WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Applied Longitudinal Analysis

GARRETT M. FITZMAURICE
NAN M. LAIRD
JAMES H. WARE
Department of Biostatistics
Harvard University
Boston, MA

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Fitzmaurice, Garrett M., 1962–
Applied longitudinal analysis / Garrett M. Fitzmaurice, Nan M. Laird, James H. Ware.
p. cm. — (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 0-471-21487-6 (cloth)
1. Longitudinal method. 2. Regression analysis. 3. Multivariate analysis. 4. Medical statistics. I. Laird, Nan M., 1943– II. Ware, James H., 1941– III. Title. IV. Series.

QA278.F575 2004
519.5'3—pcc22 2004040891

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Laura, Kieran, and Aidan

— G.M.F.

To Joel, Richard, and Lily

— N.M.L.

To Janice, Cameron, and Jake

— J.H.W.

Contents

Preface

Our goal in writing this book is to provide a rigorous and systematic description of modern methods for analyzing data from longitudinal studies. In recent years, there have been remarkable developments in methods for longitudinal analysis. Despite these important advances, the methods have been somewhat slow to move into the mainstream. *Applied Longitudinal Analysis* bridges the gap between theory and application by presenting a comprehensive account of these methods in a way that is accessible to a wide range of readers.

The impetus for this book arose from teaching a graduate-level course on "Applied Longitudinal Analysis" at the Harvard School of Public Health. As course instructors, we were frustrated by the lack of a suitable textbook that adequately covered modern statistical methods for longitudinal analysis at a level accessible to a broad audience of researchers and graduate students in the health and medical sciences. We envision this book as a textbook for such a course and, subsequently, as a reference resource for researchers and graduate students. It is also suitable for graduate students in statistics and for statisticians already working in the health sciences, governmental health-related agencies, and the pharmaceutical industry. It is intended to allow a diverse group of statisticians, researchers, and graduate students in substantive fields to master modern methods for longitudinal data analysis.

The scope of this book is broad, covering methods for the analysis of diverse types of longitudinal data arising in the health sciences. The methods are presented in the setting of numerous applications to real data sets. Our main emphasis is on the practical rather than the theoretical aspects of longitudinal analysis. Twenty-five real data sets, drawn from studies in health-related fields, are

used throughout the text and problem sets to illustrate the applications of longitudinal methods. These data sets can be downloaded from the web site for the book: `www.biostat.harvard.edu/~fitzmaur/ala`. Although the methods are applied to data sets drawn from the health sciences, they apply equally to other areas of application, for example, education, psychology, and other branches of the behavioral and social sciences.

Because longitudinal data are a special case of clustered data, albeit with a natural ordering of the measurements within a cluster, we include also a description of modern methods for analyzing clustered data, more broadly defined. Indeed, one of our goals is to demonstrate that methods for longitudinal analysis are, more or less, special cases of more general regression methods for clustered data. As a result, a comprehensive understanding of longitudinal data analysis provides the basis for a broader understanding of methods for analyzing the wide range of clustered data that commonly arises in studies in the biomedical and health sciences.

The prerequisites for a course based on this book are an introductory course in statistics and a strong background in regression analysis. Some previous exposure to generalized linear models (e.g., logistic regression) would be helpful, although these models are reviewed in the text. An understanding of matrix algebra or calculus is not assumed; the reader will be gently introduced to only those aspects of vector and matrix notation necessary for understanding the matrix representation of regression models for longitudinal data. Because vectors and matrices are used to simplify notation, the reader is required to attain some basic facility with the addition and multiplication of vectors and matrices. Although we do not assume a high level of mathematical preparation, a willingness to read and consider mathematical ideas is required. More technical or mathematical sections of the book are marked with asterisks and may be omitted at first reading without loss of continuity.

To use the methods described in this book, appropriate statistical software is required. In general, the methods available via commercially available software lag behind the recent advances in statistical methods; longitudinal data analysis is not exceptional in this regard. Recently, the introduction of new programs for analyzing multivariate and longitudinal data has made these methods far more accessible to practitioners and students. We use *SAS*, which is widely available, to perform the analyses presented throughout the text. Illustrative *SAS* commands are included at the end of many of the chapters, with basic descriptions of their usage. Programming statements and computer output for the examples, prepared using *SAS*, can be downloaded from the web site: `www.biostat.harvard.edu/~fitzmaur/ala`. We selected *SAS* because all of the analyses we discuss can be performed using its procedures. Many of the methods can be carried out using alternative software packages (e.g., *S-PLUS* and *Stata*) or special purpose programs (e.g., *BMDP5-V*) and this book can be supplemented with any one of them. Readers are encouraged to perform and verify the results of analyses using software of their choice. Because statistical software is constantly evolving, we anticipate that all of the methods we discuss will soon be available within most of the major statistical packages.

Throughout the text references have been kept to an absolute minimum. Instead, at the end of each chapter we include suggestions for further readings that provide

more in-depth coverage of certain topics. We also include "bibliographic notes" that highlight key references in the mainstream statistical literature. Although many of our readers may find the latter references to be too technical, they are included to give due credit to those who have contributed to the statistical methods described in each chapter.

Finally, we would like to thank the many friends and colleagues who have helped us to write this book. A special word of thanks to Misha Salganik, for preparation of the diagrams and many helpful suggestions for improvement of graphical displays. We are especially grateful to Joe Hogan and Russell Localio, for reading a first draft and providing invaluable feedback, comments, and suggestions that improved the book. We would also like to thank Rino Bellocco, Brent Coull, Nick Horton, Sharon-Lise Normand, Misha Salganik, Judy Singer, S.V. Subramanian, and Florin Vaida, for their insightful comments on several chapters. We are grateful to the students who have participated in the course on "Applied Longitudinal Analysis" at the Harvard School of Public Health since its inception; they have provided the impetus and motivation for writing this book. We gratefully acknowledge support from grant GM 29745 from the National Institutes of Health. The first author gratefully acknowledges support from the Junior Faculty Sabbatical Program at the Harvard School of Public Health; the support provided by a sabbatical created a unique opportunity to begin writing this book. Last, but not least, we thank Steve Quigley and Susanne Steitz of Wiley, for their advice and encouragement during all stages of this project.

GARRETT M. FITZMAURICE
NAN M. LAIRD
JAMES H. WARE

Boston, Massachusetts
March, 2004

Acknowledgments

Throughout this book we have used data sets drawn from published studies in health-related fields to exemplify important concepts in the analysis of longitudinal and clustered data. We are grateful to the following investigators for sharing their data with us: Graham Bentham, Doug Dockery, Brian Flay, Robert Greenberg, Keith Henry, Aviva Must, Elena Naumova, George Rhoads, Jan Schouten, Linda Van Marter, and Gwen Zahner.

We also thank the following publishers for permission to reproduce published data sets in print and electronic format: The American Statistical Association, Blackwell Publishing, Brooks/Cole (a division of Thomson Learning), CRC Press, Elsevier, Iowa State Press, Oxford University Press, and SAS Institute Inc.

Finally, in all data sets used throughout this book, the original subject identification (ID) numbers have been deleted and replaced with new subject ID numbers, to ensure that the data sets cannot be linked to the original records.

Part I

Introduction to Longitudinal and Clustered Data

1

Longitudinal and Clustered Data

1.1 INTRODUCTION

Research on statistical methods for the design and analysis of human investigations expanded explosively in the second half of the twentieth century. Beginning in the early 1950s, the U.S. government shifted a substantial part of its research support from military to biomedical research. The legislative foundation for the modern National Institutes of Health (NIH), the Public Health Service Act, was passed in 1944 and NIH grew rapidly throughout the 1950s and 1960s. The NIH sponsored many of the important epidemiologic studies and clinical trials of that period, including the influential Framingham Heart Study (Dawber *et al.*, 1951; Dawber, 1980).

The typical focus of these early studies was morbidity and, especially, mortality. Investigators sought to identify the causes of early death and to evaluate the effectiveness of treatments for delaying death and morbidity. In the Framingham Heart Study, participants were seen at two-year intervals. Survival outcomes during successive two-year periods were treated as independent events and modelled using multiple logistic regression. The successful use of multiple logistic regression in this setting, and the recognition that it could be applied to case-control data, led to widespread use of this methodology beginning in the 1960s. The analysis of time-to-event data was revolutionized by the seminal 1972 paper of D.R. Cox, describing the proportional hazards model (Cox, 1972). This paper was followed by a rich and important body of work that established the conceptual basis and the computational tools for modern survival analysis.

Though the design of the Framingham Heart Study and other cohort studies called for periodic measurement of the patient characteristics thought to be determinants of

chronic disease, interest in the levels and patterns of change of those characteristics over time was initially limited. As the research advanced, however, investigators began to ask questions about the behavior of these risk factors. In the Framingham Heart Study, for example, investigators began to ask whether blood pressure levels in childhood were predictive of hypertension in adult life. In the Coronary Artery Risk Development in Young Adults (CARDIA) Study, investigators sought to identify the determinants of the transition from normotensive or normocholesterolemic status in early adult life to hypertension and hypercholesterolemia in middle age (Friedman *et al.*, 1988). In the treatment of arthritis, asthma, and other diseases that are not typically life-threatening, investigators began to study the effects of treatments on the level and change over time in measures of severity of disease. Similar questions were being posed in every disease setting. Investigators began to follow populations of all ages over time, both in observational studies and clinical trials, to understand the development and persistence of disease and to identify factors that alter the course of disease development.

This interest in the temporal patterns of change in human characteristics came at a period when advances in computing power made new and more computationally intensive approaches to statistical analysis available at the desktop. Thus, in the early 1980s, Laird and Ware proposed the use of the EM algorithm to fit a class of linear mixed effects models appropriate for the analysis of repeated measurements (Laird and Ware, 1982); Jennrich and Schluchter (1986) proposed a variety of alternative algorithms, including Fisher-scoring and Newton-Raphson algorithms. Later in the decade, Liang and Zeger introduced the generalized estimating equations in the biostatistical literature and proposed a family of generalized linear models for fitting repeated observations of binary and counted data (Liang and Zeger, 1986; Zeger and Liang, 1986). Many other investigators writing in the biomedical, educational, and psychometric literature contributed to the rapid development of methodology for the analysis of these "longitudinal" data. The past 25 years have seen considerable progress in the development of statistical methods for the analysis of longitudinal data. Despite these important advances, methods for the analysis of longitudinal data have been somewhat slow to move into the mainstream. This book bridges the gap between theory and application by presenting a comprehensive description of methods for the analysis of longitudinal data accessible to a broad range of readers.

1.2 LONGITUDINAL AND CLUSTERED DATA

The defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time, thereby allowing the direct study of change over time. The primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change. With repeated measures on individuals one can capture within-individual change. Indeed, the assessment of within-subject changes in the response over time can only be achieved within a longitudinal study design. For example, in a cross-sectional study, where the response is measured at a single occasion, one can only obtain estimates of between-individual

differences in the response. That is, a cross-sectional study may allow comparisons among sub-populations that happen to differ in age, but it does not provide any information about how individuals change during the corresponding period.

To highlight this important distinction between cross-sectional and longitudinal study designs, consider the following simple example. Body fatness in girls is thought to increase just before or around menarche, levelling off approximately 4 years after menarche. Suppose investigators are interested in determining the increase in body fatness in girls after menarche. In a cross-sectional study design, investigators might obtain measurements of percent body fat on two separate groups of girls: a group of 10-year-old girls (a pre-menarcheal cohort) and a group of 15-year-old girls (a post-menarcheal cohort). In this cross-sectional study design, direct comparison of the average percent body fat in the two groups of girls can be made using a two sample (unpaired) t-test. This comparison does not provide an estimate of the change in body fatness as girls age from 10 to 15 years. The effect of growth or aging, an inherently within-individual effect, simply cannot be estimated from a cross-sectional study that does not obtain measures of how individuals change with time. In a cross-sectional study the effect of aging is potentially confounded with possible cohort effects. Put in a slightly different way, there are many characteristics that differentiate girls in these two different age groups that could distort the relationship between age and body fatness. On the other hand, a longitudinal study that measures a single cohort of girls at both ages 10 and 15 can provide a valid estimate of the change in body fatness as girls age. In the longitudinal study the analysis is based on a paired t-test, using the difference or change in percent body fat within each girl as the outcome variable. This within-individual comparison provides a valid estimate of the change in body fatness as girls age from 10 to 15 years. Moreover, since each girl acts as her own control, changes in percent body fat throughout the duration of the study are estimated free of any between-individual variation in body fatness.

A distinctive feature of longitudinal data is that they are *clustered*. In longitudinal studies the clusters are composed of the repeated measurements obtained from a single individual at different occasions. Observations within a cluster will typically exhibit positive correlation, and this correlation must be accounted for in the analysis. Longitudinal data also have a temporal order; the first measurement within a cluster necessarily comes before the second measurement, and so on. The ordering of the repeated measures has important implications for analysis. There are, however, many studies in the health sciences that are not longitudinal in this sense but which give rise to data that are clustered or cluster-correlated. For example, clustered data commonly arise when intact groups are randomized to health interventions or when naturally occurring groups in the population are randomly sampled. An example of the former is group-randomized trials. In a group-randomized trial, also known as a cluster-randomized trial, groups of individuals, rather than the individuals themselves, are randomized to different treatments or health interventions. Data on the health outcomes of interest are obtained on all individuals within a group. Alternatively, clustered data can arise from random sampling of naturally occurring groups in the population. Families, households, hospital wards, medical practices, neighborhoods, and schools are all instances of naturally occurring clusters in the population

that might be the primary sampling units in a study. Finally, clustered data can arise when data on the health outcome of interest are simultaneously obtained either from multiple raters or from different measurement instruments.

In all of these examples of clustered data, we might reasonably expect that measurements on units within a cluster are more similar than the measurements on units in different clusters. The degree of clustering can be expressed in terms of correlation among the measurements on units within the same cluster. This correlation invalidates the crucial assumption of independence that is the cornerstone of so many standard statistical techniques. Instead, statistical models for clustered data must explicitly describe and account for this correlation. Because longitudinal data are a special case of clustered data, albeit with a natural ordering of the measurements within a cluster, this book includes a description of modern methods of analysis for clustered data, more broadly defined. Indeed, one of the goals of this book is to demonstrate that methods for the analysis of longitudinal data are, more or less, special cases of more general regression methods for clustered data. As a result, a comprehensive understanding of methods for the analysis of longitudinal data provides the basis for a broader understanding of methods for analyzing the wide range of clustered data that commonly arises in studies in the biomedical and health sciences.

The examples described above consider only a single level of clustering, for example, repeated measurements on individuals. More recently, investigators have developed methodology for the analysis of multilevel data, in which observations may be clustered at more than one level. For example, the data may consist of repeated measurements on patients clustered by clinic. Alternatively, the data may consist of observations on children nested within classrooms, nested within schools. Though the analysis of multilevel data is not the primary focus of this book, multilevel data are discussed in Chapter 17.

Interest in the analysis of longitudinal and multilevel data continues to grow. New and more flexible models have been developed and advances in computation, such as Markov chain Monte Carlo (MCMC) methods, have allowed greater flexibility in model specification. Moreover, improvements in statistical software packages, especially SAS and S-PLUS, have made these models much more accessible for use in routine data analysis. Despite these advances, however, methods for the analysis of longitudinal data are not widely used and are seen to be accessible only to statisticians with specialized expertise.

We believe that the methodology for the analysis of longitudinal data can be much more widely understood and applied. It is our hope that this book will help to make that possible. It provides a comprehensive introduction to methods for the analysis of longitudinal data, written for a reader with a basic knowledge of statistics and a strong background in regression analysis. The book does not require a high level of mathematical preparation but does assume a willingness to read and consider mathematical ideas.

1.3 EXAMPLES

To highlight some of the distinctive features of longitudinal and clustered data, we introduce four examples drawn from studies in the biomedical sciences. These four examples will be used later in the book to illustrate different analytic approaches. Additional examples, also drawn from studies in the biomedical and health sciences, will be introduced in later chapters of the book.

1.3.1 Treatment of Lead-Exposed Children (TLC) Trial

Exposure to lead can produce cognitive impairment, especially among young children and infants. A young child exposed to high levels of lead may experience various adverse health effects, including hyperactivity, hearing or memory loss, learning disabilities, and damage to the brain and nervous system. Although the use of lead as an additive in gasoline has been discontinued, at least in the United States, resulting in a dramatic reduction in airborne lead levels, a small percentage of children continue to be exposed to lead at levels that can produce impairment. Much of this exposure is due to deteriorating lead-based paint (e.g., chipping and peeling paint) in older homes. Lead was used as a pigment and drying agent in "alkyd" oil-based paint. While the United States government banned the use of lead-based paint in housing in 1978, many homes built before 1978 contain lead-based paint. When lead-based paint deteriorates it becomes lead paint chips, which can be eaten by young children, and lead-contaminated paint dust, which can be ingested by young children during normal teething and hand-to-mouth behavior. The United States Centers for Disease Control and Prevention (CDC) has concluded that children with blood lead levels above 10 micrograms per deciliter (μg/dL) of whole blood are at risk of adverse health effects.

Lead poisoning in children is treatable in the sense that there are medical interventions, known as chelation treatments, that can help a child to excrete the lead that has been ingested. Until recently, chelation treatment of children with high levels of blood lead was administered by injection and required hospitalization. A new chelating agent, succimer, enhances urinary excretion of lead and has the distinct advantage that it can be given orally, rather than by injection. In the 1990s, the *Treatment of Lead-Exposed Children (TLC) Trial Group* conducted a placebo-controlled, randomized trial of succimer in children with confirmed blood lead levels of 20–44 μg/dL; levels well above the CDC's threshold for concern about the adverse health effects of exposure to lead (Treatment of Lead-exposed Children (TLC) Trial Group, 2000; Rogan *et al.*, 2001). The children were aged 12–33 months at enrollment and lived in deteriorating inner city housing. The mean age of the children at randomization was 2 years and the mean blood lead level was 26 μg/dL. Children received up to three 26-day courses of succimer or placebo and were followed for 3 years.

Table 1.1 presents data on blood lead levels at baseline, week 1, week 4, and week 6 for 10 randomly selected children from the study. The mean blood lead levels at each measurement occasion for a random subset of 100 children, broken down by treatment group, are presented in Table 1.2. As expected, due to randomization, the

Table 1.1 Blood lead levels (μg/dL) at baseline, week 1, week 4, and week 6 for 10 randomly selected children from the TLC trial.

ID	Group[a]	Baseline	Week 1	Week 4	Week 6
79	P	30.8	26.9	25.8	23.8
8	S	26.5	14.8	19.5	21.0
44	S	25.8	23.0	19.1	23.2
11	P	24.7	24.5	22.0	22.5
69	S	20.4	2.8	3.2	9.4
29	S	20.4	5.4	4.5	11.9
46	P	28.6	20.8	19.2	18.4
13	P	33.7	31.6	28.5	25.1
74	P	19.7	14.9	15.3	14.7
53	P	31.1	31.2	29.2	30.1

[a] P = Placebo; S = Succimer.

Table 1.2 Mean blood lead levels (and standard deviation) at baseline, week 1, week 4, and week 6 for children from the TLC trial.

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.5	15.5	20.8
	(5.0)	(7.7)	(7.8)	(9.2)
Placebo	26.3	24.7	24.1	23.6
	(5.0)	(5.5)	(5.8)	(5.6)

mean response at baseline is similar in the two treatment groups. However, there are discernible differences in the patterns of change in the mean response over time. A graphical presentation of the mean blood lead levels at each occasion is displayed in Figure 1.1. Note that at week 1 there appears to be a dramatic drop in initial blood lead levels among the children treated with succimer. However, this is followed by a rebound in blood lead levels, as lead stored in the bones and tissues is mobilized and a new equilibrium is achieved. In contrast, for the children treated with placebo, the trend in the mean response over time is relatively flat.

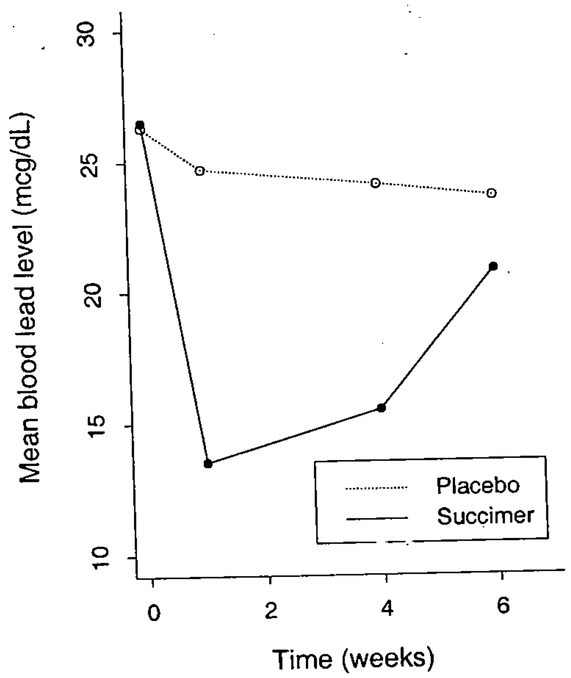

Fig. 1.1 Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

1.3.2 Muscatine Coronary Risk Factor Study

In 1998 the American Heart Association (AHA) announced that obesity had been added to the AHA's list of major preventable risk factors for coronary heart disease. These major preventable risk factors include smoking, high blood cholesterol, high blood pressure, and sedentary lifestyle. Unlike risk factors which cannot be altered, such as heredity, increasing age and being male, obesity is a risk factor that many individuals can alter and control. The medical definition of obesity is quite simple: an excess of body fat. Obesity is primarily caused by consuming too many calories and not getting enough physical exercise. Obesity can lead to higher blood cholesterol and triglyceride levels, lower HDL cholesterol (HDL cholesterol, the "good" cholesterol, has been linked to lower risk of coronary heart disease), and higher blood pressure. Thus obesity can contribute to higher coronary risk in a variety of different ways.

Public health scientists now accept that obesity is a chronic disease, just like high blood pressure or high blood cholesterol. Its causes are a complex, individualized combination of genetics, behavior and lifestyle. There is also increased awareness that obese children are at increased risk for obesity as adults.

Table 1.3 Obesity status of cohort of children, aged 7–9 at entry, from the Muscatine study.

	Child's Obesity Status[a]			
Gender	1977	1979	1981	Count
Males				
No Missing	1	1	1	20
	1	1	0	7
	1	0	1	9
	1	0	0	8
	0	1	1	8
	0	1	0	8
	0	0	1	15
	0	0	0	150
Missing Time 1	*	1	1	13
	*	1	0	3
	*	0	1	2
	*	0	0	42
Missing Time 2	1	*	1	3
	1	*	0	1
	0	*	1	6
	0	*	0	16
Missing Time 3	1	1	*	11
	1	0	*	1
	0	1	*	3
	0	0	*	38
Missing Times 1,2	*	*	1	14
	*	*	0	55
Missing Times 1,3	*	1	*	4
	*	0	*	33
Missing Times 2,3	1	*	*	7
	0	*	*	45
Females				
None Missing	1	1	1	21
	1	1	0	6
	1	0	1	6
	1	0	0	2
	0	1	1	19
	0	1	0	13
	0	0	1	14
	0	0	0	154
Missing Time 1	*	1	1	8
	*	1	0	1
	*	0	1	4
	*	0	0	47
Missing Time 2	1	*	1	4
	1	*	0	0
	0	*	0	16
	0	*	1	3
Missing Time 3	1	1	*	11
	1	0	*	1
	0	1	*	3
	0	0	*	25
Missing Times 1,2	*	*	1	13
	*	*	0	39
Missing Times 1,3	*	1	*	5
	*	0	*	23
Missing Times 2,3	1	*	*	7
	0	*	*	47

[a] 1 = Obese; 0 = Not Obese; * = Missing.

In 1970 researchers from the University of Iowa began to examine the links between child and adult coronary health. Of particular interest were the associations between coronary risk factors in youth and coronary disease in adults. The Muscatine Coronary Risk Factor (MCRF) study, a longitudinal survey of school-age children in Muscatine, Iowa, had the goal of examining the development and persistence of risk factors for coronary disease in children (Woolson and Clarke, 1984; Lauer *et al.*, 1997). In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years, were obtained biennially from 1977 to 1981. Data were collected on 4856 boys and girls. On the basis of a comparison of their weight to age-gender specific norms, children were classified as obese or not obese. One objective was to determine whether the prevalence of obesity increases with age and whether patterns of change in obesity are the same for boys and girls.

A summary of the obesity data for children in one of the five cohorts, who were 7–9 years old in 1977, is presented in Table 1.3. Because all the variables are discrete, the data can be summarized as counts in a contingency table. For example, the first 8 rows of Table 1.3 provide a count of the number of children with each of the 8 (or 2^3) possible sequences of binary responses over the three measurement occasions. A similar table could be constructed for each of the remaining four cohorts of children. Note that although each child was eligible to participate in all three surveys, the data are incomplete for many children. In fact, less than 40% of the children provided complete data at all three measurement occasions. For convenience, in Table 1.3 the missingness of obesity is treated as a third category of the obesity status variable.

1.3.3 Clinical Trial of an Anti-Epileptic Drug

Epilepsy is a chronic neurologic disorder that may result from brain injury, developmental malformation, or a genetic abnormality. It is characterized by recurrent seizures caused by sudden, excessive electrical activity in the brain. Seizures are classified as generalized, in which the electrical discharge occurs throughout the brain, and partial onset, wherein the electrical activity is localized.

Data for the third example come from a placebo-controlled clinical trial of 59 epileptics conducted by Leppik *et al.* (1987). Patients with partial seizures were enrolled in a randomized clinical trial of the anti-epileptic drug, progabide. Participants in the study were randomized to either progabide or a placebo, as an adjuvant to the standard anti-epileptic chemotherapy. Progabide is an anti-epileptic drug whose primary mechanism of action is to enhance gamma-aminobutyric acid (GABA) content; GABA is the primary inhibitory neurotransmitter in the brain.

Prior to receiving treatment, baseline data on the number of epileptic seizures during the preceding 8-week interval were recorded. Counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visits were recorded. The average rates of seizures (per week) at baseline and in the four post-randomization visits are presented in Table 1.4. A graphical presentation of the average rates of seizures at each occasion in the progabide and placebo groups is displayed in Figure 1.2. The main goal of the study was to compare the changes in the average rates of seizures in the two groups.

