into the file so that you know how many remain to be collected before further work begins.
- Once all of the forms have been collected, before you do anything else, scan through all sheets to get an impression for their completeness. If there are omissions in the data-collection sheet (*ie* forgetting to complete the last page of a questionnaire), returning to the data source to complete these data will more likely be successful if it is done soon after data were initially collected rather than weeks or months later (after data analysis has begun).

26.3 Data coding

Some issues related to data coding have already been discussed in Chapter 3, in particular, the advisability of having a space to allow for coding directly on the data-collection sheet. Some other issues to consider when coding your data are as follows.

- As noted in Chapter 3, assign a specific number to all missing values. Be sure that this specific number is not a legitimate value for any of your responses.
- If you have 'open' questions, scan the responses and develop a list of needed codes before starting coding.
- Maintain a master list of all codes assigned.
- Use numeric codes. In general, avoid the use of string variables except for rare instances where you need to capture some textual information (*eg* a comment field).
- Only code one piece of information in a single variable. Never make compound codes! For example, if you have recorded both the sex and breed of cats in a study, it might be tempting to code them as 1=male, domestic shorthair, 2=female domestic shorthair, 3= male Siamese, etc. Do not do this. Create separate variables for sex and breed. (In fact, sex might be coded in two variables male/female and neutered/intact).
- For all types of data, note any obvious outlier responses (*eg* an individual cow's milk production reported as 250 kg/day) and correct them on the datasheet.
- Use a different coloured pen so your coding notations can clearly be differentiated from anything previously recorded on the data collection sheets.

26.4 Data entry

Some of the issues to consider when entering your data into a computer file are as follows.

- Double-data entry, followed by comparison of the two files to detect any inconsistencies, is preferable to single-data entry.
- Spreadsheets are a convenient tool for initial data entry, but these must be used with extreme caution; because it is possible to sort individual columns, it is possible to destroy your entire dataset with one inappropriate 'sort' command.

- Custom data entry software programs provide a greater margin of safety and allow you to do more data verification at the time of entry. One such program in the public domain is EpiData (http://www.epidata.dk/).
- If you expect large quantities of multilevel data (*eg* every lactation for each dairy cow from several herds over several years), using hierarchical database software can make data entry and retrieval more efficient. Alternatively, you can set up separate files for data at each level (*eg* a herd file, a cow file *etc* and merge the files after data entry.
- As soon as the data-entry process has been completed, save the original data files in a safe location. In large, expensive trials it might be best to have a copy of all originals stored in another location.
- If the data entry program which you use does not have the ability to save your data in the format of the statistical package that you are going to use, there are a number of commercially available software programs geared specifically to convert data from one format to another.
- If you use a general purpose program (*eg* spreadsheet) to enter your data, as soon as the data are entered, convert them to files usable by the statistical program that you are going to use for the analysis. Do all of the analyses in that statistical program (*ie* don't start doing basic statistics in the spreadsheet). You are going to need the statistical program eventually, and it will be a lot easier to keep track of all of your analyses if they are all done there. This will also simplify the process of tracking modifications to the data.

26.5 KEEPING TRACK OF FILES

It is important that you have a system for keeping track of all your files. Some suggestions that will help you do this are:

- Assign a logical name with a two-digit numerical suffix (*eg* calf01). Having a two-digit suffix allows you to have 99 versions which still sort correctly when listed alphabetically.
- When data manipulations are carried out, save the file with a new name (*ie* the next available number). Do not change data and then overwrite the file.
- Keep a simple log of files created (Table 26.1) with some very brief information about the contents of the file.

Table 26.1 Example of data of files created in a study on calf septicemia

File Name	Date created	Description	# Obs.	# Vars.
calf01.wb3	27/09/97	original calf data entry by Glen, QP format, 1 record per calf	275	41
calf01.dta	28/09/97	original file - Stata format	275	41
calf02.dta	30/09/97	breed codes expanded	272	47
		three records with invalid IDs dropped		
etc...				

26.6 KEEPING TRACK OF VARIABLES

You are often faced with keeping track of a bewildering array of variables in a dataset from an epidemiologic study. We are not advocating studies with huge numbers of predictors (in fact we discourage such studies), but even a relatively focused study can give rise to a large number of variables once transformed variables, and/or recoded categorical variables, have been created. To help keep track of these variables, we recommend the following.

- Use short (but informative) names for variables and have all related variables start with the same name. For example, the following might be a logical set of variable names for information relating to the age of a dog.
  - age = the original data (in years)
  - age_ct = age after centring by subtraction of the mean
  - age_ctsq = quadratic term (age_ct squared)
  - age_c2 = age categorised into 2 categories (young vs old)
  - age_c3 = age categorised into 3 categories.
- Long names can often be shortened, but kept recognisable, by removing vowels (*eg* flrng as a short form for flooring).
- In some cases, adding a single letter prefix might help keep groups of variables together. For example, a series of bacteriological results might be named b_ecoli, b_staphau, b_strepag *etc.*
- If the statistics program you use is 'case sensitive' (*ie* differentiates between 'd' and 'D'), use ONLY lower-case letters.
- At some point you will want to prepare a master list of all variables with some very basic information. It should be possible to have the statistical program prepare this listing (or one similar to it).

26.7 PROGRAM MODE VERSUS INTERACTIVE PROCESSING

Some statistical programs can be used in an interactive mode where individual functions are carried out by either selecting items from menus or typing in a command. While very useful for exploring your data and trying out analyses, this interactive mode should not be used for any of the 'real' processing and/or analysis of your data because it is very difficult to keep a clear record of steps taken when using programs in this manner. Consequently, it is difficult, or impossible, to reconstruct the analyses you have completed.

The alternative is to use the program in 'program mode' in which you compile the commands necessary to carry out a series of processing steps or analyses into a program and then run the program. These program files can be saved (again, a logical naming convention is required) and used to reconstruct any analyses you have carried out. Nearly all of the programs used in the analyses presented in the examples in this text were carried out using these types of program. These programs are shown in Chapter 27.

26.8 DATA EDITING

Before beginning any analyses, it is very helpful to spend some time editing your data. The most important components of this process are labelling variables and values within variables, formatting variables and correctly coding missing values.

- All variables should have a label attached to them which more fully describes the contents of the variable. While variable names are often quite short (*eg* < 8 or <16 characters), labels can be much longer. **Note** With some computer programs, the labels are stored in a separate file.
- Categorical variables (hopefully they are all numeric) should have meaningful labels attached to each of the categories. For example, sex could be coded as 1 or 2, but should have labels for 'male' and 'female' attached to those values.
- The number that was assigned to all missing values needs to be converted into the code used by your statistics program for missing values.
- Some programs will allow you to attach 'notes' directly to the dataset (or to individual variables within the dataset). These explanatory notes can be invaluable in documenting the contents of files.

26.9 DATA VERIFICATION

Before you start any analyses, you must verify that your data are correct. This can be combined with the following two processes (processing your outcome and predictor variables) because both involve going through all of your variables, one-by-one.

- If you have a very small dataset, you might want to print the entire dataset (make sure it aligns all values for one variable in one column) and review it for obvious errors. However, this is rarely feasible for datasets from epidemiologic studies.
- For continuous variables:
  - determine the number of valid observations and the number of missing values
  - check the maximum and minimum values (or the five smallest and five largest) to make sure they are reasonable (if they are not, find the error, correct it and repeat the process)
  - prepare a histogram of the data to get an idea of the distribution and see if it looks reasonable.
- For categorical variables:
  - determine the number of valid observations and the number of missing values
  - obtain a frequency distribution to see if the counts in each category look reasonable (and to make sure there are no unexpected categories).

26.10 DATA PROCESSING – OUTCOME VARIABLE(S)

While you are going through the data verification process, you can also start the processing of your outcome variable. To do this you will need to review the stated

goals of the study to determine the format(s) of the outcome variable(s) which best suits the goal(s) of the study. For example, you might have conducted a clinical trial of a vaccine for the control of infectious salmon anemia (ISA) in farmed salmon, and have recorded daily mortalities from ISA for the duration of the study. From this single mortality variable, you could compute the mean daily mortality rate, the cumulative mortality over the study, the peak mortality observed, whether or not the sea cage in which the salmon were housed met some set of criteria for having an 'outbreak' of ISA, or the time interval from when the fish were transferred to salt water to the onset of an outbreak. Which you choose to analyse will depend on the goals of the study. Once you have identified the appropriate outcome variable(s), consider the following.

- If the outcome is categorical, is the distribution of outcomes across categories acceptable? For example, you might have planned to carry out a multinomial regression of a three-category outcome, but if there are very few observations in one of the three categories, you might want recode it to a two-category variable.
- If the outcome is continuous, does it have the characteristics necessary to support the analysis planned?
  - If linear regression is planned, is it distributed approximately normally? If not, explore transformations which might normalise the distribution. **Note** It is the normality of the residuals which is ultimately important, but if the original variable is far from normal, and there are no very strong predictors, it is unlikely that the residuals will be normally distributed.
  - If it is a rate (or count) and Poisson regression is planned, are the mean and variance of the distribution approximately equal? If not, consider negative binomial regression or alternative analytic approaches. (As above, the assumption of equality of the mean and variance applies to the residuals, but should be approximately true in the original data, unless there are one or more very strong predictors, if this is to be the case.)
  - If it is time-to-event data, what proportion of the observations are censored? You might also want to generate a simple graph of the empirical hazard function to get an idea what shape it has.

26.11 Data processing – predictor variables

It is important to go through all predictor variables in your dataset to determine how they will be handled. Some issues to consider include the following.

- Are there many missing values? If there are, you might have to abandon plans to use that predictor, or conduct two analyses, one on the subset in which the predictor is present and one on the full dataset (by ignoring the predictor).
- What is the distribution of the predictor?
  - If it is continuous, is there a reasonable representation over the whole range of values? If not, it might be necessary to categorise the variable (see comments about evaluating the relationship between predictors and outcome in section 26.13).
  - If it is categorical, are all categories reasonably well represented? If not, you might have to combine categories.

26.12 Data processing – multilevel data

If your data are multilevel (*eg* lactations within cows within herds), it is necessary to evaluate the hierarchical structure of the data.

- What is the average (and range) number of observations at one level in each higher level unit? For example, what is the mean, minimum and maximum number of lactations per cow in the dataset? Similarly, what are those values for the number of cows per herd?
- Are animals uniquely identified within a hierarchical level? It is often useful to create one unique identifier for each observation in the dataset. This will help identify specific points when evaluating outliers, influential observations *etc.* This can either be done by creating a variable that consists of a combination of the herd and animal identifiers, or simply assigning a unique sequential number to each unit in the dataset.

26.13 Unconditional associations

Before proceeding with any multivariable analyses, it is important to evaluate unconditional associations within the data.

- Associations between pairs of variables can be evaluated using the following techniques.
  - Two continuous variables – correlation coefficient, scatterplot, simple linear regression
  - One continuous and one categorical variable – one-way ANOVA, simple linear or logistic regression
  - Two categorical variables – cross-tabulation and χ^2. Cross-tabulations are particularly useful for identifying unexpected observations (*eg* cases of mastitis in males).
- Associations between predictors and the outcome variable(s) need to be evaluated to:
  - Determine if there is any association at all, as it might be possible to ignore predictors with virtually no association with the outcome at this stage (see Chapter 15)
  - Determine the functional form (*eg* is it linear?) of the relationship between any continuous predictor and an outcome (discussed in Chapters 15 and 16)
  - To get a simple picture of the strength and direction of the association between predictors and outcome, to aid in the interpretation of results of the complex statistical models you will subsequently build.
- Associations between pairs of predictors need to be evaluated to determine if there is a potential for collinearity problems (highly correlated predictors).
- Special attention needs to be paid to potential confounding variables. Evaluate the associations between these variables and the key predictors of interest and the outcome. This will provide some insight into whether or not there is any evidence of confounding in your data (*ie* particularly if there is a strong association with both the key predictor and the outcome).

26.14 KEEPING TRACK OF YOUR ANALYSES

You are now ready to proceed with the more substantial analysis of your data. However, before starting, it is wise to set up a system for keeping track of your results. A few points to keep in mind to facilitate this process are as follows.

- Carry out your analyses in substantial 'blocks'. For example, if computing descriptive statistics, do so for all variables, not just one or two. (Eventually you will need descriptive statistics for all of them, so you might as well keep them together.)
- Most statistical packages allow you to keep a 'log' file which records all of the results from a set of analyses.
  - Give these log files the same name as the program file (except with a different extension)
  - If you are doing some analyses in interactive mode, make sure you keep a complete log file as it will be the only record of what you have done.
- A 3-ring binder (2 or 4 rings in Europe) is a very convenient way to store printouts of all analytical work. Label and date all printouts and describe briefly what each contains on the first page of the printout. This will simplify finding results later.

Following the steps outlined above will not guarantee that you obtain the best possible results from your analyses. However, the process will minimise the number of mistakes and lost time that affect all researchers that are just starting to develop experience with data analysis (and some of us who have been doing it for years). As you gain experience, you might choose to modify some of the items described above as you identify more efficient ways to conduct your analyses.

Good luck!

27

DESCRIPTION OF DATASETS

All datasets used in the examples and sample problems in this text are provided for pedagogical purposes only. They are provided so that the reader can (in conjunction with the program files listed in Chapter 28) recreate the examples included in the text, or compute solutions for sample problems provided. Contributors have made data available to the readers of this text on this understanding and consequently, this is the only use for which they are provided.

In some cases, datasets have been modified since the initial publication of results from the study which generated the data. In many cases, only a subset of the original data (*ie* a subset of variables or a subset of observations) are included. Consequently, the reader should not expect to be able to duplicate results obtained in the original publication.

In the descriptions that follow, unless otherwise specified, all variables coded 0 or 1 (0/1) have the following meaning:

0 = no, absent or negative
1 = yes, present or positive

All datasets can be downloaded from the Veterinary Epidemiologic Research website (http://www.upei.ca/ver). Datasets are directly accessible to Stata using Stata's internet commands (*eg* net from http://www.upei.ca/ver/data – see Stata documentation for details) or as zip files in a variety of statistics program formats (see website for details).

The authors extend their sincere thanks and appreciation to the contributors of these datasets.

Name **ap2**

Study type single cohort
# of records 1,114
Unit of record pig
Contributor Håkan Vigre

Reference

Vigre H, Dohoo IR, Stryhn H, Busch ME. Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumonia* and *Mycoplasma hyopneumoniae* at different levels in Danish multi-site pig production facilities. Prev Vet Med 2003; accepted.

Brief description

Data were collected on 1,114 pigs from 35 batches produced on six farms that employed an 'all-in, all-out' production process. Pigs were weighed and blood sampled at the time of transfer from the weaner barn to the finisher barn (approximately 70 days of age) and again six weeks later (shortly before slaughter). Blood samples were tested for antibodies to *Actinobacillus pleuropneumonia* (Type 2), *Mycoplasma hyopnueumonia*, the influenza virus and the porcine respiratory and reproductive syndrome virus (PRRS). Two of the objectives of the study were to determine when seroconversion to the various agents occurred and at which level of the population (*eg* pig, batch or herd) most of the variation in seroconversion occurred.

(continued on next page)

Table of variables

Variable	Description	Codes/units
farm_id	farm identification	
batch_id	batch identification number	
litt_id	litter identification number	
pig_id	pig identification	
parity	farrowing number of sow	
vacc_mp	batch vaccinated against M. hyopneumoniae	0/1
seas_fin	season pigs in finishing unit	0 = summer 1 = winter
age_t	pig age at transfer from weaning to finishing unit	days
w_age_t	weight at age_t	kg
age_t6	age plus approx. 6 weeks	days
w_age_t6	weight at age_t6	kg
dwg_fin	daily weight gain between age_t and age_t6	gm
ap2_t	serological reac. against A. pleuropneumoniae serotype 2 at age_t	0/1
mp_t	serological reac. against M. hyopneumoniae at age_t	0/1
infl_t	serological reac. against influenza virus at age_t	0/1
prrs_t	serological reac. against PRRS virus at age_t	0/1
ap2_t6	serological reac. against A. pleuropneumoniae serotype 2 at age_t6	0/1
mp_t6	serological reac. against M. hyopneumoniae at age_t6	0/1
infl_t6	serological reac. against influenza virus at age_t6	0/1
prrs_t6	serological reac. against PRRS virus at age_t6	0/1
ap2_sc	seroconversion to ap2 during the finishing period	0/1

Name **beef_ultra**

Study type single cohort
# of records 487
Unit of record animal
Contributor Greg Keefe

Reference
Keefe GP, Dohoo IR, Valcour J, Milton RL. Assessment of ultrasonic imaging of marbling at entry into the feedlot as a predictor of carcass traits at slaughter. J Anim Sci 2003; submitted.

Brief description
Data were collected on 487 cattle at the time that they entered a feedlot for 'fattening' prior to slaughter. Data consisted of demographic information plus readings obtained from an ultrasonic evaluation of the animal. Ultrasound measurements of backfat thickness, loineye area and the percentage of intramuscular fat ('marbling') were obtained. The objective of the study was to determine if ultrasound examination of the animal at the time of entry into a feedlot was able to predict final carcass grade (AAA, AA or A). Carcass grade depends primarily on the amount of intramuscular fat in the carcass at the time of slaughter.

Table of variables

Variable	Description	Codes/units
farm	farm id	
id	animal id	
grade	carcass grade	1 = AAA 2 = AA 3 = A
breed	breed (known or estimated)	multiple
sex	gender	0 = female 1 = male
bckgrnd	animal backgrounded	0/1
implant	hormone implant used	0/1
backfat	backfat thickness	mm
ribeye	area of rib eye muscle	sq cm
imfat	intramuscular fat score	% of area
days	fattening period	days
carc_wt	carcass weight	kg

Name **bst_mast**

Study type meta-analysis
# of records 29
Unit of record group of cows
Contributor Ian Dohoo

Reference
Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Shewfelt W, Preston A et al. A meta-analysis review of the effects of rBST. 2. Effects on animal health, reproductive performance and culling. Can J Vet Res 2003; in press.

Brief description
On request from Health Canada, the Canadian Veterinary Medical Association established an expert panel to review the production and health effect of recombinant bovine somatotropin (rBST) in dairy cattle. The panel carried out a meta-analysis of all available literature and evaluated a wide range of production and health effects. The data in this file consist of risk ratios for clinical mastitis that were associated with the use of rBST. Data from 29 distinct groups of cows, from 20 separate studies are included. The precision of the point estimate is included in the form of 95% confidence limits.

Table of variables

Variable	Description	Codes/units
study	study number	
group	cow group number	
parity	parity group	1 = primiparous 2 = all ages combined 3 = multiparous
study_yr	year of study	
rr	risk ratio	
cilow	lower 95% confidence limit	
cihigh	upper 95% confidence limit	
dur	duration of treatment	days
dose_day	daily dosage	mg/day

Name **bst_milk**

Study type meta-analysis
# of records 28
Unit of record group of cows
Contributor Ian Dohoo

Reference
Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Dowling P, Preston A et al. A meta-analysis review of the effects of rBST. 1. Methodology and effects on production and nutrition related parameters. Can J Vet Res 2003; in press.

Brief description
On request from Health Canada, the Canadian Veterinary Medical Association established an expert panel to review the production and health effect of recombinant bovine somatotropin (rBST) in dairy cattle. The panel carried out a meta-analysis of all available literature and evaluated a wide range of production and health effects. The data in this file consist of change in level of milk production (fat-corrected milk) that were associated with the use of rBST. Data from 28 distinct groups of cows, from 19 separate studies are included. The precision of the point estimate is included both in the form of 95% confidence limits and the SE of the point estimate.

Table of variables

Variable	Description	Codes/units
study	study number	
group	cow group number	
parity	parity group	1 = primiparous 2 = all ages combined 3 = multiparous
study_yr	year of study	
diff	difference in milk production	
cilow	lower 95% confidence limit for difference	
cihigh	upper 95% confidence limit for difference	
se	standard error of difference	
ncows	number of cows in study	
dur	duration of treatment	days
dose_day	daily dosage	mg/day

Name **bvd_test**

Study type single cohort
# of records 2,162
Unit of record cow
Contributor Ann Lindberg

Reference
Lindberg A, Groenendaal H, Alenius S, Emanuelson U. Validation of a test for dams carrying foetuses persistently infected with bovine viral-diarrhoea virus based on determination of antibody levels in late pregnancy. Prev Vet Med 2003; 51: 199-214.

Brief description
Blood or milk samples were collected from 2,162 pregnant cows at various stages of lactation. Following the birth of their calf, the status of the calf with regard to persistent infection (PI) with the bovine viral diarrhea (BVD) virus was determined. The blood and milk samples were tested using an ELISA to determine levels of BVD virus antibodies in the cow. A variety of cutpoints were then examined to determine which gave the best combination of sensitivity and specificity for detecting PI+ calves. Logistic regression was used to evaluate the effects of other factors (particularly stage of lactation) on the estimated sensitivity and specificity of the test.

(continued on next page)

Table of variables

Variable	Description	Codes/units
cow_id	cow identification	
breed	breed	1 = red and white 2 = black and white 3 = beef 4 = other
parity	parity group	1 = primiparous 2 = all ages combined 3 = multiparous
pregmon	pregnancy month at test	
season	calving season	1 = winter 2 = spring 3 = summer 4 = autumn
spec	type of specimen	0 = milk 1 = blood
calfst	calf status	0/1
od	optical density	
co_5	test result dichotomised at 0.5	0/1
co_6	test result dichotomised at 0.6	0/1
	. . . etc . . .	
co_15	test result dichotomised at 1.5	0/1
co_16	test result dichotomised at 1.6	0/1

Name **calf**

Study type retrospective cohort
# of records 254
Unit of record calf
Contributor Jeanne Lofstedt

Reference
Lofstedt J, Dohoo IR, Duizer G. Model to predict septicemia in diarrheic calves. J Vet Int Med 1999; 13: 81-88.

Brief description
These data come from a retrospective analysis of the medical records from all diarrheic calves which were presented to Atlantic Veterinary College, PEI, Canada between 1989 and 1993. The ultimate objective of the study was to develop a logistic model which would predict whether or not the calf was septic at the time of admission (septic calves have a much poorer prognosis than non-septic calves and are not usually worth treating, given economic considerations).

There are 254 observations (records) and 14 variables in the dataset (calf). The original dataset had far more variables (including a lot of laboratory data) but this dataset contains only a subset of the demographic data and the physical examination data collected. All observations were determined on the day of admission, except for the outcome (sepsis) which was based on all data available at the time of death or discharge.