Table 1.4 Mean rate of seizures per week (and standard deviation) at baseline, week 2, week 4, week 6, and week 8 in the clinical trial of progabide.

Group	Baseline	Week 2	Week 4	Week 6	Week 8
Progabide	3.96	4.29	4.21	4.06	3.37
	(3.5)	(9.1)	(5.9)	(7.0)	(5.6)
Placebo	3.85	4.68	4.14	4.39	4.00
	(3.3)	(5.1)	(4.1)	(7.3)	3.8)

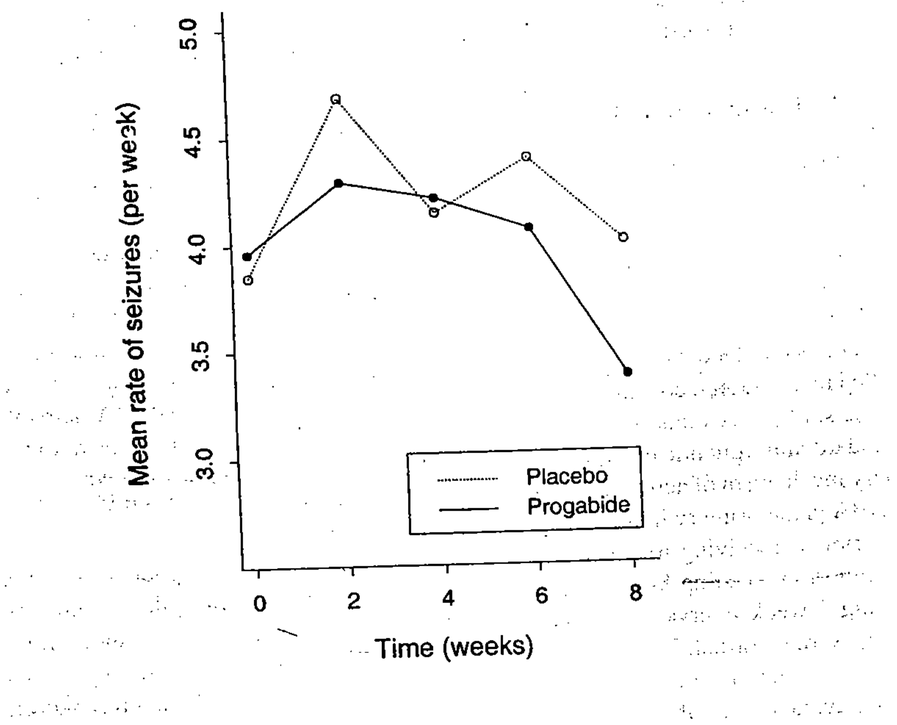

Fig. 1.2 Mean rate of seizures (per week) at baseline, week 2, week 4, week 6, and week 8 in the progabide and placebo groups.

1.3.4 Connecticut Child Surveys

There is now accumulating evidence that the rates of psychiatric disorders in children are substantial, with reported population prevalence rates of childhood psychopathology ranging from 12% to 22%. However, children are considered to be unreliable in reporting on their own psychopathology. As a result, many contemporary surveys of childhood psychopathology use proxy informants, usually a child's parent (or primary caregiver) and teacher, to report on the child's psychiatric status. In numerous studies, the agreement among multiple informant reports on the child's psychopathology has been found to be poor. It is thought that much of this disagreement is less a result of the unreliability of the informant reports than of true differences in children's behaviors and emotions across different situations and settings, most notably in the home and school. A central issue in studies of risk factors for childhood psychopathology is utilization of the information obtained about the child's mental health status from multiple sources or informants.

Data for our example come from two parallel epidemiological surveys that assessed the mental health and service needs of children, aged 6 to 11, in rural and urban communities in Connecticut (Zahner *et al.*, 1992, 1993). The first survey, the New Haven Child Survey (NHCS), was conducted in 1986–1987 in New Haven, Connecticut, a predominantly minority metropolitan center. The second survey, the Eastern Connecticut Child Survey (ECCS) was conducted in 1988–1989 and replicated the NHCS in a non-metropolitan planning region covering the eastern third of Connecticut. The two studies used comparable survey procedures. In particular, they used parallel questionnaires designed to be self-administered by the children's parents and teachers. Children's emotional and behavioral problems were assessed with the Child Behavior Checklist (CBCL) and the Teacher's Report Form (TRF), 118-item symptom inventories covering problems commonly seen in child guidance clinics. The CBCL and TRF scales do not provide diagnoses of psychiatric disorders; instead, they provide broad-band measures of emotional (or "internalizing") and behavioral (or "externalizing") disturbance. The CBCL and TRF scale scores can be dichotomized at published clinical cut-points.

Thus the New Haven Child Survey and the Eastern Connecticut Child Survey provided both a parent's and a teacher's report of psychiatric disturbance in the child as assessed by parallel forms of a standardized psychiatric symptom checklist. These data provide multiple source (here, from two sources: the parent and teacher) information on the psychiatric outcome variable of interest. Of note, these data are cross-sectional but the two sources of information about each child's psychopathology are likely to be positively correlated. Thus data from the Connecticut Child Surveys are an example of clustered, but not longitudinal, data. In this setting, unlike a typical longitudinal study, the major interest of the analysis is not in changes in the response over time. Instead, the major focus of the analysis is on the effects of subject-specific covariates on the outcome.

Table 1.5 displays social and demographic characteristics of the children and the overall rates of externalizing disturbance as determined by CBCL and TRF scale scores in the clinical range.

Table 1.5 Frequency distribution for variables from the Connecticut Child Surveys.

Variables	Count	Percent
Parent Informant (*N*=2501)		
Externalizing		
0 = Normal	2112	84
1 = Borderline/Clinical	389	16
Teacher Informant (*N*=1428)		
Externalizing		
0 = Normal	1159	81
1 = Borderline/Clinical	269	19
Area		
1 = Rural	874	35
2 = Suburban	428	17
3 = Small city	386	15
4 = Large city	813	33
Single Parent		
0 = No	1982	79
1 = Yes	519	21
Child's Health		
0 = Good health	1329	53
1 = Fair/Bad Health	1172	47
Child's Gender		
0 = Female	1284	52
1 = Male	1207	48

The four examples considered in this section differ in terms of outcome variable, study design, and goals or objectives of the analysis. In the first example from the TLC trial, the outcome variable, blood lead level, is continuous. In the second example from the MCRF study, the outcome variable, obesity status, is binary. In the third example from the clinical trial of progabide, the outcome variable is a count. These three examples illustrate the diverse types of longitudinal data that arise in the health and medical sciences. A notable feature of the second example is the amount of missing data. Missing data are a common problem in longitudinal studies in the health sciences. As we will discuss in later chapters, one will need to examine the reasons for any missingness to determine the validity of inferences about changes in the response over time. Next, consider the design of these studies. The first and third examples are experiments, where the treatments have been chosen by the investigators and randomly assigned to the study participants. The second example is an observational study where the study participants are followed forward

in time to observe the outcome variable at future time points; however, unlike the randomized clinical trial, the investigators cannot directly control the comparability of groups (here, males and females). While the first three examples involve longitudinal study designs, the fourth example is a cross-sectional observational study. In the Connecticut Child Surveys, variables are measured at a single time point on a sample of children. Because information on the outcome variable of interest is obtained from two sources (the parent and teacher), these data are also clustered. Finally, we note that the goals of the analysis are similar for the first three examples: characterize the change in the outcome variable over time and the factors that influence change. In the fourth example, however, the objective of the analysis is not to characterize change in the outcome variable over time. Instead, the goal is to examine the effects of subject-specific covariates on the outcome. In later chapters we describe modern methods for analyzing diverse types of longitudinal data arising from both experiments and observational studies. Because longitudinal data are a special case of clustered data, we also describe methods of analysis for clustered data, more broadly defined.

1.4 REGRESSION MODELS FOR CORRELATED RESPONSES

The last 25 years have seen remarkable advances in methods for analyzing longitudinal and clustered data. In particular, there now exists a broad and flexible class of models for correlated data based on a regression paradigm. Indeed all of the methods that are described in later chapters can be thought of as regression models for correlated responses. In this section we provide motivation for the regression paradigm for correlated responses.

Regression models are widely used and provide a very general and versatile approach for analyzing data. Our use of the term "regression model" here is not strictly limited to the standard linear regression model for a continuous response variable. Instead, we use this term more broadly to refer to any model that describes the dependence of the mean of a response variable on a set of covariates in terms of some form of regression equation. While the simplest case is the familiar linear regression model for a continuous response variable, there are many possible generalizations. For example, regression models have been developed for other response variables, such as binary responses or counts. For the binary response variable, linear logistic regression has been widely used for many applications. For counts, Poisson or log-linear regression is often appropriate. Another important generalization is to observations that cannot be assumed to be statistically independent of one another, that is, regression models for correlated responses. In later chapters we consider both kinds of generalizations of the standard linear regression model.

Note that the term "linear" has appeared in all three of the examples of regression models considered so far. Linearity in this setting has a very precise meaning and refers to the fact that all of these models for the mean (or some transformation of the mean) are linear in the regression parameters. For example, letting Y denote the response variable and X a covariate, the following three models for the mean response

$$E(Y) = \beta_1 + \beta_2 X,$$

$$E(Y) = \beta_1 + \beta_2 \log(X),$$

and

$$E(Y) = \beta_1 + \beta_2 X + \beta_3 X^2,$$

are all cases where the mean is linear in the regression parameters (where $E(Y)$ denotes the mean or expectation of Y). All three models are linear in the regression parameters, even if the latter two are non-linear in the covariate. In this book we only consider models where the mean response, or some suitable transformation of the mean response (e.g., log transformation in Poisson regression), is linear in the regression parameters. We do not consider models that are fundamentally non-linear in the regression parameters. For example, the following two models

$$E(Y) = \beta_1 + e^{\beta_2 X},$$

and

$$E(Y) = \beta_1/(1 + \beta_2 e^{-\beta_3 X}),$$

are cases where the mean is non-linear in the regression parameters. However, we remind the reader that our focus on models that are linear in the regression parameters does not preclude relationships between the mean response and covariates that are curvilinear or non-linear. This type of non-linearity can be accommodated by taking appropriate transformations of the mean response (e.g., log transformation in Poisson regression) and the covariates (e.g., log(dose)), and/or by including polynomials. For example, a quadratic trend in the mean response over time can be incorporated by including both time and time^2 in the regression model. The inclusion of transformed covariates in no way violates the "linearity" of the regression model, that is, the model is still linear in the regression parameters.

As noted earlier, we use the term regression model to refer to any model that describes the dependence of the response variable on a set of covariates in terms of some form of regression equation. In particular, the regression parameters express how the mean of the response variable depends on the covariates. For example, in the case of the linear regression model for a continuous response, the regression coefficients express the dependence of the mean of the outcome in terms of a linear combination of the covariates. In the linear logistic model for a binary response, the regression coefficients express the dependence of the log odds of a positive response in terms of a linear combination of the covariates. Note, however, that the log odds is simply a non-linear transformation of the mean or probability of a positive response. Thus, in both of these cases the mean of the response variable, or some appropriate transformation of the mean, is related to a linear combination of the covariates.

One appealing aspect of the regression paradigm concerns the nature of the explanatory variables. A feature of the regression modelling approach is that it can incorporate mixtures of discrete and continuous covariates in a relatively seamless fashion. That is, the covariates can be continuous (and often referred to as quantitative), such as body weight, age, time, and dose. Furthermore, the mean response, or any suitable transformation of the mean, can be related to a continuous covariate in a curvilinear or non-linear fashion by simply taking an appropriate transformation of the covariate or by the inclusion of polynomials (e.g., time and time^2). Alternatively, the covariates can be discrete (or qualitative), such as gender and treatment group. Finally, regression models can include mixtures of discrete and continuous covariates, and products among them. As a result, within a regression paradigm, it is no more difficult to analyze longitudinal data arising from a carefully designed experiment with a single qualitative covariate or factor (e.g., a randomized placebo-controlled longitudinal clinical trial) than from an observational study where there are many covariates, some of which are discrete, the others continuous. Of note, in the latter case, regression models can often be used to distinguish within- and between-subject trends in the response (e.g., "longitudinal" versus "cross-sectional" effects of age); this topic will be discussed in greater depth in later chapters.

Regression models can usually be formulated in such a way that certain regression parameters have interpretations which bear directly on the scientific question of main interest. For example, in a regression model for data from a longitudinal clinical trial, a particular regression coefficient can be given an interpretation in terms of the constant rate of change in the mean response over time in one of the treatment groups. Alternatively, the absence (or setting to zero) of a particular regression coefficient can be given an interpretation in terms of two treatment groups having the same underlying rate of change in the response variable over time.

So far, we have emphasized that it is not necessary to distinguish whether the covariates are continuous or discrete (or a mixture of the two) within a regression paradigm. However, from a purely historical perspective, linear models for a continuous response with only discrete covariates have often been referred to as *analysis of variance* (ANOVA) models. In contrast, linear models for a continuous response with only continuous covariates have often been referred to as *linear regression* models. Indeed, some textbooks and courses in statistics present linear regression and analysis of variance as almost distinct analytic procedures. A large part of the reason for this arbitrary distinction is historical. Analysis of variance had its earliest roots in agricultural applications, especially carefully designed experiments where the responses (e.g., crop yield) could be indexed by one or more classifying factors (e.g., plot, crop variety) or qualitative experimental factors (e.g., different types of fertilizers). In contrast, linear regression was initially developed for the analysis of observational data. Some of the earliest applications of linear regression can be traced back to astronomy. By their very nature, the data arising from studies in astronomy were purely observational (e.g., the positions and magnitudes of the heavenly bodies) and not the product of experimental manipulations. As a result of their somewhat different historical roots, ANOVA and linear regression have often been presented as almost distinct procedures, intended for the analysis of data arising from studies

which differ in design (experimental versus observational) and the nature of the covariates (discrete versus continuous). Later, it was recognized that linear regression is a very general model that incorporates analysis of variance as a special case.

Thus, although many of the commonly used statistical models for correlated data were originally developed for data arising from studies which differed in design, aims, and the nature of the covariates, almost all of these developments fall within the regression paradigm for correlated data. So from a purely pedagogical perspective, it is not necessary to distinguish methods for analyzing longitudinal or correlated data arising from observational studies and from studies with experimental designs. From this point of view, we have purposely chosen not to focus on many of the early developments in methodology for analyzing correlated data, for example, the repeated measures ANOVA and multivariate analysis of variance (MANOVA). Instead, we focus on a more general and versatile regression paradigm that encompasses most, if not all, of the earlier developments as special cases, but can also handle all of the complexities that arise in applications. When viewed as special cases within the regression paradigm, the underlying (and often unrealistic) assumptions made by many of the earliest methods for analyzing correlated data are more readily understood.

In summary, we view the regression paradigm as a very flexible and versatile approach for analyzing longitudinal and correlated data arising from many different types of studies. Regression models can provide a parsimonious description or explanation of how the mean response in a longitudinal study changes with time, and how these changes are related to covariates of interest. Thus, our use of regression models is primarily intended for descriptive purposes, that is, for determining the most salient aspects of patterns of change in the mean response. While this does not necessarily preclude their use as a possible explanation of the underlying probabilistic data generating mechanism that might have produced the repeated responses, the latter is not considered to be the main focus of the analysis. Instead, our primary goal is to provide a simple description of the discernible patterns of change in the response over time, and their relation to covariates, via regression coefficients that bear directly on the scientific questions of main interest.

1.5 ORGANIZATION OF THIS BOOK

The book is organized into four main parts. The first part, consisting of Chapters 1 and 2, provides the reader with an overview of the most salient aspects of longitudinal data. In Chapter 2, we introduce some notation and many of the analytic issues that arise with longitudinal data. We discuss the main features that distinguish longitudinal data from cross-sectional data. We highlight the major goals and objective of longitudinal analysis. We consider the aspect of longitudinal data that complicates their analysis, namely, the correlation among repeated measures on the same individuals. We provide some intuition for how and why the correlation arises in longitudinal data and the potential consequences of ignoring it in the analysis.

The second part, consisting of Chapters 3–9, focuses on methods for analyzing longitudinal data when the response variable is continuous and assumed to have an approximate multivariate normal (or Gaussian) distribution. In Chapter 3, we introduce a general linear regression model for longitudinal data. We present a broad overview of different approaches for modelling the mean response over time and for accounting for the correlation among repeated measures on the same individual. These topics are discussed in much greater depth in subsequent chapters. In Chapter 4 we discuss estimation, via the method of maximum likelihood (ML), and inference concerning the regression coefficients and the covariance among the repeated measures. Longitudinal data present us with two aspects of the data that require modelling: the mean response over time and the covariance among repeated measures on the same individuals. In Chapters 5 and 6, the emphasis is on modelling the mean response. Two main approaches are distinguished: the analysis of response profiles (Chapter 5) and parametric or semi-parametric curves (Chapter 6). In Chapter 7, we discuss models for the covariance in longitudinal data and develop an overall modelling strategy that takes account of the interdependence between the models for the mean and covariance. Chapter 8 introduces a very flexible class of models for analyzing longitudinal known as linear mixed effects models. These models assume that some subset of the regression parameters vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population. Specifically, the mean response is modelled as a combination of fixed effects that are assumed to be shared by all individuals, and random effects that are unique to a particular individual. In Chapter 9, we discuss residual diagnostics for assessing the adequacy of models for longitudinal data and for detecting outlying observations and/or outlying individuals.

The chapters in the second part of the book cover many of the well-established methods for the analysis of longitudinal data and provide the foundation for future chapters that focus on discrete response variables (e.g., repeated binary responses and repeated count data). The third part, consisting of Chapters 10–13, focuses on methods for analyzing longitudinal data with outcomes that are not continuous. When the response is discrete, linear models are no longer appropriate for relating the mean to covariates. Instead, we consider extensions of generalized linear models for longitudinal data. In Chapter 10 we review the most salient features of generalized linear models for a single, univariate response; in later chapters, we discuss how generalized linear models can be extended to handle longitudinal responses. In generalized linear models a suitable non-linear transformation of the mean response is related to the covariates. However, this non-linearity raises some additional issues concerning the interpretation of the regression coefficients. In Chapters 11–12 we present two classes of models for analyzing discrete longitudinal data that account for the correlation among repeated measures in fundamentally different ways. In Chapter 13 we compare and contrast these two classes of models. One of the underlying themes emphasized in Chapters 11–13 concerns how different models for discrete longitudinal data have somewhat different targets of inferences. Thus, to ensure that the regression parameters bear directly on the question of scientific interest, greater care is needed in the choice of model for discrete longitudinal data.

The final part of the book, consisting of Chapters 14–17, focuses on a number of advanced topics. In Chapter 14 we address the issue of missing data in longitudinal studies and the assumptions required to ensure that the methods discussed in earlier

chapters provide valid inferences. In Chapter 15 we consider some aspects of the design of longitudinal studies, including sample size and power, issues concerning the estimation and interpretation of time-varying covariate effects, and cross-sectional versus longitudinal information. In Chapter 16 we discuss regression models for repeated measures and related designs and emphasize how the methods discussed in earlier chapters can be applied in these settings. In Chapter 17 we present an overview of methods for analyzing multilevel data. Chapters 16 and 17 demonstrate how regression models for longitudinal data are special cases of general regression models for correlated data, more broadly defined.

1.6 FURTHER READING

The presentation of methodology for the analysis of longitudinal data in subsequent chapters assumes that the reader has a basic knowledge of statistics and a strong background in regression analysis. A useful review of introductory statistical principles and methods, targeted at applied researchers, can be found in the books by Pagano and Gauvreau (2000) and Altman (1990). A comprehensive overview of regression concepts can be found in Kleinbaum *et al.* (1999); a more advanced presentation of similar topics can be found in Neter *et al.* (1996).

2

Longitudinal Data: Basic Concepts

2.1 INTRODUCTION

In this chapter we present a broad overview of the main objectives of longitudinal analysis and some of the defining features of longitudinal data. Our primary goal is to emphasize that the major focus of the analysis of longitudinal data is on the assessment of within-individual changes in the response variable over time. That is, longitudinal analysis is concerned with estimating how individuals change throughout the duration of the study and examining the factors that influence heterogeneity among individuals in how they change over time. We also review the most salient features of longitudinal study designs, introduce some notation for longitudinal data, and highlight the main aspects of longitudinal data that complicate their analysis. Many of the concepts and issues introduced here will be discussed in much greater depth in later chapters of the book.

2.2 OBJECTIVES OF LONGITUDINAL ANALYSIS

In the health sciences, longitudinal studies play an important role in enhancing our understanding of the development and persistence of disease. There is much natural heterogeneity among individuals in terms of how diseases develop and progress. This heterogeneity is due to genetic, environmental, social and behavioral factors. A longitudinal study design permits the discovery of individual characteristics that can explain these inter-individual differences in changes in health outcomes over time.

The distinguishing feature of longitudinal studies is that the study participants are measured repeatedly throughout the duration of the study, thereby permitting the direct assessment of changes in the response variable over time. In cross-sectional studies, where measurements are obtained at only a *single* point in time, it is not possible to assess individual changes on the basis of a single snapshot of the individual's response taken at a given time. Thus, the defining feature of a longitudinal study is that two or more observations of the response variable, taken at different times, are made on at least some of the study participants. Typically, although not always, longitudinal study designs call for a fixed number of repeated measurements to be made on all study participants at a set of common time points. The occasions of measurement are not necessarily distributed evenly throughout the duration of the study.

By obtaining measurements of the same individuals repeatedly through time, longitudinal studies can address fundamental questions concerning the assessment of within-individual changes in the response variable. The main goal, indeed the *raison d'être*, of a longitudinal study, is to characterize the change in the response over time. While the measurement of within-individual changes is a fundamental objective of a longitudinal study, it is also of interest to determine whether these within-individual changes in the response are related to selected covariates. For example, in the *Treatment of Lead-Exposed Children Trial*, introduced in Chapter 1, repeated measures of blood lead levels were obtained at baseline (or week 0), week 1, week 4, and week 6, thereby allowing assessment of within-individual changes in blood lead levels over a six-week period. In this study, it was not simply of interest to describe the overall pattern of within-individual changes in blood lead levels over time but also to relate these changes to the assigned treatment (placebo versus succimer).

In its most elementary form, a measure of the observed within-individual change in the response can be conceptualized in terms of simple "change scores" or "difference scores", for example, the differences between post-treatment and pre-treatment measurements of the response. The main objective of a longitudinal analysis is to describe trends in these within-individual changes in the response and to relate these changes to selected covariates (e.g., treatment group). This simple notion of within-individual change extends naturally from "difference scores" to more general "response trajectories" over time. For example, a "difference score" happens to be proportional to the slope (or constant rate of change) of a linear response trajectory. However, other kinds of response trajectories, for example, piecewise linear or curvilinear, can be used to parsimoniously smooth and summarize within-individual changes in the response throughout the duration of the study. In either case, the fundamental ideas remain the same: we want to assess and describe within-individual changes in the response over time via comparison of measurements on the same individual taken later in time with those taken earlier.

A longitudinal analysis of within-individual changes proceeds in two conceptually distinct stages. First, within-individual change in the response is characterized in terms of some appropriate summary of the changes in the repeated measurements on each individual during the period of observation (e.g., using "difference scores" or some form of "response trajectory"). Second, these estimates of within-individual changes are then related to inter-individual differences in selected covariates. Although these two stages of the analysis are conceptually distinct, they can be combined in a statistical model for longitudinal data. That is, a single statistical model for longitudinal data can be used to both capture how individuals change over time and to relate within-individual changes in the response to selected covariates.

For example, in the *Treatment of Lead-Exposed Children Trial* the investigators were interested in assessing changes in blood lead levels over time. In particular, they wanted to determine whether chelation treatment with succimer reduced blood lead levels over time relative to any changes in the placebo group. This study question can be addressed in an analysis that compares the two treatment groups in terms of the differences between post-treatment and pre-treatment measurements of blood lead levels. Although the major objective of the analysis is quite clear, there are many ways to construct and test hypotheses concerning treatment effects on changes in blood lead levels over time. For instance, the two treatment groups can be compared in terms of all post-treatment changes in the mean blood lead levels from baseline (or pre-treatment). Alternatively, the two treatment groups can be compared in terms of the rate of decline of blood lead levels over time, where the rate of decline is expressed in terms of a slope. Thus, although the scientific question of interest has a seemingly simple formulation in terms of whether changes in blood lead levels are affected by treatment, there are many different ways to proceed with a longitudinal analysis of these data. The choice of one analytic approach over another will usually depend upon statistical considerations (e.g., issues of precision), the design of the study, and the specific scientific question of interest. These are topics that will be discussed in more detail in later chapters of the book.

Finally, it is an inescapable fact that the assessment of within-subject changes in the response over time can be achieved only within a longitudinal study design. A cross-sectional study simply cannot estimate how individuals change over time since the response is measured at only a single occasion. A longitudinal study can estimate how individuals change and also do so with great precision because each individual acts as his or her own control. By comparing each individual's responses at two or more occasions, a longitudinal analysis can remove extraneous, but unavoidable, sources of variability among individuals. The key point here is that there is natural heterogeneity among individuals in many extraneous variables. Although these extraneous variables are not of any substantive interest, they can potentially have an impact on the response variable. The beauty of a longitudinal study design is that any extraneous factors (regardless of whether they have been measured or not) that influence the response, and whose influence persists but remains relatively stable throughout the duration of the study (e.g., gender, socioeconomic status, and many genetic, environmental, social and behavioral factors), are eliminated or blocked out when an individual's responses are compared at two or more occasions. By eliminating these major sources of variability or "noise" from the estimation of within-individual change, a very precise estimate of change can often be obtained.

In summary, the fundamental objective of a longitudinal analysis is the assessment of within-individual changes in the response and the explanation of systematic differences among individuals in their changes. Given that certain individuals change more (or less) than others, the goal of a longitudinal analysis is to determine whether

these individuals have larger or smaller values on selected covariates. Finally, in some longitudinal studies, it may also be of interest to make predictions about how specific individuals change over time. In the latter case, longitudinal studies permit more reliable prediction by borrowing information from all individuals to better predict within-individual change over time for a specific individual.