(continued on next page)

Table of variables

Variable	Description	Codes/units
case	hospital case number	
age	age at admission	days
breed	breed	coded 1-9
sex	sex	0 = female 1 = male
attd	attitude of calf	0 = bright, alert 1 = depressed 2 = unresponsive, comatose
dehy	% dehydration	
eye	uveitis/hypopyon clinically evident	0/1
jnts	swollen joints clinically evident	number of joints affected
post	posture of calf	0 = standing 1 = sternal 2 = lateral
pulse	pulse rate	beats per minute
resp	respiratory rate	breaths per minute
temp	rectal temperature	°C
umb	swollen umbilicus clinically evident	0/1
sepsis	sepsis (outcome)	0/1

Name **calf_pneu**

Study type cohort
# of records 24
Unit of record calf
Contributor Iver Thysen

Reference
Thysen I. Application of event time analysis to replacement, health and reproduction data in dairy cattle research. Prev Vet Med 1988; 5: 239-250.

Brief description
These published data were used in one of the early publications in the veterinary literature discussing the use of survival analysis techniques. The data consist of mortality records from 24 calves that were housed in one of two housing systems: continuous housing, or batch (*ie* all-in all-out) housing.

Table of variables

Variable	Description	Codes/units
calf	calf id	
stock	stocking method	
days	time to death or censoring	days
died	died	

Name **colostrum**

Study type single cohort
# of records 180
Unit of record calf
Contributors Gilles Fecteau

Reference
None

Brief description
Data on the colostrum fed to 180 calves were collected from several dairy herds in Quebec. Herd identification was recoded to be the single large herd in the study compared to an amalgamation of small herds. The bacterial load in the colostrum was determined and the quantity of colostrum fed to the calf recorded. Calves were followed for three weeks and their health status over that period recorded as healthy (no illness), mild illness or serious illness (including death). The objective of the study was to determine if bacterial load in colostrum affected calf health.

Table of variables

Variable	Description	Codes/units
herd	herd of origin	0 = collection of small herds 1 = large herd
calf_id	calf identification	
health	health score	0 = healthy 1 = mild illness 2 = serious illness
qty	quantity of colostrum	litres
log_tot	natural log of total bacterial load	

Name **dairy_dis**

Study type survey (cross-sectional)
# of records 2454
Unit of record cow
Contributors John VanLeeuwen, Greg Keefe

Reference

VanLeeuwen J, Keefe GP, Tremblay R, Power C, Wichtel JJ. Seroprevalence of infection with *Mycobacterium avium* subspecies *paratuberculosis*, bovine leukemia virus, and bovine viral diarrhea virus in Maritime Canada dairy cattle. Can Vet J Res 2001; 42: 193-198.

Brief description

These data were collected as part of a prevalence survey of four infectious diseases of dairy cattle in eastern Canada. 30 herds in each of three provinces (Prince Edward Island, Nova Scotia and New Brunswick) were randomly selected from lists of all dairy herds participating on a milk-production monitoring program. Within each herd, approximately 30 animals were randomly selected and blood samples collected. These samples were tested for antibodies to: *Neospora caninum*, *Mycobacterium avium (*subsp. *paratuberculosis)* and enzootic bovine leukosis virus. In addition, a group of non-vaccinated heifers were bled and tested for bovine virus diarrhea virus, but these test results are not included in this dataset. Sampling weights were computed as the inverse of the product of the probability of a herd being selected and the probability of a cow being selected within a herd.

(*continued on next page*)

Table of variables

Variable	Description	Codes/units
prov	province	
herd	herd identification number	
cow	cow identification number	
lact	lactation number	
dim	days in milk	days
johnes	Johne's test result	0/1
leukosis	leukosis test result	0/1
neospora	neospora test results	0/1
tot_hrd	total herds in province	
prob_hrd	probability of herd being selected	
tot_cow	total cows in herd	
tot_smpl	total cows sampled in herd	
prob_cow	probability of cow being selected	
prob_smp	overall probability of a cow being selected	
weight	sampling weight	

Name **daisy**

Study type single cohort
# of records 162
Unit of record cow
Contributor Wayne Martin

Reference
None

Brief description
These data are based on real cow-reproduction data but have been modified in order to demonstrate a number of points related to linear regression. Consequently, they are now fictional data. They consist of data about the occurrence of a number of disease conditions which occur in the early post-partum period, along with measures of reproductive performance such as the interval from calving to first estrus, interval to first breeding and the calving to conception interval (all measured in days). The objective of the studies based on these data is to evaluate the effect of various diseases on reproductive performance.

Table of variables

Variable	Description	Codes/units
farmnum	farm identifier	
cownum	cow identifier	
firstest	first observed estrus	
firstbrd	postpartum to first breeding	days
calvcon	postpartum to conception	days
age	age	yrs
culled	cow removed from the herd	0/1
dayscull	postpartum to cow removal	days
endomet	endometritis	0/1
mastitis	mastitis	0/1
metritis	metritis	0/1
milkfev	milk fever	0/1
ovar	cystic ovarian disease	0/1
pyomet	pyometritis	0/1
retpla	retained fetal membranes (placenta)	0/1

Name **elisa_repeat**

Study type experimental
# of records 40
Unit of record milk sample
Contributor Javier Sanchez

Reference
Sanchez J, Dohoo IR, Markham RJF, Leslie KE, Conboy G. Evaluation of the repeatability of a crude adult indirect *Ostertagia ostertagi* ELISA and methods of expressing test results. Vet Parasitol 2002; 109: 75-90.

Brief description
Forty individual cow milk samples were repeatedly tested (six times) using a crude *Ostertagia* antigen indirect ELISA. Results were recorded both as raw optical density (OD) values and values adjusted based on the readings for the positive and negative controls in the plate.

Table of variables

Variable	Description	Codes/units
id	sample identification	
raw1	raw OD - sample #1	
raw2	raw OD - sample #2	
	. . .etc . . .	
raw6	raw OD - sample #6	
adj1	adjusted OD - sample #1	
adj2	adjusted OD - sample #2	
	. . .etc . . .	
adj6	adjusted OD - sample #6	

Name **fec**

Study type single cohort
# of records 2,250
Unit of record monthly fecal egg count
Contributors Ane Nødtvedt, Javier Sanchez, Ian Dohoo

Reference

Nødtvedt A, Dohoo IR, Sanchez J, Conboy G, DesCôteaux L, Keefe GP et al. The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. Can J Vet Res 2002; 66: 249-257.

Brief description

Monthly (in some herds less frequently) fecal egg samples were collected from lactating age dairy cows (n=313) in 38 herds over a period of 1 year. The data were collected as part of a multifaceted study into parasitism in lactating dairy cows which included a longitudinal epidemiologic investigation and a controlled trial of the effects of deworming at calving with eprinomectin. The effects of factors at the sampling-day, cow and herd levels on fecal egg counts were evaluated.

Table of variables

Variable	Description	Codes/units
province	Canadian province	1 = PEI 2 = Quebec 3 = Ontario 4 = Saskatchewan
herd	herd identifier	
cow	unique cow identifier	
visit	visit number	
tx	eprinomectin treatment at calving	0/1
fec	fecal egg count	eggs/5 gm
lact	lactation	0 = primiparous 1 = multiparous
season	season	1 = oct-dec 99 2 = jan-mar 00 3 = apr-jun 00 4 = jul-sep 00
past_lact	lactating cows have access to pasture	0/1
man_heif	manure spread on heifer pasture	0/1
man_lact	manure spread on cow pasture	0/1

Name **feedlot**

Study type case-control
# of records 588
Unit of record animal
Contributor Wayne Martin

References

1. Martin, SW, Harland, RJ, Bateman, KG, Nagy, É. The association of titres to *Haemophilus somnus* and other putative pathogens, with the occurrence of bovine respiratory disease and weight gain in feedlot calves. Can J Vet Res 1998; 62: 262-267.
2. Martin, SW, Nagy, É, Shewen, PE, Harland, RJ. The association of titres to bovine coronavirus with treatment for bovine respiratory disease and weight gain in feedlot calves. Can J Vet Res 1998; 62: 257-261.

Brief description

This dataset represents the combined data from a number of studies on the role of specific micro-organisms as causes of bovine respiratory disease (BRD). Typically these beef cattle enter feedlots in the fall of the year and approximately 30% will develop BRD. The general strategy for the studies was to bleed all of the animals on arrival at the feedlot and again 28 days later (since most of the occurrence of BRD occurs in that time period). For analyses, we used all of the samples from cases and an approximately equal number from controls. In some of the smaller groups we used all samples and hence in these the study design was essentially a single cohort. The studies were conducted at essentially the same feedlots in different years, but depending on their size, one feedlot could have numerous groups of calves on the study in any given year.

The titres were recorded in a quantitative manner but have been dichotomised in this dataset.

Note At the time these data were collected, one of the important bacteria was called Pasteurella hemolytica and it is referred to as such in this dataset. Elsewhere in the text it is referred to by its newer name Mannheima hemolytica.

(continued on next page)

Table of variables

Variable	Description	Codes/units
group	group identification	
tag	eartag number	
province	province of feedlot	1 = Alberta 2 = Ontario
brd	clinical bovine respiratory disease (case-control)	0/1
brsvpos	arrival titre to brsv	0/1
brsvsc	seroconversion to brsv during study	0/1
bvdpos	arrival titre to bvd virus	0/1
bvdsc	seroconversion to bvd during study	0/1
ibrpos	arrival titre to ibr virus	0/1
ibrsc	seroconversion to ibr virus during study	0/1
pipos	arrival titre to pi3 virus	0/1
pisc	seroconversion to pi3 virus during study	0/1
phcypos	arrival titre to Ph cytotoxin	0/1
phcysc	seroconversion to Ph cytotoxin during study	0/1
phaggpos	arrival titre to Ph agglutinins	0/1
phaggsc	seroconversion to Ph during study	0/1
hspos	arrival titre to Hs	0/1
hssc	seroconversion to Hs during study	0/1
wt0	arrival weight	kg
wt28	28-day weight	kg

Name **fish_morts**

Study type cross-sectional
# of records 236
Unit of record sea-cage
Contributors Larry Hammell, Ian Dohoo

References

1. Hammell KL, Dohoo IR The epidemiology of hemorrhagic kidney syndrome - infectious salmon anemia in Atlantic salmon in Atlantic Canada. Bristol, England: Society for Veterinary Preventive Epidemiology and Preventive Medicine, 1999.
2. Hammell KL, Dohoo IR Challenges of investigating mortality patterns and management factors associated with ISAV outbreaks in eastern Canada. Paris, France: O.I.E., Risk Analysis in Aquatic Animal Health, 2000.
3. Hammell KL, Dohoo IR. Mortality patterns in Infectious Salmon Anemia virus outbreaks in New Brunswick, Canada. Journal of Fish Diseases 2003; accepted.

Brief description

Following the introduction of infectious salmon anemia virus to the Bay of Fundy (Canada), an epidemiological investigation of risk factors for the disease was initiated. At the time the study was started, the etiology of the mortalities was not known and cages were designated as 'outbreaks' or not, based on the pattern of mortalities observed in the cage. One of the risk factors identified as being associated with an increased risk of elevated mortalities was the feeding of dry (as opposed to wet- or moist-) feed. These data are a very small subset of the original data collected. They describe mortalities in 236 cages from 16 different sites over a period of just a few days (depending on the number of days between dives for collection of dead fish).

Table of variables

Variable	Description	Codes/units
cage_id	cage id	
days	time since last dive	days
morts	number of dead fish retrieved	
fish	estimated number of fish in cage	
feed	dry feed (compared to wet)	0 = wet 1 = dry

Name **isa_risk**

Study type cross-sectional
# of records 182
Unit of record sea cage
Contributors Larry Hammell, Ian Dohoo

References

1. Hammell KL, Dohoo IR The epidemiology of hemorrhagic kidney syndrome - infectious salmon anemia in Atlantic salmon in Atlantic Canada. Bristol, England: Society for Veterinary Preventive Epidemiology and Preventive Medicine, 1999.
2. Hammell KL, Dohoo IR Challenges of investigating mortality patterns and management factors associated with ISAV outbreaks in eastern Canada. France: O.I.E., Risk Analysis in Aquatic Animal Health, 2000.
3. Hammell KL, Dohoo IR. Mortality patterns in infectious salmon anemia virus outbreaks in New Brunswick, Canada. Journal of Fish Diseases 2003; accepted.

Brief description

Following the introduction of infectious salmon anemia virus to the Bay of Fundy (Canada), an epidemiological investigation of risk factors for the disease was initiated. At the time the study was started, the etiology of the mortalities was not known and cages were designated as 'outbreaks' or not, based on the pattern of mortalities observed in the cage. A large number of risk factors were evaluated and this dataset consists of the records for 182 cages which had complete data on a subset of those factors (see list below). While the factors listed below were all fixed factors (*ie* didn't change during the study period), the data were used to compute a time-varying factor: whether or not there had been another positive cage (net-pen) at the site. This was used in survival models of the time to the occurrence of an outbreak.

Table of variables

Variable	Description	Codes/units
sitepen	(1000*site)+cage identifier	
site	site identifier	
net-pen	cage identifier	
datestrt	date fish first put in cage	
apr01_97	April 1 1997	
date	date of outbreak OR censoring	
case	case (outbreak)	0/1
cummrt96	cum. mort. during 1996	
size	cage size	0 = <10,000 1 = >10,000
par	initial population at risk in cage (number of fish)	
numcage	number of cages at site	

Name **isa_test**

Study type cross-sectional
# of records 1,071
Unit of record fish
Contributors Carol McClure, Larry Hammell

Reference
McClure C, Hammell KL, Stryhn H, Dohoo IR, Hawkins LJ. Application of surveillance data in the evaluation of infectious salmon anemia diagnostic tests. Dis of Aquatic Org 2003; submitted.

Brief description
Following the identification of the infectious salmon anemia virus in the Bay of Fundy (Canada), a lot of fish were tested using a variety of diagnostic tests. It was realised that tests often gave conflicting results and the available data were used to provide a preliminary evaluation of the operating characteristics of each test. Fish that were derived from sea cages (net-pens) that had a confirmed outbreak of ISA were considered to be 'gold standard positive.' Fish sampled from sites which did not have any outbreaks of ISA (in any cages) during the study period were considered 'gold standard negative.' Other fish sampled were not included in this study. Test results from a total of 1071 fish that had multiple tests performed and which could be classified as positive (n=264) or negative (n=807) were included in the dataset.

Table of variables

Variable	Description	Codes/units
id	case identification	
date	submission date	
site	site identification	
cage	cage identification	
subm	submission identification	
fish	fish number for each case	
dz	disease status (clinical)	
histo	histology	0 = negative 1 = suspicious 2 = positive
histo_np	histo neg/pos (pos=susp+pos)	0/1
ifat1	IFAT laboratory 1	0-4
ifat1_np	IFAT-lab1 neg/pos (pos is ≥ 1)	0/1
ifat2	IFAT - laboratory 2	0-4
ifat2_np	IFAT-lab2 neg/pos (pos is ≥ 2)	0/1
pcr	PCR	0/1
vi	virus isolation	0/1

Name **lympho**

Study type clinical trial (fictional)
# of records 300
Unit of record dog
Contributor Ian Dohoo

Reference
None

Brief description
These data are from a fictional clinical trial of two treatments for lymphosarcoma in dogs. The study was (hypothetically) conducted as a multicentre (n=7 clinics) controlled trial. Dogs meeting the eligibility criteria for entry into the trial (n=300) had the tumour surgically removed (only dogs with tumours which could be surgically removed were eligible) and then were randomly assigned to one of four treatment groups: no treatment, radiation only, chemotherapy only and both radiation and chemotherapy. Dogs were randomly assigned within each centre, so the total number of dogs on each treatment group are not exactly equal for all treatments. Each dog was followed from the time of treatment until it died from a relapse of the lymphosarcoma or was lost to follow-up (*eg* died of other causes, owner moved away from the study site) and the time to the occurrence of either of those was recorded.

Table of variables

Variable	Description	Codes/units
dogid	the dog's study identification number	
age	age of dog in years when it was diagnosed with lymphosarcoma	yrs
rad	whether or not the dog received radiation therapy	0/1
chemo	whether or not the dog received chemotherapy	0/1
died	whether the dog died or was lost to follow-up	0/1
months	the number of months after the start of therapy before the dog died or was lost to follow-up	mo

Name **nocardia**

Study type case-control
# of records 108
Unit of record herd
Contributors Lynn Ferns, Ian Dohoo

Reference
Ferns L, Dohoo IR, Donald A. A case-control study of Nocardia mastitis in Nova Scotia dairy herds. Can Vet J Res 1991; 32: 673-677.

Brief description
This dataset contains a subset of the data obtained from a case-control study of Nova Scotia dairy herds with and without Nocardia mastitis. There had been a dramatic increase in the incidence of Nocardia mastitis in Canada since 1987 and this study was carried out to identify risk factors associated with the occurrence this disease. A total of 54 case herds and 54 control herds were visited for data-collection purposes during the summer of 1989.

(continued on next page)

Table of variables

Variable	Description	Codes/units
id	herd identification number	
casecont	case/control status of herd	0 = control 1 = case
numcow	number of cows milked	
prod	average milk production for the herd	kg/cow/day
bscc	average bulk-tank SCC over the first 6 months of 1988	'000s of cells/ml
dbarn	type of barn dry cows kept in	1 = freestall 2 = tiestall 3 = other
dout	type of outdoor area used for dry cows	1 = pasture 2 = yard/drylot 3 = none 4 = other
dcprep	method of teat end preparation prior to dry cow therapy administration	1 = no prep. 2 = washed only 3 = washed and disinfected 4 = dry cow therapy not used
dcpct	percent of dry cows treated with dry-cow therapy	%
dneo	dry-cow product containing neomycin used on farm in last year	0/1
dclox	dry cow product containing cloxacillin used on farm in last year	0/1
doth	Other dry cow products used (eg penicillin or novobiocin based) used on farm in last year	0/1

Name **pgtrial**

Study type clinical trial
# of records 319
Unit of record cow
Contributor Jeff Wichtel

Reference
None

Brief description
A clinical trial of the effect of prostaglandin administration at the start of the breeding period was carried out in three North Carolina dairy herds. On each of the three farms, the producer determined when he was ready to start breeding cows in his herd and at that time, cows were randomly assigned to receive a single injection of prostaglandin or a placebo. These cows were then followed (up to a maximum of 346 days) until they conceived (confirmed by rectal examination) or were culled. In addition to evaluating the effect of treatment on reproductive performance, three other factors were considered (parity, body condition score and herd).

Table of variables

Variable	Description	Codes/units
herd	herd identification number	
cow	cow identification number	
tx	treatment	0/1
lact	lactation number	
thin	body condition	0 = normal 1 = thin
dar	days at risk	days
preg	pregnant or censored	0 = censored 1 = pregnant

Name **pig_adg**

Study type cross-sectional
# of records 341
Unit of record pig
Contributor Theresa Bernardo

References

1. Bernardo TM, Dohoo IR, Donald A, Ogilvie T, Cawthorne R. Ascariasis, respiratory disease and production indices in selected PEI swine herds. Can J Vet Res 1990; 54: 267-273.
2. Bernardo TM, Dohoo IR, Ogilvie T. A critical assessment of abattoir surveillance as a screening test for swine ascariasis. Can J Vet Res 1990; 54: 274-277.
3. Bernardo TM, Dohoo IR, Donald. Effect of ascariasis and respiratory disease on growth rates in swine. Can J Vet Res 1990; 54: 278-284.

Brief description

These are data on the growth performance and abattoir findings of pigs from a selection of Prince Edward Island, Canada farms. The data were collected to study the inter-relationships among respiratory diseases (atrophic rhinitis and enzootic pneumonia), ascarid levels and daily weight gain. Atrophic rhinitis score was determined by splitting the snout and measuring the space ventral to the turbinates. An adjustment to the score was made if the nasal septum was deviated. Lung scores were recorded on a scale of 0 to 3 (negative to severe pneumonia) and then converted to either the presence or absence of pneumonia. Parasite burdens were evaluated using fecal egg counts, counts of adult worms in the intestine and visual assessment of the liver for ascarid tracks. Production data were recorded by monitoring the pigs on the farms of origin from birth through to slaughter.

(*continued on next page*)

Table of variables

Variable	Description	Codes/units
farm	farm identification number	
pig	pig identification number	
sex	sex of the pig	0 = female 1 = castrate
dtm	days to market (ie from birth to slaughter)	days
adg	average daily weight gain	gm
mm	measurement of snout space	mm
ar	atrophic rhinitis score	0-5
lu	lung score for enzootic pneumonia	0 = negative 1 = mild 2 = moderate 3 = severe
pn	pneumonia (lu>0)	0/1
epg5	fecal gastrointestinal nematode egg count at time of slaughter	eggs/5 gm
worms	count of nematodes in small intestine at time of slaughter	
li	liver score (based on number of parasite induced 'white spots')	0 = negative 1 = mild 2 = severe
ar2	severe atrophic rhinitis (ar>4)	0/1

Name **pig_farm**

Study type cross-sectional
# of records 69
Unit of record farm
Contributor Dan Hurnik

Reference

1. Hurnik D, Dohoo IR, Donald AW, Robinson NP. Factor analysis of swine farm management practices on Prince Edward Island. Prev Vet Med 1994; 20: 135-146
2. Hurnik D, Dohoo IR, Bate LA. Types of farm management as risk factors for swine respiratory disease. Prev Vet Med 1994; 20: 147-157.

Brief description

A cross-sectional study of pig farms in Prince Edward Island (Canada) was carried out to investigate risk factors for respiratory diseases (enzootic pneumonia and pleuritis). The prevalence of each disease was determined at slaughter from routine evaluations of thoracic viscera. Data on risk factors were collected by the investigator during visits to each farm. Data on a wide variety of factors were collected and the challenge was to sort out relationships among these factors and between them and the respiratory diseases given a very limited sample size.

(continued on next page)

Table of variables

Variable	Description	Codes/units
farm_id	farm identification	
pneu	pneumonia prevalence	
pncode	pneumonia - categorical (3 levels)	0 < 10% 1 = 10-40% 2 > 40%
pleur	pleuritis prevalence	
plcode	pleuritis - categorical (3 levels)	0 = 0% 1 = 0-8% 2 > 8%
num	number of pigs examined at slaughter	
size	herd size	
growth	average daily gain	gm/day
cmpfd	pigs fed complete mixed feed	0/1
suppl	supplement added to feed	0/1
prmx	premix fed	0/1
strmed	starter ration medicated	0/1
selenium	selenium added to feed	0/1
dryfd	feed fed dry (vs wet)	0 = wet 1 = dry
flrfd	pigs fed on floor	0/1
rooms	number of separate rooms in barn	
m3pig	air volume per pig	m^3
shipm2	density (pigs shipped per m^2)	pigs/m^2
exhaust	exhaust fan capacity (proportion of recommendation)	
inlet	air inlet size (proportion of recommendation)	
maninlt	manual adjustment of air inlets	0/1
mixmnr	manure mixed between pens	0/1
straw	straw bedding used	0/1
washpns	frequency of pen washings (per yr)	
strdnst	floor space - starter hogs (sq m)	m^2
grwdnst	floor space - grower hogs (sq m)	m^2
fnrdnst	floor space - finishing hogs (sq m)	m^2

(*continued on next page*)

Table of variables

Variable	Description	Codes/units
lqdmnr	manure handled as a liquid	0/1
floor	floor slatted	0/1
sldprtn	solid partitions between some pens	0/1
hlfsld	half-solid partitions between some pens	0/1
pigwtr	pigs per water nipple	
numpen	number of pens	
mixgrp	pigs from multiple groups mixed	0/1
hldbck	slow growing pigs held back from slaughter	0/1
dstfrm	distance (km) to nearest hog farm	km
hmrsd	all pigs home raised	0/1
nmbsrc	number of sources of pigs	
mnlds	only minimal disease pigs raised	0/1
vet	veterinary visits per year	
feedsls	feed salesman visits per year	
neighbr	neighbour visits per year	
pigprdc	pig producer visits per year	
trucker	trucker visits per year	
you	owner works in barn	0/1
family	family members work in barn	0/1
hrdhlp	hired help works in barn	0/1
exprnce	years of experience	yrs

Name **prew_mort**

Study type cross-sectional
# of records 6552
Unit of record litter
Contributor Jette Christensen

Reference
Christensen J, Svensmark B. Evaluation of producer-recorded causes of preweaning mortality in Danish sow herds. Prev Vet Med 1997; 32: 155-164.

Brief description
These data are a subset of 16 herds from a dataset collected by Jette Christensen in Denmark to study factors affecting preweaning mortality in pigs. These data have three levels in the hierarchy (litters (n=6552) within sows (n=3162) within farms (n=16)):

The key outcome of interest is preweaning mortality with a litter classified as having preweaning mortality or not if one or more piglets died before weaning.

Table of variables

Variable	Description	Codes/units
herd	unique herd id	
sowid	unique sow id	
litter	unique litter id	
lmort	prewmort in litter	0/1
herdtype	herd type	0 = production I = breeding herd
year		
month	month	jan = I dec = 12
quarter	quarter of year	I = jan-mar 2 = apr-jun 3 = jul-sept 4 = oct-dec
sow_parity	parity of sow	
sow_tx	sow required treatment (2d before to 7d after farrowing)	0/1
dead	number of dead piglets in litter	
lsize	litter size	
n	number at risk in litter	
stillb	number stillborn	

Name **reu_cc**

Study type single cohort
# of records 2509
Unit of record lactation
Contributors Emmanuel Tillard, Ian Dohoo

Reference
Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. Prev Vet Med 2001; 50: 127-144.

Brief description
These data were collected as part of an ongoing research programme into dairy cattle fertility being carried out on Reunion Island (a French overseas department located in the Indian ocean) by researchers with CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement). Two separate datasets have been compiled. This one contains data about the calving to conception interval, while the second had data on the interval from calving to first service. The data have a 4-level hierarchy (lactations (n=2509) within cows (n=1345) within herds (n=50) within geographic regions (n=5)).

Table of variables

Variable	Description	Codes/units
region	geographic region	
herd	herd number	
cow	unique cow number	
obs	unique observation number	
lact	lactation number	
cc	calving to conception interval	days
lncc	calving to conception interval - log transformed	
lncfs_ct	calving to first service interval - log transformed and centred	
heifer	age	0 = multiparous 1 = primiparous
ai	type of insemination at first service	0 = natural 1 = ai

Name **reu_cfs**

Study type single cohort
# of records 3027
Unit of record lactation
Contributors Emmanuel Tillard, Ian Dohoo

Reference
Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. Prev Vet Med 2001; 50: 127-144.

Brief description
These data were collected as part of an ongoing research program into dairy cattle fertility being carried out on Reunion Island (a French overseas department located in the Indian ocean) by researchers with CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement). Two separate datasets have been compiled. This one contains data on the interval from calving to first service. The data have a 4-level hierarchy lactations (n=3027) within cows (n=1575) within herds (n=50) within geographic regions (n=5)].

A second dataset containing only the first recorded lactation within each cow was saved as reu_cfs_1lact.

Table of variables

Variable	Description	Codes/units
region	geographic region	
herd	herd number	
cow	unique cow number	
obs	unique observation number	
lact	lactation number	
cfs	calving to first service interval	days
lncfs	calving to first service interval - log transformed	
fscr	first service conception	0/1
heifer	age	0 = multiparous 1 = primiparous
ai	type of insemination at first service	0 = natural 1 = ai

Name **sal_outbrk**

Study type matched case-control
# of records 112
Unit of record individual (person)
Contributor Tine Hald

References

1. Molbak K, Hald DT. An outbreak of Salmonella typhimurium in the county of Funen during late summer. A case-controlled study. Ugeskr Laeger 1997; 159(36): 5372-5377.
2. Hald DT. *Salmonella* in pork: Epidemiology, control and the public health impact. Copenhagen: Royal Veterinary & Agric. Univ. 2001.

Brief description

The data are from an investigation of an outbreak of Salmonella in Funen County of Denmark in 1996. The data consisted of 39 cases of Salmonella typhimurium phage type 12 and 73 controls matched for age, sex and municipality of residence. Data on numerous food exposures were recorded and a small subset of those data are included in the dataset -sal_outbrk-.

(*continued on next page*)

Table of variables

Variable	Description	Codes/units
match-grp	case-control pair identifier	
date	interview date	
age	age	yrs
gender	gender	0 = male 1 = female
casecontrol	case-control status	0/1
eatbeef	ate beef in previous 72 hours	0/1
eatpork	ate pork in previous 72 hours	0/1
eatveal	ate veal in previous 72 hours	0/1
eatlamb	ate lamb in previous 72 hours	0/1
eatpoul	ate poultry in previous 72 hours	0/1
eatcold	ate cold sliced meats in previous 72 hours	0/1
eatveg	ate vegetables in previous 72 hours	0/1
eatfruit	ate fruit in previous 72 hours	0/1
eateggs	ate eggs in previous 72 hours	0/1
slt_a	ate pork processed at slaughterhouse A	0/1
dlr_a	ate pork marketed by wholesaler A	0/1
dlr_b	ate pork marketed by wholesaler B	0/1

Name **scc_40**

Study type longitudinal
# of records 14,357
Unit of record test-day observations
Contributors Jens Agger and Danish Cattle Organization, Paul Bartlett, Henrik Stryhn

Reference

Stryhn H, Andersen JS, Bartlett PC, Agger JFA. Milk production in cows studies by linear mixed models. Proc. of symposium in applied statistics, Copenhagen, January 2001. Proceedings (ed. Jensen NE. Linde P): 1-10.

Brief description

These data are a very small subset of a large mastitis dataset collected by Jens Agger and the Danish Cattle Organization. This dataset contains records from 14,357 test-day observations in 2,178 cows from 40 herds. Milk weights (production records) were collected approximately monthly, and only records from a single lactation for each cow were included in this dataset. Factors that may have affected the somatic cell count (SCC) were also recorded. The major objective of this study was to determine if the relationship between the somatic cell count and milk production varies for cows with different characteristics (age, breed, grazing or not *etc*).

A subset of these data called -scc40_2level- was created by only taking the first observation for each cow, thereby reducing the dataset to two levels (herds and cows).

Table of variables

Variable	Description	Codes/units
herdid	herd id	
cowid	cow id	
test	approximate month of lactation	0 to 10
h_size	average herdsize	
c_heifer	parity of the cow	1 = heifer 0 = multiparous
t_season	season of test day	1 = jan-mar 2 = apr-jun 3 = jul-sep 4 = oct-dec
t_dim	days in milk on test-day	days
t_lnscc	log somatic cell count on test day	

Name **smpltype**

Study type longitudinal
# of records 1114
Unit of record pig
Contributor Håkan Vigre

Reference
Vigre H, Dohoo IR, Stryhn H, Busch ME. Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumonia* and *Mycoplasma hyopneumoniae* at different levels in Danish multisite pig production facilities. Prev Vet Med 2003; submitted.

Brief description
These data are derived from the -ap2- dataset. In addition to the original data, this dataset contains indicator variables (made up) that identify those pigs that were part of a: simple random sample, systematic random sample, stratified random sample, cluster sample and multistage sample.

Table of variables

Variable	Description	Codes/units
farm_id	farm identification	
pig_id	pig identification	
barn_ord	order of pigs in barn (as they were walked down an alley)	
parity	farrowing number of the sow (categorised)	1 = 1 2 = 2 3 = 3-4 4 = 5+
dwg_t	daily weight gain to approx 65 days	gm/day
smp_srs	pig in the simple random sample	0/1
smp_syst	pig in the systematic random sample	0/1
smp_strat	pig in the stratified (by parity) sample	0/1
smp_clust	pig in the cluster (by herd) sample	0/1
smp_ms	pig in the multistage sample (psu=herd)	0/1

Name **tb_real**

Study type retrospective cohort
# of records 134
Unit of record animal groups
Contributors Ian Dohoo, Fonda Munroe

References

1. Munroe FA, Dohoo IR, Mcnab WB. Estimates of within-herd incidence rates of *Mycobacterium bovis* in Canadian cattle. Prev Vet Med 2000; 45: 247-256.
2. Munroe FA, Dohoo IR, Mcnab WB, Spangler L. Risk factors for the between-herd spread of *Mycobacterium bovis* in Canadian cattle and cervids between 1985 and 1994. Prev Vet Med 1999; 41: 119-133.

Brief description

A retrospective evaluation of all (n=9) outbreaks of tuberculosis in domestic animals (dairy and beef cattle, cervids and bison) in Canada between the years of 1985 and 1994 was carried out to investigate risk factors for the spread of tuberculosis within and between herds. Detailed records from the epidemiologic investigation of all outbreaks (including records on all contact herds) were reviewed and a summary of each outbreak prepared. This dataset contains data only from herds in which tuberculosis was observed. In each herd, the most probably date on which the infection entered the herd was determined and the number of new cases arising within the herd determined from the herd testing results. The number of animals in each age, sex and type group was determined and the number of animal days at risk was computed. The effects of age (three groups), sex (two groups), and animal type (five groups) on the incidence rate of new infections was investigated. **Note** To meet confidentiality and regulatory concerns, these data have deliberately been falsified.

Table of variables

Variable	Description	Codes/units
obs	observation number	
farm_id	farm identification	
type	type of animal	1 = dairy cattle 2 = beef cattle 3 = cervid 4 = other
sex	sex	1 = female 2 = male
age	age category	1 = 0-12 mo 2 = 12-24 mo 3 = >24 mo
reactors	number of pos/reactors in the group	
par	animal days at risk in the group	

28

PROGRAM FILES

This chapter contains the program files used for the analyses presented in all of the examples in this text. With the exception of a few examples in Chapters 21 and 22, all examples were worked out using Stata®, Version 7. Consequently, virtually all of the program files are Stata -do- files. Most examples in Chapter 21 were analysed using SAS® (primarily Proc Mixed) and the SAS program files are presented. Some examples in Chapters 21 and 22 were analysed using MLwiN® but this program was used interactively, so program files are not available for those exercises.

During the production of this text, Version 8 of Stata was released. However, to maintain consistency throughout the book, Version 7 was used for all examples. In a few instances, the syntax of Stata commands changed between Versions 7 and 8. Consequently, if you use Version 8 (or subsequent versions) to run these programs, you may need to add the statement:

```
version 7
```

at the start of each program. In the same light, the graphics capabilities of Stata were extended greatly between Versions 7 and 8. If you are running Version 8, you will either need to add the version statement shown above to any program containing the -graph- command, or replace it with the command -gr7-.

All of the -do- files (and SAS program files) used in the book will be available from the textbook's website (http://www.upei.ca/ver . In addition, -do- files used to generate many of the Figures which are not contained in Examples will be available through the Veterinary Epidemiologic Research website.

All of the program files assume that the data are stored in a folder on your c: drive called:

```
c:\ver\data
```

The data files are also all available from the Veterinary Epidemiologic Research website. They can either be downloaded from the site, or accessed directly from within Stata using Stata's -net from- capabilities (see Stata documentation for details).

Most of the -do- files are relatively short and straightforward to follow. A few, which simulate data prior to analysing them, are much longer and more complex. In addition, these programs save data files in a specified location which you may have to change if the specified folder does not exist on your computer.

1 Introduction and causal concepts

There are no program files for this chapter.

2 Sampling

```
  * Example 2.2 Analysis of stratified data
  * open the Dairy Disease dataset
use "c:\ver\data\dairy_dis.dta", clear
  * compute the prevalence treating the sample as a simple random sample
svyset, clear
svyprop neospora
  * compute the overall prevalence treating the sample as a stratified sample
svyset strata prov
svyprop neospora
svyprop neospora, by(prov)

  * Example 2.3 Analysis of weighted data
  * open the Dairy Disease dataset
use "c:\ver\data\dairy_dis.dta", clear
  * compute the prevalence treating the sample as stratified and taking weights into account
svyset, clear
svyset strata prov
svyset pweight weight
svyprop neospora

  * Example 2.4 Analysis of multistage sampled data
  * open the Dairy Disease dataset
use "c:\ver\data\dairy_dis.dta", clear
  * compute the prevalence treating the sample as stratified and taking weights into account
  *  and considering the primary sampling unit.
svyset, clear
svyset strata prov
svyset pweight weight
svyset psu herd
svyprop neospora

  * Example 2.5/2.6 Sample size calculations
  * sample size if steers randomly allocated to treatments
sampsi .15 .10,  p(.8)
  * sample size if treatment applied at the farm level
sampsi .15 .10,  p(.8)
sampclus, rho(0.3) obs(50)
  * note this program generates slightly different answers than the hand calculations in the
  * text due to using a slightly different formula
```

```
  * Example 2.7 Power calculation by simulation
set more off
set seed 123456
  * open the pig_adg dataset
use "c:\ver\data\pig_adg.dta", clear
  * generate a new variable for the presence/absence of worms
gen w2=worms>0
  * regress adg on w2, se and farm
xi:regress adg w2 sex i.farm
  * compute predicted values for each observation and determine the standard error of
  * prediction
predict pred, xb
  * set up a file to hold the results from the simulation
tempname memhold
postfile `memhold' beta z pval using "c:\ver\ch2\ex2_7_rslt.dta", replace
  * set up loop to repeat the analyses many times
local i=0
while `i'<1000 {
  * generate a new outcome variable adg_new that is normally distributed with a mean at the
  * predicted value and a std. dev. of 46.905
gen rand=invnorm(uniform())
gen adg_new=pred+46.905*rand
  * regress adg_new on the same set of predictors
quietly xi:regress adg_new w2 sex i.farm
  * obtain the regression coefficient and compute its z statistic and P-value
matrix V=e(V)
matrix B=e(b)
scalar beta=B[1,1]
scalar z=B[1,1]/sqrt(V[1,1])
scalar pval=2*(1-normprob(abs(z)))
  * post the beta, z and pval to the results file
post `memhold' (beta) (z) (pval)
  * repeat the loop
drop rand adg_new
local i = `i'+1
}
  * stop the process of posting results to a file
postclose `memhold'
  * open the file that captured the results
use "c:\ver\ch2\ex2_7_rslt.dta", clear
  * compute and display the power
count if pval<0.05
scalar power=r(N)/1000
display " "
display "Power is " %8.3f scalar(power)
  * compare the computed power to a simple calculation based on the mathematical formulae
  * presented
sampsi 500 507.7, sd(46.9) n1(114) n2(227)
```

3 Questionnaire design

There are no program files for this chapter

4 Measures of disease frequency

```
  * Example 4.3 Confidence intervals for risks and rates
  * open the dairy disease dataset
use c:\ver\data\dairy_dis, clear
  * drop all but the herd=1 data
keep if herd==1
  * reformat the Johne's and leukosis variables
format johnes leukosis %6.4f
  * compute the approximate confidence intervals
ci leukosis
ci johnes
  * compute the exact binomial confidence intervals
ci leukosis, binomial
ci johnes, binomial
  * compute incidence rates and exact confidence intervals
gen age=lact+2
ci leukosis, poisson exposure(age)
ci johnes, poisson exposure(age)
  * clear the data to prevent the original file being overwritten
clear

  * Example 4.4 Indirect standardisation
  * open the standardisation dataset
use "c:\ver\data\stdize.dta", clear
  * compute indirect standardised rates
istdize totcases hy type using "c:\ver\data\stdize_ind.dta", by(region) rate(rate 0.06) print

  * Example 4.5 Direct standardisation
  * open the standardisation dataset
use "c:\ver\data\stdize.dta", clear
  * compute indirect standardised rates
dstdize  cases hy type, by(region) using("c:\ver\data\stdize_dir.dta")
```

5 Screening/Diagnostic tests

```
  * Example 5.1 Measuring agreement - quantitative test results
  * open the elisa_repeat dataset
use "c:\ver\data\elisa_repeat.dta", clear
  * coefficient of variation
  * compute the mean and SD and coef. of variation of the 6 raw and 6 adjusted values
egen raw_avg=rmean(raw1-raw6)
egen raw_sd=rsd(raw1-raw6)
gen raw_cv=raw_sd/raw_avg
```

```
egen adj_avg=rmean(adj1-adj6)
egen adj_sd=rsd(adj1-adj6)
gen adj_cv=adj_sd/adj_avg
  * compute the average cv for the raw and adjusted values
sum raw_cv adj_cv
  * Pearson correlation coefficient
corr raw1 raw2 adj1 adj2
  * concordance
concord raw1 raw2
concord adj1 adj2, g(ccc) gap(6) saving("c:\ver\ch5\fig5_1.gph", replace)
  * Limits of Agreement plot
concord adj1 adj2, g(loa) l1(" ") l2("Difference Between Values") /*
    */ saving("c:\ver\ch5\fig5_2.gph", replace) pen(11121)

  * Example 5.2 Agreement between two dichotomous tests
  * open the isa-test dataset
use "c:\ver\data\isa_test.dta", clear
  * compute McNemar's chi-square for assessing the positive proportions
mcc ifat1_np ifat2_np
  * compute kappa using the dichotomised IFAT results (2 labs)
kap ifat1_np ifat2_np
  * compute the confidence interval for kappa
kapgof ifat1_np ifat2_np

  * Example 5.3 Agreement among ordinal test results
  * open the isa-test dataset
use "c:\ver\data\isa_test.dta", clear
  * set up the agreement (weight) matrix
kapwgt isa_wt 1 \ 0.7 1 \ 0.3 0.7 1 \ 0 0.3 0.7 1 \ 0 0 0.3 0.7 1
kapwgt isa_wt
  * compute the weighted kappa
kap ifat1 ifat2, wgt(isa_wt) tab

  * Example 5.4 Sensitivity, specificity and predictive values
  * open the bvd-test dataset
use "c:\ver\data\bvd_test.dta", clear
  * make a copy of the data, keep only the blood samples
preserve
keep if spec==1
  * dichotomise the test results
gen pos=od>0.92 if od~=.
  * compute the relevant test characteristics
diagt calfst pos
  * restore the original (full) dataset
restore

  * Example 5.7 Parallel and serial interpretation
  * open the isa-test dataset
use "c:\ver\data\isa_test.dta", clear
  * cross-tabulate IFAT and PCR by VI
```

```
bysort dz:tab ifat1_np pcr, row col
  * generate results for series interpretation
gen series=0 if pcr~=. & ifat1_np~=.
replace series=1 if ifat1_np==1 & pcr==1
tab dz series, row
  * generate results for parallel interpretation
gen parallel=0 if pcr~=. & ifat1_np~=.
replace parallel=1 if (ifat1_np==1 | pcr==1) & parallel==0
tab dz parallel, row
drop series parallel

  * Example 5.9 Cutpoints
  * open the bvd-test dataset
use "c:\ver\data\bvd_test.dta", clear
  * limit the data to blood samples only
preserve
keep if spec==1
  * compute Se and Sp at various cutpoints
diagt calfst co_5
diagt calfst co_6
diagt calfst co_7
diagt calfst co_8
diagt calfst co_9
diagt calfst co_10
diagt calfst co_11
diagt calfst co_12
diagt calfst co_13
diagt calfst co_14
diagt calfst co_15
  * restore the data
restore

  * Example 5.10 ROC and sensitivity specificity curve
  * generating a graph of sensitivity and specificity vs cutpoints
  * open the BVD_test dataset
use "c:\ver\data\bvd_test.dta", clear
  * ROC curve
  * generate parametric and non-parametric ROC curves
rocfit calfst od, cont(15)
  * graph the two curves
rocplot, conf pen(12222) saving("c:\ver\ch5\fig5_4.gph", replace) gap(6)
  * sensitivity vs specificity curve
  * determine the range of OD values and determine the width of an interval that is 1% or the
range
sum od
scalar min=r(min)
scalar max=r(max)
scalar intvl=(max-min)/100
  * set up loop to compute the sensitivity and specificity at each of the 100 cutpoints
gen pos=.
```

```
gen cp=.
gen se=.
label var se "Sensitivity"
gen sp=.
label var sp "Specificity"
local i=0
while `i'<100 {
  local cut=(min + `i'*intvl)
  quietly replace pos=od>`cut'
  quietly diagt calfst pos
  local sen=r(sens)
  local spe=r(spec)
  quietly replace cp=`cut' if _n==`i'+1
  quietly replace se=`sen' if _n==`i'+1
  quietly replace sp=`spe' if _n==`i'+1
  local i = `i'+1
}
graph se sp cp, ylab(0 20 40 60 80 100) xlab(0 .5 1 1.5 2 2.5) b2("Cut-point") /*
    */ l1("Sensitivity/Specificity") gap(6) saving("c:\ver\ch5\fig5_5.gph", replace)
```

*** Example 5.11 Generating likelihood ratios at a number of cutpoints**

```
  * open the BVD_test dataset
use "c:\ver\data\bvd_test.dta", clear
  * obtain the categories used for cutpoints
egen od_cat=cut(od), at(0 .5 .7 .9 1.1 1.3 1.5 1.7 1.9 2.1 4)
  * derive likelihoods
tab od_cat calfst , col
```

*** Example 5.12 Estimating Se and Sp using logistic regression**

```
  * open the bvd-test dataset
use "c:\ver\data\bvd_test.dta", clear
  * fit the logistic regression model in D+ animals
xi: logit co_10 pregmon i.season if calfst==1
  * compute predicted values
predict xb, xb
  * list predicted values for cow calving in fall, and  7 pregnant
sort season pregmon calfst
quietly by season pregmon calfst: gen temp=_n
list calfst pregmon season xb if pregmon==7 & season==4 & temp==1 & calfst==1
  * fit the logistic regression model in D- animals
xi: logit co_10 pregmon i.season if calfst==0
  * compute predicted values
quietly drop xb temp
predict xb, xb
  * list predicted values for cow calving in fall, and  7 pregnant
sort season pregmon calfst
quietly by season pregmon calfst: gen temp=_n
list calfst pregmon season xb if pregmon==7 & season==4 & temp==1 & calfst==0
```

6 Measures of association

```
  * Example 6.2 Confidence intervals for measures of association
  * compute direct and test-based estimates of CI for incidence rate data
iri 60 157 284 1750
iri 60 157 284 1750, tb
  * compute exact and test-based CI for incidence risk data
csi 60 157 41 359, exact
csi 60 157 41 359, tb
  * compute Cornfield's test-based and Woolf's CI for odds ratios
csi 60 157 41 359, or
csi 60 157 41 359, or woolf
csi 60 157 41 359, or tb
cci 60 157 41 359, exact
```

7 Introduction to observational studies

There are no program files for this chapter.