2.3 DEFINING FEATURES OF LONGITUDINAL DATA

At this point we need to introduce some terminology that will be used throughout the remainder of the book. We also introduce some notation for longitudinal data and highlight the main aspects of longitudinal data that complicate their analysis, namely, the correlation among repeated observations obtained on the same individual.

2.3.1 Terminology

In a longitudinal study the participants, or, more generally, the units being studied, are referred to as *individuals* or *subjects*. In many, but certainly not all, longitudinal studies, the individuals are human subjects. In other longitudinal studies, the individuals may be animals (e.g., laboratory mice or rats). Depending upon the specific context, we use the terms *individuals* and *subjects* interchangeably to refer to the participants in a longitudinal study. As mentioned earlier, in a longitudinal study individuals are measured repeatedly at different *occasions* or *times*. Later we will introduce some notation that can distinguish the responses from different individuals in a longitudinal study as well as the repeated measurements on any particular individual. Thus, adopting the terminology introduced so far, the defining feature of a longitudinal study design is that measurements of the response variable are taken on the same *individuals* at several *occasions*.

The number of repeated observations, and their timing, can vary widely from one longitudinal study to another. For example, a clinical trial designed to examine the efficacy of a new analgesic agent may take repeated measures of a self-reported pain scale at baseline and at the end of six 15-minute intervals. This would result in seven repeated measures that are equally separated in time. On the other hand, an observational study of human growth may take measurements of height and weight at 3-month intervals from birth to age 2 years, followed by yearly observations from infancy through young adulthood. By design, the latter study would result in a sequence of repeated measures of height and weight that are unequally separated in time. In both of these examples, the number and the timing of the repeated measurements are the same for all individuals, regardless of whether the occasions of measurement are equally or unequally distributed throughout the duration of the study. Loosely borrowing statistical terminology from the field of experimental design, we refer to the latter studies as being "balanced" over time, that is, all individuals have the same number of repeated measurements obtained at a common set of occasions.

It is an almost inescapable feature of longitudinal studies in the health sciences, especially those where the repeated measurements extend over a relatively long duration, that some individuals will miss their scheduled visit or date of observation. In some studies this may necessitate that observations be made some time before or after the scheduled time. Consequently, the sequence of observation times is no longer common to all individuals in the study due to mistimed measurements. In that case, we refer to the data as being "unbalanced" over time, that is, the repeated measurements are not obtained at a common set of occasions. Unbalanced longitudinal designs are commonplace when the longitudinal study involves retrospectively collected data (e.g., longitudinal data obtained from medical record databases). Alternatively, highly unbalanced longitudinal data can arise when it is of interest to define the timings of the measurements relative to some benchmark event that occurs during the follow-up period. For example, in a study examining changes in body fat in girls before and after menarche (to be discussed in Section 8.8), the study was designed to begin annual follow-up measurements of body fat prior to menarche and continue for four years after menarche. Although this study design is balanced if the timing of measurements is defined as the time since the baseline measurement, the data are inherently unbalanced if the timing of measurements is defined as the time since an individual experienced menarche. Thus longitudinal studies that are balanced over time when the timing of measurements is defined according to one origin can become highly unbalanced when time is defined in terms of a different origin.

Although longitudinal designs that are unbalanced over time often arise due to happenstance, they are sometimes planned by the investigators. In a "rotating panel" study design, which is commonly used in health surveys to reduce response burden, individuals rotate in and out of the study after providing a pre-determined number of repeated measures. For example, two or more "panels" of individuals are measured repeatedly for a restricted number of occasions, with the first measurement for each "panel" of individuals being staggered. Thus, some individuals rotate out (either temporarily or permanently) of the sample, whereas other individuals rotate in to the sample. The primary motivation for this type of study design is to reduce costs and the overall burden of participating in the study for any individual, while providing observations at every occasion for some pre-determined proportion of the sample. An important characteristic of the rotating panel design is that the number and timing of the measurements is pre-determined and by design. Furthermore, the decision about whether to obtain a measurement on an individual at any specific occasion is pre-determined *a priori* by the investigators and is not related to the response variable.

Missing data are a common problem in longitudinal studies. Indeed, missing data are the rule, not the exception, in longitudinal studies in the health sciences. Study participants do not always appear for a scheduled observation or simply leave the study before its completion. When some observations are missing, the data are necessarily unbalanced over time since not all individuals have the same number of repeated measurements obtained at a common set of occasions. However, to distinguish missing data in a longitudinal study from other kinds of unbalanced data, such data sets are often referred to as being "incomplete". This distinction is important

and emphasizes the fact that an intended measurement on an individual could not be obtained.

One of the consequences of lack of balance and/or missing data is that it requires some care to recover within-individual change. For example, consider a setting where each individual is measured on each of n occasions. Then consider plotting the mean response at each occasion. Differences in the mean response over time measure the within-individual change. This is because the difference in the means is also the mean of the differences when each subject is measured at every occasion. When data are missing, and especially when there is attrition of subjects whose responses are different from those who remain in the study, then a plot of the mean response over time can be misleading; changes over time may reflect the pattern of missingness or the attrition, and not within-individual change. As we will discuss in later chapters, one will need to examine assumptions and the appropriateness of the analysis carefully to determine the validity of the inferences with unbalanced designs and/or missing data. Although the methods discussed in this book are designed to handle unbalanced designs and missing data, it is worth keeping in mind that it is always preferable to have balanced designs, because these designs can only capture within-individual change.

When longitudinal data are incomplete there are ramifications for their analysis that go beyond whether a particular statistical method can handle unbalanced longitudinal data. First, when there are missing data it should be intuitively clear that there must necessarily be some loss of information. Thus there is a price to be paid in terms of efficiency or the precision with which changes over time can be estimated. However, besides causing inefficiency, in some circumstances missing data can introduce bias in the estimates of change. As a result, when longitudinal data are incomplete the reasons for any missingness must be carefully considered. In Chapter 14 we discuss some of the consequences of incomplete data in longitudinal studies. In all subsequent chapters we allow for missing data but implicitly make assumptions about the reasons for any missingness. These assumptions are discussed in Section 4.3 and spelled out in greater detail in Chapter 14.

In summary, longitudinal data can be balanced and complete when all individuals are measured at a common set of occasions and there are no missing data. In our experience, longitudinal data in the health sciences are rarely balanced and complete unless the subjects lack human volition (e.g., laboratory rats) or the length of the study is relatively short (e.g., a longitudinal study of the efficacy of an analgesic where the repeated measurements can be obtained in a single study visit). It is far more common to have longitudinal data that are unbalanced and/or incomplete. As a result, to be of real practical use, methods for the analysis of longitudinal data must be able to handle data that are unbalanced over time and possibly incomplete.

Finally, an aspect of longitudinal data that features prominently in their statistical analysis is that repeated measures on the same individual are usually positively correlated. As mentioned earlier, correlated observations are a positive feature of longitudinal data because they provide more precise estimates of the rate of change or the effect of covariates on that rate of change than would be obtained from an equal number of independent observations of different individuals. Nevertheless, the correlation among repeated measures violates the fundamental assumption of independence that

is the cornerstone of so many standard regression techniques. In later sections, we consider the different sources and nature of the correlation among longitudinal data, and the potential consequences of not accounting for it in the analysis.

2.3.2 Notation

Next we introduce some notation that will be used extensively throughout the book. Let Y_{ij} denote the response variable for the i^{th} individual $(i = 1, ..., N)$ at the j^{th} occasion $(j = 1, ..., n)$. If the repeated measures are assumed to be equally separated in time, this notation will be sufficient. Later, however, we will need to refine the notation to handle the case where the repeated measures are unequally separated and unbalanced over time.

In the statistical literature, the usual convention is to denote a random variable by an upper-case letter (e.g., Y_{ij} is the response variable for the i^{th} individual at the j^{th} occasion) and the realized value of a random variable by the corresponding lower-case letter (e.g., y_{ij} denotes the realized value of Y_{ij}). For the most part, we adopt this convention throughout the book. However, whenever we deviate from this convention, it should be clear from the context whether we are referring to a random variable or to its realized value. In Table 2.1 we represent the n observations (or realized values of Y_{ij}) on the N individuals in a two-dimensional array, with rows corresponding to individuals and columns corresponding to the responses at each occasion. Given that we have n repeated measures of the response variable on the same individual, we can group these into a $n \times 1$ response vector, denoted by

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix}.$$

For notational convenience, we can denote the response vectors Y_i in a completely equivalent way as

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})'.$$

(Readers unfamiliar with vectors and matrices should take this opportunity to review the *Gentle Introduction to Vectors and Matrices* in Appendix A; because vectors and matrices are used extensively throughout this book to simplify notation, the reader is required to have some basic facility with the addition and multiplication of vectors and matrices.)[1]

In the analysis of data from a longitudinal study, the main interest is in the mean response, in particular, changes in the mean response over time and how these changes depend upon covariates (e.g., treatment group, exposures). We denote the mean or

[1] Another common convention in the statistical literature is the use of bold type for vectors (and sometimes for matrices). As it will be clear from the context, we do not do so throughout this book.

Table 2.1 Tabular representation of longitudinal data, with n repeated observations on N individuals.

Individual	Occasion				
	1	2	3	...	n
1	y_{11}	y_{12}	y_{13}	...	y_{1n}
2	y_{21}	y_{22}	y_{23}	...	y_{2n}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
N	y_{N1}	y_{N2}	y_{N3}	...	y_{Nn}

expectation of each response Y_{ij} by

$$\mu_j = E(Y_{ij}),$$

where $E(\cdot)$ can be loosely thought of as denoting a long-run average over a large population of subjects at the j^{th} occasion. A somewhat more precise definition of the expectation of Y_{ij} (and of expectation more generally) is that it is a *weighted* average of all the possible values of Y_{ij}, with weights being the probabilities of occurrence of each possible value. So far, our discussion of the mean of Y_{ij} has assumed that the mean response can change over time; this is reflected in our use of a single-letter subscript for the mean, μ_j. In many longitudinal studies, the main goal is to relate changes in the mean response over time to covariates. To additionally allow the mean response and, in particular, changes in the mean response, to vary from individual to individual as a function of individual-level covariates, we require the use of double-letter subscripts,

$$\mu_{ij} = E(Y_{ij}).$$

Here, expectation denotes a long-run average over a large subpopulation of subjects who share similar values of the covariates (e.g., subjects assigned to the active treatment group, unexposed subjects) at the j^{th} occasion. In this notation, the mean response can change over time (denoted by the dependence of μ_{ij} on the subscript j) and changes in the mean response can be related to individual-level covariates (denoted by the dependence of μ_{ij} on the subscript i). A simple illustration of a model for the mean response that depends upon time, and that allows changes in the mean response to also depend upon covariates, is presented in Section 2.4. In Chapters 5 and 6, we present a detailed discussion of two broad approaches for modelling changes in the mean response over time and for relating these changes to covariates.

Next we consider the correlation or dependence among the n responses on the same individual. The notions of dependence and independence have precise meanings in statistics. Specifically, two variables are said to be *independent* if the conditional distribution of one of them does not depend on the other. For example, LDL cholesterol level would be considered independent of gender if the distribution of LDL cholesterol level were the same for males and females. Many standard statistical techniques (e.g., linear regression and analysis of variance for a single, univariate response) make the assumption that the study observations are realizations of random variables that are independent of one another. This assumption will be quite reasonable when the study design calls for one observation to be obtained from each individual and individuals are randomly selected from a larger population. Strictly speaking, the observations are independent only when the sampling of individuals is done with replacement; however, when the population is large relative to the sample size, any dependence induced by sampling without replacement is negligible and can be ignored for all practical purposes. The independence assumption is also justified when the study calls for one observation to be obtained from each individual and individuals are randomly assigned to different treatment conditions. Moreover, the assumption of independent observations can often be justified on purely physical or scientific grounds when the responses from distinct individuals in the study are considered to be completely unrelated to each other. That is, the response of one individual neither influences or is influenced by the response of another. However, in the case where more than a single observation is obtained on the *same* individual, the assumption of independent observations is simply untenable. That is, the response of an individual on one occasion is very likely to be predictive of the response of the same individual at a future occasion. For example, an individual with a high LDL cholesterol level on one occasion is very likely to also have a high LDL cholesterol level on the next occasion. Similarly, an individual with a low LDL cholesterol level on one occasion is likely to have a low LDL cholesterol level on the next occasion. Put simply, with repeated observations on the same individual, past responses are predictive of future responses. Moreover, with a quantitative response variable, this dependence among the repeated measures on the same individual can be characterized by their correlation. As mentioned earlier, the correlation among repeated measures is a positive feature of longitudinal data because correlated observations provide more precise estimates of the rate of change or the effect of covariates on that rate of change than would be obtained from an equal number of independent observations of different individuals. As we will see in later chapters, models for longitudinal data put the correlation among repeated measures to good advantage when estimating changes in the response over time.

2.3.3 Dependence and Correlation

In Section 2.5 we discuss the different sources of correlation among longitudinal data. Before doing so, we must define the term "correlation". To simplify the discussion of correlation, we consider a simple longitudinal design that is balanced and complete, with n repeated measurements of the response variable made at a common set of occasions on N individuals.

Before we can give a formal definition of correlation we need to introduce the notions of *variance* and *covariance*. If we denote the expectation or mean of Y_{ij} by

$$\mu_{ij} = E(Y_{ij}),$$

then the variance of Y_{ij} is defined as

$$\sigma_j^2 = E\{Y_{ij} - E(Y_{ij})\}^2 = E(Y_{ij} - \mu_{ij})^2.$$

While the mean, μ_{ij}, provides a measure of the location of the center of the distribution of Y_{ij}, the variance provides a measure of the spread or dispersion of the values of Y_{ij} around their mean. The positive square-root of the variance, σ_j, is known as the *standard deviation*. (Readers unfamiliar with expectations and variances are encouraged to take this opportunity to review the *Properties of Expectations and Variances* in Appendix B.) Note that in our discussion of the variance we have implicitly assumed that it can vary from occasion to occasion (reflected in our use of a single-letter subscript, σ_j^2). In principle, the variance can also be allowed to depend on individual-level covariates; this would require the use of double-letter subscripts. For ease of exposition we have chosen not to do so; in later chapters we will discuss how the variances can vary not only from one occasion to another, but also as a function of selected covariates.

Next we consider the dependence among the responses in a longitudinal study. The *covariance* between the responses at two different occasions, say Y_{ij} and Y_{ik}, is denoted by

$$\sigma_{jk} = E\left\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right\},$$

and provides a measure of the *linear* dependence between Y_{ij} and Y_{ik}. The covariance between Y_{ij} and Y_{ik} can take on both positive and negative values. When the covariance is zero, there is no linear dependence between the responses at the two occasions. The magnitude of the covariance depends not only on the degree of dependence between the two variables but also on their units of measurement. Any changes in the measurement scales will result in a change in the value of the covariance. For example, the covariance between body weight and LDL cholesterol level will be different if body weight is measured in kilograms rather than pounds. Of note, the covariance of a variable with itself (e.g., the covariance between Y_{ij} and Y_{ij}) is simply the variance of the variable.

While the sign (positive or negative) of the covariance indicates whether there is positive or negative dependence between the two variables, the magnitude of the covariance is somewhat difficult to interpret without comparison to the underlying variability of the two variables. For example, if $\sigma_{jk} = 10$, this information alone indicates that there is dependence between Y_{ij} and Y_{ik} (since $\sigma_{jk} \neq 0$) and that the dependence is positive (i.e., Y_{ij} increases as Y_{ik} increases and vice versa). However, depending on the magnitude of the variances of Y_{ij} and Y_{ik}, $\sigma_{jk} = 10$ may indicate weak or strong dependence. As a result, the covariance alone is not too informative; it must be interpreted relative to the magnitude of the variances of the two variables. To provide a measure of linear dependence between Y_{ij} and Y_{ik} that is in some sense

free of the units of measurement (or variability) of the two variables, the correlation is widely used.

The correlation between Y_{ij} and Y_{ik} is denoted by

$$\rho_{jk} = \frac{E\left\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right\}}{\sigma_j \sigma_k},$$

where σ_j and σ_k are the standard deviations of Y_{ij} and Y_{ik}, respectively. The correlation, unlike the covariance, is a measure of dependence that is unitless or free of the scales of measurement of Y_{ij} and Y_{ik}. This is achieved by dividing each variable by its respective standard deviation. As a result, the correlation between body weight and LDL cholesterol level is the same regardless of whether body weight is measured in kilograms or pounds. This makes it a more readily interpretable measure of linear dependence between two variables. Note that when the covariance is zero, so too is the correlation.

By definition, correlation must take values between -1 and 1. Recall that a correlation of 1 or -1 is obtained when there is a perfect linear relationship between the two variables. That is, if pairs of values of Y_{ij} and Y_{ik} were plotted as points on a two-dimensional scatter-plot, the resulting points would lie perfectly along a straight line when $\rho_{jk} = \pm 1$. As the points depart from a perfect straight-line relationship, the correlation moves closer to zero. A positive correlation implies that one variable increases as the other variable increases. On the other hand, a negative correlation implies that one variable decreases as the other increases. Although two variables that are statistically independent of one another will necessarily be uncorrelated, variables can be uncorrelated without being independent (since correlation only measures *linear* dependence). Statistical independence is a stronger condition than zero correlation; it implies no dependence whatsoever, that is, no *linear* or *non-linear* dependence between the variables. On the other hand, correlation quantifies the degree to which two variables are related or dependent, provided that the dependence is *linear*.

With longitudinal data the repeated measures on the same individual are anticipated to be positively correlated. When the n repeated measures are collected into a vector $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})'$ we can define the variance-covariance matrix to be the following two-dimensional array of variances and covariances:

$$\mathrm{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Y_{i1}) & \mathrm{Cov}(Y_{i1}, Y_{i2}) & \cdots & \mathrm{Cov}(Y_{i1}, Y_{in}) \\ \mathrm{Cov}(Y_{i2}, Y_{i1}) & \mathrm{Var}(Y_{i2}) & \cdots & \mathrm{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Y_{in}, Y_{i1}) & \mathrm{Cov}(Y_{in}, Y_{i2}) & \cdots & \mathrm{Var}(Y_{in}) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix},$$

where $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk}$ (and we have implicitly assumed that the variances and covariances are constant across individuals). Note that there is a symmetry to this matrix in the sense that $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_{ik}, Y_{ij})$. Also, recall that the covariance of a variable with itself is the variance. Thus we can denote

$$\sigma_{kk} = \text{Cov}(Y_{ik}, Y_{ik}) = \text{Var}(Y_{ik}) = \sigma_k^2.$$

For the remainder of the book, and to avoid any potential confusion, we denote the standard deviation and variance of Y_{ik} by σ_k and σ_k^2, respectively. Also, we often refer to the variance-covariance matrix of Y_i as the covariance (matrix) of Y_i or simply $\text{Cov}(Y_i)$. Thus

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

We can also define the correlation matrix, $\text{Corr}(Y_i)$, in terms of a similar two-dimensional array,

$$\text{Corr}(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}.$$

This matrix is also symmetric in the sense that $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = \text{Corr}(Y_{ik}, Y_{ij})$. Also, the diagonal elements of the matrix are all equal to 1 since they denote the correlation of a variable with itself.

With longitudinal data the usual assumptions for standard regression analysis do not hold. In particular, repeated observations on the same individual are not independent and the variance of the repeated measurements is not usually constant over the duration of the study (e.g., the variance of baseline measurements is often discernibly smaller than post-baseline measurements). With longitudinal data, heterogeneity of variance over time can be accounted for by allowing the elements on the main diagonal of the covariance matrix to differ. The lack of independence among the repeated measurements is accounted for by allowing the off-diagonal elements of the covariance and correlation matrices to be non-zero. Moreover, with longitudinal data, the correlations are expected to be positive and the sequential nature of longitudinal data implies that there may be a pattern to the correlations. For example, a pair of repeated measures that have been obtaining closely together in time are expected to be more highly correlated than a pair of repeated measures further separated in time. In general, with longitudinal data the correlation among the repeated measures is expected to decline with increasing time separation. In later chapters of this book we will discuss models for the covariance matrix that attempt to capture this structure or pattern in the correlations and that allow the variances to change over time.

In the following section we consider a simple example to highlight the main objectives of a longitudinal analysis and to reinforce the concepts of covariance and correlation that were introduced earlier.

2.4 EXAMPLE: TREATMENT OF LEAD-EXPOSED CHILDREN TRIAL

In this simple illustration we consider data from the *Treatment of Lead-Exposed Children Trial*. The TLC trial was a placebo-controlled, randomized study of succimer in children with blood lead levels of 20–44 μg/dL. Recall that the data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or to placebo. Although there were some minor departures from the measurement schedule (e.g., due to mistimed measurements), for the purposes of illustration, we regard these data as arising from a balanced design.

Objectives of Analysis

In general, the main objective of a longitudinal analysis is to describe changes in the mean response over time, and how these changes are related to covariates of interest. In the TLC trial, the investigators were interested in determining whether chelation treatment with succimer reduces blood lead levels over time relative to any changes observed in the placebo group. Although the scientific objective of this study is clear, there are many possible ways to express this question in terms of within-individual changes in blood lead levels. For instance, the null hypothesis of no treatment effect on changes in blood lead levels over time could be expressed as:

$$\text{H}_0\text{: } \mu_j(S) = \mu_j(P), \quad \text{for all} \quad j = 1, ..., 4,$$

where the notation $\mu_j(S)$ and $\mu_j(P)$ is used to denote the mean response at the j^{th} occasion in the succimer and placebo groups, respectively. This null hypothesis states that the mean responses *at every time point* coincide or are equal in the two treatment groups. As we mentioned earlier, the regression approach to modelling longitudinal data can be formulated in such a way that certain regression parameters correspond to the scientific question of interest. Here, a regression model for the blood lead level data might include main effects for treatment group and time, in addition to their interaction. The null hypothesis given above can then be expressed in terms of the regression parameters for both the main effect of treatment group and the time by treatment group interaction.

Alternatively, the null hypothesis of no treatment effect on change in blood lead levels over time could be expressed as:

$$\text{H}_0\text{: } \mu_j(S) - \mu_1(S) = \mu_j(P) - \mu_1(P), \quad \text{for all} \quad j = 2, ..., 4.$$

This null hypothesis states that all changes in the mean response from baseline are equal in the two treatments groups. Of note, this second version of the null hypothesis

is implied by the first. The second version is somewhat less restrictive in that the treatment groups could have differences in means at baseline but identical changes from baseline over time. As we shall see later in this book, there are a variety of ways to handle baseline measurements in the analysis of longitudinal data. Once again, a regression model can be formulated corresponding to this second version of the null hypothesis. Specifically, the null hypothesis can be expressed in terms of the regression parameters for the treatment group by time interaction (in a model that includes main effects for treatment group and time).

Finally, a third possibility is to express the null hypothesis in terms of the rate of decline of blood lead levels in the two treatment groups, where the rate of decline or trajectory over time is defined parametrically (e.g., in terms of the slope of a linear response trajectory). However, before we can express and test this null hypothesis, we need to specify more precisely what we mean by rate of decline. In Chapter 6 we will describe how simple parametric (e.g., linear or quadratic) or semi-parametric curves (e.g., piecewise linear) can be used to describe trajectories of the mean response changes over time. From a statistical perspective, expressing the null hypothesis in terms of simple parametric curves can result in tests of treatment effects that have greater statistical power.

Although there are a number of different ways to express the null hypothesis, in all three instances the main scientific goal is to establish whether changes in the mean response are affected by treatment. More generally, in most longitudinal studies the primary focus is on determining whether the mean response changes over time and whether the changes are related to covariates. The statement of the study hypothesis will depend to a certain extent on the design of the study and the specific goals of the analysis. In Chapters 5 and 6 we will consider this issue in much greater detail.

Correlation and Covariance

In our discussion of correlation, for ease of exposition, we restrict attention to the longitudinal data from the placebo treated group in this trial. Recall that the data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6. Thus, for the subset of 50 children who were randomly assigned to the placebo group, we let Y_{ij} denote the blood lead level for the i^{th} individual ($i = 1, ..., 50$) at the j^{th} occasion ($j = 1, ..., 4$).

The inter-dependence (or time-dependence) among the four repeated measures of blood lead level can be examined by constructing a scatter-plot of each pair of repeated measures. Figure 2.1 displays scatter-plots constructed for all six possible pairings of the four repeated measures. Figure 2.1 indicates that there is a relatively strong positive relationship between repeated measures of blood lead levels over time.