8 Cohort studies

There are no program files for this chapter.

9 Case-control studies

There are no program files for this chapter.

10 Hybrid study designs

There are no program files for this chapter.

11 Controlled trials

There are no program files for this chapter.

12 Validity in observational studies

There are no program files for this chapter.

13 Confounder bias: analytic control and matching

There are no program files for this chapter.

14 LINEAR REGRESSION

```
  * Example 14.1 Simple linear regression
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * show ANOVA table from regressing calvcon on age
reg calvcon age

  * Example 14.2 Multiple linear regression
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * regress calvcon on age, metritis and ovar
reg calvcon age metritis ovar

  * Example 14.3 Testing the significance of multiple predictors
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * regress calvcon on age, retpla, metritis and ovar
reg calvcon age retpla metritis ovar
test retpla metritis ovar

  * Example 14.4 Rescaling predictor variables
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * rescale age using age-2
gen age_sc = age - 2
  * regress calvcon on re-scaled age to make constant interpretable
reg calvcon age_sc

  * Example 14.6 Hierarchical indicator (dummy) variables
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * change all ages greater than 8 to 8
replace age=8 if age>8
  * regress calvcon on age coded using regular dummies
xi: reg calvcon i.age
  * create hierarchical dummy variables (this will be done one at a time)
drop if age==.
gen hage3=age>=3
gen hage4=age>=4
gen hage5=age>=5
gen hage6=age>=6
gen hage7=age>=7
gen hage8=age>=8
  * regress calvcon on age coded using hierarchical dummies
reg calvcon hage3 hage4 hage5 hage6 hage7 hage8

  * Example 14.7 Centring variables (collinearity)
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
```

```
  * generate a quadratic term without centring
gen age_sq=age^2
  * build the quadratic model and compute VIF
reg calvcon age
reg calvcon age age_sq
vif
  * generate a quadratic term using centred variable
sum age
gen age_ct= age-4.34
gen age_ct_sq=age_ct^2
  * build the quadratic model and compute VIF
reg calvcon age_ct
reg calvcon age_ct age_ct_sq
vif

  * Example 14.8 Interaction
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * create an interaction term between ovar and metritis
gen ovarmet=ovar*metritis
reg calvcon metritis ovar ovarmet

  * Example 14.9 Interaction - dichotomous variables
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * create an interaction term between retpla and metritis
gen rpmet=retpla*metritis
reg calvcon retpla metritis rpmet

  * Example 14.10 Interaction - dichotomous and continuous variables
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * create the interaction term for ovar and firstest
gen ovarest=ovar*firstest
  * fit the regression model
reg calvcon ovar firstest ovarest
  *obtain the predicted means
predict pcalvcon, xb
  * create a separate outcome for those with and without ovar
gen ovar1=pcalvcon if ovar==1
gen ovar0=pcalvcon if ovar==0
  * graph the results
graph ovar0 ovar1 firstest, xlab ylab c(.llll) l1t("Days to Conception") gap(4) /*
    */ sa("c:\ver\ch14\fig14_2.gph", replace)

  * Example 14.11 Interaction - two continuous variables
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
  * create the interaction term to demonstrate how firstbrd effect varies by age
gen agebrd=age*firstbrd
```

```
reg calvcon age firstbrd agebrd
  * obtain the predicted means
predict p, xb
  * obtain a new categorical variable for age
egen age_c4=cut(age), at(2, 4, 6, 8, 15)
  * plot ages 2-3 4-5 6-7 and 8 or more versus p
  * shows how effect of firstbrd varies by age
gen p2=p if age_c4==2
gen p4=p if age_c4==4
gen p6=p if age_c4==6
gen p8=p if age_c4==8
sort firstbrd
  * produce the graph
graph p2 p4 p6 p8 firstbrd if firstbrd<220, xlab ylab c(llll) pen(1111) /*
   */ l1t("Days to Conception") gap(4) sa("c:\ver\ch14\fig14_3.gph", replace)

  * Example 14.12 Inferring causation
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * fit a regression model
reg calvcon farm1 age metritis ovar firstbrd
  * refit the model without intervening variables
reg calvcon farm1 age metritis

  * Example 14.13 Examining homoscedasticity
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * fit a regression model
reg calvcon farm1 age metritis ovar firstbrd
  * obtain predicted calvcon
predict pcalvcon, xb
  * obtain standardised residuals
predict stdres, rstan
  * produce graph and test for heteroscedasticity
ksm stdres pcalvcon, lowess t1title("Homoscedasticity Plot, Lowess residual smoother") /*
   */ l1title("   ") l2title("Standardised residuals") ylab gap(4) /*
   */ b2title("   ") b1title("Predicted Calving-Conception Interval (days)") xlab /*
   */ sa("c:\ver\ch14\fig14_4.gph", replace)
reg calvcon farm1 age metritis ovar firstbrd
hettest

  * Example 14.14 Assessing normality of residuals
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * fit a regression model
reg calvcon farm1 age metritis ovar firstbrd
  * obtain predicted calvcon
```

```
predict pcalvcon, xb
  * obtain standardised residuals and test them statistically
predict stdres, rstan
swilk stdres
  * generate a normal probability plot
qnorm stdres, t1title("Q-Q Plot of Residuals") l1title(" ") xlab ylab gap(4) /*
   */ l2title("Standardised residuals") b1title("Inverse normal distribution") /*
   */ b2title(" ") sa("c:\ver\ch14\fig14_5a", replace)
  * generate a histogram
graph stdres, histogram bin(12) t1title("Histogram of Studentized Residuals") /*
   */ l1title(" ") l2title("Proportion of cases") ylab gap(4)xlab /*
   */ sa("c:\ver\ch14\fig14_5b", replace)
  * generate a combined graph
graph using "c:\ver\ch14\fig14_5a" "c:\ver\ch14\fig14_5b", sa("c:\ver\ch14\fig14_5.gph",
replace)

  * Example 14.15 Assessing linearity of age-calvcon relationship
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * Fit a regression model
reg calvcon farm1 age metritis ovar firstbrd
  * obtain standardised residuals
predict stdres, rstan
  * produce graph with lowess smoother
ksm stdres age,lowess ylab gap(4) l1title(" ") l2title("Standardised residuals") /*
   */ b2title(" ") b1title("age") xlab sa("c:\ver\ch14\fig14_6.gph", replace)

  * Example 14.16 Transformations
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * first evaluate Box-Cox transformations
boxcox calvcon farm1 age metritis ovar firstbrd
  * generate residuals from the model fit
predict resid, residuals
  * graph and test the residuals (note: the graphs and the test results are not displayed
  * in the text - they are included here for additional information)
qnorm resid, t1title("Q-Q Plot of Residuals") l1title(" ") ylab xlab gap(4) /*
   */ l2title("Raw residuals") b2title(" ") b1title("Inverse normal distribution")
swilk resid
  * determine the optimal skewness correction
lnskew0 c=calvcon

  * Example 14.17 Robust standard errors
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * fit the regression model with robust SE
reg calvcon farm1 age metritis ovar firstbrd, robust
```

```
  * Example 14.18 Other diagnostic parameters
  * open the dataset daisy.dta
use "c:\ver\data\daisy.dta", clear
gen farm1=farmnum==1
  * fit the regression model with robust SE
reg calvcon farm1 age metritis ovar firstbrd, robust
  * obtained desired statistics
predict pcalvcon, xb
predict residuals, re
predict stdres, rsta
predict studres, rstu
predict cooks, cook
predict lever, lever
predict dfits, dfits
  * obtain descriptive summary statistics
summ stdres studres cooks lever dfits
  * drop the cases with no 'calvcon'
drop if calvcon==.
  * listing the 5 biggest and 5 smallest residuals
sort stdres
list cownum calvcon age firstbrd residuals studres stdres in 1/5
list cownum calvcon age firstbrd residuals studres stdres in -5/-1
  * listing the leverage cases
sort lever
list cownum calvcon age metritis ovar firstbrd lever stdres in -5/-1
  * listing the influential cases
sort dfits
list cownum calvcon age metritis ovar firstbrd cooks dfits in 1/5
list cownum calvcon age metritis ovar firstbrd cooks dfits in -5/-1
  * checking for influential values of firstbrd
predict dbfirbrd, dfbeta(firstbrd)
sort dbfirbrd
list cownum calvcon firstbrd dbfirbrd in 1/5
list cownum calvcon firstbrd dbfirbrd in -5/-1
```

15 MODEL-BUILDING STRATEGIES

```
  * Example 15.1 Correspondence analysis
  * Risk factors for swine respiratory disease
  * open the calf septicemia dataset
use "c:\ver\data\pig_farm.dta", clear
  * categorise several predictor variables
egen inlet_c=cut(inlet), at(0 1 2 50) icodes
egen size_c=cut(size), at(0 1000 1500 3000) icodes
egen exprnce_c=cut(exprnce), at(0 10 20 60) icodes
egen exhaust_c=cut(exhaust), at(0 1 1.3 2.5) icodes
  * carry out a multiple correspondence analysis of 8 variables
mca pncode inlet_c size_c exprnce_c exhaust_c hldbck floor hmrsd, d(2)
```

```
  * Example 15.2 Regression model validation - 1
  * open the calf septicemia dataset
use "c:\ver\data\pig_farm.dta", clear
set more off
  * log transform -pneu- and rescale size
gen lnpneu=log(pneu)
replace size=size/1000
 * fit a forward selection model
sw regress lnpneu num-exprnce, pe(0.05)
fitstat
  * fit a backward elimination model
sw regress lnpneu num-exprnce, pr(0.05)
fitstat

  * Example 15.3 Regression model validation - 2
  * open the calf septicemia dataset
use "c:\ver\data\pig_farm.dta", clear
  * log transform -pneu-
gen lnpneu=log(pneu)
  * generate a random number
set seed 153
gen rand=uniform()
  * fit a model based using a forward selection procedure
  * and using only 60% of the data
sw regress lnpneu num-exprnce if rand<=0.75, pe(0.05)
  * compute predicted values
predict pv
  * compute the correlations between predicted values and observed
  * for the estimation subset and the validation subset
corr pv lnpneu if rand<=0.75
corr pv lnpneu if rand>0.75

  * Example 15.4 Regression model validation - 3
  * open the calf septicemia dataset
use "c:\ver\data\pig_farm.dta", clear
  * log transform -pneu-
gen lnpneu=log(pneu)
  * fit the backward's selection model from Example 15.2
regress lnpneu hldbck size exhaust exprnce floor
  * compute the IQR for each of the continuous predictors
codebook size exhaust exprnce
```

16 LOGISTIC REGRESSION

```
  * Example 16.1 Comparing logistic regression models
  * open the Nocardia dataset
use c:\ver\data\nocardia, clear
  * fit the full model and save its log-likelihood value
logit casecont dcpct dneo dclox
```

```
lrtest, saving(full)
  * fit the reduced model and save its log-likelihood
logit casecont dcpct
lrtest, saving(red)
  * compare the reduced to the full model
lrtest, using(full)
  * fit the null model and save its log-likelihood
logit casecont
lrtest, saving(null)
  * compare the full and null models
lrtest, using(full)
  * refit the full model and compute a multiple Wald's test
  * of dneo and dclox
logit casecont dcpct dneo dclox
test dneo dclox
```

```
  * Example 16.2 Interpreting logistic regression coefficients
  * open the Nocardia dataset
use c:\ver\data\nocardia, clear
  * fit a model containing dcpct, dneo dclox and dbarn (2 levels)
xi:logit casecont dcpct dneo dclox i.dbarn
  * refit the same model to get odds ratio estimates
xi:logistic casecont dcpct dneo dclox i.dbarn
lrtest, saving(full)
test _Idbarn_2 _Idbarn_3
xi:logistic casecont dcpct dneo dclox
lrtest, using(full)
```

```
  * Example 16.3 Effects of factors on the probability scale
  * open the Nocardia dataset and fit the model
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * create predicted probabilities as -dcpct- goes from 0 to 100%
  * do separately for herds using neomycin and cloxacillin
prgen dcpct, gen(pneo)  ncases(101) x(dneo 1 dclox 0)
prgen dcpct, gen(pclox)  ncases(101) x(dneo 0 dclox 1)
  * plot the probability of Nocardia mastitis as a function of -dcpct-
graph  pneop1 pcloxp1 pneox, s(op) b2("Dry-cow treatment (%)") /*
   */ l2("Prob. of Nocardia mastitis") xlab(0 25 50 75 100) ylab (0 .3 .6 .9) key1(symbol(o) /*
   */ "neomycin") key2(symbol(p) "cloxacillin") saving(fig16_2, replace)
```

```
  * Example 16.4 Assessment of confounding
  * open the Nocardia dataset and fit a full model
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * refit the model without -dcpct- to see if it is a "confounder" for -dneo-
logit casecont dneo dclox
```

```
  * Example 16.5 Assessment of interaction
  * open the Nocardia dataset and fit a full model
```

```
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * refit the model including an interaction term between -dneo- and -dclox-
gen neoclox=dneo*dclox
logit casecont dcpct dneo dclox neoclox

  * Example 16.6 Evaluating continuous independent variables - 1
  * open the Nocardia dataset and fit simple model with -numcow-
use c:\ver\data\nocardia,clear
logit casecont numcow
  * generate Pearson residuals and plot them against numcow
predict r , res
graph r numcow, xlab ylab l1(" ") l2("Pearson Residuals") gap(4) /*
   */ saving(c:\ver\ch16\fig16_4, replace)

  * Example 16.7 Evaluating continuous independent variables - 2
  * open the Nocardia dataset and fit simple model with -numcow-
use c:\ver\data\nocardia,clear
  * create a categorical variable for numcow (4 levels)
  * cutpoints chosen to match those used by -lintrend- (below)
egen numcow4=cut(numcow), at(0,32,42,55,999)
xi:logit casecont i.numcow4
  * generate a plot of log odds of outcome vs categories of -numcow-
lintrend casecont numcow, g(4) plot(log) saving(c:\ver\ch16\fig16_5, replace) xlab ylab

  * Example 16.8 Evaluating continuous independent variables - 3
  * open the Nocardia dataset
use "c:\ver\data\nocardia.dta", clear
  * fit a quadratic model of -numcow- to -casecont-
gen numcow_ct = numcow-75
gen numcow_ct_sq=numcow_ct^2
corr numcow_ct numcow_ct_sq
logit casecont numcow_ct numcow_ct_sq
  * create orthogonal polynomials (power=2) and use them
orthpoly numcow, degree(2) gen(numcow_op1 numcow_op2)
sum numcow numcow_op1 numcow_op2
corr numcow_op1 numcow_op2
logit casecont numcow_op1 numcow_op2
  * use fractional polynomials in the model
fracpoly logit casecont numcow, degree(2)

  * Example 16.9 Evaluating continuous independent variables - 4
  * open the Nocardia dataset
use "c:\ver\data\nocardia.dta", clear
  * generate a smoothed curve of log odds of
  * -casecont- on -numcow-
ksm casecont numcow, logit ylab xlab l2("Log odds of -casecont-") /*
     */ saving(c:\ver\ch16\fig16_5, replace)
```

```
  * Example 16.10 Evaluating continuous independent variables - 5
  * open the Nocardia dataset
use "c:\ver\data\nocardia.dta", clear
  * fit linear splines for the herd sizes 0-70 >70
mkspline numcow_sp1 70 numcow_sp2=numcow
sort numcow
list numcow numcow_sp1 numcow_sp2 in 88/93
logit casecont numcow_sp1 numcow_sp2

  * Example 16.11 Residuals and covariate patterns
  * open the Nocardia dataset and fit a model with -dcpct-, -dneo-, and -dclox-
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * compute residuals on the basis of 1 per covariate pattern
predict pv, p
predict cov, num
predict res, res
  * use -glm- to fit the same model and
  * compute residuals on the basis of 1 per individual observation
glm casecont dcpct dneo dclox, family(binomial) link (logit)
predict glmres, p
listblck id casecont dcpct dneo dclox cov pv res glmres if id==86
listblck id casecont dcpct dneo dclox cov pv res glmres if id==22

  * Example 16.12 Goodness-of-fit tests
  * open the Nocardia dataset and fit a model with -dcpct-, -dneo-, and -dclox-
use "c:\ver\data\nocardia.dta", clear
logistic casecont dcpct dneo dclox
  * compute the Pearson and deviance chi-square statistics
lfit
ldev
  * compute the Hosmer-Lemeshow chi-square with 5 groups
lfit, g(5) table

  * Example 16.13 Predictive ability of the model
  * open the Nocardia dataset and fit a model with -dcpct-, -dneo-, and -dclox-
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * compute the sensitivity and specificity of the model at a cutoff of 0.5
lstat
  * generate a graph of sens. and spec. vs cutoffs
lsens, s(op) saving("c:\ver\ch16\fig16_6", replace)
  * generate an ROC curve for the model
lroc, saving("c:\ver\ch16\fig16_7", replace)

  * Example 16.14 Identifying important observations - 1
  * open the Nocardia dataset and fit a model with -dcpct-, -dneo-, and -dclox-
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * compute residuals on the basis of 1 per covariate pattern
```

```
predict pv, p
predict cov, num
predict res, res
predict rst, rst
  * graph standardised residuals vs herd id
graph rst cov, s([cov]) xlab(1 5 10 15 20 25 30) ylab saving("c:\ver\ch16\fig16_8",replace)
  * list the observations with the 2 large positive residuals
sort rst
listblck cov casecont dcpct dneo dclox pv rst in -2/-1

  * Example 16.15 Identifying important observations - 2
  * open the Nocardia dataset and fit a model with -dcpct-, -dneo-, and -dclox-
use "c:\ver\data\nocardia.dta", clear
logit casecont dcpct dneo dclox
  * compute residuals and other diagnostics on the basis of 1 per covariate pattern
predict pv, p
predict cov, num
predict res, res
predict rst, rst
predict lev, hat
predict db, db
predict dx2, dx
predict ddev, ddev
  * list the 4 covariate patterns with the largest leverage
preserve
collapse  (mean) casecont dcpct dneo dclox pv lev db dx2 ddev (count) id, by(cov)
sort lev
format casecont %6.2f
listblck cov id casecont dcpct dneo dclox pv lev in -4/-1
  * list the 4 covariate patterns with the largest delta betas
sort db
listblck cov id casecont dcpct dneo dclox pv db in -4/-1
  * list the 4 covariate patterns with the largest dx2 and ddev values
sort dx2
listblck cov id casecont dcpct dneo dclox pv dx2 in -4/-1
sort ddev
listblck cov id casecont dcpct dneo dclox pv ddev in -4/-1

  * Example 16.16 Conditional logistic regression
  * open the Salmonella outbreak dtaset dataset
use "c:\ver\data\sal_outbrk.dta", clear
  * fit conditional and ordinary logistic regressions
  * with -slt_a_ as the sole predictor
clogit casecontrol slt_a, group(match_grp) or
logistic casecontrol slt_a
```

17 MODELLING MULTINOMIAL DATA

```
  * Example 17.1 Simple multinomial regression model
  * open the beef ultrasound dataset
use c:\ver\data\beef_ultra.dta, clear
  * fit a simple multinomial logistic model
mlogit grade sex, b(1)
mlogit grade sex, b(1) rrr
tab grade sex, col chi

  * Example 17.2 Multiple  multinomial regression model
  * open the beef ultrasound dataset
use "c:\ver\data\beef_ultra.dta", clear
  * fit a multiple multinomial logistic model and test the significance of sex
mlogit grade sex backfat ribeye imfat carc_wt , b(1)
test sex
lrtest, saving(full)
mlogit grade backfat ribeye imfat carc_wt ,b(1) rrr
lrtest, using(full)
  * compute predicted probabilities
mlogit grade sex backfat ribeye imfat carc_wt , b(1)
predict pA pAA pAAA, p
sort id
list id grade sex imfat pA pAA pAAA in 1/20

  * Example 17.3 Adjacent category model
  * open the beef ultrasound dataset
set more off
use "c:\ver\data\beef_ultra.dta", clear
  * first define constraints
constraint define 1 [AA]sex=0.5*[A]sex
constraint define 2 [AA]backfat=0.5*[A]backfat
constraint define 3 [AA]ribeye=0.5*[A]ribeye
constraint define 4 [AA]imfat=0.5*[A]imfat
constraint define 5 [AA]carc_wt=0.5*[A]carc_wt
  * fit constrained model
mlogit grade sex backfat ribeye imfat carc_wt , b(1) constraint(1-5)
  * LRT of unconstrained vs constrained models
mlogit grade sex backfat ribeye imfat carc_wt , b(1)
lrtest , saving(unc)
mlogit grade sex backfat ribeye imfat carc_wt , b(1) constraint(1-5)
lrtest, using(unc)

  * Example 17.3b Continuation ratio model
  * open the beef ultrasound dataset
use c:\ver\data\beef_ultra.dta, clear
  * first create the two new outcome variables
gen gr2=(grade==2)
replace gr2=. if grade>2
gen gr3=(grade==3)
```

```
  * fit the logistic models and store the log-likelihoods
logit gr2 sex backfat ribeye imfat carc_wt
scalar ll2=e(ll)
  * fit the logistic models and store the log-likelihoods
logit gr3 sex backfat ribeye imfat carc_wt
scalar ll3=e(ll)
display ll2 + ll3
```