The estimated covariances and correlations among the four repeated measures are displayed in Tables 2.2 and 2.3. Examination of the main diagonal of the covariance matrix reveals that the variances increase over time. In our experience, increasing variance over time is a very common characteristic of longitudinal data. Thus the changing variance of longitudinal data is another type of nuisance problem that is non-standard in most regression settings. Examination of the correlations in Table

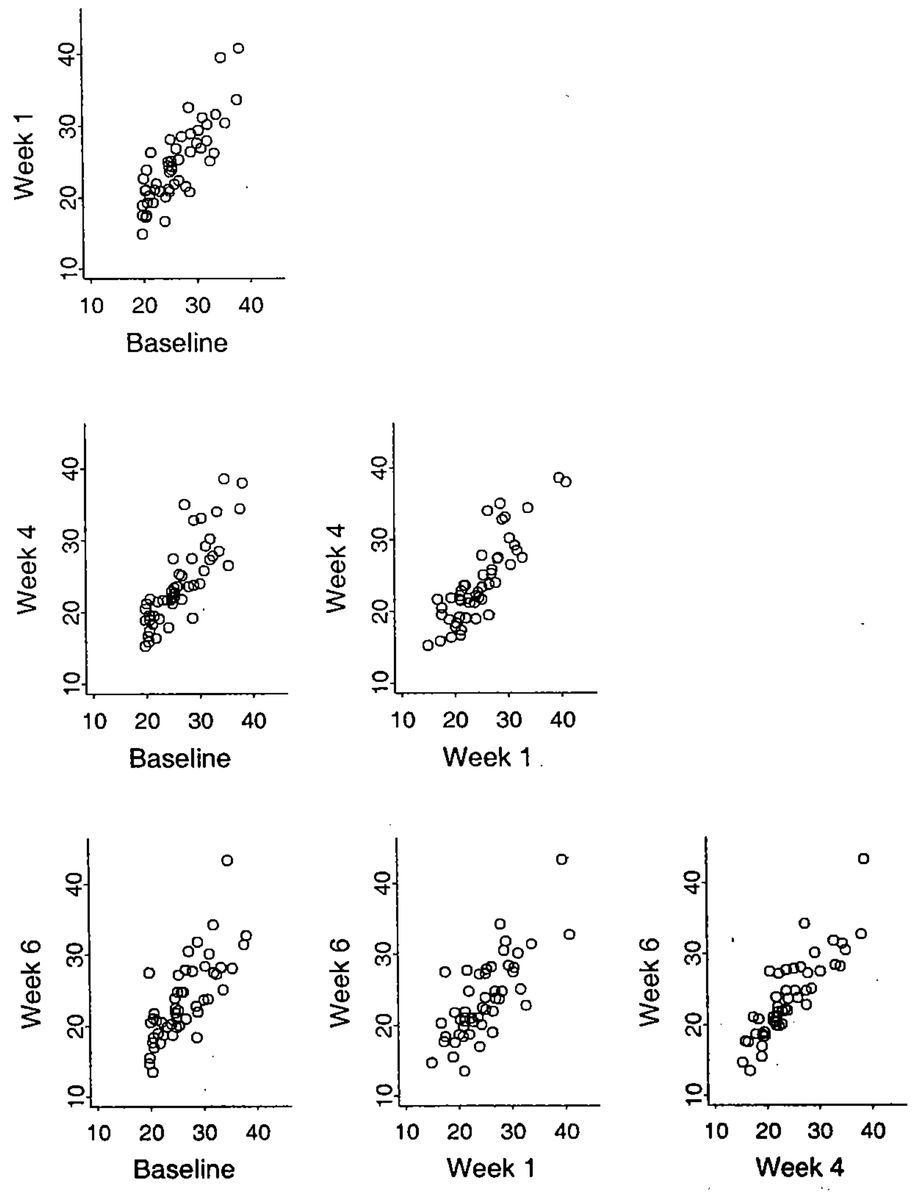

Fig. 2.1 Pairwise scatter-plots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Table 2.2 Estimated covariance matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Covariance Matrix			
25.2	22.8	24.3	21.4
22.8	29.8	27.0	23.4
24.3	27.0	33.1	28.2
21.4	23.4	28.2	31.8

Table 2.3 Estimated correlation matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Correlation Matrix			
1.00	0.83	0.84	0.76
0.83	1.00	0.86	0.76
0.84	0.86	1.00	0.87
0.76	0.76	0.87	1.00

2.3 confirms that the correlations are all positive and that the correlation shows a tendency to decrease with increasing time separation.

While the scatter-plots in Figure 2.1 provide a clear indication of the positive correlation among the repeated measures, there is another, albeit less obvious, way to graphically assess the dependence among the repeated measures. This can be achieved using a single scatter-plot that plots the responses on the vertical axis and the times of measurements on the horizonal axis, with successive repeated measures on the same individual joined with straight lines; we refer to the resulting display as a *time plot*. The dependence among the repeated measures is assessed by comparing the relative amount of between-subject and within-subject variability. It is usually sufficient, and generally more informative, to produce this scatter-plot for only a few randomly selected individuals; it can be very difficult to discern the two distinct sources of variability in a scatter-plot based on all of the individuals in the study. In Figure 2.2, based on 4 randomly selected individuals from the placebo group in the TLC trial, we see that there is very substantial within-subject variability in blood lead levels. This can be discerned from the somewhat jagged appearance of the line

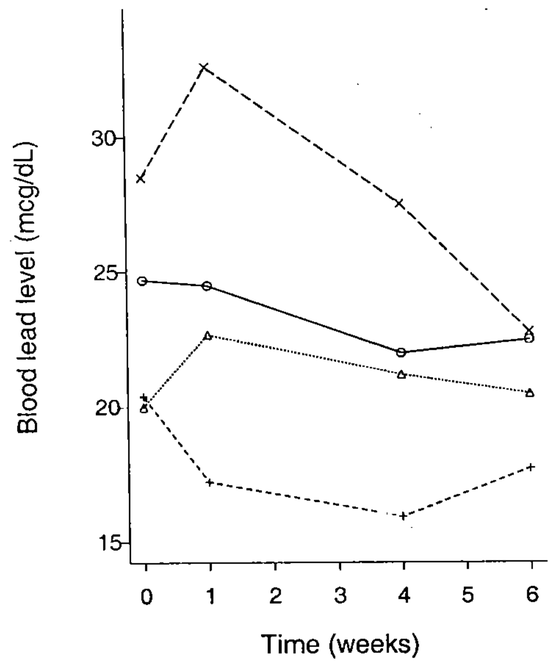

Fig. 2.2 Time plot of blood lead levels at baseline, week 1, week 4, and week 6 for four randomly selected children from the placebo group of the TLC trial.

segments that join the repeated measures on any individual. In addition, there is also substantial between-subject variability. This can be discerned from the fact that some of the individuals have consistently high blood lead levels at all four occasions, while others have consistently low blood lead levels. At first glance this appears to be a very indirect way to assess the degree of dependence among repeated measures and, in our experience, it is not usually the most satisfactory or informative graphical display of that dependence. Nonetheless, it does provide a direct explanation for one of the major sources of the correlation among repeated measures, namely, between-individual heterogeneity. In the next section we examine the three major sources of the correlation among repeated measures in a longitudinal study.

We have seen that with longitudinal data the usual assumptions for standard regression analysis do not hold. In particular, repeated observations on the same individual are not independent and the variance of the repeated measurements is not usually constant over the duration of the study (e.g., the variance of baseline measurements is often discernibly smaller than post-baseline measurements). The correlation among repeated measures is a positive feature of longitudinal data because correlated obser-

vations provide more precise estimates of the rate of change than would be obtained from an equal number of independent observations of different individuals. Although it is important to take this correlation into account in the analysis, the correlations may not be of substantive interest in their own right. If so, we need to accommodate the correlation in an analysis of longitudinal data, but the correlation is not the main focus of the analysis *per se*. Instead, the main interest in any longitudinal study is in describing changes in the mean response over time, and how these changes are related to covariates of interest. For example, in the TLC trial, the main interest is in determining whether chelation treatment with succimer reduces blood lead levels over time relative to any changes in the placebo group. There is no substantive interest in the correlation among the four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6.

2.5 SOURCES OF CORRELATION IN LONGITUDINAL DATA

In this section we consider some of the potential sources of the correlation within longitudinal data. While it is almost an article of faith that longitudinal data are correlated, it is worth pausing to consider why this is the case and, moreover, why longitudinal data are usually positively correlated. Our practical experience with many longitudinal studies in the biological and health sciences has led to the following empirical observations about the nature of the correlation among repeated measures in longitudinal studies: (i) the correlations are positive, (ii) the correlations often decrease with increasing time separation, (iii) the correlations between repeated measures rarely ever approach zero, even in cases where they are taken many years apart, and (iv) the correlation between a pair of repeated measures taken very closely together in time rarely approaches one. These empirical observations have led us to conclude that there are generally three potential sources of variability that have an impact on the correlation among repeated measures on the same individual: (i) between-individual heterogeneity, (ii) within-individual biological variation, and (iii) measurement error. Next, we examine each of these sources of variability in turn and discuss their impact on the correlation among repeated measures.

Between-Individual Heterogeneity

The first source of variability is between-subject heterogeneity and this reflects natural variation in individuals' propensity to respond. In any longitudinal study, some individuals consistently respond higher than the average, while others consistently respond below the average. Thus one source of the positive correlation among repeated measures is the heterogeneity or variability in the response variable between individuals in the population. For almost every health outcome that might be of interest, we can expect to find some degree of heterogeneity. In a certain sense, there are always some individuals who are "high respondents" (e.g., individuals with high blood pressure), some who are "low respondents" (e.g., individuals with low blood pressure) and the remainder who are "medium respondents" (e.g., individuals with blood pressure within the so-called normal range).

The central idea that has been introduced here is that each individual's underlying propensity to respond, whether it be "high", "medium" or "low", and whether it be due to genetic, environmental, social or behavioral factors (or some combination of these factors), is shared by all of the repeated measures obtained on that individual. As a result, an individual with a high value for the response variable at one occasion will be expected to have a relatively high value at subsequent occasions. Consequently, a pair of repeated measures on the same individual will be expected to be more similar than single observations obtained from two randomly selected individuals. That is, part of our intuition for why there is a positive correlation among longitudinal responses is that we expect the repeated responses from the same individual to be more similar than the responses across different individuals.

There can also be heterogeneity among individuals in their response trajectories over time. That is, given a treatment or intervention that should lead to an improvement or increase in the response variable, different individuals will invariably show different gains over time. Changes in the response over time, due to the effects of treatments, interventions or exposures of some kind, are not expected to be completely uniform across all individuals. There will be some individuals whose gains will be above average, while there will be others whose gains are below average. In cases where there is variability in individuals' response trajectories over time, this can account for not only the positive correlation among repeated measures, but also the pattern of decreasing correlations with increasing time separation.

In statistical models for longitudinal data, between-individual variability can be accounted for by the introduction of individual-specific "random effects" (e.g., randomly varying intercepts and slopes). That is, to account for between-individual heterogeneity in propensity to respond, some of the effects or regression coefficients in statistical models are assumed to vary randomly. This topic will be discussed in much greater detail in Chapter 8.

In summary, one important source of variability in longitudinal data that has a direct impact on the correlation among the repeated measures is between-subject variation in the response. Another important source of variability is within-subject variation. The notion here is that even in the absence of any treatment, exposure, or intervention, many health-related outcomes are in a state of so-called *dynamic constancy*. That is, although an individual's underlying propensity to respond may be "high", and this propensity to respond remains relatively fixed over extended periods of time, the observed sequence of repeated measures on this individual will vary in a random manner around this underlying response level. These random fluctuations can be accounted for by at least two main factors: inherent within-individual biological variation in the response over time and measurement error. Next, we examine each of these sources of variability in turn.

Within-Individual Biological Variation

The inherent biological variability of many health outcomes is an important source of variability that has an impact on the correlation among longitudinal responses. Many health-related variables, for example, blood pressure and self-reported pain, fluctuate considerably even over relatively short intervals of time. These fluctuations may be due to circadian rhythms or perhaps influenced by temperature, light, season, diet or infection. Of the many health-related variables that change over time, a small number vary in quite predictable cyclical rhythms that may be daily (e.g, body temperature), monthly (e.g., estrogen levels in pre-menopausal women), or seasonal in nature. However, most health-related variables do not have such predictable cyclical rhythms. Instead, a sequence of repeated measures on any particular individual will vary around some homeostatic set point in a random manner. Many of these variable can be thought of as realizations of some biological process or combination of biological processes operating within the individual that vary over time. This variability is sometimes referred to as the *inherent within-individual biological variability*. Inherent biological variability of this kind is evident in almost all measured biological parameters, for example, serum cholesterol, blood pressure, heart rate, and so on.

The notion here is that there is some underlying biological process (or combination of processes) that changes through time in a relatively smooth and continuous fashion. As a result, random deviations or departures from an individual's underlying response trajectory are likely to be more similar (e.g., both positive or both negative) when measurements are obtained very close together in time. That is, successive random deviations cannot be assumed to be independent. One consequence of this type of variation is that measurements taken very closely together will typically be more highly correlated than measurements that are further separated in time. That is, all others things being equal, measurements on the same individual will be more alike the closer in time they are taken, and will be less similar the further apart in time. For example, when blood pressure is measured repeatedly at 30-minute intervals, adjacent measurements will be more highly correlated than when the repeated measurements are taken weeks or months apart. Thus, inherent within-individual biological variability in the response variable over time introduces serial correlation among repeated measures and results in the correlation matrix having a distinctive structure, with the correlation decreasing as the time separation between repeated measures increases.

Another conceptualization of the within-individual biological variation is in terms of the failure to precisely specify each individual's response trajectory over time. If each individual has a slightly different response trajectory over time, then any misspecification of these response trajectories will induce correlation among the repeated measures. Recall from the definitions of variances and covariances that they are measures of deviations from some model for the mean response. To the extent that the model does not hold for individuals, as, for example, when the true trend is quadratic but linear is fitted, the repeated observations will be correlated due to model misspecification. The interdependence between the models for the mean and covariance is a topic that will be discussed at greater length in Chapter 7.

Measurement Error

A final source of variability in longitudinal data is random measurement error. For some health outcomes, for example, height and weight, variation due to measurement error can be almost negligible (or can be made negligible with the use of more sophisticated measurement instruments). However, for many other outcomes, the variability due to measurement error can be quite substantial. Although this source of variability can account for some of the within-subject variation in many health outcomes, it should not be confused with the inherent (within-individual) biological variability of these outcomes. That is, where it is possible to take two measurements of the response simultaneously on the same individual, thus ruling out the possibility of any inherent biologic variability, the values would not be expected to agree due to the imprecision of the measurement procedure. For example, suppose the variable of interest is nutrient intake, as determined by a particular biomarker in the blood. Furthermore, suppose that a blood sample is drawn on each individual and the vial of blood is divided into two sub-samples that are each subjected to laboratory measurement of the biomarker of interest. In general, these two replicate measures of the biomarker are not expected to agree due to random measurement error.

Measurement error is a ubiquitous component of almost all studies, longitudinal or not. Virtually all measurements are fallible and so it is useful to be able to quantify the relative magnitude of the errors associated with a given measurement procedure. Commonly, the precision of the measurement procedure is expressed in terms of a coefficient of *reliability*. The term reliability has a very precise statistical definition and refers to the extent to which replicate measurements, taken under the same conditions, are similar. In the example of the measurement of nutrient intake via a biomarker in the blood, the extent to which these two replicate measures of the biomarker are similar is taken to be an index of the reliability of the biomarker measurement. Note that, at least hypothetically, we could imagine obtaining many such replicate measurements on an individual under as near uniform conditions as possible. In that case an individual's "true" or error-free score is defined as the average of all the (hypothetical) replicate measurements. The statistical definition of reliability then expresses the relative magnitude of the variability of the true scores to the overall variability of the data. That is, reliability is defined as the proportion of the total or overall variability that is due to individual-to-individual variability in the true scores. The reliability of measurements of height is approximately 0.98, while the reliability of measurements of LDL cholesterol can be as low as 0.85. In certain populations, self-reported measures of well-being and quality-of-life can have reliabilities of less than 0.5. In this definition, reliability implicitly depends upon the heterogeneity of the true scores in the population of scientific interest. Thus reliability is not a fixed characteristic of the measurement. Because of this, it is preferable to express the precision of a measurement directly in terms of the variance of the measurement errors or alternatively its square-root. (The latter is commonly referred to as the *standard error of measurement*.)

Given that the response variable in most longitudinal studies will be measured with error, what is the potential impact of this variability on the correlation among

repeated measures? In general, the effect of unreliability is to "attenuate" or shrink the correlation among the repeated measures closer to zero. For example, if a fallible measure of the response variable has a reliability of 0.8 in the population of interest, the correlation among any pair of repeated measures will be attenuated by a factor of 0.8. In general, the larger the variance of the measurement errors, the greater is the attenuation of the correlation among repeated measures in a longitudinal study. Hence, use of a less reliable measurement procedure or instrument will result in repeated measurements with smaller correlations than if a more reliable measurement procedure or instrument had been used.

Although we have distinguished two conceptually distinct sources of within-subject variation, within-individual biological variation in the response over time and measurement error, many longitudinal studies will not have sufficient data to estimates these separate sources of variability. That is, for many longitudinal designs it may not be possible to estimate both sources of variability from the data at hand. Instead, for purposes of estimation, both sources may need to be combined into a single component of within-subject variance.

Before concluding our discussion of these three sources of variability in longitudinal data, it is worth pausing to consider the distinctions between them. These three distinct sources of variability can be characterized in a graphical display of longitudinal data on two hypothetical individuals (see Figure 2.3). Figure 2.3 displays six repeated measurements of the response (denoted by empty circles) on two individuals, say individual A and B. The three sources of variability in the response can be distinguished by considering the additional variability in the response each source produces. The contributions of these three sources of variability are highlighted in Figure 2.4. In Figure 2.4(a) the between-subject variation is reflected in the degree of separation among the true underlying "response profiles". These two hypothetical response profiles can be thought of as being representative of the response trajectory for "high" and "low" respondents. If between-subject heterogeneity were the only source of variability in longitudinal data, then the six repeated measures on these two hypothetical individuals would fall along the corresponding response profiles in Figure 2.4(a). However, in addition to between-individual variation, there is also within-individual variation. Figure 2.4(b) illustrates how a sequence of repeated measures on any individual might vary in a random manner around their long-run average (or "true" underlying response) due to inherent within-individual biological variation in the response over time. In the absence of any measurement error, repeated measures on these two hypothetical individuals would fall along the corresponding jagged curves; these error-free repeated measures are denoted by the solid circles. However, due to the imprecision of the measurement procedure, the actual repeated measures on these two hypothetical individuals (denoted by empty circles) vary in a random manner around the corresponding jagged lines (see Figure 2.4(c)). The relative magnitude of the between-individual and within-individual sources of variability will be different from one health outcome to another. Their relative magnitude is an important determinant of the degree of correlation among repeated measures.

Finally, we consider the impact of these three sources of variability on the correlation among repeated measures and briefly discuss the potential consequences of

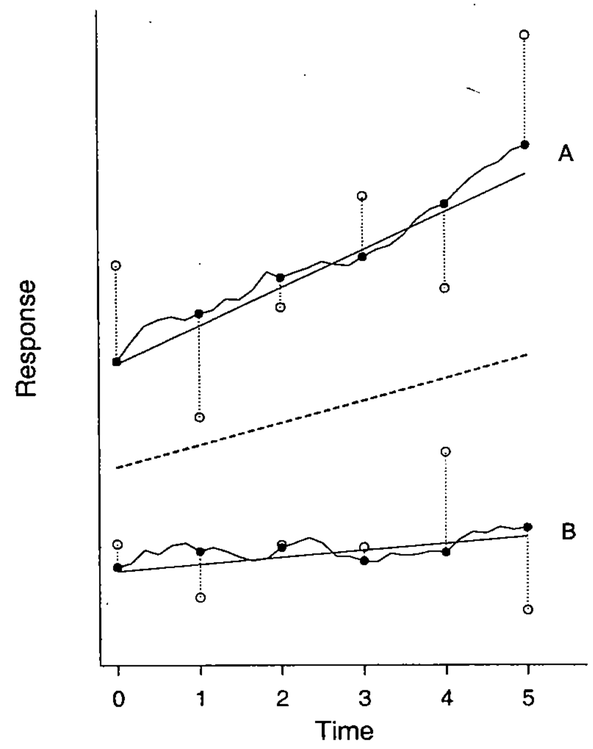

Fig. 2.3 Graphical representation of the three sources of variability in longitudinal data for two hypothetical individuals: • denotes repeated measure free of measurement error, o denotes observed repeated measure with measurement error.

ignoring the correlation. Earlier, we described four empirical observations about the correlation among repeated measures in longitudinal studies. Here we consider how the three sources of variability in longitudinal data can account for these empirical observations. First, we noted that the correlations among repeated measures are positive. The positive correlation among repeated measures is a direct consequence of both between-individual heterogeneity and within-individual biological variation in the response over time. These two sources of variability act in union to induce positive correlation among the repeated measures. Second, we noted that the correlation tends to decrease with increasing time separation. This is a direct consequence of the inherent within-individual biological variation in the response over time and/or between-individual heterogeneity of response trajectories over time. Third, it was noted that the correlations between repeated measures rarely approach zero, even in cases where the repeated measures are taken many years apart. This is a direct consequence of between-subject heterogeneity in the underlying propensity to respond.

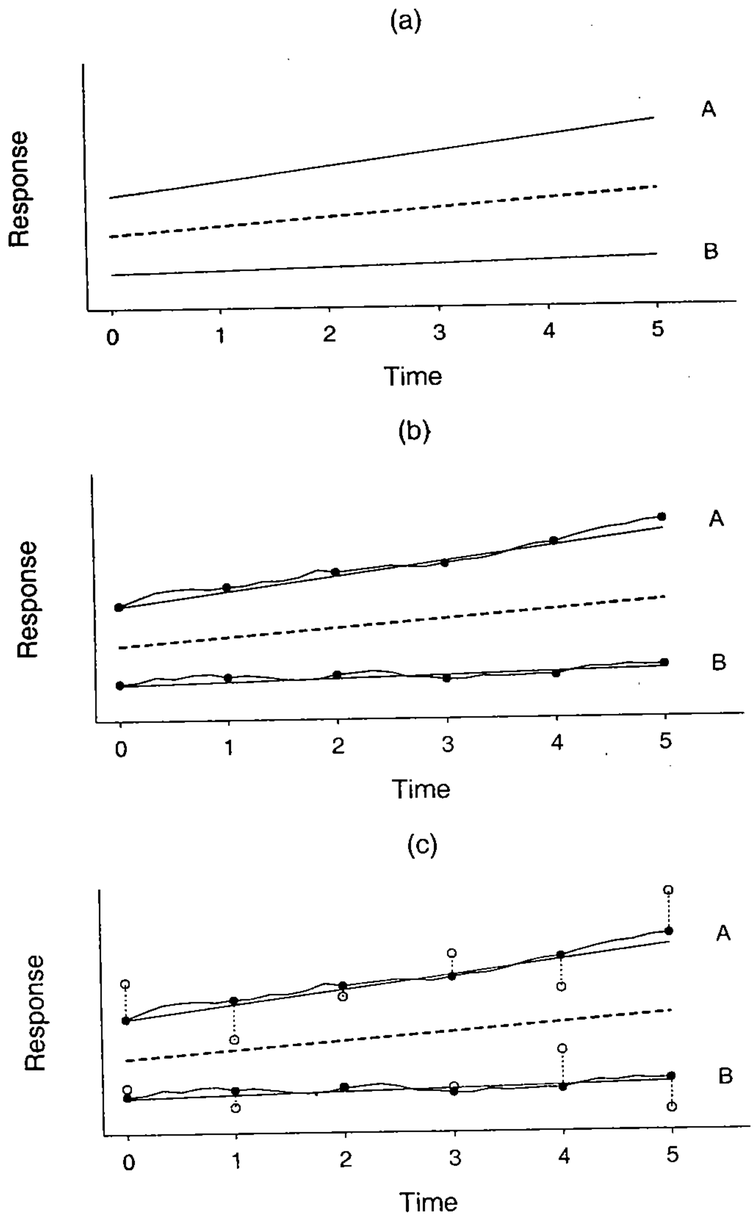

Fig. 2.4 Graphical representation of the cumulative impact of three sources of variability in longitudinal data: (a) between-individual heterogeneity, (b) within-individual biological variation (where • denotes repeated measure free of measurement error), and (c) measurement error (where o denotes observed repeated measure with measurement error).

That is, an individual's propensity to respond persists across all repeated measures on that individual, regardless of how far apart the measurements are in time. Finally, we noted that the correlation between a pair of repeated measures taken very closely together in time rarely approaches one. This final observation is a direct consequence of measurement error. The correlation between any pair of repeated measurements, regardless of how close the measurement occasions, is constrained by the reliability of the measurement procedure.

While it is likely that all three sources of variability contribute to the variability of longitudinal data, one may be more dominant than another. It may not always be necessary, or indeed possible, to separately estimate these three unique sources of variability. This issue will be examined more closely in Chapter 8. Finally, we remind the reader of the definitions of variance and covariance given in Section 2.3. The variance and covariance are measures defined in terms of a particular model for the mean response over time. As a result, there is a subtle interdependence between the model for the mean response and the model for the covariance. To the extent that the model for the mean response does not fit the data well, the observations will be correlated and overdispersed due to misspecification of the model for the mean response. The interdependence between the models for the mean and covariance, and the ramifications of this interdependence for model selection, are discussed in greater depth in Chapter 7.

Consequences of Ignoring Correlation among Longitudinal Data

We have seen that longitudinal data are usually positively correlated, and that the strength of the correlation is often a decreasing function of the time separation. Next we consider the potential implications of ignoring the correlation among the repeated measures. In later chapters of this book we will discuss this topic in greater detail. Here we provide a hint of the potential impact of ignoring the correlation with a simple illustration using data from the *Treatment of Lead-Exposed Children Trial*. Consider only the first two repeated measures from this study, taken at baseline (or week 0) and week 1. Suppose it is of interest to determine whether there has been a change in the mean response over time. A very natural estimate of the change in the mean response over time is

$$\widehat{\delta} = \widehat{\mu}_2 - \widehat{\mu}_1,$$

where

$$\widehat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}.$$

For the data from the TLC trial, the estimate of the change in the mean response over time in the succimer group is -13.0 (or $13.5 - 26.5$). Of course, this estimate is not of much use without some estimate of its sampling variability. To obtain the standard error (SE) we need to estimate the variability of this estimator of change.

An expression for the variance of $\widehat{\delta}$ is given by

$$\text{Var}(\hat{\delta}) = \text{Var}\left\{\frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1})\right\} = \frac{1}{N}\left(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\right).$$

It is the inclusion of the last term, $-2\sigma_{12}$, in the expression above that accounts for the correlation among the first two repeated measures. For the data at hand, we can substitute estimates of the variance and covariances into this expression to obtain the following estimate of the variance of $\widehat{\delta}$

$$\widehat{\text{Var}}(\hat{\delta}) = \frac{1}{50}\{25.2 + 58.9 - 2(15.5)\} = 1.06.$$

If we had simply ignored the fact that the data are correlated and proceeded with an analysis assuming that all observations are independent (and hence uncorrelated, with zero covariance), we would instead have obtained the following (incorrect) estimate of the variance of $\widehat{\delta}$

$$\frac{1}{50}(25.2 + 58.9) = 1.68,$$

which is approximately 1.6 times larger. Thus, in this very simple illustration, ignoring the correlation leads to quite discernible overestimation of the variability of the estimate of change. This in turn would lead to an overly pessimistic estimate of precision, resulting in standard errors that are too large, confidence intervals that are too wide, and p-values for the test of H_0: $\delta = 0$ that are too large. In summary, failure to take account of the correlation among the repeated measures will, in general, result in incorrect estimates of the sampling variability and can lead to quite misleading scientific inferences. This topic will be discussed at much greater length in later chapters of the book.

2.6 FURTHER READING

A compelling illustration of the strengths of a longitudinal study design can be found in Chapter 1, Section 1.1, of Diggle *et al.* (2002).

Problems

2.1 The *Treatment of Lead-Exposed Children* (TLC) trial was a placebo-controlled, randomized study of succimer (a chelating agent) in children with blood lead levels of 20–44 micrograms/dL. Recall that the data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or to

placebo. For this problem set we focus only on the 50 children assigned to chelation treatment with succimer.

The raw data are stored in an external file: `lead.dat`

Each row of the data set contains the following 5 variables:

ID Y_1 Y_2 Y_3 Y_4

2.1.1 Read the data from the external file and calculate the sample means, standard deviations, and variances of the blood lead levels at each occasion.

2.1.2 Construct a time plot of the mean blood lead levels versus time (in weeks). Describe the general characteristics of the time trend.

2.1.3 Calculate the 4×4 covariance and correlation matrices for the four repeated measures of blood lead levels.

2.1.4 Verify that the diagonal elements of the covariance matrix are the variances by comparing to the descriptive statistics obtained in Problem 2.1.1.

2.1.5 Verify that the correlation between blood lead levels at baseline (week 0) and week 1 is equal to the covariance between blood lead levels at baseline and week 1, divided by the product of the standard deviations of the blood lead levels at baseline and week 1.

Part II

Linear Models for Longitudinal Continuous Data

3

Overview of Linear Models for Longitudinal Data

3.1 INTRODUCTION

In Part II the focus is exclusively on linear models for longitudinal data with response variables that are continuous and have distributions that are approximately symmetric, without excessively long tails (or skewness) or outliers. The models for longitudinal data presented in Part II also provide the foundations for more general models for longitudinal data when the response variable is discrete or a count. In this chapter we introduce some vector and matrix notation and present a general linear regression model for longitudinal data. A specific feature of the model is that the mean response is linear in the regression parameters. We present a broad overview of different approaches for modelling the mean response over time and for accounting for the correlation among repeated measures on the same individual. These topics are discussed in much greater depth in subsequent chapters. We also consider some elementary descriptive methods for exploring longitudinal data, especially trends in the mean response over time. We conclude the chapter with an historical survey of some of the earliest developments in methods for analyzing longitudinal and repeated measures data.

We must emphasize at the outset that the statistical methods presented in Part II use the assumption that the longitudinal responses have an approximate multivariate normal distribution to derive estimates and statistical tests, but do not require it. That is, the methods discussed in Part II are based on, but do not require, the assumption that the responses have a multivariate normal distribution. Given this distributional assumption, the method of maximum likelihood, presented in Chapter 4, provides a very general technique for estimation and for inference. In Chapter 4 we briefly

discuss statistical methods for constructing confidence intervals, testing hypotheses, and assessing the adequacy of models. In Chapter 11, where we discuss alternative methods of estimation, it will become more apparent that we do not require the assumption of multivariate normality.

3.2 NOTATION AND DISTRIBUTIONAL ASSUMPTIONS

In this section we introduce some vector and matrix notation that will be used extensively throughout the remainder of the book. Readers without any prior exposure to matrix algebra are encouraged to review the introduction to vectors and matrices presented in Appendix A; we guarantee that the small investment involved in mastering the material in Appendix A will pay handsome dividends later. Throughout this book we do not presume that the reader has a profound understanding of matrix algebra, however, some basic facility with the addition and multiplication of vectors and matrices is required. As will soon become apparent, our primary motivation for the use of vectors and matrices is the compactness with which multivariate statistical techniques can be presented and described when expressed in vector and matrix notation.

Notation

In Chapter 2 we assumed that a sample of N subjects are measured repeatedly over time. We let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. As was mentioned in Chapter 2, either by design or happenstance, subjects may not have the same number of repeated measures and may not be measured at the same set of occasions. To accommodate both of these features, we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. For example, a study may be designed to take repeated measurements on all subjects at the same set of n occasions. However, missing data are a common problem in almost all longitudinal studies and some subjects may not have observations at all n occasions (i.e., n_i denotes the number of *observed* responses on the i^{th} subject, where $n_i \leq n$). Missing data not only produce a varying number of repeated measurements of subjects in a longitudinal study, but also have important consequences for the validity of any method of analysis. In Section 4.3 we outline some of the key issues and assumptions required for valid analyses when there are missing data; this topic is discussed in greater detail in Chapter 14. In addition to missing data, there may be mistimed measurements in the sense that measurements are not obtained at the planned n occasions; instead they are obtained some time before or after the intended measurement occasions. Thus both the number and the timing of the repeated measurements may not be common for all subjects.

It is convenient to group the response variables for the i^{th} subject into an $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, ..., N.$$

Note that the vector Y_i is simply a time-ordered collection of the n_i response variables for the i^{th} subject. The Y_{ij}'s are often called the components, entries, or elements of Y_i. The vector Y_i is said to be of *order* $n_i \times 1$, meaning that it consists of n_i rows and 1 column of elements.

The vectors of responses, Y_i, for the N subjects are assumed to be independent of one another. Note, however, that while the vectors of responses obtained on different subjects can usually be assumed to be independent of one another (e.g., repeated measures of a health outcome for one patient in a clinical trial are not expected to predict or influence the health outcomes for another patient in the same trial), the repeated measures on the same subject are emphatically not assumed to be independent observations.

When the number of repeated measures is the same for all subjects in the study (and there are no missing data) it is not necessary to include the index i in n_i (since $n_i = n$, for $i = 1, ..., N$). Similarly, if the repeated measures are observed at the same set of occasions it is not necessary to include the index i in t_{ij} (since $t_{ij} = t_j$, for $i = 1, ..., N$). For example, in the *Treatment of Lead-Exposed Children Trial* all subjects had the same number of repeated measures, $n = 4$, and were measured at the same set of occasions, $\{t_1 = 0, t_2 = 1, t_3 = 4, t_4 = 6\}$.

Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \;\; j = 1, ..., n_i.$$

Note that X_{ij} is a vector of covariates associated with Y_{ij}, the response variable for the i^{th} subject at the j^{th} occasion. The p rows of X_{ij} correspond to different covariates. There is a corresponding vector of covariates associated with each of the n_i repeated measurements on the i^{th} subject. That is, X_{i1} is a $p \times 1$ vector whose elements are the covariate values associated with the response variable for the i^{th} subject at the 1^{st} measurement occasion, X_{i2} is a $p \times 1$ vector whose elements are the covariate values associated with the response variable for the i^{th} subject at the 2^{nd} measurement occasion, and so on. The vector X_{ij} may include two main types of covariates: covariates whose values do not change throughout the duration of the study and covariates whose values change over time. Examples of the former include gender and fixed experimental treatments. Examples of the latter include time since baseline, current smoking status, and environmental exposures. In the former case,

the same values of the covariates are replicated in the corresponding rows of X_{ij}, for $j = 1, ..., n_i$. In the latter case, the values taken by the covariates can vary over time (for at least some individuals) and the values in the corresponding rows of X_{ij} can be different at each measurement occasion. The inclusion of time-varying covariates whose values at any occasion cannot be predicted (e.g., current smoking status) can raise subtle issues concerning the interpretation and estimation of the resulting models. A discussion of these issues is deferred until Chapter 15.

We can group the vectors of covariates into an $n_i \times p$ matrix of covariates:

$$X_i = \begin{pmatrix} X_{i1}' \\ X_{i2}' \\ \vdots \\ X_{in_i}' \end{pmatrix}, \quad i = 1, ..., N;$$

where X_{ij}' denotes the *transpose* of the vector of covariates, X_{ij}. Recall that the *transpose* is a function that interchanges the rows and columns of a matrix (see Appendix A); thus X_{ij}' denotes a $1 \times p$ row vector of covariates for the i^{th} subject at the j^{th} occasion. The matrix X_i is simply an ordered collection of the values of the p covariates for the i^{th} subject at each of the n_i measurement occasions. That is,

$$X_i = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix},$$

where the rows of X_i correspond to the covariates associated with the responses at the n_i different measurement occasions, and the columns of X_i correspond to the p distinct covariates.

By now it should be apparent that the use of vectors and matrices can greatly facilitate exposition by allowing the repeated measurements on the response variable and the covariates to be expressed in a succinct manner. So far, we have assumed that each subject in the study has a vector of repeated responses, denoted by Y_i, and associated with each repeated measure, a vector of p covariates which can be collectively grouped into a matrix, X_i. Later, we will present a simple numerical example to reinforce the reader's understanding of the vector and matrix notation used so far.

Next, we consider a linear regression model for changes in the mean response over time and for relating the changes to the covariates,

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij}, \quad j = 1, ..., n_i; \tag{3.1}$$

where $\beta_1, ..., \beta_p$ are unknown regression coefficients relating the mean of Y_{ij} to its corresponding covariates. This regression model describes how the responses at every occasion are related to the covariates. That is, there are n_i separate regression

equations for the response variable at each of the n_i occasions

$$\begin{aligned} Y_{i1} &= \beta_1 X_{i11} + \beta_2 X_{i12} + \cdots + \beta_p X_{i1p} + e_{i1} = X_{i1}'\beta + e_{i1} \\ Y_{i2} &= \beta_1 X_{i21} + \beta_2 X_{i22} + \cdots + \beta_p X_{i2p} + e_{i2} = X_{i2}'\beta + e_{i2} \\ &\vdots \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \vdots \quad\quad \vdots \\ Y_{in_i} &= \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + \cdots + \beta_p X_{in_ip} + e_{in_i} = X_{in_i}'\beta + e_{in_i} \end{aligned} \tag{3.2}$$

where the unknown regression parameters are grouped together into a $p \times 1$ vector, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$. Here the e_{ij} are random errors, with mean zero, representing deviations of the responses from their corresponding predicted means

$$E(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}.$$

Typically, although not always, $X_{ij1} = 1$ for all i and j, and then β_1 is the intercept term in the model. Our use of β_1, rather than β_0 or α, to denote the intercept is somewhat arbitrary but does lead to minor simplification of the notation used throughout the book.

Finally, using vector and matrix notation, the regression model given by (3.1) or (3.2) can be expressed in an even more compact form,

$$Y_i = X_i\beta + e_i, \tag{3.3}$$

where $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ is an $n_i \times 1$ vector of random errors associated with the corresponding elements of the vector of responses on the i^{th} subject. The regression model given by (3.3) is simply a shorthand representation for

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}.$$

By comparing (3.2) and (3.3), it should now be apparent to the reader that one of the chief advantages of using vector and matrix notation is that regression models relating longitudinal responses to multiple predictors can be expressed in a very economical fashion.

Thus far we have made no distributional assumptions about Y_i. The only assumption made is that the mean of the longitudinal response vector is related to the covariates via the linear regression model given above. Before discussing assumptions about the distribution of Y_i, let us return to the *Treatment of Lead-Exposed Children Trial* in order to reinforce understanding of the notation introduced so far and to clarify how the regression parameters in (3.3) describe pattern of change in the mean response and their relation to covariates.

Illustration: Treatment of Lead-Exposed Children Trial

Recall that in the *Treatment of Lead-Exposed Children Trial* there are 100 study participants who have blood lead levels measured at the same set of 4 occasions: baseline (or week 0), week 1, week 4 and week 6. Since all subjects have the same number of repeated measures observed at the same set of occasions, the index i can be dropped from both n_i and t_{ij}. That is, $n_1 = n_2 = \cdots = n_N = n$ and similarly $t_{1j} = t_{2j} = \cdots = t_{Nj} = t_j$, for $j = 1, ..., 4$. In the TLC trial the response vector is of length 4 ($n = 4$) and all subjects are measured at the same set of occasions: $t_1 = 0, t_2 = 1, t_3 = 4$, and $t_4 = 6$.

Next, suppose that it is of interest to fit a model to the mean response that assumes that the mean blood lead level changes linearly over time, but at a rate that might be different for the two treatment groups. In particular, we might want to fit a model where the two treatment groups have the same intercept (or mean response at baseline) but different slopes. This can be represented in the following regression model

$$\begin{aligned} Y_{ij} &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + e_{ij} \\ &= X_{ij}'\beta + e_{ij}, \end{aligned}$$

where $X_{ij1} = 1$ for all i and all j. That is, $X_{ij1} = 1$ for all subjects and at all measurement occasions, and thus β_1 is an intercept term. The second covariate, $X_{ij2} = t_j$, represents the week in which the blood lead level was obtained. Finally, $X_{ij3} = t_j$ if the i^{th} subject is assigned to the succimer group and $X_{ij3} = 0$ if the i^{th} subject is assigned to the placebo group. As we will show, this coding of X_{ij2} and X_{ij3} allows the slopes for time to differ for the two treatment groups. The three covariates can be grouped into a 3×1 vector of covariates X_{ij}. Thus, for children in the placebo group

$$E(Y_{ij}|X_{ij}) = \beta_1 + \beta_2 t_j,$$

where β_1 represents the mean blood lead level at baseline (week = 0) and β_2 has interpretation as the change in mean blood level (in $\mu g/dL$) per week. For example, the expected change in mean blood level, from baseline to 6 weeks, is $\beta_1 \times 6$ for children in the placebo group. Similarly, for children in the succimer group

$$E(Y_{ij}|X_{ij}) = \beta_1 + (\beta_2 + \beta_3)t_j,$$

where β_1 represents the mean blood level at baseline (assumed to be the same as in the placebo group since the trial randomized subjects to the two groups) and $\beta_2 + \beta_3$ has interpretation as the change in mean blood level per week. Thus, if the two treatment groups differ in their rates of decline in blood lead levels, then $\beta_3 \neq 0$. The regression parameters have useful interpretations that bear directly on questions of scientific interest. Moreover, hypotheses of interest can be expressed in terms of the absence (or setting to zero) of certain regression parameters. For example, the hypothesis that the two treatments are equally effective in reducing blood lead levels corresponds to a hypothesis that $\beta_3 = 0$.

To reinforce the vector and matrix notation we have introduced in this section, it is instructive to examine the matrix of covariates, X_i, and the realized values of Y_i, for any particular subject in the trial. For example, for the study participant with ID = 79 (see Table 1.1), the realized values of Y_i are

$$\begin{pmatrix} 30.8 \\ 26.9 \\ 25.8 \\ 23.8 \end{pmatrix}.$$

This individual[1] was assigned to treatment with placebo and thus has the following matrix of covariates, X_i,

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix};$$

the latter is often referred to as the *design matrix*. The four rows of X_i correspond to the covariates associated with the blood lead levels at the four measurement occasions (weeks 0, 1, 4 and 6). The elements of the first column are all ones (and multiply the intercept term, β_1). The second column contains values that denote the week in which the blood lead level was obtained. All of the elements of the third column are zero (for subjects assigned to the placebo group). On the other hand, for the study participant with ID = 8, the realized values of Y_i are

$$\begin{pmatrix} 26.5 \\ 14.8 \\ 19.5 \\ 21.0 \end{pmatrix}.$$

This individual was assigned to treatment with succimer and thus has the following design matrix or matrix of covariates, X_i,

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix}.$$

Finally, using vectors and matrices, the model for the mean blood lead levels can be represented as

$$E(Y_i) = X_i\beta,$$

[1] In all data sets used throughout this book, the original subject IDs have been replaced with new subject ID numbers to ensure that the data sets cannot be linked to the original records.

where

$$E(Y_i) = E\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + 4\beta_2 \\ \beta_1 + 6\beta_2 \end{pmatrix}$$

for children in the placebo group, and

$$E(Y_i) = E\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + (\beta_2 + \beta_3) \\ \beta_1 + 4(\beta_2 + \beta_3) \\ \beta_1 + 6(\beta_2 + \beta_3) \end{pmatrix}$$

for children in the succimer group.

Distributional Assumptions

So far, the only assumptions made concern patterns of change in the mean response over time and their relation to covariates. Specifically, given that the vector of random errors, e_i, is assumed to have mean zer(, the regression model given by (3.3) implies that

$$E(Y_i|X_i) = \mu_i = X_i\beta, \tag{3.4}$$

where $\mu_i = (\mu_{i1}, ..., \mu_{in_i})'$ is the $n_i \times 1$ vector of means for the i^{th} individual, with $\mu_{ij} = E(Y_{ij}|X_i) = E(Y_{ij}|X_{ij})$. This model describes how the vector of mean responses is related to the covariates.

Next we consider distributional assumptions concerning the vector of random errors, e_i. The response vector Y_i in (3.3) is assumed to be comprised of two components, a "systematic component", $X_i\beta$, and a random component, e_i. The systematic component implies that the mean response can be expressed as a simple weighted sum of the fixed, but unknown, regression coefficients, β. The random variability of Y_i arises from the addition of e_i, the "random component." This implies that assumptions made about the shape of the distribution of the random errors translate into assumptions about the shape of the distribution of Y_i given X_i. Thus, in a certain sense, we can almost interchangeably refer to the distribution of either the errors, e_i, or the responses, Y_i; their respective distributions differ only in terms of a shift in location. That is, the errors have a distribution with a mean that is centered at zero, while Y_i has the same distributional form except that the mean is centered at $X_i\beta$. As a result, throughout this book we will interchangeably refer to the distributions of Y_i and e_i and, more specifically, the covariance matrix of Y_i and e_i. (Recall that variances and covariances are invariant to changes in location.)

Next, Y_i, the vector of continuous responses, is assumed to arise from a multivariate normal distribution, with mean response vector

$$E(Y_i) = \mu_i = X_i\beta,$$

and covariance matrix,

$$\Sigma_i = \text{Cov}(Y_i).$$

The multivariate normal distribution is completely specified by the vector of means, μ_i, and the covariance matrix, Σ_i. The multivariate normal distribution can be considered to be the multivariate analogue of the univariate normal distribution. Indeed, if Y_i has a multivariate normal distribution, then each of its components, Y_{ij}, has a corresponding univariate normal distribution, with mean μ_{ij} and variance σ_j^2.

Recall that, while observations from different individuals are assumed to be independent of one another, repeated measurements of the same individual are not assumed to be independent. This lack of independence is captured by the off-diagonal elements of the covariance matrix, Σ_i. The covariance matrix has been indexed by i and this allows, in principle, the covariance matrix to depend upon the covariates, X_i (e.g., on the times of the repeated measures). In the case where all individuals have the same number of repeated measures, obtained at a common set of occasions, and where there is no dependence of the covariance matrix on the covariates, we can drop the index i and simply denote the covariance matrix by Σ. This would be analogous to the assumption of homogeneity of variance in linear regression for a univariate response, that is, for the vector of responses, it is assumed that there is homogeneity of covariance. However, when individuals have unequal numbers of repeated measures and/or when the repeated measures are obtained at different occasions, the covariance matrix will typically depend upon the number and timing of the measurements. In principle, the covariance can also depend upon covariates other than time, for example, the covariance could depend upon treatment group. However, in practice, this type of dependency of the covariance on covariates is very rarely ever assumed; this is analogous to the ordinary univariate regression setting, where we usually do not allow the error variance to depend upon covariates.

Multivariate Normal Distribution

So far, we have discussed the multivariate normal distribution in a very general way, noting how it can be seen as a very natural, multivariate extension of the univariate normal distribution. Next, we present a more detailed description of the multivariate normal distribution since it forms the basis for a general method of estimation that will be described in Chapter 4. However, we remind the reader that the statistical methods presented in later chapters use the assumption that the longitudinal responses have an approximate multivariate normal distribution to derive estimates and statistical tests, but do not require it when data are complete (i.e., no missing data) or when there is missingness but the observed data can be regarded as a random sample of the complete data.

The foundation for much of statistics is based on probability theory. Indeed, the formal basis for many statistical methods is an assumed probability distribution for the response variable. Broadly speaking, a probability distribution describes the likelihood or relative frequency of occurrence of particular values of the response variable. In particular, the probability density function for Y, denoted hereafter by $f(y)$, de-

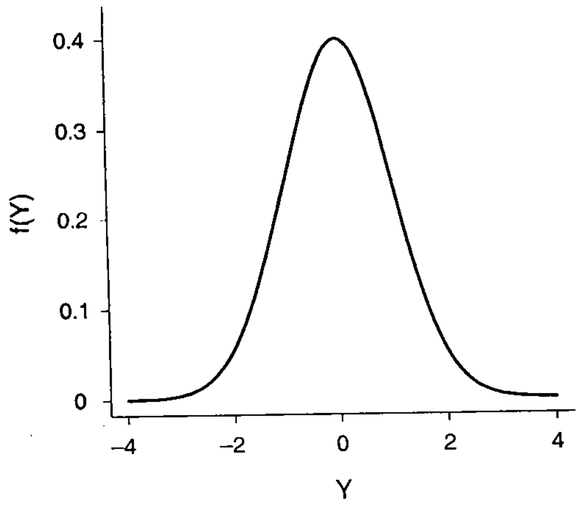

Fig. 3.1 Plot of univariate normal density function with zero mean and unit variance.

scribes the probability or relative frequency of occurrence of particular values of Y. Before we describe some of the properties of the multivariate normal distribution, we first review the univariate normal distribution.

Consider a single univariate response from a longitudinal study at a particular occasion, say Y_{ij}. We assume that the mean of Y_{ij} is related to the covariates by the following linear regression model:

$$Y_{ij} = X_{ij}'\beta + e_{ij},$$

where the errors, e_{ij}, have a *univariate* normal distribution with mean zero and constant variance σ_j^2; we denote this by $e_{ij} \sim N(0, \sigma_j^2)$. Recall that if the e_{ij}'s have a normal distribution with mean zero and constant variance σ_j^2, then Y_{ij} also has a normal distribution, except with mean $\mu_{ij} = X_{ij}'\beta$ and constant variance σ_j^2. Mathematically, the univariate normal (or Gaussian) probability density function for Y_{ij} can be expressed as

$$f(y_{ij}) = \left(2\pi\sigma_j^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(y_{ij} - \mu_{ij}\right)^2 / \sigma_j^2\right\},$$

where $-\infty < y_{ij} < \infty$. Specifically, $f(y_{ij})$ describes the familiar bell-shaped curve illustrated in Figure 3.1. Note that the area under the curve between any two values represents the probability of Y_{ij} taking a value within that range.

The normal distribution has some notable features. First, the distribution is completely determined by two parameters, the mean μ_{ij} and variance σ_j^2 (or standard deviation σ_j). Also, note that the expression for the normal probability density given above depends to a very large extent on

$$\frac{\left(y_{ij} - \mu_{ij}\right)^2}{\sigma_j^2} = (y_{ij} - \mu_{ij})(\sigma_j^2)^{-1}(y_{ij} - \mu_{ij}).$$

The latter is the squared distance of y_{ij} from μ_{ij}, but expressed in standard deviation units. Thus, it can be interpreted as the standardized distance of y_{ij} from its mean, relative to the variability or spread of values around the mean, μ_{ij}.

In the context of a longitudinal study, with n_i repeated measures on the i^{th} individual, we have a vector of responses and need to consider their *joint* probability distribution. While a univariate probability density function describes the probability or relative frequency of occurrence of particular values of a single random variable, a joint probability density function describes the probability or relative frequency with which the vector of responses take on a particular set of values. The multivariate normal distribution is a natural extension of the univariate normal distribution for a single response to a vector of responses. The multivariate normal joint probability density function for Y_i can be expressed as

$$f(y_i) = f(y_{i1}, y_{i2}, \ldots, y_{in_i})$$

$$= (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}\left(y_i - \mu_i\right)' \Sigma_i^{-1} \left(y_i - \mu_i\right)\right\},$$

where $-\infty < y_{ij} < \infty$ for $j = 1, ..., n_i$, $\mu_i = E(Y_i) = (\mu_{i1}, ..., \mu_{in_i})'$, $\Sigma_i = \mathrm{Cov}(Y_i)$, and $|\Sigma_i|$ denotes the *determinant* of Σ_i. The determinant of Σ_i is also known as the *generalized variance*. The determinant of Σ_i summarizes the salient features of the variation expressed by Σ_i in a single number; a more detailed definition of $|\Sigma_i|$ requires a greater understanding of matrix algebra than is assumed in this book.

Note the remarkable similarity between the expressions for the univariate and multivariate normal probability density functions. In some sense, the multivariate normal joint probability density function simply replaces the expression for the standardized distance of y_{ij} from μ_{ij},

$$(y_{ij} - \mu_{ij})(\sigma_j^2)^{-1}(y_{ij} - \mu_{ij}),$$

with a multivariate analogue for the standardized distance of the vector y_i from the vector μ_i,

$$\left(y_i - \mu_i\right)' \Sigma_i^{-1} \left(y_i - \mu_i\right),$$

where Σ_i^{-1} denotes the inverse of the matrix Σ_i (inversion for matrices is the analogue of the reciprocal for numbers in the sense that multiplication by the matrix Σ_i^{-1} can be thought of as division by the matrix Σ_i). Although the latter expression is somewhat more complicated than in the univariate case, it does nonetheless have interpretation in terms of a standardized measure of distance in multivariate space. For example, if Y_i is bivariate, with $Y_i = (Y_{i1}, Y_{i2})'$, then

$$(y_i - \mu_i)' \, \Sigma_i^{-1} \, (y_i - \mu_i)$$

$$= (1 - \rho_{12}^2)^{-1} \left\{ \frac{(y_{i1} - \mu_{i1})^2}{\sigma_1^2} + \frac{(y_{i2} - \mu_{i2})^2}{\sigma_2^2} - 2\rho_{12} \frac{(y_{i1} - \mu_{i1})(y_{i2} - \mu_{i2})}{\sqrt{\sigma_1^2 \sigma_2^2}} \right\},$$

where

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}.$$

Although this is a somewhat more complex expression, since it must account for the correlation between Y_{i1} and Y_{i2}, it does nonetheless provide a single measure of distance that (i) adjusts for differences in the variances of Y_{i1} and Y_{i2}, by effectively down-weighting deviations from the mean when the variance is larger, and (ii) adjusts for the magnitude of the correlation (or overlapping information) between Y_{i1} and Y_{i2}. When there is no correlation between Y_{i1} and Y_{i2} (and $\rho_{12} = 0$), the distance of y_i from μ_i is simply the sum of the component standardized distances. On the other hand, when there is strong positive correlation, the distance of y_i from μ_i also includes a component that factors in whether or not y_{i1} and y_{i2} are *both* larger (or smaller) than μ_{i1} and μ_{i2}, respectively. The latter adjustment is made because part of the standardized distance of y_{i2} from μ_{i2} is predictable from Y_{i2}'s correlation with Y_{i1}, and vice versa.