*** Example 17.4 Proportional odds model**

```
  * open the beef ultrasound dataset
use c:\ver\data\beef_ultra.dta, clear
  * fit a proportional odds model
ologit grade sex backfat ribeye imfat carc_wt , table
  * compute predicted values for all individuals
capture drop pA pAA pAAA
predict pAAA pAA pA, p
  * list the relevant variables for the first 5 observations
sort id
listblck id grade sex backfat ribeye imfat carc_wt in 1/5
listblck id pAAA pAA pA in 1/5
  * generate smoothed values of predicted probabilities
capture drop pAsm pAAsm pAAAsm
ksm pAAA imfat if imfat>3 & imfat<6, bw(.3) gen(pAAAsm)
ksm pAA imfat if imfat>3 & imfat<6, bw(.3) gen(pAAsm)
ksm pA imfat if imfat>3 & imfat<6, bw(.3) gen(pAsm)
lab var pAAAsm "Grade AAA"
lab var pAAsm "Grade AA"
lab var pAsm "Grade A"
  * graph the smoothed predicted values against imfat
sort imfat
graph pAAAsm pAAsm pAsm imfat, c(L[l] L[-] L[.]) s(iii) gap(3) pen(111) /*
   */ ylab xlab l2("Probability of grade") saving(c:\ver\ch17\fig17_1.gph, replace)
  * testing the proportional odds assumption with likelihood ratio and Wald tests
omodel logit grade sex backfat ribeye imfat carc_wt
brant, detail
```

18 Modelling count and rate data

*** Example 18.1 Poisson regression - fitting the model**

```
  * open the tb_real dataset
use c:\ver\data\tb_real.dta, clear
  * fit the Poisson model and conduct overall goodness-of-fit tests
xi: poisson reactors i.type sex i.age, exp(par)
poisgof
poisgof, pearson
xi: poisson reactors i.type sex i.age, exp(par) irr
  * compute the mean number of cases per 10000 days for the whole population
gen ir=10000*reactors/par
sum ir
```

```
  * compute the number of fewer cases expected in males than females
di 2.898 * 0.696

  * Example 18.2 Poisson regression - examining the model
  * open the tb_real dataset
use c:\ver\data\tb_real.dta, clear
  * fit the Poisson model and conduct overall goodness-of-fit tests
xi: poisson reactors i.type sex i.age, exp(par)
poisgof
poisgof, pearson
  * refit the model using -glm- to get a wider range of diagnostic parameters
xi: glm reactors i.type sex i.age, lnoff(par) fam(poisson) link(log)
  * compute predicted values, residuals and Cook's distance
predict mu
predict dev, dev
predict pear, pearson
predict cook, c
predict ans, a
format mu dev pear cook %6.3f
sort dev
  * list the largest and smallest residuals and the largest Cook's distances
listblck obs type sex age par reactors mu dev cook in 1/5, noobs
listblck obs type sex age par reactors mu dev cook in -5/-1, noobs
sort cook
listblck obs type sex age par reactors mu dev cook in -5/-1, noobs
  * generate a normal probability plot of Anscombe residuals
qnorm ans

  * Example 18.3 Negative binomial regression
  * open the tb_real dataset
set more off
use c:\ver\data\tb_real.dta, clear
  * fit the negative binomial model and conduct overall goodness-of-fit tests
xi: nbreg reactors i.type sex i.age, exp(par)
  * refit the model using -glm- to get a wider range of diagnostic parameters
xi: glm reactors i.type sex i.age, lnoff(par) fam(nbinomial 1.740375) link(log)
  * estimate the amount of overdispersion
sum reactors
di 1.46*1.74
  * compute predicted values, residuals and Cook's distance
predict mu
predict dev, dev
predict pear, pearson
predict cook, c
predict ans, a
format mu dev pear cook %6.3f
sort dev
  * list the largest and smallest residuals and the largest Cook's distances
listblck obs type sex age par reactors mu dev cook in 1/5, noobs
listblck obs type sex age par reactors mu dev cook in -5/-1, noobs
```

```
sort cook
listblck obs type sex age par reactors mu dev cook in -5/-1, noobs
  * compute the deviance chi-square goodness-of-fit (sort-of by manually)
gen dev_sq=dev^2
sum dev_sq
di 134*0.7415
di chi2(127, 99.36)

  * Example 18.4 Zero inflated negative binomial regression
  * open the fecal egg count dataset
use "c:\ver\data\fec.dta", clear
set more off
  * fit a zero-inflated model
xi:zinb fec i.lact i.season i.province  past_lact man_heif man_lact if tx==0, /*
   */ inflate(i.lact i.herd) cl(cow) nolog
  * refit the model with regular standard errors to allow the computation of the Vuong statistic
xi:zinb fec i.lact i.season i.province  past_lact man_heif man_lact if tx==0, /*
   */ inflate(i.lact i.herd) nolog vuong
```

19 Modelling survival data

```
  * Example 19.2 Actuarial and Kaplan-Meier survival functions
  * open the calf pneumonia dataset
use "c:\ver\data\calf_pneu.dta", clear
  * generate an actuarial life table (survival and hazard)
ltable days died, interval(15)
ltable days died, interval(15) hazard
  * generate a Kaplan-Meier estimate of the survival function
stset days, f(died)
stslist

  * Example 19.3 Comparing survival functions
  * open the calf pneumonia dataset
use "c:\ver\data\calf_pneu.dta", clear
  * compute and graph the Kaplan-Meier survival functions
stset days, f(died)
sts graph, by(stock) noborder sa(c:\ver\ch19\fig19.6.gph, replace)
sts test stock, detail
sts test stock, w
sts test stock, tw
sts test stock, peto

  * Example 19.4 Cox proportional hazards model
  * open the prostaglandin trial dataset
use "c:\ver\data\pgtrial.dta", clear
  * fit a Cox proportional hazards model
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr
  * refit the model to obtain hazard ratios
stcox
```

```
  * Example 19.5 Time varying covariates
  * open the ISA risk factor dataset
use "c:\ver\data\isa_risk.dta", clear
  * compute a time variable for # of days since Apr. 1st
gen days=date-apr01_97
stset days , f(case) id(sitepen)
  * compute the days to the first outbreak at each site
gen days2=days if case==1
sort site
quietly by site: egen first_da=min(days2)
  * list the data for 3 netpens before modification
sort site netpen _d _t0
listblck site netpen _t0 _t _d if site==19 & netpen>=39 & netpen<=56
  * list the data after modification for a time varying covariate
stsplit pos, after(time=first_da) at(0)
replace pos=pos+1
sort site netpen _d _t0
listblck site netpen _t0 _t _d pos if site==19 & netpen>=39 & netpen<=56
stcox pos

  * Example 19.6 Time varying covariates - continuous
  * open the prostaglandin dataset
use "c:\ver\data\pgtrial.dta", clear
  * fit a model with treatment as a tvc
stset dar, f(preg)
stcox tx, tvc(tx) texp(log(_t))

  * Example 19.7 Schoenfeld residuals
  * open the pgtrial dataset
use "c:\ver\data\pgtrial.dta", clear
  * fit the Cox proportional hazards model and save the Schoenfeld and scaled /*
    */ Schoenfeld residuals
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr sca(sca*) sch(sch*)
  * generate a smoothed graph of scaled Schoenfeld residuals vs log(time)
stphtest, log plot(lact) yline(0) l2("Scaled Schoenfeld - lactation") /*
   */ l1(" ") b2(" ") b1("Time (log scale)") xlab(1 5 20 50 100 300) /*
   */ gap(4) sa(c:/ver/ch19/fig19_14.gph, replace)
stphtest, detail log

  * Example 19.8 Independence of censoring
  * open the pgtrial dataset
use "c:\ver\data\pgtrial.dta", clear
  * fit the Cox proportional hazards model with -tx- as a time varying covariate
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr tvc(tx) texp(ln(_t))
  * recode censoring to assume complete positive correlation
preserve
replace preg=1 if preg==0
stset dar, f(preg)
```

```
xi: stcox i.herd tx lact thin, nohr tvc(tx) texp(ln(_t))
restore
  * recode censoring and dar to assume complete negative correlation
preserve
replace dar=400 if preg==0
replace preg=1 if preg==0
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr tvc(tx) texp(ln(_t))
restore

  * Example 19.9 Cox goodness-of-fit tests
  * open the prostaglandin trial dataset
set more off
use "c:\ver\data\pgtrial.dta", clear
  * fit a Cox proportional hazards model
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr mg(mgale)
stcoxgof, gr(5)
drop mgale
  * repeat for model with tx as TVC
xi: stcox i.herd tx lact thin, nohr mg(mgale) tvc(tx) texp(ln(_t))
stcoxgof, gr(5)
  * Fig 19.15 a and b  Evaluating overall fit - Cox model (2)
  * open the prostaglandin trial dataset
set more off
use "c:\ver\data\pgtrial.dta", clear
  * fit a Cox proportional hazards model
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr mg(mgale)
  * compute modified Cox-Snell residuals
predict cs, csnell
replace cs=cs+0.693 if preg==0
label var cs "Modified Cox-Snell residuals"
  * re "stset" the data and fit the new Cox model
stset cs, f(preg)
sts gen ch=na
sort cs
graph ch cs cs, c(ll) s(..) xlab(0 1 2 3 4) ylab(0 1 2 3 4) t1("no time varying covariate") /*
   */ l2("Cumulative hazard") gap(4) pen(11) sa("c:\ver\ch19\fig19_16a.gph", replace)
  * repeat for model with tx as TVC
use "c:\ver\data\pgtrial.dta", clear
  * fit a Cox proportional hazards model
stset dar, f(preg)
xi: stcox i.herd tx lact thin, nohr mg(mgale) tvc(tx) texp(ln(_t))
  * compute modified Cox-Snell residuals
predict cs, csnell
replace cs=cs+0.693 if preg==0
label var cs "Modified Cox-Snell residuals"
  * re "stset" the data and fit the new Cox model
stset cs, f(preg)
```

```
sts gen ch=na
sort cs
graph ch cs cs, c(ll) s(..) xlab(0 1 2 3 4) ylab(0 1 2 3 4) /*
   */ t1("Treatment as a time varying covariate") l2("Cumulative hazard") gap(4) pen(11) /*
   */ sa("c:\ver\ch19\fig19_16b.gph", replace)
  * combine the two graphs
graph using "c:\ver\ch19\fig19_16a.gph" "c:\ver\ch19\null.gph" /*
   */ "c:\ver\ch19\fig19_16b.gph", sa("c:\ver\ch19\fig19_16.gph", replace)

  * Example 19.10 Exponential model
  * open the prostaglandin trial dataset
set more off
use "c:\ver\data\pgtrial.dta", clear
  * rescale the lactation variable
gen lact2=lact-1
  * fit an exponential model
stset dar, f(preg) id(cow)
xi: streg i.herd tx lact2 thin, nohr dist(exp)
  * derive a step function estimate of the baseline hazard
sum _t
stsplit day, at(1(5)346)
gen day1= 0<=day & day<20
gen day2= 20<=day & day<40
gen day3= 40<=day & day<80
gen day4= 80<=day & day<120
gen day5= 120<=day
xi: streg i.herd tx lact2 thin day2 day3 day4 day5, nohr dist(exp)
gen h0=exp(_b[_cons] + _b[day2]*day2 + _b[day3]*day3 +/*
   */ _b[day4]*day4 + _b[day5]*day5)
graph h0 day if day<=200, c(J) s(i) sort l1(" ") l2("Baseline hazard") /*
   */ b2(" ") b1("Time (days)") xlab ylab sa("c:\ver\ch19\fig19_20.gph", replace)

  * Example 19.11 Weibull model
  * open the prostaglandin trial dataset
set more off
use c:\ver\data\pgtrial.dta, clear
  * rescale the lactation variable
gen lact2=lact-1
  * fit a Weibull model (no time varying covariates)
stset dar, f(preg) id(cow)
xi: streg i.herd tx lact2 thin, nohr dist(weib)
  * modify the model making -tx- a TVC with log(time)
  * results not shown in book
stsplit day, at(1(1)346)
stset dar, f(preg) id(cow)
gen tx_day=tx*day
gen tx_lnday=tx*(log(day))
xi: streg i.herd tx lact2 thin tx_lnday, dist(weib) tr
```

```
  * Example 19.12 Lognormal model
  * open the prostaglandin trial dataset
use "c:\ver\data\pgtrial.dta", clear
  * rescale the lactation variable
gen lact2=lact-1
  * fit a lognormal model (results as TR)
stset dar, f(preg) id(cow)
xi: streg i.herd tx lact2 thin, dist(lognormal) tr
  * results as coefficients
streg
  * modify the model making -tx- a TVC with log(time)
  * results not shown in book
stsplit day, at(1(1)346)
stset dar, f(preg) id(cow)
gen tx_day=tx*day
gen tx_lnday=tx*(log(day))
xi: streg i.herd tx lact2 thin tx_lnday, dist(lognormal) tr

  * Example 19.13 Multiple failure time data
  * the first section of this program generates hypothetical data
  * the second section analyses those data
  * generate some hypothetical multiple failure time data
  * first generate the basic dataset with survival times
  * capture log close
  * log using "c:\ver\ch19\ex19_13gen.log", text replace
clear
set obs 2000
set more off
gen id=_n
gen e=1.5
rndpoix e
replace xp=5 if xp>5
gen X=(_n>1000)
  * expand the dataset so there is 1 record per event
expand xp
sort id
quietly by id: gen evnum=_n
gen outcome=(xp>0)
  * make the last event a censoring event (if more than 1 event)
  * sort id evnum
  * quietly by id: replace outcome=0 if _n==_N & _n>1
  * generate days to outcome
gen days=round((100*uniform()),1)
quietly by id:replace days=days[_n-1]+round((40*uniform()),1) if _n==2
quietly by id:replace days=days[_n-1]+round((30*uniform()),1) if _n==3
quietly by id:replace days=days[_n-1]+round((20*uniform()),1) if _n==4
quietly by id:replace days=days[_n-1]+round((10*uniform()),1) if _n==5
replace days=round((0.8*days ),1) if X==1
replace days=days+1 if days==0
quietly by id: replace days=days+1 if days==days[_n-1] & _n>1
```

```
  * compute the final exit time for each subject
sort id evnum
quietly by id: gen exit=days[_N]
  * display a bit of the data and save the file
format id evnum days exit outcome %5.0f
listblck id evnum days exit outcome in 1/10
  * save "c:\ver\ch19\ex19_13.dta", replace
  * generate Anderson-Gill data
use "c:\ver\ch19\ex19_13.dta", clear
sort id evnum
gen start=0
quietly by id: replace start=days[_n-1] if _n>1
gen end=days
stset end, f(outcome) id(id) enter(start) exit(exit)
format id evnum days exit outcome _t0 _t _d %5.0f
listblck id evnum start end exit outcome _t0 _t _d in 1/10, noobs
save "c:\ver\ch19\ex19_13ag.dta", replace
  * generate PWP data
use "c:\ver\ch19\ex19_13.dta", clear
sort id evnum
gen start=0
gen end=days
quietly by id: replace end=days-days[_n-1] if _n>1
stset end, f(outcome) enter(start) exit(exit)
format id evnum days exit outcome _t0 _t _d %5.0f
listblck id evnum start end exit outcome _t0 _t in 1/10, noobs
save "c:\ver\ch19\ex19_13pwp.dta", replace
  * generate WLW data
use "c:\ver\ch19\ex19_13.dta", clear
sort id evnum
quietly by id: gen last=_n==_N
expand (6-evnum) if last
sort id evnum
quietly by id: replace evnum=_n
gen start=0
gen end=days
replace outcome=0 if evnum>xp
stset days, f(outcome)
format id evnum days exit outcome _t0 _t _d %5.0f
listblck id evnum start end outcome _t0 _t _d in 1/20, noobs
save "c:\ver\ch19\ex19_13wlw.dta", replace

  * Example 19.13 Multiple failure time data - analysis of data
  * open the hypothetical dataset
capture log close
log using "c:\ver\ch19\ex19_13mev.log", text replace
  * fit an Anderson-Gill model to the data
use "c:\ver\ch19\ex19_13ag.dta", clear
stset end, f(outcome) id(id) enter(start) exit(exit)
stcox X, nohr efron robust nolog
```

```
  * fit a Prentice Williams Peterson (conditional risk set) model
use "c:\ver\ch19\ex19_13pwp.dta", clear
stset end, f(outcome) enter(start) exit(exit)
stcox X, nohr efron robust nolog
  * fit Wei, Lin, Weissfeld model to the data
use "c:\ver\ch19\ex19_13wlw.dta", clear
stset days, f(outcome)
stcox X, nohr efron strata(evnum) cluster(id) nolog
log close
```

```
  * Example 19.14 Individual frailty models - hypothetical data
  * the first part of this program generates hypothetical data the second part analyses those data
  * part 1 - generating survival data constant hazard with individual frailty
clear
set obs 1000
set seed 12345
set more off
gen id=_n
gen haz=0.01
gen day=.
gen fail=0
gen failday=.
local i = 1
while `i' <= 100 {
display day
gen rand=uniform()
quietly replace day=`i'
quietly replace haz=.02 if day>=20 & day<40
quietly replace haz=0.005 if day>=40 & day<60
quietly replace haz=0.0025 if day>=60
quietly replace failday=`i' if rand<haz & fail==0
quietly replace fail=1 if rand<haz & fail==0
drop rand
local i = `i' + 1 }
replace failday=100 if fail==0
save "c:\ver\ch19\ex19_14.dta", replace
```

```
  * Example 19.14 Individual frailty models - analysis of data
  * open the hypothetical dataset
use "c:\ver\ch19\ex19_14.dta", clear
stset failday, f(fail)
sts gen bh=h
ksm bh failday, gen(bhsm) bw(0.1)
label var bhsm "Empirical smoothed hazard"
  * fit a Weibull model and generate the hazard function
streg, dist(weib)
stcurve, haz outfile("c:\ver\ch19\temp.dta", replace)
merge using "c:\ver\ch19\temp.dta"
rename haz1 haz_weib
label var haz_weib "Weibull hazard"
```

```
drop _merge
  * fit a log-normal model and generate the hazard function
streg, dist(lognormal)
stcurve, haz outfile("c:\ver\ch19\temp.dta", replace)
merge using "c:\ver\ch19\temp.dta"
rename haz1 haz_logn
label var haz_logn "Log-normal hazard"
drop _merge
  * fit a Weibull - gamma model and generate the uncond. hazard
streg, dist(weib) fr(gamma)
stcurve, haz outfile("c:\ver\ch19\temp.dta", replace) uncond
merge using "c:\ver\ch19\temp.dta"
rename haz1 haz_fr
label var haz_fr "Weibull / Gamma hazard"
drop _merge
  * graph the various frailty functions
sort _t
graph bhsm haz_weib haz_logn haz_fr _t, pen(1111) xlab ylab /*
   */ l2("Hazard") gap(5) b2(" ") b1("Time") s(iiii) c(J[l] l[_] l[-] l[.]) /*
   */ sa("c:\ver\ch19\fig19_21.gph", replace)

  * Example 19.15 Weibull model - Gamma frailty
  * open the prostaglandin trial dataset
set more off
use "c:\ver\data\pgtrial.dta", clear
  * rescale the lactation variable
gen lact2=lact-1
  * fit a Weibull model (no time varying covariates)
stset dar, f(preg) id(cow)
xi: streg i.herd tx lact2 thin, nohr dist(weib) fr(gamma)

  * Example 19.16 Weibull model - shared gamma frailty
  * open the prostaglandin trial dataset
set more off
use "c:\ver\data\pgtrial.dta", clear
  * rescale the lactation variable
gen lact2=lact-1
  * fit a Weibull model (no time varing covariates)
stset dar, f(preg) id(cow)
streg tx lact2 thin, nohr dist(weib) fr(gamma) shared(herd)
```

CHAPTER 20 INTRODUCTION TO CLUSTERED DATA

```
  * Example 20.1 Clustering in a continuous data model
  * the first part of this program file generates the simulated data
  * the second part analyses the data
  * Part 1 - simulation
clear
pause on
```

```
set seed 12345
  * create 100 herds and set herd sizes
set obs 100
gen nc=round((40+10*invnorm(uniform())),1) in 1/50
replace nc=round((200+50*invnorm(uniform())),1) in 51/100
sort nc
gen herd=_n
  * for each herd compute herd average milk production compute herd level milk production
gen milk_h= 30 + 7*invnorm(uniform())
save "c:\ver\ch20\ex20_1_herds.dta", replace
  * build a dataset with -X- as a herd level variable first assign -X- as a herd level variable to
  * 1/2 the herds
gen rand=uniform()
sort rand
gen X=(_n>50)
drop rand
  * expand the dataset to 1 record per cow
expand nc
  * generate individual cow milk production values
gen milk=milk_h + (5*X) + 8*invnorm(uniform())
replace milk=(milk*-1) if milk<0
save "c:\ver\ch20\ex20_1_herd.dta", replace
  * build a dataset with -X- as a cow level variable 1/2 animals in each herd have X=1
use "c:\ver\ch20\ex20_1_herds.dta", clear
  * expand the dataset to 1 record per cow
expand nc
  * assign cows to be -X- (0/1)
gen rand=uniform()
sort herd rand
drop rand
quietly by herd: gen X=(_n>0.5*nc)
  * compute cow level milk production (5kg higher if X = 1)
gen milk=milk_h + (5*X) + 8*invnorm(uniform())
save "c:\ver\ch20\ex20_1_cow.dta", replace
  * Part 2 - analysis
  * open the simulated data with -X- as a herd level variable
use "c:\ver\ch20\ex20_1_herd.dta",clear
  * fit an ordinary linear regression
reg milk X
  * fit a linear mixed model with herd as the cluster
xtreg milk X, i(herd) re
  * collapse the data to the herd level and fit an ordinary reg.
collapse (mean) milk X, by(herd)
reg milk X
  * open the simulated data with -X- as a cow level variable
use "c:\ver\ch20\ex20_1_cow.dta",clear
  * fit an ordinary linear regression
reg milk X
  * fit a linear mixed model with herd as the cluster
xtreg milk X, i(herd) re
```

```
  * Example 20.2 Clustering in a binomial model
  * the first part of this program file generates the simulated data
  * the second part analyses the data
  * Part 1 - simulation
clear
pause on
set seed 12345
  * create 100 herds and set herd sizes
set obs 100
gen nc=round((40+10*invnorm(uniform())),1) in 1/50
replace nc=round((200+50*invnorm(uniform())),1) in 51/100
sort nc
gen herd=_n
  * for each herd compute logit of disease prevalence
gen logit_dis_p=(-1.4) + invnorm(uniform())
save "c:\ver\ch20\ex20_2_herds.dta", replace
  * build a dataset with -X- as a herd level variable
gen rand=uniform()
sort rand
gen X=(_n>50)
drop rand
  * add the effect of -X-
replace logit_dis_p=logit_dis_p + (0.69*X)
  * expand the dataset to 1 record per cow
expand nc
  * assign cows to be diseased (0/1)
gen dis_p=exp(logit_dis_p)/(1+exp(logit_dis_p))
gen Y=uniform()<dis_p
save "c:\ver\ch20\ex20_2_herd.dta", replace
  * build a dataset with -X- as a cow level variable 1/2 animals in each herd have X=1
use "c:\ver\ch20\ex20_2_herds.dta", clear
  * expand the dataset to 1 record per cow
expand nc
  * assign cows to be -X- (0/1)
gen rand=uniform()
sort herd rand
quietly by herd: gen X=(_n>0.5*nc)
drop rand
  * add the effect of -X-
replace logit_dis_p=logit_dis_p + (0.69*X)
  * assign cows to be diseased or not (0/1)
gen dis_p=exp(logit_dis_p)/(1+exp(logit_dis_p))
gen Y=uniform()<dis_p
save "c:\ver\ch20\ex20_2_cow.dta", replace
  * Part 2 - analysis
  * open the dataset -X- as a herd level variable
use "c:\ver\ch20\ex20_2_herd.dta", clear
preserve
collapse (mean) Y X nc, by(herd)
graph Y,bin(20) xlab ylab l1(" ") l2("Proportion of herds") /*
```

```
   */ gap(4) b1("Prevalence of disease") b2(" ") /*
   */ sa("c:\ver\ch20\fig20.2.gph", replace)
restore
  * fit an ordinary logistic regression
logit Y X
  * fit a random effects model with herd as the cluster (note results in text differ slightly since
  * they were obtained using -gllamm- in Version 8 of Stata)
xtlogit Y X, i(herd)
  * open the dataset -X- as a cow level variable
use "c:\ver\ch20\ex20_2_cow.dta", clear
  * fit an ordinary logistic regression
logit Y X
  * fit a a random effects model with herd as the cluster (note results in text differ slightly since
  * they were obtained using -gllamm- in Version 8 of Stata)
xtlogit Y X, i(herd)

  * Example 20.3 Fixed effects linear regression model
set more off
  * open the simulated data from Example 20.1
use "c:\ver\ch20\ex20_1_cow.dta", clear
  * fit a fixed effects linear regression model
xi:reg milk X i.herd

  * Example 20.4 Stratified analysis and fixed effects logistic regression model
set more off
  * open the simulated data from Example 20.2
use "c:\ver\ch20\ex20_2_cow.dta", clear
  * carry out a stratified analysis
cc Y X, by(herd)
  * fit a fixed effects logistic regression model
xi:logit Y  X i.herd
```