In addition, many of the properties of the multivariate normal distribution are similar to the univariate normal distribution. First, it is completely determined by the mean response vector, μ_i and by the covariance matrix, Σ_i. Also, as mentioned already, $f(y_i)$ depends to a very large extent on the standardized distance

$$(y_i - \mu_i)' \, \Sigma_i^{-1} \, (y_i - \mu_i).$$

While the multivariate normal distribution shares many of the properties of the univariate normal distribution, the assumption of multivariate normality is much more difficult to verify from the data at hand. Unlike the univariate case, where there are simple graphical tools for assessing the validity of the assumption of normality (e.g., histograms and normal quantile plots), for most practical purposes, it is quite difficult to assess whether a vector of responses has a multivariate normal distribution. Although statistical tests of multivariate normality have been developed, in general, they are not very helpful since they will often detect departures from normality that are of no real substantive importance.

Perhaps the most useful assessment of the validity of the assumption of multivariate normality is through the use of graphical displays. For example, histograms and box and whisker plots of the responses at each occasion can be used to detect gross departures of Y_{ij} from univariate normality. These simple graphical displays can also be used to determine an appropriate transformation of the response variable so that the marginal distributions of the Y_{ij}'s more closely approximate normal distributions. However, a caveat of the use of this technique is that, although a multivariate normal distribution for Y_i implies that each of the separate Y_{ij} has a univariate normal

distribution, univariate normal distributions for the Y_{ij} do not necessarily imply that Y_i has a multivariate normal distribution. Therefore, the assumption of multivariate normality cannot be formally verified by examination of each of the component variables separately. However, gross departures from univariate normality can be taken to indicate that the multivariate normal assumption is not tenable.

Another property of the multivariate normal distribution for Y_i is that the association between any pair of responses is *linear*. Consequently, if Y_i has a multivariate normal distribution then scatter-plots of all possible pairs of the responses should provide no evidence of discernible departures from a linear trend among the pairs of variables. Once again, a caveat of this simple graphical technique is that it cannot be used to establish that Y_i has a multivariate normal distribution; it can only provide evidence of discernible departures from multivariate normality.

At this point the reader may have some concerns about making the assumption of multivariate normality, especially given the inherent difficulties of verifying this assumption from the longitudinal data at hand. Fortunately, as will be discussed in later chapters, the assumption of multivariate normality is not so critical for estimation and valid inferences about β when data are complete (i.e., no missing data). Moreover, this property extends to the setting of incomplete data if the observed data can be regarded as a random sample of the complete data. Some hints for why the normality assumption is not so critical can be found in the literature on linear regression for a single response variable. We remind the reader that there are some well-known results from linear regression models for a univariate response concerning the impact of departures from a (univariate) normal distribution. In that setting, the assumption of univariate normality has been found to be not quite so critical as the assumptions made about the independence of the errors and homogeneity of the variance of the errors. That is, in linear regression for a single response it is departures from the assumption about the independence of the observations and the assumption of constant variance of the errors that have a major impact on the analysis. Departures from normality, unless they are very extreme (e.g., highly skewed response data), are not so critical. In the longitudinal data setting there are very similar results, which suggest that it is the assumptions about the dependence among the errors and assumptions about the variances and covariances that have the greatest impact on statistical inference. Departures from multivariate normality, unless they are very extreme, are not so critical. In later chapters we will discuss this topic at greater length and also describe how the assumption that Y_i has a multivariate normal distribution can be relaxed or avoided altogether.

In summary, in longitudinal studies the repeated measurements on the same individual are inherently dependent or correlated. This lack of independence can be accounted for by considering the multivariate distribution of the entire vector of repeated measurements. Note that while the repeated measurements are correlated, we implicitly assume that the vectors of observations from different individuals are independent of one another. In Part II of this book we are primarily concerned with longitudinal data that are continuous and we make the assumption that their joint distribution is multivariate normal for the purpose of deriving estimates and statistical tests. However, the methods that are discussed do not require the assumption

that the responses have a multivariate normal distribution. In practice, longitudinal data are not anticipated to have a joint distribution that is *exactly* multivariate normal. The multivariate normal distribution is adopted as an approximation, but one that has many convenient statistical properties. In Chapter 4 we present a method for estimating β and Σ_i, and for making inferences about β and Σ_i, which is derived from the multivariate normal assumption for the longitudinal responses. In later chapters we discuss how the assumption that Y_i has a multivariate normal distribution can be avoided altogether.

3.3 SIMPLE DESCRIPTIVE METHODS OF ANALYSIS

Next we consider some simple graphical tools for describing the most salient features of longitudinal data. The formal statistical analysis of longitudinal data should always be preceded by simple graphical displays of the data. A natural way to display longitudinal data is through the use of a *time plot*. A time plot is simply a scatter-plot, with the responses on the vertical axis and the measurement times on the horizonal axis. For a variety of reasons, the time plot of the raw longitudinal data is not always very helpful or readily interpretable. First, in most longitudinal studies the set of measurement occasions is common to many, if not all, of the study participants. As a result, a time plot will result in many overlapping data points at each measurement occasion. The most extreme example of this problem arises in the time plot of binary data; it is impossible to discern any information about time trends from the resulting time plot due to the overlapping data points (e.g., 0's and 1's) at each measurement occasion.

Also, note that the time plot does not indicate which data points represent repeated measurements on the same individual. To circumvent the latter problem, the time plot can be supplemented by joining or connecting successive repeated measures on the same individual with straight lines. However, the resulting line segments do not necessarily enhance the time plot; more often than not, the result is a "spaghetti" plot that is not very informative about trends in the mean response over time. Perhaps the only useful source of information provided by the simple time plot concerns the presence of extreme outliers in the data and whether the variability in the data changes discernibly with time.

Some of the problems with the time plot of longitudinal data can be illustrated using data from the *Treatment of Lead-Exposed Children Trial*. Figure 3.2 displays a time plot of the blood lead level data for the group treated with succimer. Recall that, in this study, repeated measurements were taken at the same set of occasions for all subjects in the trial. Because of the resulting overlap of data points at the four measurement occasions, it is difficult to discern any pattern in the mean response trend over time. As noted earlier, the most extreme case of this problem arises when the response variably is binary; then it is impossible to discern any information about time trends from the resulting time plot due to the completely overlapping data points. Figure 3.3 displays a time plot of the repeated binary response, indicating whether each child has a blood lead level below 10 $\mu g/dL$. (The U.S. Centers for Disease

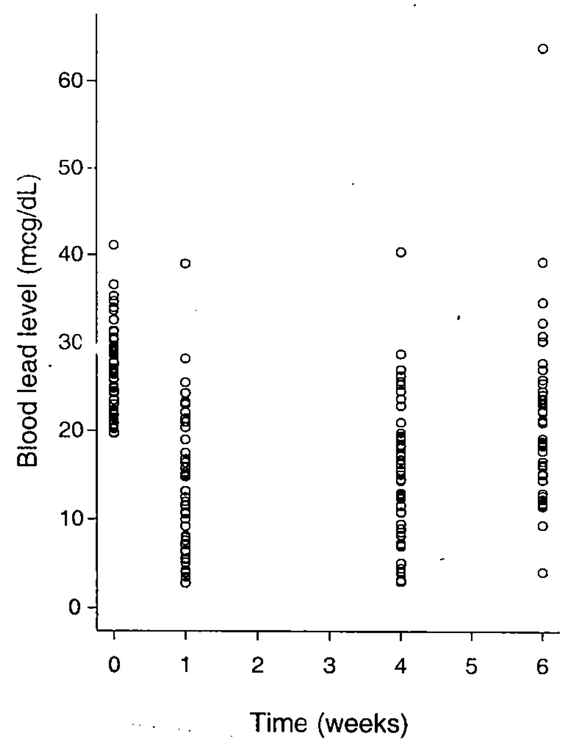

Fig. 3.2 Time plot of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

Control defines 10 $\mu g/dL$ as the threshold for concern about exposure to lead.) Here, the binary response $Y_{ij} = 1$ if the i^{th} child's blood lead level is above 10 $\mu g/dL$ at the j^{th} occasion, and $Y_{ij} = 0$ otherwise. Due to the overlapping 0's and 1's, the time plot provides no information about the trend in the mean response (or probability that a blood lead level is above 10 $\mu g/dL$) over time.

In Figure 3.4 the time plot of blood lead levels is supplemented with line segments joining successive measures on the same individual. However, Figure 3.4 is only a little more informative about trends in the mean response over time than Figure 3.2. Although, in principle, Figure 3.4 distinguishes two sources of variability in the data, between-subject variability and within-subject variability, in practice, it is difficult to assess their relative magnitude from the time plot. However, Figure 3.4 does reveal an observation at week 6 that is a potential outlier, given the previous measurements of blood lead levels for that child. In summary, time plots of the raw data, with or without joined line segments for successive repeated measurements on the same individual, can reveal important features of the data. However, time plots of the raw

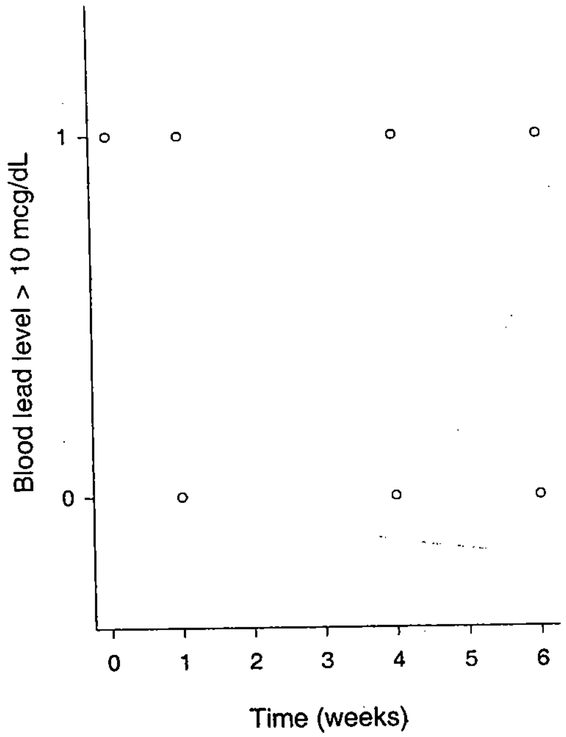

Fig. 3.3 Time plot of blood lead levels > 10 mcg/dL at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

data are not always the most informative displays of longitudinal data, especially when the data are balanced over time. With time plots of balanced longitudinal data it can be difficult to discern the "signal" (i.e., the trend in the mean response over time) from the "noise" in the data and the between-subject and within-subject sources of variability are often almost completely obscured. With highly unbalanced data, time plots of the raw data, with joined line segments, are easier to interpret.

In general, it is usually more informative to display a time plot of the average or mean response, with successive points on the graph joined by straight lines. In addition, time plots of the mean response for different levels of discrete covariates (e.g., different treatment or exposure groups) can be overlayed on the same graph. The construction of these plots is relatively straightforward when the timing of the repeated measures is the same for all individuals. The time plots can also be enhanced by including standard error bars for the mean response at each occasion. For example, Figure 3.5 displays the mean blood lead levels in the succimer and placebo groups at

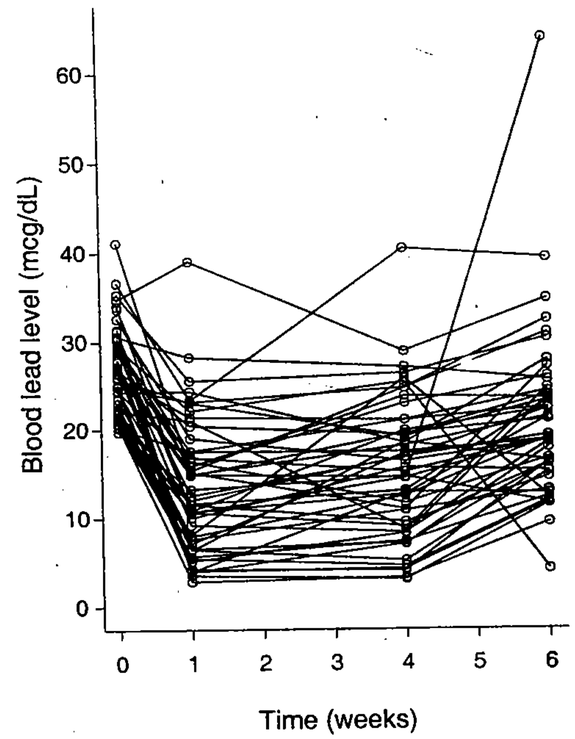

Fig. 3.4 Time plot, with joined line segments, of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

weeks 0, 1, 4, and 6. From this simple display it is readily apparent that the effect of succimer is greater after one week of treatment and that there appears to be a rebound effect thereafter.

Overall, a graphical display of the mean response can be quite enlightening and can provide the basis for choosing an appropriate model for the analysis of change over time. For example, the time plot of the mean response in Figure 3.5 suggests that the analysis of the blood lead levels at all four occasions may require non-linear (e.g., quadratic) or perhaps piecewise linear trends over time.

Simple time plots of the mean response are less straightforward when a covariate of interest is quantitative (e.g., dose of drug). For the purposes of producing a graphical display of the mean response trend, one simple, but often quite effective, approach is to construct a small number of groupings or "reference categories" for the quantitative covariate in question. For ease of exposition, we consider three groupings that can be generically denoted as "low", "medium", and "high". Then, given this set of

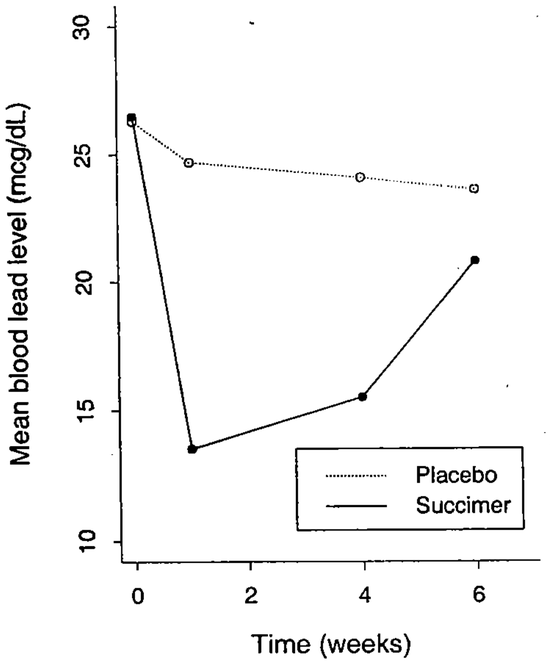

Fig. 3.5 Time plot of the mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups.

reference categories, the construction of the time plot of the mean response trend can proceed along exactly the same lines as for the case of a truly discrete covariate having only three levels. That is, we can simply plot the mean response trends overlayed for the different values of the reference categories. Thus, for all practical purposes, the graphical display of the mean response trends is no more difficult when the covariate of interest is quantitative. The only question that remains is how best to choose appropriate reference categories for a quantitative covariate.

Ideally, at least two or three reference categories for a quantitative covariate should be chosen and in such a way that investigators in the field can readily appreciate the substantive importance of going from one level to the other. For example, the change in going from "low" to "medium" to "high", or vice versa, should have some subject-matter meaning. For some quantitative covariates there may be natural choices for the reference levels (e.g., corresponding to intervals that represent "normal" and "abnormal" ranges). For other quantitative covariates, especially those that are less well established or unfamiliar to investigators in the field, there may not be an obvious choice for the reference categories. In the latter case, the choice can be made on the basis of the data at hand. For example, one possible choice is to group the covariate at

the 25th and 75th percentiles. This will produce "low" (or lowest quartile), "medium" (2nd or 3rd quartiles) and "high" (or highest quartile) reference categories. It must be acknowledged, though, that the number and choices of reference groups is, to some extent, arbitrary; reference groups that are more or less extreme than those suggested here could equally be chosen (e.g., tertiles or quartiles).

So far, our discussion has assumed that many, if not all, individuals are measured at the same set of occasions. When the times of measurement are not the same for all individuals, construction of time plots of the mean response can pose difficulties due to sparseness of data at any particular occasion. For example, Figure 3.6 displays a time plot of longitudinal data on lung function growth in children and adolescents from the Six Cities Study of Air Pollution and Health. The data are from a cohort of 300 school-age girls living in Topeka, Kansas, who, in most cases, were enrolled in the first or second grade (between the ages of six and seven). The girls were measured annually until high school graduation (approximately at age eighteen) or loss to follow-up, and each girl provided a minimum of one and a maximum of twelve observations. At each examination, pulmonary function measurements were obtained from simple spirometry. The basic maneuver in simple spirometry is maximal inspiration followed by forced exhalation as rapidly as possible into a closed chamber. A widely used measure computed from simple spirometry is the volume of air exhaled in the first second of the maneuver, FEV_1. Figure 3.6 displays a time plot of log(FEV_1/height) versus age for the 300 girls. Although children were measured approximately annually, the data are highly unbalanced when age, rather than chronological time, is used as the metameter for lung function growth. Figure 3.7 displays a time plot, with joined line segments, of log(FEV_1/height) versus age for 50 randomly selected girls. Because each girl is not measured at the same age, construction of plots of the mean response versus age can pose difficulties due to sparseness of data at any particular age.

In cases where the occasions of measurement are different, it is helpful to produce a "smoothed" plot of the mean response trend over time. A smooth plot of the trend can be obtained using a variety of different approaches that can be generically referred to as "smoothing techniques". Many of these smoothing techniques approach the estimation of the mean response at any time by considering not only the observations at that occasion but also the "neighboring" observations. That is, the estimated mean is based on observations taken before, at, and after the time of interest. The mean response at any time, say t, is taken to be a weighted average of the observations in some close proximity or neighborhood of time t.

One well-known special case of this approach is the so-called "running average" or "moving average". For longitudinal data that are balanced and complete (no missing data), the moving average at time t, denoted S_t, is given by

$$S_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=-k}^{k} w_j y_{i,t+j}, \qquad t = k+1, ..., n-k;$$

where k is some positive integer (e.g., $k = 1$ or $k = 2$) and we refer to $2k+1$ as being the *order* of the moving average. This expression for the moving average assumes that

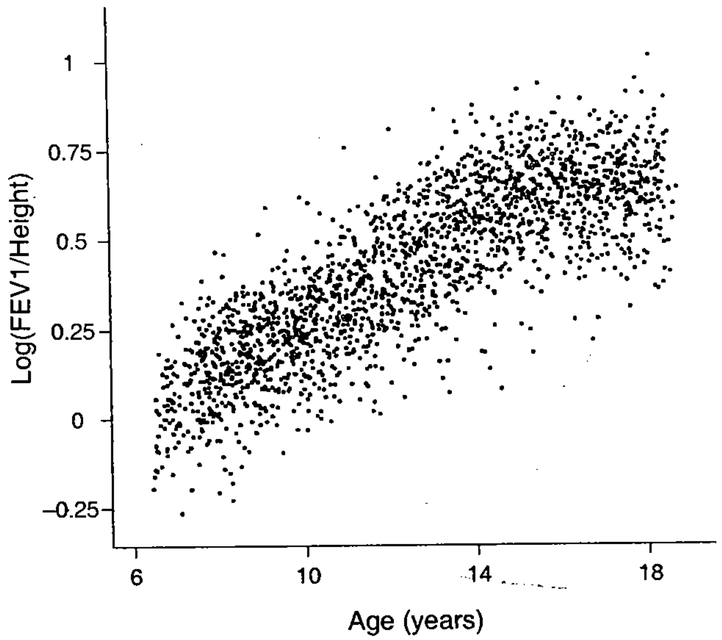

Fig. 3.6 Time plot of log(FEV_1/height) versus age in years for girls from Topeka.

all N individuals are measured at the same set of occasions. With highly unbalanced and/or incomplete longitudinal data, a similar expression for the moving average can be derived. The order of the moving average determines a symmetric neighborhood of values used to estimate the mean response at time t. The higher the order of the moving average the greater the smoothness of the resulting estimate of the mean time trend. Correspondingly, the lower the order of the moving average the greater the roughness of the resulting estimate of the time trend, often producing a curve that has many "wiggles" (for lack of a better term) and/or a somewhat jagged appearance. The w_j are a set of weights whose only restriction is that they must sum to one (i.e., $\sum_{j=-k}^{k} w_j = 1$). Ordinarily the w_j are positive and in cases where they are unequal they are chosen so that they decrease symmetrically about some maximum value, that is, $w_j = w_{-j}$, and $w_0 > w_1 > \cdots > w_k$. As a result, observations obtained in close proximity (in a temporal sense) to time t have the greatest impact or "weight" in the calculation of the mean or average response at time t. This definition of the moving average will be somewhat problematic at the beginning and end of the time plot since the "neighborhood" of values near the end points is necessarily smaller. This problem can be rectified by altering the summation to range from $j = \max(-k, 1-t)$ to $j = \min(k, n-t)$ and dividing by the corresponding sum of the included weights.

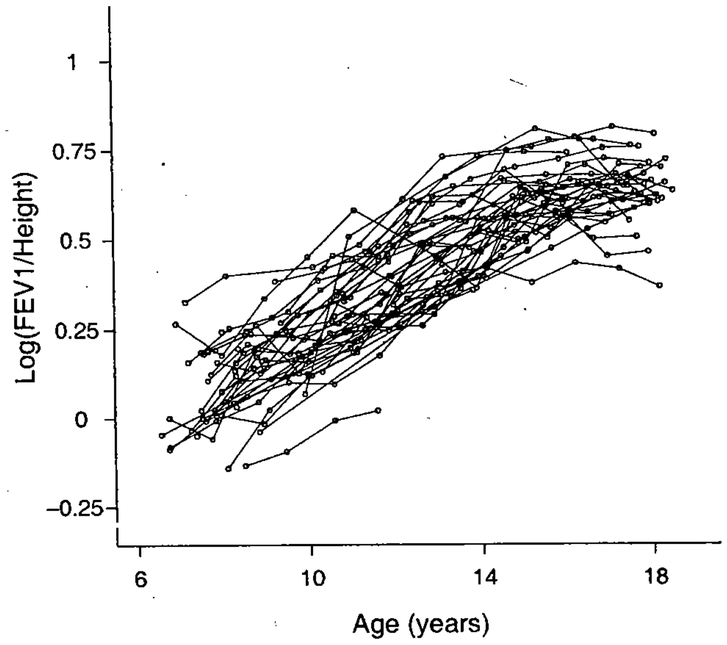

Fig. 3.7 Time plot, with joined line segments, of log(FEV_1/height) versus age in years for 50 randomly selected girls from Topeka.

A simple example of a "moving average" is

$$S_t = \frac{1}{N}\sum_{i=1}^{N} \frac{y_{i,t-1} + y_{it} + y_{i,t+1}}{3}.$$

In this example the weights are all equal (i.e., $w_{-1} = w_0 = w_1 = 1/3$).

Moving averages are best suited to smoothing observations that are approximately equally separated in time. They are not ideal for handling completely irregularly spaced observations. When longitudinal data are irregularly spaced and unbalanced over time, other nonparametric regression methods can be used to estimate the mean response trend over time. One popular method available in most standard statistical software packages is locally weighted regression or *lowess*. The basic idea behind most of the nonparametric regression methods is very similar. They attempt to trace the salient features of the mean response as a function of time while making only minimal assumptions about the form of the relationship. For example, the lowess estimate at time t is best understood by imagining that there is a "window" centered at time t. One estimate of the mean at time t is obtained by taking some weighted average of all the observations that fall within the window. The lowess estimate,

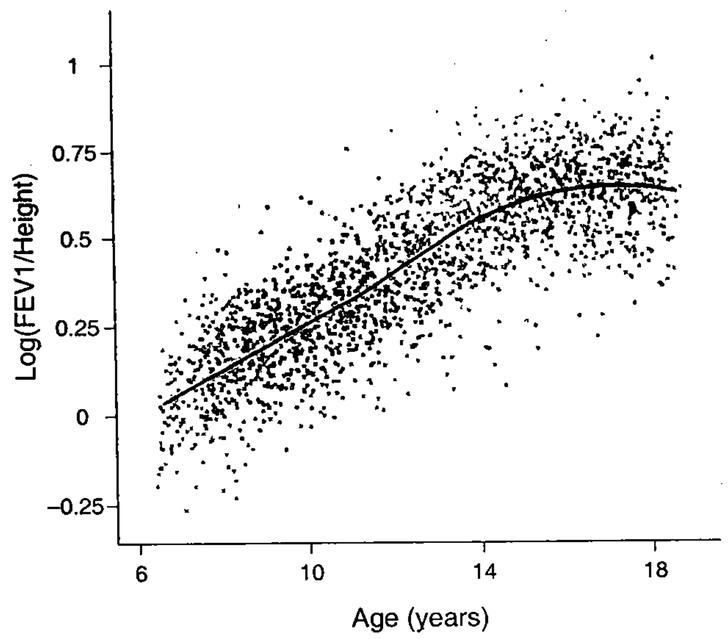

Fig. 3.8 Time plot of log(FEV_1/height) versus age in years, with *lowess* smoothed curve superimposed, for girls from Topeka.

however, is not based on a simple weighted average of the observations within the window. Instead, it is determined by fitting a straight line to the data within the window using a robust regression technique that gives more weight to observations close to the center of the window and that also down-weights potential outliers. The lowess estimate of the mean at time `t` is simply the predicted value at time `t` from the fitted regression line. The entire lowess curve is obtained by moving a window of fixed width from the first measurement occasion to the last, and repeating the process at every time. Figure 3.8 displays a lowess curve for the lung function data described earlier. Unlike the time plot of the raw data in Figure 3.6, the lowess curve is informative about changes in lung function as the children grow older. The smooth curve produced by the lowess procedure indicates ages where lung function appears to develop more rapidly.

All smoothing techniques require that a smoothing parameter, often referred to as the *bandwidth* parameter, be specified. This parameter controls the amount of smoothing. For example, the width of the window in the lowess procedure determines how jagged or smooth the resulting plot appears; the wider the window, the smoother the resulting curve will be. The choice of smoothing or bandwidth parameter involves

the classical tradeoff between bias and precision. Excessive smoothing decreases the variance of the estimate of the mean trend but at the risk of introducing bias. Insufficient smoothing is unlikely to introduce bias but will result is a quite variable estimate of the mean response trend. All smoothing techniques must compromise in some way, and the goal is to find an appropriate tradeoff between these two competing forces: increased bias versus decreased variance of the estimated mean response trend over time.

Finally, we note that standard applications of nonparametric smoothing techniques to longitudinal data ignore the correlation among repeated measures on the same individual. On the whole, the correlation among repeated measures is not likely to grossly distort the estimated mean response trend when standard smoothing techniques are applied to longitudinal data. Correlation is likely to have a much greater impact on the construction of confidence bands for the smoothed curve. As a result, we caution the reader that the confidence bands produced by standard statistical software for lowess, and other nonparametric smoothing techniques, are likely to be optimistically biased. That is, confidence bands constructed under the assumption that the correlation among repeated measures is zero will be too narrow and could potentially lead to misleading inferences. In summary, the routine application of nonparametric smoothing methods to longitudinal data can be useful for exposing trends in the mean response over time, with the caveat that confidence bands produced by standard statistical software packages should be ignored. A final note concerns attrition over time. If attrition or dropout occurs in a substantial number ($> 5\%$) of subjects, then the end of the estimated mean curve can be distorted if subjects who leave the study differ from those who remain; the impact of dropout, and of missingness more generally, is discussed in greater detail in Section 4.3 and Chapter 14.

3.4 MODELLING THE MEAN

In this section we introduce several approaches for modelling the mean of a vector of longitudinal responses. Two main approaches are distinguished: the analysis of response profiles and parametric or semi-parametric curves. Both of these approaches are discussed in greater detail in Chapters 5 and 6.

As mentioned earlier, the analysis of longitudinal data focuses on changes in the mean response over time, and on the relation of these changes to covariates. For example, in the *Treatment of Lead-Exposed Children Trial*, the investigators were primarily interested in how blood lead levels changed over time and whether these changes were related to the treatment assigned. The fact that measurements obtained on the same individual are not independent, but are positively correlated, is an important consideration in their analysis, but for most longitudinal studies the correlation is not usually of scientific interest *per se*.

In Section 2.4 we mentioned that regression models for longitudinal data can usually be formulated to encapsulate the main research questions of interest in terms of a set of regression parameters. That is, certain regression parameters will have interpretations which bear directly on the scientific question or questions of interest.

For example, in a regression model for the longitudinal blood lead level data from the TLC trial, treatment group interaction effects have direct interpretation in terms of how the underlying rate of change in mean blood lead levels differs between the two treatment groups. Before discussing approaches for modelling the mean response over time, it is important to clarify the distinction between so-called *substantive* and *nuisance* parameters in the context of a longitudinal study.

Substantive and Nuisance Parameters for Longitudinal Data

In regression models for longitudinal data, the regression parameters, β, relate changes in the mean response over time to covariates and are usually considered to be of primary or intrinsic interest. The regression parameters β can be defined so as to summarize important aspects of the research questions. As a result we often refer to these parameters as the *substantive* parameters. On the other hand, in many applications parameters that summarize aspects of the covariance or correlation among the repeated measures are considered to be of secondary interest. In statistics, parameters that are associated with these secondary aspects of the data are often referred to as *nuisance* parameters. Thus for the analysis of longitudinal data the correlation or covariance parameters are often thought of as nuisance parameters since there is no intrinsic interest in them.

By making this distinction between substantive and nuisance parameters, the covariance among longitudinal responses is, in a certain sense, regarded as a secondary aspect of the data (relative to the mean response over time). However, we must emphasize that this distinction does not imply that the covariance can be disregarded or simply ignored. Indeed, the covariance among repeated measures must be properly acknowledged to assure an appropriate method of analysis. The distinction between substantive and nuisance parameters has some important ramifications for the types of statistical methods that are adopted. For example, there will be a high premium attached to methods that yield valid estimates of the substantive parameters across a broad range of different assumptions about the nuisance parameters. In the context of longitudinal data, this implies that there will be a premium attached to methods that yield unbiased estimates of change in the mean response over time under a broad range of assumptions about the structure of the covariance among longitudinal responses.

Finally, we must also emphasize that the distinction between substantive and nuisance parameters should be determined only on subject-matter grounds. In the context of analyzing longitudinal data, the regression parameters, β, are typically the substantive parameters since the primary focus is on characterizing changes in the mean response over time. The elements of β have this interpretation. The covariances among the repeated measures are nuisance parameters. However, in some settings where correlated data arise there can be a complete reversal of roles. For example, with clustered data arising from a study of the familial aggregation of a disease-related outcome, parameters that summarize the dependence of the mean on certain risk factors, say β, are usually considered to be nuisance parameters, while the correlations among the responses for different family members are the substantive parameters of direct scientific interest. In family studies the goal is to determine if the presence of disease in a family member increases the risk of disease to relatives. The correlations among siblings and between parent-child pairs are of main interest because their relative magnitudes can be used to provide indirect evidence of genetic risk for the disease due to the sharing of the same gene pool.

Modelling the Mean Response over Time

Much of the focus in the analysis of longitudinal data is on the mean response. There are two broad approaches for modelling the mean response over time: the analysis of response profiles and parametric or semi-parametric curves. The first approach allows arbitrary patterns in the mean response over time; it is related to a more traditional approach known in the statistical literature as "profile analysis". In the analysis of response profiles no specific time trend is assumed. Instead, the times of measurement are regarded as levels of a discrete factor. This approach to the analysis of longitudinal data is only applicable when all individuals are measured at the same set of occasions and the number of occasions is usually small. We describe the main features of the analysis of response profiles in Chapter 5.

A second approach is to assume a parametric curve (e.g., linear or quadratic trend) for the mean response over time. This approach can dramatically reduce the number of model parameters. By their very nature, parametric curves provide a very parsimonious description of trends in the mean response over time, and of covariate effects on the mean response over time. For example, a linear trend in the mean response can be characterized by a single regression parameter that has interpretation in terms of the constant rate of change in the mean response over time. In addition, parametric curves describe the mean response as an explicit function of time. As a result, and in contrast to profile analysis, there is no necessity to require that all individuals in the study have the same set of measurement times, nor even the same number of repeated measurements.

Note that, while the analysis of response profiles allows for an arbitrary pattern of mean responses over time, parametric curves impose an explicit structure on the mean responses. Although it will not always be possible to fit longitudinal data adequately with parametric curves, our experience with data from longitudinal studies suggests that in many cases the trends over time for the duration of the study are relatively simple (e.g., linear or quadratic trends in time). Alternatively, semi-parametric curves (e.g., piecewise linear) can be adopted. A more detailed discussion of modelling the mean using parametric and semi-parametric curves is presented in Chapter 6.

3.5 MODELLING THE COVARIANCE

The defining feature of longitudinal data is that repeated responses are obtained on the same individuals over time and the resulting responses on the same individual are correlated. Although the correlation, or more generally, the covariance among the

repeated responses, is not usually of intrinsic interest, it cannot simply be ignored. Instead, the covariance among repeated measures is an important aspect of the data that must be properly accounted for to yield valid inferences about the regression parameters of primary interest. Accounting for the correlation among repeated measures completes the specification of any regression model for longitudinal data and usually increases efficiency or the precision with which the regression parameters can be estimated. When an appropriate model for the covariance has been adopted, correct standard errors are obtained and valid inferences about the regression parameters can be made. In addition, when there are missing data correct modelling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In Chapter 11 we also consider a method for analyzing longitudinal data that ignores the correlation among the repeated measures for the purposes of estimation of the regression parameters, but makes an appropriate adjustment to the standard errors for the purposes of inference.

Three broad approaches to modelling the covariance among repeated measures can be distinguished: (1) unstructured covariance, (2) covariance pattern models, and (3) random effects covariance structures. The first is to allow any arbitrary pattern of covariance among the repeated measures. This results in what is ordinarily referred to as an "unstructured" covariance. That is, no explicit structure is assumed for the covariance among the repeated measures (other than the homogeneity of covariance across different individuals). Thus, when there are n repeated measures, the n variances at each occasion and the $n \times (n-1)/2$ pairwise covariances (or correlations) are estimated. Historically, the unstructured covariance matrix has been the model of choice for the covariance in the analysis of response profiles. That is, the analysis of response profiles assumes arbitrary patterns for both the mean response over time (and their relation to covariates) and for the variances and covariances. This approach to modelling the covariance, however, is not limited to the analysis of response profiles and could equally be adopted when the mean response is modelled with parametric or semi-parametric curves. There are two potential drawbacks with this approach. The first is that the number of covariance parameters can be quite large. If there are n measurement occasions, the $n \times n$ covariance matrix has $n \times (n+1)/2$ unique parameters. Thus, in a longitudinal study with 10 measurement occasions, an unstructured covariance has 55 parameters (10 variances and 45 covariances). When the number of covariance parameters to be estimated is large relative to the sample size, then estimates are likely to be unstable. The second drawback of this approach is that it is only applicable when all individuals are measured at the same set of occasions. That is, it cannot accommodate mistimed measurements or, more generally, irregularly timed measurements.

Alternative approaches to modelling the covariance place structure on the covariance matrix. There are two main strategies. The first approach borrows ideas from the statistical literature on time series analysis. Time series data, in contrast to longitudinal data, arise from studies with a small number of replications or individuals (often only a single replication) and a large number of repeated measures. That is, in times series data N, the number of replications is small (often $N = 1$) relative to the number of repeated measures, n. With longitudinal data, it is the reverse situation,

with N being large relative to the number of repeated measures, n. Thus time series data consist of a small number of long sequences of repeated measurements, whereas longitudinal data consist of a large number of relatively short sequences of repeated measurements. Although times series data and longitudinal data are dissimilar in structure, and the data analytic goals are usually quite different, they do share one common feature: the repeated measures are correlated. Most of the models for the covariance in the time series literature incorporate at least one important aspect of longitudinal data: repeated measures taken closer together in time are expected to be more highly correlated than repeated measures further apart in time. This implies that the correlations decay as the time separation increases. Quite often, the correlation among repeated measures is expressed as an explicit function of the time separation. In the latter case, these models can be used with unequally spaced observations. In addition, many of the models for the variance assume *stationarity*, namely, that the variance does not change as a function of time. Much of the statistical literature on the analysis of time series data has focused on parametric models that can adequately describe the covariance structure among the repeated measures with only a few parameters. These parsimonious models for the covariance can also be adopted for longitudinal data and are discussed in Chapter 7.

An alternative, and somewhat indirect, strategy for imposing structure on the covariance is through the introduction of *random effects*. Historically, simple random effects models were one of the earliest approaches for analyzing repeated measures data. In the so-called univariate repeated measures ANOVA model, the correlation among repeated measurements is accounted for by the inclusion of a single individual-specific random effect. This effect can be thought of as a randomly varying intercept, representing an aggregation of all the unobserved or unmeasured factors that make some individuals "high responders" and some individuals "low responders". The consequence of adding a single individual-specific random effect to every measurement on any given individual is that the resulting repeated measurements will be positively correlated. Thus the inclusion of random effects imposes structure on the covariance.

The univariate repeated measures ANOVA model has a very long history and has enjoyed widespread use in many fields of application; this model is discussed in greater detail in Section 3.6. Although the introduction of a single individual-specific random effect induces correlation among repeated measures, a feature of the model is that the resulting positive correlation is constant, and does not vary as a function of the time between any pair of repeated measurements. In addition, the variance is constant over time. These constraints on the covariance structure are somewhat unappealing for longitudinal data. However, this problem can be easily remedied by the inclusion of more than one random effect. That is, the constraints on the covariance induced by the repeated measures ANOVA model can be relaxed by assuming that a subset of the regression parameters (e.g., intercepts and slopes) vary randomly across individuals. If the inclusion of a single individual-specific random effect induces positive correlation among repeated measures, albeit with a somewhat unappealing structure on the correlations, it should not come as a surprise that the inclusion of additional randomly varying coefficients induces patterns of correlation

among the repeated measures that are somewhat less restrictive. In addition, these models permit the variance to change over time in a smooth fashion. Indeed, random effects models provide both very flexible and parsimonious models for the covariance and are particularly well suited to handling longitudinal data that are irregularly timed. These models are discussed at length in Chapter 8.

3.6 HISTORICAL APPROACHES

We conclude this chapter with a brief survey of some of the earliest developments in methods for analyzing longitudinal and clustered data. Historically, a variety of relatively simple methods have been developed for the analysis of repeated measures data. Some, but not all, of these happen to be special cases of the regression models for longitudinal data that are the focus of later chapters of this book. As a result, in this section we provide only a brief historical survey of some of these approaches, highlighting their relation to more general models, and noting some of their potential limitations. Many of the shortcomings of these methods alluded to here will be more readily apparent when the methods are viewed as special cases of the regression models considered in later chapters.

From a historical perspective, three methods for the analysis of repeated measures data can be distinguished: (1) univariate repeated measures analysis of variance (ANOVA), (2) multivariate repeated measures analysis of variance (MANOVA), and (3) methods based on summary measures. All three of these approaches have enjoyed varying degrees of popularity, and some are still in widespread use, in different areas of application. Many of these approaches are unnecessarily restrictive in their assumptions and their analytic goals. For example, ANOVA and MANOVA focus on comparing groups in terms of their mean response trend over time but provide little information about how individuals change over time. Also, as we will see later, ANOVA and MANOVA have numerous features that limit their usefulness for the analysis of longitudinal data. In contrast, the regression models that are discussed throughout the remainder of this book make more realistic assumptions and can address the major scientific questions of interest in a longitudinal study. For all of the reasons that were outlined in Section 1.4, we view the regression paradigm as being the most useful, general, and versatile approach for analyzing longitudinal data arising from the health sciences.

Repeated Measures Analysis by ANOVA

One of the earliest proposals for analyzing correlated responses was the repeated measures analysis of variance (ANOVA), sometimes referred to as the "univariate" or "mixed-model" analysis of variance. The analysis of variance paradigm was developed in the early part of the twentieth century by R. A. Fisher. Although many of the early applications of ANOVA were to designed experiments in agriculture, since then it has found widespread application in many other disciplines. In the repeated

measures ANOVA model, the correlation among repeated measurements is assumed to arise from the additive contribution of an individual-specific random effect to each measurement on any given individual. Thus the model assumes the correlation between repeated measurements arises because each subject has an underlying (or latent) level of response which persists over time and influences all repeated measurements on that subject. This individual-specific effect is regarded as a random variable.

A notable feature of ANOVA models is that the response is related to a set of discrete covariates or factors. In the ANOVA paradigm the occasions of measurement are treated as an additional, within-subject, factor. Thus, if we let X_{ij} denote the vector of indicator variables for the study factors (e.g., treatment group, time, and their interaction), the repeated measures ANOVA model can be expressed as

$$Y_{ij} = X_{ij}'\beta + b_i + e_{ij},$$

where b_i is a random individual-specific effect and e_{ij} is a within-individual measurement error (it is implicitly assumed that $X_{ij1} = 1$ for all i and all j). Although both the b_i and e_{ij} are random, they are assumed to be independent of each other. Specifically, the b_i are assumed to have a normal distribution, with mean zero and variance, $\text{Var}(b_i) = \sigma_b^2$. The errors, e_{ij}, are assumed to also have a normal distribution with mean zero, but with variance, $\text{Var}(e_{ij}) = \sigma_e^2$.

Since both b_i and e_{ij} have mean zero, the model for the mean response, averaged over both sources of variability, is given by

$$E(Y_{ij}) = \mu_{ij} = X_{ij}'\beta.$$

Thus, in the repeated measures ANOVA model, the response for the i^{th} individual is assumed to differ from the population mean, μ_{ij}, by an individual-specific random effect, b_i, that persists throughout all measurement occasions, and a within-subject measurement error, e_{ij}. That is, the repeated measures ANOVA model distinguishes two main sources of variation in the data: between-subject variation, σ_b^2, and within-subject variation, σ_e^2. The between-subject variation acknowledges the simple fact that subjects respond differently; some are "high" responders, some are "low" responders, and some are "medium" responders. The within-subject variation acknowledges that there are random fluctuations that arise from the process of measurement, for example, due to measurement error and/or sampling variability.

Given these assumptions about the two main sources of variation, the covariance matrix of the repeated measurements has the following structure:

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_e^2 \end{pmatrix}.$$

The derivation of the variances and covariances is not important; a more detailed account will be given in Chapter 8 (see Section 8.1). What is important to note is

that the variances at every occasion are equal, $(\sigma_b^2 + \sigma_e^2)$, as are the covariances, σ_b^2. Consequently, the correlation among any pair of repeated measures,

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2},$$

is positive (by virtue of the fact that the variances, σ_b^2 and σ_e^2, must be positive) and constant, regardless of the time that has elapsed between the measurement occasions.

This particular covariance structure is also known as *compound symmetry* and has a randomization justification in certain repeated measures designs (e.g., split-plot designs). In an experiment where the within-subject factor is randomly allocated to subjects, randomization arguments can be made to show that the constant variance and constant correlation conditions hold (see Chapter 16 for a more detailed discussion of the randomization argument). Historically, this provided an attractive justification for using the repeated measures analysis by ANOVA in randomized experiments. The randomization argument is simply not justifiable in the longitudinal data setting; measurement occasions cannot be randomly allocated to subjects. As a result, the compound symmetry assumption for the covariance is often inappropriate for longitudinal data. That is, the constraint on the correlation among repeated measurements is somewhat unappealing for longitudinal data, where the correlations are expected to decay with increasing separation in time. Also, the assumption of constant variance across time is often unrealistic. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. Finally, as originally conceived, the repeated measures ANOVA model was developed for the analysis of data from designed experiments, where the repeated measures are obtained at a set of occasions common to all individuals, the covariates are discrete factors (e.g., treatment group and time), and the data are complete. Thus the repeated measures ANOVA could not be readily applied to longitudinal data that were irregularly spaced, incomplete, or when it was of interest to include quantitative covariates in the analysis.

Despite the somewhat unappealing structure imposed on the covariance, the requirement of a longitudinal design balanced on time, and the restriction to discrete covariates, the repeated measures ANOVA was nonetheless widely adopted for the analysis of longitudinal data. Perhaps one of the major reasons for its widespread use was because the ANOVA formulation led to relatively simple computational formulas that could be performed with a desk or pocket calculator (or indeed with pen, paper and a good deal of perseverance). Historically, the repeated measures ANOVA was probably one of the few models that could realistically be fit to longitudinal data at a time when computing was in its infancy. However, with modern computing, and the widespread availability of statistical software for fitting a broader class of models for correlated data, there is little reason to analyze longitudinal data under the inherent limitations and constraints imposed by the repeated measures ANOVA model.

Repeated Measures Analysis by MANOVA

Previously, we described the repeated measures ANOVA for longitudinal data and noted in passing that it is sometimes referred to as the "univariate" or "mixed-model" analysis of variance. Analysis of variance was originally developed as a statistical model for independent observations, for example, observations on a single response variable obtained from independent subjects. By regarding the measurement occasions as levels of a within-subject factor, and by including a randomly varying individual-specific effect, the ANOVA model can be formulated in a way that allows for the possibility that repeated measures of the response obtained on the same individual are positively correlated. However, the repeated measures ANOVA is nevertheless conceptualized as a model for a single or univariate response variable.

In contrast, "multivariate" analysis of variance (MANOVA) is an extension of the analysis of variance model to handle cases where there are multiple response variables. That is, where ANOVA focuses on the analysis of a single response variable, MANOVA focuses on the analysis of a multivariate vector of response variables. In a certain sense, MANOVA is a multivariate analogue of ANOVA.

Since MANOVA was originally developed for the analysis of a multivariate vector of response variables, it is worth emphasizing some of the distinctions between longitudinal responses and more general cases of multivariate responses. Recall that longitudinal data give rise to a vector of responses. Thus the responses in a longitudinal study are inherently multivariate. On the other hand, the multivariate responses arising from a longitudinal study are commensurate, being repeated measures over time of the same response variable. With longitudinal data, the repeated measures represent selected observations of the main features of some underling continuous process that is potentially changing over time. This is in contrast to having a single measure of multiple, but substantively different or distinct, response variables (e.g., blood pressure, blood glucose, and LDL cholesterol levels). With more general multivariate data, where we have single measurements of multiple, but distinct, response variables, there is no notion of an underlying continuum. Finally, with longitudinal data the covariance among the repeated measures can be expected to have certain features or patterns; with more general multivariate data there is rarely any indication of structure to the covariance matrix.

Thus, MANOVA was developed to allow investigators to simultaneously analyze a single measure of multiple response variables (e.g., blood pressure, blood glucose, and LDL cholesterol levels), each of which is of interest in its own right. MANOVA also allowed investigators to examine linear combinations of the response variables, rather than the original variables themselves. Although MANOVA was developed for multiple, but substantively different, response variables, statisticians soon recognized that such data share a common feature with longitudinal data, namely, that they are correlated. This led to the development of a very specific variant of MANOVA, known as repeated measures analysis by MANOVA (or sometimes referred to as multivariate repeated measures ANOVA), for the analysis of longitudinal data.

The repeated measures analysis by MANOVA is a special case of a more general approach known as *profile analysis*. The analysis of response profiles will be discussed

in much greater detail in Chapter 5. Here we describe the basic idea underlying the repeated measures analysis by MANOVA, but without much technical detail. In Chapter 5 we will illustrate how the repeated measures analysis by MANOVA relates to profile analysis and highlight the potential limitations of this approach for analyzing longitudinal data.

The main idea underlying the repeated measures analysis by MANOVA can be best understood be considering a simple example. Suppose that we have two treatment groups (e.g., placebo and active treatment) and subjects are measured repeatedly on n occasions. In such a study design, three fundamental questions can be considered:

1. Are the trends in the mean response over time the same in the two groups?
2. Averaged over the two groups, is the overall trend in the mean response over time flat?
3. Are the overall mean responses, averaged over occasions, the same in the two groups?

Note that the first question is equivalent to asking whether there is a "group × time interaction". Ordinarily, this first question must be addressed before consideration of the remaining questions, since it rarely makes sense to examine group or time main effects when there is an interaction. The second and third questions are equivalent to asking whether there are main effects of "time" and "group", respectively.

To address each of these questions the repeated measures analysis by MANOVA proceeds by constructing a new set of variables, derived from the original set of repeated measures. The new set of derived variables, numbering as many as the number of repeated measures, then form the basis of a MANOVA. That is, the repeated measures analysis by MANOVA proceeds by constructing a set of derived variables and uses relevant subsets of these to address each of the three questions posed above.

A simple example will help to motivate the main ideas. Suppose that in a longitudinal clinical trial, designed to compare a new treatment to placebo, repeated measures of the response variable are obtained on three occasions ($n = 3$). A repeated measures analysis by MANOVA proceeds by constructing three derived variables, say V_{i1}, V_{i2}, and V_{i3}. The first derived variable is simply the sum (or average) of the responses. That is, for each individual we can construct

$$V_{i1} = (Y_{i1} + Y_{i2} + Y_{i3}).$$

This derived variable provides no information about within-individual changes in the response over time. Instead, it provides information about the mean level of the response, averaged over all three occasions.

The two remaining derived variables are constructed to provide information about possible within-individual changes in the response over time. For example, the following two derived variables,

$$V_{i2} = (Y_{i2} - Y_{i1}), \text{ and } V_{i3} = (Y_{i3} - Y_{i1}),$$

provide information about changes in the response (from time 1) at times 2 and 3, respectively. The set of three derived variables can be obtained by applying the following transformation matrix,

$$\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix},$$

to the original vector of responses. That is,

$$\begin{pmatrix} V_{i1} \\ V_{i2} \\ V_{i3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix}.$$

Thus, in the repeated measures analysis by MANOVA, a transformation matrix takes a sequence of repeated measures and produces an equal number of derived variables that are used in subsequent analyses. The first row of the transformation matrix creates the sum (or average) of the repeated measures (it makes no difference whether the sum or average is used since the latter is proportional to the former). The first derived variable provides information about the mean level of the response, averaged over all measurement occasions, and can be used to address the third question concerning whether there is a "group" effect. The remaining rows of the transformation matrix construct derived variables that provide information about change over time. There are many different ways to obtain a set of derived variables that describe change over time and so there are many possible choices of values for the remaining rows of the transformation matrix. For example, if it is of interest to construct derived variables that represent linear and quadratic contrasts of time the following transformation matrix can be used

$$\begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}.$$

It can be shown that the multivariate statistics for tests of "time" effects and their interactions produced by the repeated measures analysis by MANOVA are invariant to how change over time is characterized in the transformation matrix.

Given the set of derived variables, the repeated measures analysis by MANOVA proceeds as follows. First, the two derived variables representing contrasts of time, V_{i2} and V_{i3}, are analyzed using MANOVA. For example, the first question can be addressed by comparing the groups in terms of these two derived variables. Specifically, in our simple example with two groups, this is achieved by using a multivariate extension of the two-sample t-test which is known as Hotelling's T^2 test. A test of no differences between groups on these two derived variables is equivalent to a test of the "group × time interaction". Next, and assuming that there is no "group × time interaction", the second question can be addressed. The second question is concerned

with the shape of the overall (i.e., averaged over groups) trend in the mean response over time. If the mean response trend over time is flat, then the two derived variables have expectation zero. As a result, the second question can be addressed by using a multivariate extension of the single sample t-test, the single sample Hotelling's T^2 test, of the hypothesis that V_{i2} and V_{i3} have mean zero. Finally, the third question can be addressed by focusing on the first derived variable, V_{i1}. This variable is proportional to the mean of the repeated measures and can be used to assess whether there is a "group" effect. Specifically, group difference in the overall mean response, averaged over measurement occasions (or time), can be examined using a simple two-sample t-test, or, in the case of more than two groups, using ANOVA. A standard ANOVA of the first derived variable (the sum or average of the repeated measures) is performed and provides a test of the group or between-subject factor. Note that there is nothing intrinsically multivariate in this last part of the analysis since the analysis is based on a single derived variable. It is the first part of the analysis that is intrinsically multivariate.

In summary, the underlying idea behind repeated measures analysis by MANOVA is to obtain a new set of derived variables, based on a linear combination of the original sequence of repeated measures. The derived variables can be partitioned into a set that provides information about change, and a single derived variable that provides information about overall level of response. The latter can be analyzed using univariate ANOVA and the results of this analysis determine whether there are "group" or between-subject effects. This analysis addresses the question of whether the groups differ in their overall level of response. The remaining derived variables are analyzed using MANOVA and these analyses determine whether there are time effects and group $\times$ time interactions. A test of the group $\times$ time interactions addresses the question of whether the changes in the mean response over time are different in the groups. If there are no differences between groups in these derived variables, it is then of interest to ask whether the combined average of the derived variables, where the averaging is over groups, is different from zero. This addresses the question of whether there is any overall change in the mean response over time. In effect, this can be considered a test of the main effect of the time factor.