21 MIXED MODELS FOR CONTINUOUS DATA

The program files for this chapter are SAS® program files (all other chapters are Stata® program (-do-) files.

```
  * Example 21.1 Variance components and random effects
libname ver 'c:\ver\data';
proc mixed data=ver.scc40_2level;
 class herdid;
 model t_lnscc=/s;
 random herdid;
run;

  * Example 21.2 Mixed model estimates for 2-level somatic cell count data
libname ver 'c:\ver\data';
proc mixed data=ver.scc40_2level covtest;
```

```
 class t_season herdid;
 model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
 random herdid;
run;
  * calculation of ICC from formula 21.10;
proc glm data=ver.scc40_2level;
 class herdid;
 model t_lnscc=herdid;
run;

  * Example 21.4 Random slopes of -t_dim- for somatic cell count data
libname ver 'c:\ver\data';
proc mixed data=ver.scc40_2level covtest;
 class t_season herdid;
 model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s;
 random intercept t_dim / type=un subject=herdid;
run;

  * Example 21.5 Random slopes of -c_heifer- for somatic cell count data
libname ver 'c:\ver\data';
proc mixed data=ver.scc40_2level covtest;
 class t_season herdid;
 model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s;
 random intercept c_heifer / type=un subject=herdid;
run;

  * Example 21.7 Herd random effect for 2-level somatic cell count data
libname ver 'c:\ver\data';
  * model with no herd effect;
proc mixed data=ver.scc40_2level;
 class t_season herdid;
 model t_lnscc=c_heifer t_season t_dim h_size;
run;
  * analysis to demonstrate 0.25 within the profile likelihood CI;
proc mixed data=ver.scc40_2level covtest;
 class t_season herdid;
 model t_lnscc=c_heifer t_season t_dim h_size;
 random herdid;
 parms (0.25) (1) / eqcons=1;
run;

  * Example 21.9 Box-Cox analysis for somatic cell count data
libname ver 'c:\ver\data';
data boxcox;
 set ver.scc40_2level;
 scc=exp(t_lnscc);
 do lambda=-1,-0.5,-0.33,-0.25,-0.1,0,0.25,0.33,0.5,1;
```

```
   if lambda ne 0 then y=(scc**lambda-1)/lambda;
   else y=t_lnscc;
   output;
  end;
proc sort;
  by lambda;
run;
ods listing close; *remove to see all the analyses;
proc mixed data=boxcox;
  class t_season herdid;
  model y=c_heifer t_season t_dim h_size;
  random herdid;
  ods output fitStatistics=fits;
  by lambda;
run;
ods listing;
data profile;
  set fits;
  if index(descr,"-2")>0;
  n=2178;
  lnsccmean=4.7569865;
  logl=-0.5*value;
  prof_logl=logl+n*(lambda-1)*lnsccmean;
  drop descr;
proc print data=profile;
  var lambda logl prof_logl;
run;
```

*** Example 21.10 Repeated measures analysis for somatic cell count data**

```
libname ver 'c:\ver\data';
proc mixed data=ver.scc_40 covtest noclprint=15;
  class t_season herdid cowid test;
  model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
  random herdid;
  repeated test / subject=cowid type=cs;
run;
proc mixed data=ver.scc_40 covtest noclprint=15;
  class t_season herdid cowid test;
  model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
  random herdid;
  repeated test / subject=cowid type=ar(1);
run;
proc mixed data=ver.scc_40 covtest noclprint=15;
  class t_season herdid cowid test;
  model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
  random herdid;
  repeated / subject=cowid type=sp(pow)(t_dim);
```

```
run;
proc mixed data=ver.scc_40 covtest noclprint=15;
 class t_season herdid cowid test;
 model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
 random herdid;
 repeated test / subject=cowid type=arma(1,1);
run;
proc mixed data=ver.scc_40 covtest noclprint=15;
 class t_season herdid cowid test;
 model t_lnscc=c_heifer t_season t_dim h_size / ddfm=satterth s cl;
 random intercept / subject=herdid;
 repeated test / subject=cowid(herdid) type=toep;
  * note: more efficient coding, to save computing time;
run;
```

22 MIXED MODELS FOR DISCRETE DATA

```
  * Example 22.1 Random effects logistic
  * open the pig respiratory disease dataset
use c:\ver\data\pig_adg, clear
  * compute a dichotomous variable for ar greater than 1
gen ar_g1=ar>1
  * determine the unconditional association between pn and ar_g1
cc pn ar_g1
  * logistic regression with random farm effects
gllamm pn ar_g1, fam(binom) link(logit) i(farm) trace adapt

  * Example 22.3 Random effects Poisson
  * open the tb_real dataset
set more off
use "c:\ver\data\tb_real.dta", clear
  * fit a random effects model - normal herd variance - gllamm
gen logpar=log(par)
xi:gllamm reactors i.type sex i.age, fam(pois) link(log) off(logpar) adapt trace i(farm_id) dots

  * Example 22.4 Random effects logistic regression
  * open the reu_cfs dataset
set more off
use "c:\ver\data\reu_cfs.dta", clear
  * fit a random effects logistic model - 3 levels (herd, cow, lactation) using gllamm
gllamm fscr ai heifer, fam(bin) link(logit) adapt trace i( cow herd) dots

  * Example 22.6 Checking ML estimation
  * open the reu_cfs dataset
set more off
capture log close
log using "c:\ver\ch22\ex22_6.log", text replace
use "c:\ver\data\reu_cfs.dta", clear
```

```
  * fit a random effects logistic model - varying # of quadrature points
gllamm fscr ai heifer , fam(bin) link(logit) adapt trace i( cow herd) dots nip(6)
gllamm fscr ai heifer , fam(bin) link(logit) adapt trace i( cow herd) dots nip(8)
gllamm fscr ai heifer , fam(bin) link(logit) adapt trace i( cow herd) dots nip(10)
gllamm fscr ai heifer , fam(bin) link(logit) adapt trace i( cow herd) dots nip(12)
```

```
  * Example 22.7 Likelihood ratio tests for random effects logistic regression
  * open the reu_cfs dataset
set more off
use "c:\ver\data\reu_cfs.dta", clear
  * fit reduced random effects logistic models using gllamm
gllamm fscr ai heifer, fam(bin) link(logit) adapt trace i(herd) dots
gllamm fscr ai heifer, fam(bin) link(logit) adapt trace i(cow) dots
```

```
  * Example 22.9 Overdispersion parameter for artificial binomial data
  * generate the data
clear
set obs 10
generate pos=0
replace pos=20 if _n>5
  * analyse data by generalised linear model
glm pos, fam(binomial 20) link(logit)
```

23 Alternative approaches to dealing with clustered data

```
  * Example 23.1 Fixed effects model - pig respiratory diseases
  * open the pig respiratory disease dataset
use c:\ver\data\pig_adg, clear
  * compute a dichotomous variable for ar score greater than 1
gen ar_g1=ar>1
  * fit a fixed effects model for pn
xi: logit pn ar_g1 i.farm
```

```
  * Example 23.2 Overdispersion factor - pig respiratory diseases
  * open the pig respiratory disease dataset
use c:\ver\data\pig_adg, clear
  * compute a dichotomous variable for ar score greater than 1
gen ar_g1=ar>1
  * compute group identifiers
gen group=ar_g1*100+farm
  * collapse data to group level
collapse ar_g1 (sum) pn (count) n=pn, by(group)
  * fit a generalised linear model for pn with overdispersion factor
glm pn ar_g1, fam(bin n) link(logit) scale(x2)
```

```
  * Example 23.3 Robust standard errors - pig respiratory disease
  * open the pig respiratory disease dataset
use c:\ver\data\pig_adg, clear
  * compute a dichotomous variable for ar greater than 1
```

```
gen ar_g1=ar>1
  * fit the simple model with robust SE clustered on farm
logit pn ar_g1, cluster(farm)

  * Example 23.4 Robust standard errors - somatic cell counts
  * open the somatic cell count dataset
use "c:\ver\data\scc_10.dta", clear
  * compute a dichotomous variable for heifer
gen c_heifer=(c_prty3==1)
  * fit a multiple linear regression (ignoring clustering)
xi: reg t_lnscc h_sz c_heifer i.t_seas t_dim
  * fit a multiple linear regression with robust SE clustered on cows
xi: reg t_lnscc h_sz c_heifer i.t_seas t_dim, robust cluster(cowid)
  * fit a multiple linear regression with robust SE clustered on herds
xi: reg t_lnscc h_sz c_heifer i.t_seas t_dim, robust cluster(herdid)

  * Example 23.6 - GEE pig respiratory disease data
  * open the pig respiratory disease dataset
use c:\ver\data\pig_adg, clear
  * compute a dichotomous variable for ar greater than 1
gen ar_g1=ar>1
  * run GEE estimation with compound symmetry working correlation structure
xtgee pn ar_g1, i(farm) fam(bin) link(logit) robust
xtcorr , compact

  * Example 23.7 GEE - somatic cell count data
  * open the SCC dataset
use c:\ver\data\scc_10, clear
  * compute a dichotomous variable for heifer
gen c_heifer=(c_prty3==1)
  * run GEE estimation with compound symmetry working correlation structure
xi: xtgee t_lnscc h_sz c_heifer i.t_seas t_dim, i(cowid) fam(gaus) link(ident) corr(exch) robust
xtcorr, compact
  * run GEE estimation with autoregressive (ar1) working correlation structure
xi: xtgee t_lnscc h_sz c_heifer i.t_seas t_dim, i(cowid) t(test) fam(gaus) link(ident) corr(ar1)
robust
xtcorr, compact
  * run GEE estimation with unstructured working correlation structure
xi: xtgee t_lnscc h_sz c_heifer i.t_seas t_dim, i(cowid) t(test) fam(gaus) link(ident) corr(unst)
robust
xtcorr, compact
```

24 META-ANALYSIS

```
  * Example 24.3 - Fixed and random effects meta-analyses
  * open the rBST - milk dataset
use "c:\ver\data\bst_milk.dta", clear
  * meta-analyses of milk production
meta diff se
```

```
  * open the rBST - mastitis dataset
use "c:\ver\data\bst_mast.dta", clear
  * meta analyses of mastitis data
meta rr cilow cihigh, ci eform

  * Example 24.4 - Forest plot
  * open the rBST - mastitis dataset
use "c:\ver\data\bst_mast.dta", clear
  * create a study label variable
gen str20 sl="Group = "+string(group)
  * fit both fixed and random effects models and
  * generate a forest plot from the fixed effects model
meta rr cilow cihigh, ci eform graph(f) id(sl) cline xline(1) xlab(0.33 0.5 0.75 1 2 4) /*
   */ ltrunc(0.3) b2("Risk ratio - clinical mastitis") rtrunc(6) /*
   */ sa("c:\ver\ch24\fig24_1.gph", replace)

  * Example 24.5 - Stratified meta-analysis
  * open the rBST - milk dataset
use "c:\ver\data\bst_milk.dta", clear
  * meta-analyses of milk production - primiparous
meta diff se if parity==1
  * meta-analyses of milk production - all
meta diff se if parity==2
  * meta-analyses of milk production - multiparous
meta diff se if parity==3

  * Example 24.6 - Meta-regression
  * open the rBST - milk dataset
use "c:\ver\data\bst_milk.dta", clear
  * meta-regression analysis
xi: metareg diff i.parity if se~=. , wsse(se) bsest(mm)
test _Iparity_2 _Iparity_3
metareg diff dur if se~=. , wsse(se) bsest(mm)
metareg diff dose_day if se~=. , wsse(se) bsest(mm)

  * Example 24.7 - Publication bias
  * open the rBST - mastitis dataset
use "c:\ver\data\bst_mast.dta", clear
  * generate a funnel plot
  * compute Begg's and Egger's tests for publication bias
metabias rr cilow cihigh, ci gr(b) t2("Clin Mast - Risk") /*
   */ saving("c:\ver\ch24\fig24_2.gph",replace)

  * Example 24.8 - Influential studies
  * open the rBST - mastitis dataset
use "c:\ver\data\bst_mast.dta", clear
  * create a study label variable
gen str20 sl="Group = "+string(group)
  * derive estimates of the log(rr) and its SE
gen lnrr=ln(rr)
```

```
gen selow=(lnrr-ln(cilow))/1.96
gen sehigh=(ln(cihigh)-lnrr)/1.96
gen seavg=(selow+sehigh)/2
  * generate an influence plot
metainf lnrr seavg, id(sl) t2("Clin Mast - Risk") /*
   */ saving("c:\ver\ch24\fig24_3.gph",replace)
```

25 ECOLOGIC AND GROUP LEVEL STUDIES

There are no program files for this chapter.

26 A STRUCTURED APPROACH TO DATA ANALYSIS

There are no program files for this chapter.

GLOSSARY AND TERMINOLOGY

There is considerable variation in terminology and methods of presenting data among epidemiology texts and other information sources. In general, the terminology and data layouts used in this book will conform to those used in Modern Epidemiology, 2d edition (Rothman and Greenland, 1998).

GT.1 DATA LAYOUT

The outcome variable is listed in the rows of the table, the predictor variable is listed in the columns.

Risk calculations (2X2 table)

	Exposure		
	Exposed	Non-exposed	
Diseased	a_1	a_0	m_1
Non-diseased	b_1	b_0	m_0
	n_1	n_0	n

where:

a_1 = the number of subjects that have both the disease and the risk factor.
a_0 = the number of subjects that have the disease but not the risk factor.
b_1 = the number of subjects that have the risk factor but do not have the disease.
b_0 = the number of subjects that have neither the disease nor the risk factor.
m_1 = the number of diseased subjects.
m_0 = the number of non-diseased subjects.
n_1 = the number of exposed subjects.
n_0 = the number of non-exposed subjects.
n = the number of study subjects.

In general, no distinction is made between values derived from a sample and population values as it is usually easy to determine what is being referred to from the context. In select situations where the distinction is necessary, upper-case letters (*eg* A_1) will be used for population values and lower case (*eg* a_1) for sample values.

Rate calculations (2X2 table)

Here, subject-time replaces the number of non-diseased.

	Exposure		
	Exposed	Non-exposed	
Number of cases	a_1	a_0	m_1
Animal-time at risk	t_1	t_0	t

where:

a_1 = the number of cases of disease in the exposed group.
a_0 = the number of cases of disease in the non-exposed group.
t_1 = the animal-time accumulated in the exposed group.
t_0 = the animal-time accumulated in the non-exposed group.
t = the total animal-time accumulated by the study subjects.

Diagnostic tests (2X2 tables)

Gold standard layout

	Test result		
	Positive	Negative	
Disease positive	a	b	m_1
Disease negative	c	d	m_0
	n_1	n_0	n

Note The marginals are the same as for risk calculations; the inner cell values are denoted as a, b, c, d.

Test comparison layout

	Test 2 positive	Test 2 negative	Total
Test 1 positive	n_{11}	n_{12}	$n_{1.}$
Test 1 negative	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Correlated data

Matched-pair case-control data layout

		Control pair		Case totals
		Exposed	Non-exposed	
Case pair	Exposed	t	u	$t+u = a_1$
	Non-exposed	v	w	$v+w = a_0$
	Control totals	$t+v = b_1$	$u+w = b_0$	

Note If pair-matching is used in a cohort study, the same format is used but the case (rows)-control(columns) status is replaced by exposed (rows) non-exposed (columns) and the exposure status is replaced by disease status.

Significant digits

Throughout the text, data are often presented with more significant digits than normally would be warranted. This is done for clarity and to avoid rounding errors.

GT.2 Multivariable models

In general, multivariable models will be presented as follows, with explicit subscripting (*eg* for observation number) used only if absolutely necessary for clarity:

$$\text{outcome} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where the outcome may be a variety of parameters (*eg* for logistic regression outcome = ln(p/1-p) and k is the number of parameters in the model (excluding the intercept).

In some situations, βX or μ will be used to represent the entire right-hand side of the model (*ie* the linear predictor) to simplify presentation:

$$\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

The terms predictor, exposure, risk factor and independent variable will all be used to designate factors that 'cause' the outcome of interest, although in general we prefer to use one of the first two terms. These will be designated X.

The terms outcome and dependent variable will both be used for the response, but the former term is used most commonly. These will be designated Y.

GT.2.1 Multilevel models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{herd}(i)} + \varepsilon_i$$

Note For the sake of simplicity, a single index notation will be used for all multilevel data. The subscript i denotes the individual (lowest level) observation. In the example above, $u_{\text{herd}(i)}$ refers to the herd containing the i^{th} individual. If there are 40 herds, u could have one of 40 values. An alternative notation, used in some texts, has multiple indices such as $u_j + \varepsilon_{ij}$ where j refers to the herd and i to the i^{th} individual in the j^{th} herd.

GT.3 Glossary

Terms related to formulae and methods

a	number of cases
AF_e	attributable fraction in the exposed group
AF_p	attributable fraction in the population
AFT	accelerated failure time
AIC	Akaike's Information Criteria
AP	apparent prevalence
AUC	area under ROC curve
BIC	Bayesian Information Criteria (Schwartz Bayesian Criteria)
BLUP	best linear unbiased predictor

BUGS	Bayesian analysis using Gibbs sampling
c	constant (*eg* baseline hazard)
	cost of sampling
CCC	concordance correlation coefficient
chi2	chi-square (χ^2)
CI	confidence interval
corr (*Y*)	correlation matrix of *Y*
cov (*Y*)	covariance matrix of *Y*
covar	covariance
covar (+)/covar(-)	covariance in test positive (+)/negative (-) sample results
cp	cutpoint
Cp	Mallow's statistic
CV	coefficient of variation
$d_{ii'}$	distance between point i and i′
d_j	outcome events (failures) during the interval (actuarial life table) or number of events at time t_j (KM life table)
D	deviance statistic (-2*lnL)
	minimum number diseased
	duration
	disease
D-	subjects not having a specified disease/condition
D+	subjects having a specified disease/condition
DB	delta-beta
deff	design effect
df	degrees of freedom
e	2.71828 (natural number)
E	expected value (*eg* *E*(*Y*) = expected value of *Y*)
	exposure factor
E-	subjects not exposed
E+	exposed subjects
EV	extraneous variable
exp	expected cell number
	exponential function (*ie* $\exp(x) = e^x$)
F(*t*)	failure function
f(*t*)	probability density function
$f(\theta)$	prior distribution for θ (Bayesian analysis)
$f(\theta\|Y)$	posterior distribution for θ (Bayesian analysis)

FNF	false negative fraction
FPC	finite population correction
FPF	false positive fraction
G^2	likelihood ratio statistic
GEE	generalised estimating equations
GLM	generalised linear model
GLMM	generalised linear mixed model
h_i	leverage
$h(t)$	hazard function
$h_0(t)$	baseline hazard function
H*j*	distribution of host factor and/or subject time in stratum *j*
$H(t)$	cumulative hazard function
HR	hazard ratio
Hs	standard population distribution of host factor
HSe	herd sensitivity
HSp	herd specificity
i	observation counter
I	incidence rate
*I*e	expected incidence rate
ICC	intra-class correlation coefficient
ID	incidence rate difference
ID_G	incidence rate difference based on group means
I_{dir}	directly standardised rate
I_{ind}	indirectly standardised rate
IQR	interquartile range
IR	incidence rate ratio
IR_G	incidence ratio based on group-level data
Is	standard population incidence rates
j	designated for strata
	designator for categories
	designator for covariate patterns in a dataset
	designator for time intervals (actuarial life table) or time points (KM life table)
	sampling interval in systematic random sample
J	total number of *j*
k	cutpoint for herd-level testing (number of positives required for positive herd classification)
	number of predictors in a model (not including intercept)

KM	Kaplan-Meier (life table or survival model)
l_j	subjects at risk of failure at the start of the time interval (actuarial life table)
L	allowable error (one-half the length of a confidence interval)
L	likelihood function (*eg* $L(Y\|\theta)$
L_0	null or baseline likelihood function
L_{full}	likelihood function from full model
L_{red}	likelihood function from reduced (smaller) model
ln	natural log
lnL	ln (likelihood function)
log	natural log (also ln)
LR	likelihood ratio
LR_{cp}	likelihood ratio based on defined cutpoint(s)
LR_{cat}	likelihood ratio for defined category of result
LRT	likelihood ratio test
m	number of matched controls per case
	number of observations in a covariate pattern
	number of samples in a pooled sample
	number of subjects per cluster (group)
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
MSE	mean square error
n	number
	sample size
n'	adjusted sample size
N	population size
o	odds
obs	observed cell number
OR	odds ratio
OR_a	odds ratio - adjusted
OR(ABC)	odds ratio for factor ABC
OR(ABC|D)	odds ratio for factor ABC conditional on D
OR_c	odds ratio - crude
OR_j	stratum-specific odds ratio
OR_{MH}	Mantel-Haenszel adjusted odds ratio
OR_{sf}	odds ratio of sampling fractions
p	probability as in $p(D+\|E+)$ or $p(Y=1)$
p	proportion as in $\ln(p/1-p)$

p	shape parameter for Weibull distribution
p_j	probability of surviving interval j (actuarial life table) or survival at time t_j (KM life table)
P	P-value
P	prevalence
PA	population average
par	population at risk
par	parameter
PAR	population attributable risk
PD	prevalence difference
PE	prediction error
$\text{pl}(\lambda)$	profile likelihood function
PlSe	pooled-sample sensitivity
PlSp	pooled-sample specificity
PPV-	positive predictive value of a negative test
PR	prevalence ratio
PSU	primary sampling unit
PV	predictive value
PV-	negative predictive value
PV+	positive predictive value
q	$1\text{-}p$
q_j	risk of event during interval j (actuarial life table) or at time t_j (KM life table)
QIC	quasi-likelihood under the independence model information criterion
r	correlation coefficient (ρ also used)
r^2	squared correlation (R^2 also used)
r_i	raw residual
r_j	average number of subjects at risk during a time interval (actuarial life table) or at time t_j (KM life table)
r_{si}	standardised residual
r_{ti}	studentised residual
R	incidence risk
$R2$	coefficient of determination (r^2 also used)
RD	risk difference (also know as attributable risk)
REML	restricted maximum likelihood
res_p	Pearson residual
Rs	standard population incidence risk
ROC	receiver operating characteristics
RR	risk ratio (alternatively known as relative risk)

SD	standard deviation
SE	standard error
Se	sensitivity
Se_{corr}/Sp_{corr}	corrected *Se*/*Sp* based on cross-sectional validation
Se_{new}/Sp_{new}	*Se*/*Sp* of current test adjusted for *Se*/*Sp* of referent test
Se_p/Sp_p	*Se*/*Sp* in parallel interpretation of test results
Se_s/Sp_s	*Se*/*Sp* in series interpretation of test results
sf	sampling fraction
sf_{T+}/sf_{T-}	sampling fractions for cross-sectional validation
S_i	value of latent variable for individual *i*
SMR	standardised morbidity/mortality ratio
so	sampling odds
Sp	specificity
sr	sampling risk
SS	subject specific
$S(t)$	survivor function
Δt	length of period
t_j	time of event (KM life table)
t_{j-1}, t_j	time span in the interval (actuarial life table)
t or *T*	animal-time
Ts	standard population animal-time at risk
TR	time ratio
TP	true prevalence
TVC	time varying covariate
var	variance
VIF	variance inflation factor
w_j	subjects withdrawn during interval (censored observations) (actuarial life table) or censored observations at time t_j (KM life table)
X	predictor variable or design matrix of predictors
Y	outcome variable or vector of outcome values
Z	design matrix for random effects
	extraneous variable, factor or confounder
	standard normal deviate
Z_α	standard normal deviate for $\alpha/2$ Type I error
Z_β	standard normal deviate for one-tailed β Type II error

Note Acronyms are not italicised in Arial font (tables and figures).

Symbols

α	level of significance (Type I error)
	frailty factor
β	Type II error (power=1-β)
	regression coefficient or vector (1*n) of coefficients
β_{ph}	coefficient from proportional hazards model
β_{aft}	coefficient from accelerated failure time model
χ^2	chi-square statistic
χ^2_{homo}	χ^2 test for homogeneity
χ_{Wald}	Wald chi statistic
ε	error (or vector (1*n) of error values
ϕ	dispersion parameter in GLM(M)
λ	power transformation
	hazard
μ	mean
	random group effect
π	3.14159 (natural number)
θ	a specified or assumed value
θ_0	null specified value
ρ	correlation - intra-class correlation coefficient (r also used)
ρ_{ce}	confounder-exposure correlation
σ	standard deviation
σ^2	variance
σ_1^2	random slope variance for β_1
σ_h^2	herd variance
σ_r^2	regional variance
τ	distribution of survival times
	cutpoint for proportional odds
~	approximate symbol or distributed as (*eg* Y~N(0,1))
≈	approximately equal to
/	division
*	multiplication symbol
#	number

Terms related to location and animal-health problems

AID	autoimmune disease
AVC	Atlantic Veterinary College, at University of Prince Edward Island, Canada
BRD	bovine respiratory disease
BRSV	bovine respiratory syncytial virus
BSE	bovine spongiform encephalopathy
BVD	bovine viral diarrhea
BVDV	bovine viral diarrhea virus
d	day(s)
EBL	enzootic bovine leukosis
ELISA	enzyme-linked immunosorbent assay
IBR	infectious bovine rhinotracheitis (Herpes 1)
IFAT	indirect fluorescent antibody test
ISA	infectious salmon anemia
Map	*Mycobacterium avium subsp paratuberculosis*
Mh	*Mannheimia hemolytica*
mo	month(s)
MUN	milk urea nitrogen
OD	optical density
Ont.	Ontario (large province in Canada)
OVC	Ontario Veterinary College at University of Guelph, Ontario
PCR	polymerase chain reaction
PEI	Prince Edward Island (smallest province in Canada)
PI	persistently infected (*eg* with BVDV)
ppb	parts per billion
ppm	parts per million
yr	year(s)

GT.4 Probability notation

$E(Y)$ = expected value of Y

$p(D+)$ = probability of having the disease of interest

$p(T+|D+)$ = probability of being test positive given the animal had the disease of interest

$p(D+|E+)$ = probability of having the disease of interest in an exposed group

$p(D+|T+)$ = probability of having the disease of interest given the animal was test positive

c_k^n = the number of combinations of k items from n items

GT.5 NAMING VARIABLES

Variable names in the text will be set between pairs of dashes (*eg* -varname-) but the dashes will not be included in tables and figures or if the variable is used in an equation.

Modifications of variables will generally (but not always – you wouldn't expect us to be totally consistent, would you?) be named by adding a suffix to the original variable name. For example:

varname_ct	centred version of the variable
varname_sq	squared version of the variable
varname_c#	a categorical version of -varname- with n = # categories
varname_ln	log transformed version of the variable

Indicator variables will usually be named by appending the category value (or left-hand end of the category range if it is a continuous variable). For example, a variable representing herd size (-numcow-) broken into four categories (0-29, 30-59, 60-89, 90+) would result in the following four variables:

-numcow_0-
-numcow_30-
-numcow_60-
-numcow_90-

Note Unless otherwise specified, values falling exactly on the dividing point will fall in the upper category.

COMBINED REFERENCE LIST

Agresti A. Categorical data analysis. New York: John Wiley and Sons, 1990.

Agresti A. Modelling ordered categorical data: recent advances and future challenges. Stat Med 1999; 18: 2191-2207.

Aiello AE, Larson EL. Causal inference: the case of hygiene and health. Am J Infect Control 2002; 30: 503-511.

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 2001; 15: 687-693.

Alford P, Geller S, Richardson B, Slater M, Honnas C, Foreman J, Robinson J, Messer M, Roberts M, Goble D, Hood D, Chaffin M. A multicenter, matched case-control study of risk factors for equine laminitis. Prev Vet Med 2001; 49: 209-222.

Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol 1997; 26: 1323-1333.

Arunvipas P, VanLeeuwen J, Dohoo IR, Keefe GP. Evaluation of the reliability and repeatability of automated milk urea nitrogen testing. Can J Vet Res 2002; 67: 60-63.

Asch DA, Jedrziewski MK, Christakis NA. Response rates to mail surveys published in medical journals. J Clin Epidemiol 1997; 50: 1129-1136.

Atwill ER, Mohammed HO, Lopez JW, McCulloch CE, Dubovi EJ. Cross-sectional evaluation of environmental, host, and management factors associated with risk of seropositivity to *Ehrlichia risticii* in horses of New York State. Am J Vet Res 1996; 57: 278-285.

Barber S. Comment on 'A comparison of persistent anthelmintic efficacy of topical formulations of doramectin, eprinomectin, ivermectin and moxidectin against naturally acquired nematode infections of beef calves' and problems associated with the mechanical transfer (licking) of endectocides in cattle. (Letter to the editor.) Vet Parasitol, 2003; 112: 255-257.

Barbosa MF, Goldstein H. Discrete response multilevel models. Quality and quantity 2000; 34: 323-330.

Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. J Clin Epidemiol 1999; 51: 1165-1172.

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994; 50: 1088-1101.

Bell EM, Hertz-Picciotto I, Beaumont JJ. Case-cohort analysis of agricultural pesticide applications near maternal residence and selected causes of fetal death. Am J Epidemiol 2001; 154: 702-710.

Belsley DA, Kuh E, Welsch RE. Regression diagnostics. New York: Wiley, 1980.

Bendixen PH. Notes about incidence calculations in observational studies. Prev Vet Med 1987; 5: 151-156.

Bertone ER, Snyder LA, Moore AS. Environmental tobacco smoke and risk of malignant lymphoma in pet cats. Am J Epidemiol 2002; 156: 268-273.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310.

Bloch DA, Kraemer HC. 2x2 kappa coefficients: measures of agreement or association. Biometrics 1989; 45: 269-287.

Brant R. Assessing proportionality in the proportional-odds model for ordinary logistic regression. Biometrics 1990; 46: 1171-1178.

Brenner H, Greenland S, Savitz DA. The effects of non-differential confounder misclassification in ecologic studies. Epidemiology 1992; 3: 456-459.

Breslow NE Day NE. Statistical methods in cancer research Volume I The analysis of case-control studies. IARC Lyon France, 1980.

Breslow NE, Day NE. Statistical methods in cancer research Volume II The design and analysis of cohort studies. IARC Lyon France, 1987.

Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. 2003 submitted. Available from http://www.maths.nottingham.ac.uk/personal/pmzwjb/index.html

Browne WJ. MCMC Estimation in MLwiN. Notes 2002. Available from http://multivel.ioe.ac.uk/dev/devleop.html

BrowneWJ, Subramanian SV, Jones K, Goldstein H. Variance partitioning in multilevel logistic models that exhibit over-dispersion. 2003 submitted. Available from http://www.maths.nottingham.ac.uk/personal/pmzwjb/index.html

Brumbaugh GW, Edwards JF, Roussel AJ Jr., Thomson D. Effect of monensin sodium on histologic lesions of naturally occurring bovine paratuberculosis. J Comp Path 123: 22-28, 2000.

Bruun J, Ersbøll AK, Alban L. Risk factors for metritis in Danish dairy cows. Prev Vet Med 2002; 54: 179-190.

Buck C, Llopis A, Najera E, Terris M. The challenge of epidemiology: issues and selected readings. Pan American Health Organization, Washington, 1988.

Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage samples. Am J Epidemiol 1988; 128: 1198-1206.

Cameron AC, Trivedie PK. Regression analysis of count data. Cambridge: Cambridge University Press, 1998.

Campbell PJ, Hammell KL, Dohoo IR. Historical control clinical trial to assess the effectiveness of teflubenzuron to treat sea lice on Atlantic salmon. Dis Aquat Org 2002a; submitted.

Campbell PJ, Hammell KL, Dohoo IR. Randomized control clinical trial to investigate the effectiveness of teflubenzuron to treat sea lice on Atlantic salmon. Dis Aquat Org 2002b; submitted.

Carey V, Zeger SL, Diggle PJ. Modelling multivariate binary data with alternating logistic regressions. Biometrika 1993; 80: 517-526.

Carrier TK, Estberg L, Stover SM, Gardner IA, Johnson BJ, Read DH, Ardans AA. Association between long periods without high-speed workouts and risk of complete humeral or pelvic fracture in thoroughbred racing horses: 54 cases. J Am Vet Med Assoc 1998; 212: 1582-1587.

Carver DK, Fetrow J, Gerig T, Krueger T, Barnes HJ. Hatchery and transportation factors associated with early poult mortality in commercial turkey flocks. Poult Sci 2002; 81: 1818-1825.

Chatterjee N, Wacholder S. Validation studies: Bias, efficiency, and exposure assessment. Epidemiology 2002; 13: 503-506.

Cheung YB. Adjustment for selection bias in cohort studies: An application of a probit model with selectivity to life course epidemiology. Epidemiology 2001; 12: 1238-1243.

Christensen J, Gardner IA. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Prev Vet Med 2000; 45: 83-106.

Cleveland, WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 1979 74: 829-836.

Cleves M. Analysis of multiple failure-time data with Stata. Stata Tech Bull 1999; 49: 30-39.

Cleves MA, Gould WW, Gutierrez RG. An introduction to survival analysis using Stata. College Station, TX: Stata Press, 2002.

Cochrane Collaboration. The reviewer's handbook, 2001. Available from http://www.cochrane.org/cochrane/hbook.htm

Cohen ND, Mundy GD, Peloso JG, Carey VJ, Amend NK. Results of physical inspection before races and race-related characteristics and their association with musculoskeletal injuries in Thoroughbreds during races. J Am Vet Med Assoc 1999; 215: 654-661.

Collett D. Modelling binary data, 2d ed. New York: Chapman and Hall, 2002.

Collett D. Modelling survival data in medical research: texts in statistical science. New York: Chapman and Hall, 1994.

Congdon P. Applied Bayesian modelling, Wiley, 2003.

Congdon P. Bayesian statistical modelling, Wiley, 2001.

Connell FA, Koepsell TD. Measures of gain in certainty from a diagnostic test. Am J Epidemiol 1985; 121: 744-753.

Converse JM, Presser S. Survey questions: handcrafting the standardized questionnaire, Sage Publications, 1986.

Cornfield J. A statistical problem arising from retrospective studies. Berkeley CA: Third Berkeley Symp, 1956.

Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. J Clin Epidemiol 1992; 45: 1-7.

Cox DR. Regression models and life-tables (with discussion). J R Stat Soc B 1972; 34: 187-220.

Creswell JW. Qualitative inquiry and research design - choosing among five traditions. London: Sage Publications, 1998.

Dargatz DA, Hill GW. Analysis of survey data. Prev Vet Med 1996; 28: 225-237.

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7: 177-188.

Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: The impact of ignoring exposure measurement uncertainty. Stat Med 1998; 17: 1375-1389.

Dickersin K, Berlin JA. Meta-analysis : state-of-the-science. Epidemiol Rev 1992; 14: 154-176.

Diehr P, Martin DC, Koepsell T, Cheadle A, Psaty BM, Wagner EH. Optimal survey design for community intervention evaluations: cohort or cross-sectional? J Clin Epidemiol 1995; 48: 1461-1472.

Diez-Roux AV. Bringing context back into epidemiology: Variables and fallacies in multilevel analyses. Am J Pub Hlth 1998a; 88: 216-222.

Diez-Roux AV. On genes, individuals, society and epidemiology. Am J Epidemiol 1998b; 148: 1027-1032.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of longitudinal data, 2d ed. Oxford: Oxford University Press, 2002.

Dillman DA. Mail and internet surveys: the tailored design method, 2d ed. London: John Wiley & Sons, 1999.

Dingwell RT, Leslie KE, Duffield TF, Keefe GP, Kelton DF. Management strategies influencing drying-off efficiency and development of new intramammary infections in the dry period. J Dairy Sci 2003; 86: 159-168.

Dohoo IR, DesCôteaux L, Dowling P, Fredeen A, Leslie KE, Preston A et al. Report of the Canadian Veterinary Medical Association Expert Panel on rBST. Health Canada, 1-420, 1999, Ottawa, Canada, Health Canada. Available from http://www.hc-sc.gc.ca/english/archives/rbst/animals/

Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med 1997; 29: 221-239.

Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Dowling P, Preston A et al. A meta-analysis review of the effects of rBST. 1. Methodology and effects on production and nutrition related parameters. Can J Vet Res 2003a; in press.

Dohoo IR, Leslie KE, DesCôteaux L, Fredeen A, Shewfelt W, Preston A et al. A meta-analysis review of the effects of rBST. 2. Effects on animal health, reproductive performance and culling. Can J Vet Res 2003b; in press.

Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. Prev Vet Med 2001; 50:127-144.

Dohoo SE, Dohoo IR. Factors influencing the post-operative use of analgesics in dogs and cats by Canadian veterinarians. Can Vet J Res 1996a; 37: 552-556.

Dohoo SE, Dohoo IR. Post-operative use of analgesics in dogs and cats by Canadian veterinarians. Can Vet J Res 1996b; 37: 546-551.

Donald A, Gardner IA, Wiggins AD. Cut-off points for aggregate testing in the presence of disease clustering and correlation of test errors. Prev Vet Med 1994; 19: 167-187.

Donner A. The comparison of proportions in the presence of litter effects. Prev Vet Med 1993; 18: 17-26.

Dorfman A, Kimball AW, Friedman LA. Regression modeling of consumption or exposure variables classified by type. Am J Epidemiol 1985; 122: 1096-1107.

Draper NR, Smith H. Applied regression analysis 3d ed Toronto: John Wiley and Sons, 1998.

Ducrot C, Calavas D, Sabatier P, Faye B. Qualitative interaction between the observer and the observed in veterinary epidemiology. Prev Vet Med 1998; 34: 107-113.

Ducrot C, Legay J, Grohn Y, Envoldsen C, Calavas D. Approach to complexity in veterinary epidemiology; example of cattle reproduction. Natures-Sciences-Societies, 1996; 4: 23-33.

Ducrot C, Roy P, Morignat E, Baron T, Calavas D. How the surveillance system may bias the results of analytical epidemiological studies on BSE: prevalence among dairy versus beef suckler cattle breeds in France. Vet Res 2003; 34: 185-192.

Egger M, Davey Smith G, Altman DG. eds. Systematic reviews in health care. Meta-analysis in context. London: BMJ Books, 2001.

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 1997; 315: 629-634.

Enoe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev Vet Med 2000; 45: 61-81.

Estberg L, Gardner I, Stover SM, Johnson BJ. A case-crossover study of intensive racing and training schedules and risk of catastrophic musculoskeletal injury and lay-up in California thoroughbred racehorses. Prev Vet Med 1998; 33: 159-170.

Evans A. Causation and disease: a chronological journey. Am J Epidemiol 1978; 108: 249-258.

Feivesen AH. Power by simulation. The Stata Journal 2002; 107-124.

Fleiss JL. Statistical methods for rates and proportions, 2d ed. John Wiley and Sons, New York, 1981.

Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology - the essentials, 3d ed. Philadelphia: Lippincott Williams & Wilkins, 1996.

Fourichon CSH, Seegers H, Malher X. Effect of disease on reproduction in the dairy cow: a meta-analysis. Theriogenology 2000; 53: 1729-1759.

Freemantle N, Cleland J, Young P, Mason J, Harrison J. Blockade after myocardial infarction: systematic review and meta regression analysis. British Medical Journal 2002; 318: 1730-1737.

Frerichs RR. 2003 http://www.ph.ucla.edu/epi/snow.html

Friedman LM, Furberg CD, Demets DL. Monitoring response variables. Fundamentals of clinical trials. New York: Springer-Verlag, 1998.

Gardner IA, Stryhn H, Lind P, Collins MT. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev Vet Med 2000; 45: 107-122.

Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol 2003; 83: 593-602.

Giard RWM, Hermans J. The diagnostic information of tests for detection of cancer: the usefulness of the likelihood ratio concept. Eur J Cancer 1996; 32: 2042-2048.

Glass GV. Primary, secondary and meta-analysis of research. Educ Res 1976; 5: 3-8.

Goldstein H, Browne WJ, Rasbash J. Multilevel modelling of medical data. Stat Med 2002; 21: 3291-3315.

Goldstein H. Multilevel statistical models, 2d ed, London: Arnold, 1995.

Goldstein H, Browne, WJ, Rasbash, J. Partitioning variation in multilevel models. Understanding Statistics 2002; 1: 223-232.

Gotway CA, Wolfinger RD. Spatial prediction of counts and rates. Stat Med 2003; 22: 1415-1432.

Green SB. Design of randomised trials. Epidemiol Rev 2002; 24: 4-11.

Greenland S, Morgenstern H. Confounding in health research. Ann Rev Pub Hlth 2001; 22: 189-212.
Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. Int J Epidemiol 1989; 18: 269-274.
Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48.
Greenland S, Robins JM. Ecologic studies: Biases, misconceptions, and counter examples. Am J Epidemiol 1994; 139: 747-760
Greenland S, Robins JM. Confounding and misclassification. Am J Epidemiol 1985; 122: 495-505.
Greenland S. Alternative models for ordinal logistic regression. Stat Med 1994; 13: 1665-1677.
Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med 1992; 11: 1209-1223.
Greiner M, Gardner IA. Application of diagnostic tests in veterinary epidemiologic studies. Prev Vet Med 2000a; 45: 43-59.
Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev Vet Med 2000b; 45: 3-22.
Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver operating characteristic analysis for diagnostic tests. Prev Vet Med 2000; 45: 23-41.
Guan W. Bootstrapping standard errors. The Stata Journal 2003; 3: 71-80.
Gutierrez RG. Parametric frailty and shared frailty survival models. The Stata Journal 2002; 2: 22-44.
Hammell KL, Dohoo IR. Mortality patterns in infectious salmon anemia virus outbreaks in New Brunswick, Canada. Journal of Fish Diseases 2003; accepted.
Hammond RF, McGrath G, Martin SW. Irish soil and land-use classifications as predictors of numbers of badgers and badger setts. Prev Vet Med 2001; 51: 137-148.
Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol 2003; 157: 364-375.
Hanley JA, Negassa A, Edwardes MD. GEE: Analysis of negatively correlated binary responses: a caution. Stat Med 2000; 19: 715-722.
Hanson TE, Johnson WO, Gardner IA. Hierarchical models for the estimation of disease prevalence and the sensitivity and specificity of dependent tests in the absence of a gold-standard. J Agr Biol Env Sci 2003 (in press).
Hardin J, Hilbe J. Generalized linear models and extensions. College Station: Stata Press, 2001
Hardin JW, Hilbe JM. Generalized estimating equations. Boca Raton: Chapman & Hall/CRC, 2003.
Hastie T, Tibshirani R. Generalized additive models for medical research. Stat Methods Med Res 1995; 4: 187-196.
Heitjan DF. Causal inference in a clinical trial: a comparative example. Control Clin Trials 1999; 20: 309-318.
Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as prerequisite for confounding evaluation: An application to birth defects epidemiology. Am J Epidemiol 2002; 155: 176-184.
Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965; 58: 295-300.
Holman CD, Arnold-Reed DE, deKlerk N, McComb C, English DR. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. Epidemiology 2001; 12: 246-255.
Hosmer DW, Lemeshow S. Applied logistic regression, 2d ed. Toronto: John Wiley and Sons, 2000.
Hosmer DW, Lemeshow S. Applied survival analysis. Regression modelling of time to event data. New York: John Wiley & Sons, 1999.
Houe H, Ersbøll AE, Toft N, Agger JF. eds. Veterinary epidemiology from hypothesis to conclusion. Copenhagen: Samfundslitteratur KVL Bogladen, 2002.

Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998; 1623-1634.

Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-171.

Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research, 2d ed. Philadelphia: Lippincott Williams & Wilkins, 2001.

Hurnik D, Dohoo IR, Bate LA. Types of farm management as risk factors for swine respiratory disease. Prev Vet Med 1994; 20: 147-157.

Hurnik D, Dohoo IR, Donald AW, Robinson NP. Factor analysis of swine farm management practices on Prince Edward Island. Prev Vet Med 1994; 20: 135-146.

Ingram DG, Mitchell WR, Martin SW eds. Animal Disease Monitoring. CC Thomas, Springfield, Illinois, 1975.

Irvine RJ, Stein A, Halvorsen O, Langvatn R, Albon SD. Life-history strategies and population dynamics of abomasal nematodes in Svalbard reindeer (Rangifer tarandus platyrhunchus). Parasitology 2000; 120: 297-311.

Jacob M. Extra-binomial variation in logistic multilevel models - a simulation. Multilevel Modelling Newsletter 2000; 12: 8-14.

Jacobson RH. Validation of serological assays for diagnosis of infectious diseases. Rev sci tech 1998; 17: 469-486.

Kalsbeek W, Heiss G. Building bridges between populations and samples in epidemiological studies. Annu Rev Public Health 2000; 21: 147-169.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53: 457-481.

Kaufman JS, Poole C. Looking back on "causal thinking in the health sciences". Anu Rev Pub Hlth 2000; 21: 101-119.

Keefe GP, Dohoo IR, Valcour J, Milton RL. Assessment of ultrasonic imaging of marbling at entry into the feedlot as a predictor of carcass traits at slaughter. J Anim Sci 2003; submitted.

King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Stat Med 2001; 20: 2131-2147.

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. London: Lifetime Learning Publications, 1982.

Kleinbaum DG, Kupper LL, Mullen KE. Applied regression analysis and other multivariable methods, 2d ed. Boston: PWS-Kent Publishing Co, 1988.

Kleinbaum DG, Morgenstern H, Kupper LL. Selection bias in epidemiologic studies. Am J Epidemiol 1981: 113: 452-463.

Koopman JS, Weed DL. Epigenesis theory: a mathematical model relating causal concepts of pathogenesis in individuals to disease patterns in populations. Am J Epidemiol 1990; 132: 366-390.

Kraemer HC, Bloch DA. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med 1994; 13: 876-880.

Krieger N. Epidemiology and the causal web: Has anyone seen the spider? Soc Sci Med 1994; 39: 887-903.

Kutner MH, Nachtschiem CJ, Wasserman W, Neter J. Applied linear statistical models, 4th ed Boston: McGraw-Hill/Irwin, 1996.

Langford IH, Lewis T. Outliers in multilevel models (with Discussion). J R Stat Soc A 1998; 161: 121-160.

Lavori PW, Kelsey J eds. Clinical trials. Epidemiologic reviews 2002; 24: 1-90.

Lavori PW, Kelsey J. Introduction and overview. Epidemiologic reviews 2002; 24: 1-3.

Leech FB. A critique of the methods and results of the British national surveys of disease in farm animals. II. Some general remarks on population surveys of farm animal disease. Brit Vet J 1971; 127: 587-592.

Leech FB, Sellers KC. Statistical epidemiology in veterinary science. New York: MacMillan Publishing Co. Inc., 1979.

Levy PS, Lemeshow S. Sampling of populations: methods and applications, 3d ed. New York: John Wiley & Sons, Inc., 1999.

Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. J Am Stat Assoc 1989; 84: 1074-1078.

Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255-268.

Lindberg A, Groenendaal H, Alenius S, Emanuelson U. Validation of a test for dams carrying foetuses persistently infected with bovine viral-diarrhoea virus based on determination of antibody levels in late pregnancy. Prev Vet Med 2001; 51: 199-214.

Linnet K. A review on the methodology for assessing diagnostic tests. Clin Chem 1988; 34: 1379-1386.

Lofstedt J, Dohoo IR, Duizer G. Model to predict septicemia in diarrheic calves. J Vet Int Med 1999; 13: 81-88.

Long JS, Freese J. Regression models for categorical dependent variables using Stata. College Station: Stata Press, 2001.

Long JS. Regression models for categorical and limited dependent variables. London: Sage Publications, 1997.

Longford N. Standard errors in multilevel analysis. Multilevel Newsletter 1999; 11: 10-13.

Maclure M. Taxonomic axes of epidemiologic study designs: a refutationist perspective. J Clin Epidemiol 1991; 44: 1045-1053.

MacMahon B, Pugh TF. Epidemiology: principles and methods. Little Brown, Boston, 1970.

Mangione-Smith R, Elliott MN, McDonald L, McGlynn EA. An observational study of antibiotic prescribing behavior and the Hawthorne effect. Health Serv Res 2002; 37: 1603-1623.

Manske T, Hultgren J, Bergsten C. The effect of claw trimming on the hoof health of Swedish dairy cattle. Prev Vet Med 2002; 54: 113-129.

Marshall JR, Hastrup JL. Mismeasurement and the resonance of strong confounders: Uncorrelated errors. Am J Epidemiol 1996; 143: 1069-1078.

Martin SW, Kirby K, Pennock PW. Canine hip dysplasia: breed effects. Can Vet J Res 1980; 21: 293-296.

Martin SW, Meek AH, Willeberg P. Veterinary epidemiology: principles and methods. Ames: Iowa State Press, Ames, 1987.

Martin SW, Nagy E, Armstrong D, Rosendal S. The associations of viral and mycoplasmal antibody titres with respiratory disease and weight gain in feedlot calves. Can Vet J Res 1999; 40: 560-570.

Martin SW, Shoukri MM, Thoburn MA. Evaluating the health status of herds based on tests applied to individuals. Prev Vet Med 1992; 14: 33-44.

Martin SW. Estimating disease prevalence and the interpretation of screening test results. Prev Vet Med 1984; 2: 463-472.

Martin W. If multivariable modelling is the answer, what is the question? Dutch Society of Veterinary Epidemiology and Economics, Wageningen, 1996.

McCullagh P, Nelder JA. Generalized linear models, 2d ed. London: Chapman and Hall, 1989.

McDermott JJ, Deng KA, Jayatileka TN, El Jack MA. A cross-sectional cattle disease study in Kongor Rural Council, Southern Sudan. I. Prevalence estimates and age, sex and breed associations for brucellosis and contagious bovine pleuropneumonia. Prev Vet Med 1987a; 5: 111-123.

McDermott JJ, Deng KA, Jayatileka TN, El Jack MA. A cross-sectional cattle disease study in Kongor Rural Council, Southern Sudan. II. Brucellosis in cows: associated factors, impact on production and disease control considerations. Prev Vet Med 1987b; 5: 125-132.

McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol 1999; 149: 887-897.

McMichael AJ. The health of persons, populations, and planets: epidemiology comes full circle. Epidemiology 1995; 6: 633-636.

Meek AH. Veterinary epidemiology: challenges and opportunities in research. Prev Vet Med 1993; 18: 53-60.

Meinert CL. Clinical Trials: design, conduct and analysis. Oxford Oxford Univ. Press, 1986.
Mittlbock M, Schemper M. Explained variation for logistic regression. Stat Med 1996; 15: 1987-1997.
Morgenstern, H. Ecologic studies in Rothman KJ and Greenland S. Modern epidemiology, 2d ed. Philadelphia: Lippincott-Raven, 1998.
Navidi W, Weinhandl E. Risk set sampling for case-crossover designs. Epidemiology 2002; 13: 100-105.
Nespeca R, Vaillancourt JP, Morrow WE. Validation of a poultry biosecurity survey. Prev Vet Med 1997; 31: 73-86.
Nødtvedt A, Dohoo IR, Sanchez J, Conboy G, DesCôteaux L, Keefe G. Increase in milk yield following eprinomectin treatment at calving in pastured dairy cattle. Vet Parasitol 2002; 105: 191-206.
Nødtvedt A, Dohoo IR, Sanchez J, Conboy G, DesCôteaux L, Keefe GP et al. The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. Can J Vet Res 2002; 66: 249-257.
Noordhuizen JPTM, Frankena K, van der Hoofd,CM, Graat EAM. Application of quantitative methods in veterinary epidemiology. Wageningen: Wageningen Pers, 1997.
Normand SLT. Meta-analysis: formulating, evaluating, combining and reporting. Stat Med 1999; 18: 321-359
O'Callaghan CJ, Medley GF, Peter TF, Mahan SM, Perry BD. Predicting the effect of vaccination on the transmission dynamics of heartwater (cowdria ruminatium infection). Prev Vet Med 1999; 42: 17-38.
Osmond C, Gardner MJ. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. Am J Epidemiol 1989; 129: 31-35.
Otte MJ, Gumm ID. Intra-cluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys. Prev Vet Med 1997; 31: 147-150.
Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001; 57: 120-125.
Pearce N. Analytical implications of epidemiological concepts of interaction. Int J Epidemiol 1989; 18: 976-980.
Peduzzi P, Henderson W, Hartigan P, Lavori PW. Analysis of randomized controlled trials. Epidemiol Rev 2002; 24: 26-38.
Petitti DB. Meta-analysis, decision analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine, 2d ed. London: Oxford University Press, 1999.
Petitti DB. The implications of alternative views about causal inference for the work of the practicing epidemiologist. Proc of Society for Epidemiologic Research (US). Causal inference. Chapel Hill, NC, 1985.
Piantadosi S. Clinical Trials: A methodological perspective. New York John Wiley and Sons, 1997.
Poole C. Causal values. Epidemiology 2001; 12: 139-141.
Pouillot R, Gerbier G, Gardner IA. "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. Prev Vet Med 2002; 53: 67-81.
Pregibon, D. Logistic regression diagnostics. The Annals of Statistics 1981 9: 705-724.
Prentice RL. Design issues in cohort studies. Stat Methods Med Res 1995; 4: 273-292.
Priester WA. Collecting and using veterinary clinical data. In Ingram DG, Mitchell WR, Martin SW, eds. Animal Disease Monitoring. CC Thomas, Springfield Illinois, 1975.
Rabe-Hesketh S, Pickles A, Taylor C. Generalized linear latent and mixed models. Stata Technical Bulletin 2000; 53: 47-57.
Raftery AE. Bayesian model selection in social research. In: Marsden PV, editor. Sociological Methodology. Oxford: Basil Blackwell, 1996; 111-163.
Reeve-Johnson LG. Assessment of the efficacy of a novel intramammary antibiotic for the treatment of mastitis caused by *Staphylococcus aureus* during the non-lactating period in United States dairy herds. Thesis for the Royal College of Veterinary Surgeons for the Diploma of Fellowship. Royal College of Veterinary Surgeons, London, England, 2001.

Reilly M. Optimal sampling strategies for two-stage studies. Am J Epidemiol 1996; 143: 92-100.

Risks and benefits of estrogen plus progestin in healthy post-menopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002; 288: 321-333.

Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. J Clin Epidemiol 1999; 52: 569-583.

Robins JM, Greenland S. The role of model selection in causal inferences from non-experimental data. Am J Epidemiol 1986; 123: 392-402.

Robins JM, Hsieh F, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured data. J Royal Stat Soc B 1995; 57: 409-424.

Robins JM. Data, design and background knowledge in etiologic inference. Epidemiology 2001, 12: 313-320.

Rodriguez G, Elo I. Intra-class correlation in random-effects models for binary data. The Stata Journal 2003; 3: 32-46.

Rose G. Sick individuals and sick populations. Int J Epidemiol 1985; 14: 32-38.

Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992; 136: 1400-1413.

Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990; 132: 734-745.

Rothman K, Greenland S. Modern epidemiology 2d ed Philadelphia: Lippincott-Raven, 1998.

Rothman KJ. Causes. Am J Epidemiol 1976; 104: 587-592.

Royston P, Babiker A. A menu driven facility for complex sample size calculation in randomized controlled trials with a survival or binary outcome. The Stata Journal 2002; 2: 151-163.

Saah AJ, Hoover DR. Sensitivity and specificity reconsidered: the meaning of these terms in analytical or diagnostic settings. Am Intern Med 1997; 126: 91-94.

Sackett DL, Haynes RB, Guyatt GG, Tugwell P. Clincal epidemiology: a basic science for clinical medicine, 2d ed. Philadelphia: Lippincott Williams & Wilkins, 1991.

Salant P, Dillman DA. How to conduct your own survey. London: John Wiley & Sons, 1994.

Samet JM, Munoz A, eds. Cohort studies. Epidemiol Rev 1998; 20: 1-136.

Sanchez J, Nødtvedt A, Dohoo IR, DesCôteaux L. The effect of eprinomectin at calving on reproduction parameters in adult dairy cows in Canada. Prev Vet Med 2002; 56: 165-177.

Sandler DP. On revealing what we would rather hide: The problem of describing study participation. Epidemiology 2002; 13: 117.

Schaubel D, Hanley J, Collet JP, Bolvin JF, Sharpe C, Morrison HI, Mao Y. Two-stage sampling for etiologic studies. Am J Epidemiol 1997; 146: 450-458.

Schwabe CW, Riemann HP, Franti CE. Epidemiology in Veterinary Practice. Lea and Febiger, Philadelphia, 1977.

Schwabe CW. The current epidemiological revolution in veterinary medicine. Part I. Prev Vet Med 1982; 1: 5-15.

Schwabe CW. The current epidemiological revolution in veterinary medicine. Part II. Prev Vet Med 1993; 18: 3-16.

Schwabe CW. Veterinary Medicine and Human Health. Williams and Wilkins, Baltimore 3d ed, 1984.

Schwartz S. The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. Am Jour Pub Hlth 84: 819-824, 1994.

Seiler RJ. The non-diseased reactor: considerations on the interpretation of screening test results. Vet Rec 1979; 105: 226-228.

Shy C. The failure of academic epidemiology: witness for the prosecution. Am J Epidemiol 1997; 145: 479-484.

Singer RS, Atwill ER, Carpenter TE, Jeffrey JS, Johnson WO, Hirsh DC. Selection bias in epidemiological studies of infectious disease using *Escherichia coli* and avian cellulitis as an example. Epidemiol Infection 2001; 126: 139-145.

Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology 1992; 3: 449-452.

Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 1984; 13: 356-365.

Smith RD. Veterinary clinical epidemiology: a problem-oriented approach, 2d ed. Boca Raton: CRC Press, 1995.

Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Iowa State Press, Ames, Iowa, 1989.

Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modelling. London: Sage Publications, 1999.

Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard". Am J Epidemiol 1997;145: 184-196.

Sribney B, Harrell F, Conroy R. Problems with stepwise regression. Stata, Frequently Asked Questions, 1998.

Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. J Chron Dis 1981; 34: 599-610.

Streiner DL, Norman GR. Health measurement scales; a practical guide to their development and use, 2d ed. Oxford University Press, 1995.

Stryhn H, Dohoo IR, Tillard E, Hagedorn-Olsen T. Simulation as a tool of validation in hierarchical generalized linear models. Proc of Intl Symp on Vet Epidem and Econom. Breckenridge, Colorado, 2000.

Stürmer T, Thürigen D, Spiegelman D, Blettner M, Brenner H. The performance of methods for correcting measurement error in case-control studies. Epidemiology 2002; 13: 507-516.

Susser M. Causal Thinking in the Health Sciences: concepts and strategies of epidemiology. Oxford University Press, Toronto (out of print), 1973.

Susser M. Judgement and causal inference: criteria in epidemiologic studies. Am J Epidemiol 1977; 105: 1-15

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. New York: John Wiley & Sons, 2000.

Taubes G. Epidemiology faces its limits. Science 1995; 269: 164-169.

Thomas D. New techniques for the analysis of cohort studies. Epidemiol Rev 1998; 20: 122-134.

Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 1991; 44: 221-232.

Thrushfield MV. Veterinary epidemiology, 2d ed. Oxford: Blackwell Science Ltd, 1995.

Thurmond MC, Hietala SK. Effect of congenitally acquired *Neospora caninum* infection on risk of abortion and subsequent abortions in dairy cattle. Am J Vet Res 1997; 58: 1381-1385.

Thysen I. Application of event time analysis to replacement, health and reproduction data in dairy cattle research. Prev Vet Med 1988; 5: 239-250.

Tinker MK, White NA, Lessard P, Thatcher CD, Pelzer KD, Davis B, Carmel DK. Prospective study of equine colic incidence and mortality. Equine Vet J 1997; 29: 448-453.

Tukey J. The future of data analysis. Ann Math Stat 1962; 33: 1-67.

Tyler JW, Cullor JS. Titers, tests, and truisms: Rational interpretation of diagnostic serologic testing. J Am Vet Med Assoc 1989; 194: 1550-1558.

Vaarst M, Paarup-Laursen B, Houe H, Fossing C, Andersen HJ. Farmers' choice of medical treatment of mastitis in Danish dairy herds based on qualitative research interviews. J Dairy Sci 2002; 85: 992-1001.

Vaillancourt JP, Martineau GP et al. Construction of questionnaires and their use in veterinary medicine. Proc of Soc Vet Epidem and Prev Med, 1991.

Vandenbroucke JP. On the rediscovery of a distinction. Am J Epidemiol 1985; 121: 627-628.

VanLeeuwen J, Keefe G, Tremblay R, Power C, Wichtel JJ. Seroprevalence of infection with *Mycobacterium avium* subspecies *paratuberculosis*, bovine leukemia virus, and bovine viral diarrhea virus in Maritime Canada dairy cattle. Can Vet J Res 2001; 42: 193-198.

Veierod MB, Laake P. Exposure misclassification: bias in category specific Poisson regression coefficients. Stat Med 2001; 20: 771-784.

Veling J, Wilpshaar H, Frankena K, Bartels C, Barkema HW. Risk factors for clinical Salmonella enterica subsp. enterica serovar Typhimurium infection on Dutch dairy farms. Prev Vet Med. 2002; 54: 157-168.

Vigre H, Dohoo IR, Stryhn H, Busch ME. Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumoniae* and *Mycoplasma hyopneumoniae* at different levels in Danish multisite pig production facilities. Accepted Prev Vet Med 2003.

Wacholder S, Carroll RJ, Pee D, Gail MH. The partial questionnaire design for case-control studies. Stat Med 1994; 13: 623-634.

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls incase-control studies. II. Types of controls. Am J Epidemiol 1992a; 135: 1029-1041.

Wacholder S. Design issues in case-control studies. Stat Methods Med Res 1995; 4: 293-309.

Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J of Epidemiology 1986; 123: 174-184.

Waldmann MR, Hagmayer Y. Estimating causal strength: the role of structural knowledge and processing effort. Cognition 2001; 82: 27-58.

Walker C, Canfield PJ, Love DN, McNeil DR. A longitudinal study of lymphocyte subsets in a cohort of cats naturally infected with feline immunodeficiency virus. Aust Vet Jour 1996; 73: 218-224.

Walter SD, Feinstein AR, Wells CK. Coding ordinal independent variables in multiple regression analyses. Am J Epidemiol 1987; 125: 319-323.

Waltner-Toews D, Martin SW, Meek AH, McMillan I, Crouch CF. A field trial to evaluate the efficacy of a combined rotavirus-coronavirus/*Escherichia coli* vaccine in dairy cattle. Can J Comp Med 1985; 49: 1-9.

Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. Academic Emergency Medicine 2002; 9: 330-341.

Webster T. Commentary: Does the spectre of ecologic bias haunt epidemiology? Int J Epidemiol 2002; 31: 161-162.

Weed DL, Hursting SD. Biologic plausibility in causal inference: current method and practice. Am J Epidemiol 1998; 147: 415-425.

Weed DL. Environmental epidemiology: basics and proof of cause-effect. Toxicology 2002 181-182: 399-403.

Weed DL. Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol 2000; 29: 387-390.

Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. Stat Med 1997; 16: 833-839.

Weisberg S. Applied linear regression, 2d ed. New York: Wiley, 1985

Wells SJ, Garber LP, Wagner BA. Papillomatous digital dermatitis and associated risk factors in US dairy herds. Prev Vet Med 1999; 38: 11-24.

West AM, Martin SW, McEwen SA, Clarke RC, Tamblyn SE. Factors associated with the presence of *Salmonella* spp in dairy farm families in Southwestern Ontario. Can Jour Pub Hlth 1988; 79: 119-123.

White PA. Causal attribution and Mill's methods of experimental inquiry: past, present and prospect. Br J Soc Psychol 2000; 39: 429-447.

White PA. Causal judgments about relations between multilevel variables. J. Exp Psychol Learn Mem Cogn 2001; 27: 499-513.

Williams D. Modelling binary data, 2d ed. Boca Raton: Chapman and Hall/ CRC, 2002.

Wing S. Whose epidemiology, whose health? Int Jour Hlth Serv 1998; 28: 241-252.

Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002; 24: 39-53.

Wolfe R, Gould W. An approximate likelihood-ratio test for ordinal response models. Stata Tech Bull 1998; 42: 24-27.

Yang M, Heath A, Goldstein H. Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. J R Stat Soc A 2000; 163: 49-62.

Yates WDG. A review of infectious bovine rhinotracheitis, shipping fever pneumonia and viral-bacterial synergism in respiratory disease of cattle. Can J Comp Med 1982; 46: 225-263.

Zhang J. A method of correcting the odds ratio in cohort studies of common outcomes. J Am Med Assoc, November 18, 1998, 280: 1690-1691.

Zhou XH, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. Am Stat 1999; 53: 282-290.

INDEX

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

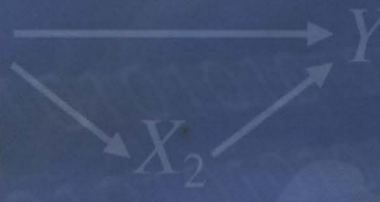
Y

Z

"A comprehensive text for the discipline"

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$AF_e = (RR\text{-}1)/RR$



ISBN 0-919013-41-4

www.upei.ca/ver