The repeated measures analysis by MANOVA has a number of features that make it unappealing for the analysis of longitudinal data. In particular, the MANOVA formulation forces the within-subject covariates to be the same for all individuals in the study. There are at least two practical consequences of this constraint. First, repeated measures MANOVA cannot be used when the design is unbalanced over time, that is, when the vectors of repeated measures are of different lengths and/or obtained at different sequences of time. Second, the repeated measures MANOVA (as implemented in existing statistical software packages) does not allow missing data. If any individual has a single missing response at any occasion, the entire data vector from that individual is excluded from the analysis. This so-called "listwise" deletion of missing data from the analysis can result in dramatically reduced sample size and very inefficient use of the available data. Listwise deletion of missing data can also produce biased estimates of change in the mean response over time when the so-called "completers" (i.e., those with no missing data) are not a random sample from the target population. When the "completers" are a biased sample from the target population, the sample means, variances, and covariances are biased estimates of the corresponding parameters in the target population. Some additional drawbacks of the repeated measures analysis by MANOVA will be discussed in Chapter 5, where a more detailed exposition on the analysis of response profiles is given.

Summary Measure Analysis

A common approach to the analysis of longitudinal data still in widespread use reduces the sequence of repeated measures for each individual to a small set of summary values. The major motivation behind this approach is that if the sequence of repeated measures can be reduced to a single number summary then standard parametric or nonparametric methods for the analysis of a univariate response can be applied to the derived measures.

For example, the area under the curve (AUC) is one common measure that is often used to summarize the sequence of repeated measures on any individual. The use of AUC is appropriate when the repeated measures for each individual are obtained at the same set of occasions. The AUC can be especially appealing in pharmacological studies where the response, or some transformation of the response, measures the absorption, concentration, or clearance of drugs. For example, the AUC can be used to estimate the clearance rate or plasma concentration of a particular dose of a drug or substance over time. The AUC can be approximated for each individual by joining adjacent measurements by line segments and summarizing the area under the curve by the sum of the areas of the resulting trapezoids (see Figure 3.9). The resulting AUC's can then be related to covariates (e.g., treatment or exposure group) using standard methods for the analysis of a univariate response (e.g., t-test, ANOVA, Wilcoxon rank sum test, or Kruskal-Wallis test).

When the covariates are discrete, and the repeated measures for each individual are obtained at the same set of occasions, the AUC analysis can also be based on the results of an analysis of response profiles (that assumes arbitrary patterns for the mean responses over time). Given the covariates, the AUC for the mean response over time is the same as the average (or mean) of the individual-specific AUC's. That is, for the case of linear models for continuous longitudinal responses, the AUC for the mean response over time coincides with the mean of the AUC's for the individuals in the population of interest. In a limited context, reducing the sequence of repeated measures for each individual to an AUC can provide a useful basis for the analysis of longitudinal data. However, the analysis of AUC's is problematic with unbalanced longitudinal data.

Another measure commonly used to summarize the sequence of repeated measures is the slope or constant rate of change in the response over time. For example, it might be assumed that a straight line (the simplest possible curve) fits the observed responses for each subject. If Y_{ij} is the response of the i^{th} individual measured at time t_{ij}, it might be assumed that

$$Y_{ij} = b_{1i} + b_{2i}\, t_{ij} + e_{ij},$$

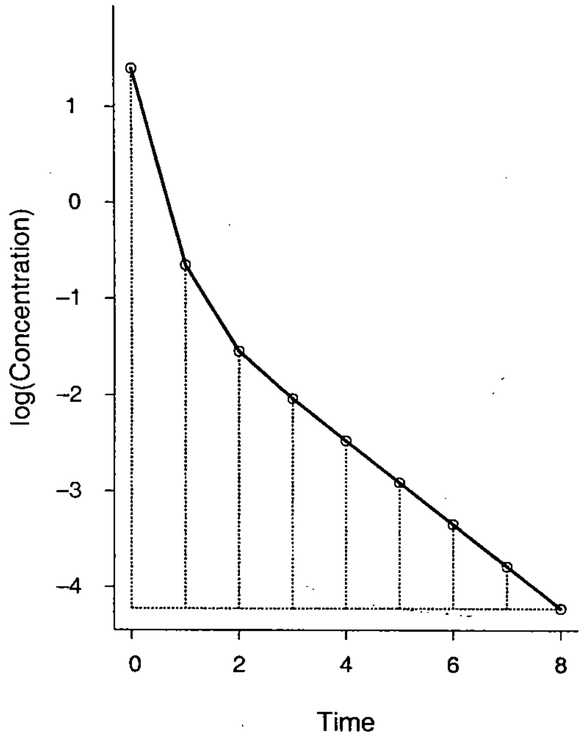

Fig. 3.9 Time plot of log(concentration) versus time, illustrating how the area under the curve (AUC) can by calculated using the trapezoidal rule.

where b_{1i} and b_{2i} are regression parameters specific to the i^{th} individual and the errors, e_{ij}, are implicitly assumed to be independent within any individual. Estimates of the individual-specific slopes (and intercepts) can then be obtained from a linear regression line fit to each individual's repeated measures. The resulting slopes can then be related to the covariates using standard parametric or nonparametric methods for the analysis of a univariate response. This approach does not require that the repeated measures for each individual be obtained at the same set of occasions. Finally, we note that extensions of this particular summary measure analysis approach lead naturally to a class of models referred to as "growth curve models". Studies of growth and aging are classic examples of observational longitudinal studies. In these studies the goal is to describe naturally occurring changes in the response over time, due to developmental or aging processes, and to compare these growth curve profiles in different groups (e.g., males and females). To meet this goal, growth curve models have been developed to summarize the pattern of response over time and allow for the possibility that individuals may belong to or be drawn from different groups. Growth

curve models can be motivated in terms of a two-stage model. Indeed, growth curve models are sometimes referred to as "two-stage" growth models. At the first stage, it is assumed that a parametric curve (e.g., linear or quadratic trend in time) fits the observed responses for each subject. In the second stage, these individual-specific growth parameters are then related to covariates that describe the different groups from which the individuals have been drawn. Growth curve models will be discussed in greater detail in Chapter 8.

Before the advent of modern computing and readily available statistical software for analyzing correlated data, summary measure analysis of longitudinal data had some very obvious appeal. First, the summary measures and their subsequent analysis can readily be understood by investigators with limited training in statistics. Also, once a summary measure has been derived, standard methods for the analysis of a univariate response (e.g., t-test, ANOVA, linear regression, Wilcoxon rank sum test, Kruskal-Wallis test) can be validly applied since issues of correlation among the observations no longer arise. That is, the summary measures on different individuals are independent of one another. Summary measure analysis can also be appealing when sample sizes are not sufficiently large for estimation of the correlation among the repeated measures. However, despite the simplicity of the method, it does have a number of distinct drawbacks. One drawback is that it forces the data analyst to focus on only a single aspect of the repeated measures over time. It should be intuitively clear that when n repeated measures are replaced by a single number summary, there must necessarily be some loss of information. Furthermore, individuals with discernibly different response profiles can have the same summary measure. For example, individual-specific response profiles with quite distinct shapes can result in the same AUC. Another potential drawback of the summary measure approach is that the covariates must be time-invariant. Thus, if one of the key covariates is time-varying, the method cannot be applied. Finally, we note that some of the summary measures that have been proposed are not well defined when there are missing data or irregularly spaced repeated measures. Even when they can be defined, these simple methods lose efficiency.

In those cases where the summary measures are well defined when individuals have missing data or different numbers of repeated measures, the analysis becomes more complicated because the derived summary measures no longer have the same variance. Similarly, if the repeated measures are taken at irregular times for different individuals, the resulting summary measures may also have different variances. In all of these cases the variance of the derived summary measures is not constant, violating a fundamental assumption made by many standard statistical methods for univariate responses. Thus, in general, the standard parametric methods for the analysis of a univariate response (e.g., t-tests, ANOVA, linear regression) cannot be validly applied to the summary measures when the design is unbalanced over time due to missing data, different numbers of repeated measures, or sequences of repeated measures taken at irregular times for different individuals.

When longitudinal data are unbalanced over time, a proper analysis of the summary measures would require that each summary measure be weighted differently. However, the chief complication here is that the specific weights given to each sum-

mary measure will in general depend implicitly upon the covariance among the repeated measures. Thus, a simple univariate analysis cannot proceed without proper consideration of the covariance, the very feature of the data that these methods were developed to avoid having to specify. In conclusion, in limited contexts summary measure analysis of longitudinal data can be useful, but it should be avoided when the data are unbalanced. When it is desirable to base analysis on a single aspect of the repeated measures over time, the regression models that are the focus of later chapters can be used. The regression modelling approach is more efficient than the summary measure analysis and can also handle unbalanced data.

3.7 FURTHER READING

Winer (1971) provides a very accessible discussion of the application of repeated measures analysis by ANOVA. A comprehensive description of repeated measures analysis by MANOVA, targeted at applied researchers, can be found in the book by Hand and Taylor (1987). Finally, for a non-technical discussion of the analysis of summary measures, readers are referred to the review articles by Matthews *et al.* (1990) and Everitt (1995).

Bibliographic Notes

An excellent discussion of scatter-plot smoothing techniques can be found in Chapter 3 of Ruppert *et al.* (2003).

Ware and Liang (1996) provide an interesting historical perspective on the development of statistical methods for the analysis of longitudinal data, with emphasis on the contributions that have been made in the biostatistical literature.

The foundations for the repeated measures analysis of variance can be found in the seminal monograph by Fisher (1925) and in the method for analyzing split-plot experiments proposed by Yates (1935); also see Scheffé (1959). Greenhouse and Geisser (1959) described an adjustment to the repeated measures analysis by ANOVA when the required assumption about the covariance matrix (compound symmetry) does not hold. The repeated measures analysis by MANOVA was introduced in the statistical literature by Box (1950); also see Danford *et al.* (1960), Geisser (1963), Cole and Grizzle (1966), and Morrison (1972). A discussion of repeated measures analyses by ANOVA and MANOVA, and the relationship between the two methods, can be found in Chapters 2, 3, and 11 of Hand and Crowder (1996).

Finally, the analysis of summary measures has a long history, dating back to the early contributions to growth curve analysis by Wishart (1938), Box (1950), and Rao (1958). Rowell and Walters (1976), in a classic paper on the analysis of longitudinal agricultural experiments, describe how linear regressions can be fitted to longitudinal data on each subject, followed by an analysis of the values of the resulting regression coefficients. The article by Rowell and Walters (1976) is widely cited for popularizing the analysis of summary measures of growth in many different disciplines.

4

Estimation and Statistical Inference

4.1 INTRODUCTION

So far, our discussion of models for longitudinal data has been very general, with no mention of methods for estimating the regression coefficients or the covariance among the repeated measures. In Chapters 5–8 we will consider models for longitudinal data where the response variable is continuous and assumed to have an approximate multivariate normal distribution. In these chapters, the main focus is on various aspects of modelling longitudinal data, with particular emphasis on models for the mean and covariance. All of the models presented in Chapters 5–8 can be expressed in terms of a general linear regression model for the mean response vector

$$E(Y_i) = X_i\beta, \tag{4.1}$$

where the response vector, Y_i, is assumed to arise from a multivariate normal distribution with covariance matrix

$$\text{Cov}(Y_i) = \Sigma_i = \Sigma_i(\theta), \tag{4.2}$$

where θ is a $q \times 1$ vector of covariance parameters. For example, with balanced longitudinal data ($n_i = n$), where an "unstructured" covariance matrix has been assumed, the elements of θ are simply the n variances and $\frac{n(n-1)}{2}$ pairwise covariances stacked in a single $q \times 1$ vector (where $q = \frac{n(n+1)}{2}$). On the other hand, if the covariance is assumed to have a "compound symmetry" pattern, then $q = 2$ and the two elements of θ represent the common value of the variances and common value of the pairwise covariances. In this section we consider a framework for estimation of the unknown parameters, β and θ (or, equivalently, Σ_i).

4.2 ESTIMATION: MAXIMUM LIKELIHOOD

Given that full distributional assumptions have been made about the vector of responses, Y_i, since the multivariate normal distribution is entirely specified by the mean vector and covariance matrix, a very general approach to estimation is the method of *maximum likelihood* (ML). The fundamental idea behind ML estimation is really quite simple and is conveyed by its name: use as estimates of β and θ the values that are most probable (or most "likely") for the data that have actually been observed. The maximum likelihood estimates of β and θ are those values of β and θ that maximize the joint probability of the response variables evaluated at their observed values. The probability of the response variables evaluated at the fixed set of observed values, and regarded as functions of β and $\Sigma_i(\theta)$, is known as the *likelihood function*. Thus estimation of β and θ proceeds by maximizing the likelihood function. In a certain sense, the method of maximum likelihood chooses values of β and θ that best explain the observed data. The values of β and θ that maximize the likelihood function are called the *maximum likelihood estimates* of β and $\Sigma_i(\theta)$, and are usually denoted $\widehat{\beta}$ and $\widehat{\Sigma}_i$ (or $\Sigma_i(\widehat{\theta})$).

Before we present any more details concerning maximum likelihood estimation of β and θ, it will be informative to consider this method of estimation in the simpler case where all observations can be assumed to be independent, that is, in the standard linear regression model with independent (and hence uncorrelated) errors that are assumed to have a univariate normal distribution.

Independent Observations

Suppose that the data arise from a series of cross-sectional studies that are repeated at n different occasions. At each occasion, data are obtained on a sample of N individuals. Here it is reasonable to assume that the observations are independent of one another, since each individual is measured at only one occasion. Also, for ease of exposition, we assume that the variance is constant, say σ^2. The mean response is related to the covariates via the following linear regression model:

$$E(Y_{ij}) = X_{ij}'\beta.$$

To obtain maximum likelihood estimates of β, we must find the values of the regression parameters that maximize the joint normal probability density function of all the observations, evaluated at the observed values of the response, and regarded as a function of β (and σ^2). Recall from Section 3.2 that the univariate normal (or Gaussian) probability density function for Y_{ij} can be expressed as

$$f(y_{ij}) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(y_{ij} - \mu_{ij}\right)^2/\sigma^2\right\},$$

where $-\infty < y_{ij} < \infty$. When all the responses are independent of one another, the likelihood function is simply the product of the individual univariate normal

probability density functions for Y_{ij},

$$\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij}).$$

It is more common to work with the log-likelihood function which will involve sums, rather than products, of the individual univariate normal probability density functions for Y_{ij}. Note that maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood; the latter is denoted by l. Hence, the goal is to maximize

$$l = \log\left\{\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij})\right\} = -\frac{K}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(y_{ij} - X_{ij}'\beta\right)^2/\sigma^2,$$

evaluated at the observed numerical values of the data, with respect to the regression parameters, β. Here, $K = n \times N$, the total number of observations. Note that β does not appear in the first term in the log-likelihood; as a result, this term can be ignored when maximizing the log-likelihood with respect to β. Furthermore, since the second term has a negative sign, maximizing the log-likelihood with respect to β is equivalent to minimizing the following function:

$$\sum_{i=1}^{N}\sum_{j=1}^{n}\left(y_{ij} - X_{ij}'\beta\right)^2.$$

Maximizing or minimizing a function is a common mathematical problem that can be solved using calculus. Specifically, the maximum likelihood estimate of β can be obtained by equating the derivative of the log-likelihood, often called the score function, to zero and finding the solution to the resulting equation. However, in the example considered here, there is no real need to resort to calculus. Obtaining the maximum likelihood estimate of β is equivalent to finding the ordinary least squares (OLS) estimate of β, that is, the value of β that minimizes the sum of the squares of the residuals. Using vector notation, the least squares solution can be written as

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}X_{ij}'\right)\right\}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}y_{ij}\right).$$

This least squares estimate is the value produced by any standard statistical software for linear regression (e.g., PROC GLM or PROC REG in SAS). In the next section, we consider how these ideas can be extended to the setting of correlated data. Also, the alert reader may have noticed that we have thus far only focused on estimation of β, ignoring estimation of σ^2; in the next section, we also consider estimation of the covariance matrix.

Correlated Observations

When there are n_i repeated measures on the same individual, it cannot be assumed that these repeated measures are independent. As a result, we need to consider the joint probability density function for the vector of repeated measures. Note, however, that the vectors of repeated measures are assumed to be independent of one another. Thus, the log-likelihood function, l, can be expressed as a sum of the individual multivariate normal probability density functions for Y_i.

To find the maximum likelihood estimate of β in the repeated measures setting we first assume that Σ_i (or θ) is *known* (and therefore does not need to be estimated); later, we will relax this very unrealistic assumption. To obtain the maximum likelihood estimate of β, we must find the value of β that maximizes the log-likelihood function. Given that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ is assumed to have a multivariate normal distribution, we must maximize the following log-likelihood function:

$$l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\,\Sigma_i^{-1}\,(y_i - X_i\beta)\right\},$$

where $K = \left(\sum_{i=1}^{N} n_i\right)$ is the total number of observations. Note that β does not appear in the first two terms in the log-likelihood; as a result, these two terms can be ignored when maximizing the log-likelihood with respect to β. Furthermore, since the third term has a negative sign, maximizing the log-likelihood with respect to β is equivalent to minimizing

$$\sum_{i=1}^{N}(y_i - X_i\beta)'\,\Sigma_i^{-1}\,(y_i - X_i\beta)\,. \tag{4.3}$$

The estimator of β that minimizes this expression is known as the *generalized least squares* (GLS) estimator of β and can be expressed as

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}(X_i'\Sigma_i^{-1}X_i)\right\}^{-1}\sum_{i=1}^{N}(X_i'\Sigma_i^{-1}y_i)\,. \tag{4.4}$$

Recall that so far we have made the somewhat unrealistic assumption that Σ_i, or θ, is *known*. Before considering how to proceed when we must relax this assumption, it is worth discussing some of the properties of the GLS estimator of β when Σ_i is known. The first very notable property is that for any choice of Σ_i, the GLS estimate of β is unbiased; that is,

$$E(\widehat{\beta}) = \beta.$$

In addition, in large samples (or asymptotically), the sampling distribution of $\widehat{\beta}$ can be shown to have a multivariate normal distribution with mean, β, and covariance,

$$\mathrm{Cov}(\widehat{\beta}) = \left\{\sum_{i=1}^{N}(X_i'\Sigma^{-1}X_i)\right\}^{-1}. \tag{4.5}$$

This is true exactly when Y_i has a multivariate normal distribution, and true in large samples even when Y_i does not have a multivariate normal distribution. (By "large samples" we mean that the sample size, N, grows larger while the number of repeated measures and model parameters remains fixed.) Thus, an important property of the GLS estimator of β, derived under the assumption of a multivariate normal distribution for Y_i, is that it provides a valid estimate of β even when the multivariate normal distribution assumption does not hold. Also, note that if Σ_i is assumed to be a diagonal matrix, with constant variance σ^2 along the diagonal (i.e., the correlations are zero and the variances are constant), the GLS estimator reduces to the ordinary least squares (OLS) estimator considered earlier. Finally, although the GLS estimator of β is unbiased for any choice of Σ_i, it can be shown that the most efficient GLS estimator of β (i.e., the estimator having smallest variance or greatest precision) is the one that uses the true value of Σ_i.

Before the reader becomes exasperated, we must now address the nagging concern that we usually do not know Σ_i (or θ). Instead, we typically must estimate $\Sigma_i(\theta)$ from the data at hand. Maximum likelihood estimation of θ proceeds in the same way as with estimation of β. That is, the maximum likelihood estimate of θ is obtained by maximizing the log-likelihood with respect to θ. As mentioned earlier, the problem of maximizing a function is a common mathematical problem that can be solved using calculus. Specifically, the maximum likelihood estimate of θ can be obtained by equating the derivative of the log-likelihood with respect to θ, also known as the score function, to zero and finding the solution to the resulting equation. However, in general, this equation is non-linear and it is not possible to write down simple, closed-form expressions for the ML estimator of θ. Instead, the ML estimate must be found by solving these equation using an iterative technique. Fortunately, computer algorithms have been developed to find the solution. Once the ML estimate of θ has been obtained, we then simply substitute the estimate of $\Sigma_i(\theta)$, say $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$, into the generalized least squares estimator of β given by (4.4) to obtain the maximum likelihood (ML) estimate of β:

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}_i^{-1}X_i\right)\right\}^{-1}\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}_i^{-1}y_i\right). \tag{4.6}$$

Interestingly, in large samples (or asymptotically), the resulting estimator of β that substitutes the ML estimate of Σ_i has all of the same properties as when Σ_i is actually known (the case we first considered). That is, in large samples:

1. $\widehat{\beta}$ is a consistent estimator of β; this property can be loosely interpreted to mean that there is very high probability that $\widehat{\beta}$ is close to the population regression parameters β for increasing sample size N. If the distribution of the errors, e_i, is assumed to be normal, or even under the weaker assumption that the distribution of e_i is symmetric, then $\widehat{\beta}$ is also an unbiased estimator of β,

$$E(\widehat{\beta}) = \beta.$$

2. The sampling distribution of $\widehat{\beta}$, when Σ_i is estimated from the data, is approximately multivariate normal with mean, β, and covariance

$$\text{Cov}(\widehat{\beta}) = \left\{ \sum_{i=1}^{N} \left(X_i^t \Sigma_i^{-1} X_i \right) \right\}^{-1}.$$

Furthermore, these properties of $\widehat{\beta}$ hold in large samples even when the assumption that Y_i has a multivariate normal distribution is not valid, provided the data are complete. Thus, an important property of the ML estimator of β, derived under the assumption of a multivariate normal distribution for Y_i, is that it provides a valid estimate of β even when the multivariate normal distribution assumption does not hold. Moreover, this appealing property of the ML estimator of β, and of any GLS estimator of β (recall, the ML estimator of β is also the GLS estimator with the ML estimate of $\Sigma_i(\theta)$ substituted), extends to the incomplete data setting when certain assumptions about missingness hold.

Thus, in terms of properties of the sampling distribution of $\widehat{\beta}$, there is no penalty for actually having to estimate Σ_i from the longitudinal data at hand. However, as comforting as this result may appear to be, it must be kept in mind that this is a large sample (i.e., as N approaches infinity) property of $\widehat{\beta}$. With sample sizes of the magnitude often encountered in many fields of application, the properties of the sampling distribution of $\widehat{\beta}$ can be expected to be adversely influenced by the estimation of a very large number of covariance parameters. This is an important issue that we will return to in Chapter 7.

4.3 MISSING DATA ISSUES

Although most longitudinal studies are designed to collect data on every individual in the sample at each time of follow-up, many studies have some missing observations. In longitudinal studies in the health sciences, missing data are the rule, not the exception. Missing data have three important implications for longitudinal analysis. First, when longitudinal data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. As a result, methods of analysis need to be able to handle the unbalanced data without having to discard data on individuals with any missing data. This feature of missingness will not be of any concern for the methods described in later chapters of the book. Second, when there are missing data, there will be a loss of information and a reduction in the precision with which changes in the mean response over time can be estimated. This reduction in precision is directly related to the amount of missing data and will also be influenced to a certain extent by how the analysis handles the missing data. For example, using only the complete cases (i.e., those individuals with no missing data) will usually be the least efficient method. Finally, when there are missing data, the validity of any method of analysis will require that certain assumptions about the reasons for any missingness, often referred

to as the *missing data mechanism*, are tenable. Consequently, when data are missing we must carefully consider the reasons for missingness.

In this section we review two general types of missing data mechanisms. The two mechanisms differ in terms of assumptions concerning whether missingness is related to responses that have been observed. The distinctions between different types of missing data mechanisms and alternative methods for handling missingness in longitudinal studies will be discussed in greater detail in Chapter 14.

The missing data mechanism can be thought of as a model that describes the probability that a response is observed or missing at any occasion. We make an important distinction between missing data mechanisms that are referred to as *missing completely at random* (MCAR) and *missing at random* (MAR). The distinction between these two mechanisms determines the appropriateness of maximum likelihood estimation under the assumption of a multivariate normal distribution for the responses and GLS without requiring assumptions about the shape of the distribution.

Data are said to be missing completely at random (MCAR) when the probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained (the *missing* responses) or the set of observed responses. That is, longitudinal data are MCAR when missingness in Y_i is simply the result of a chance mechanism that does not depend on either observed or unobserved components of Y_i. The essential feature of MCAR is that the observed data can by thought of as a random sample of the complete data. As a result, the moments (e.g., the means, variances, and covariances) and, indeed, the distribution of the observed data do not differ from the corresponding moments or distribution of the complete data.

An MCAR mechanism has important consequences for the analysis of longitudinal data. In particular, any method of analysis that yields valid inferences in the absence of missing data will also yield valid inferences when missing data are MCAR and the analysis is based on all available data, or even when it is restricted to the so-called "completers" (i.e., those with no missing data). Given that valid estimates of the means, variances, and covariances can be obtained, GLS provides valid estimates of β without requiring any distributional assumptions for Y_i. The GLS estimator of β is valid provided the model for the mean response has been correctly specified; it does not require any assumptions about the joint distribution of the longitudinal responses. The maximum likelihood (ML) estimator of β, under the assumption that the responses have a multivariate normal distribution, is also the GLS estimator (with the ML estimate of $\Sigma_i(\theta)$, say $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$ substituted). Thus in this setting the ML and GLS estimators have exactly the same properties regardless of the true distribution of Y_i.

In contrast to MCAR, data are said to be missing at random (MAR) when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained. Put another way, if subjects are stratified on the basis of similar values for the responses that have been observed, missingness is simply the result of a chance mechanism that does not depend on the values of the unobserved responses. However, because the missingness mechanism now depends upon observed responses, the distribution of Y_i in each of the distinct strata defined by the patterns of missingness is not the same

as the distribution of Y_i in the target population. This has important consequences for analysis. One is that an analysis restricted to the "completers" is not valid. Put another way, the "completers" are a biased sample from the target population. Furthermore, the distribution of the observed components of Y_i, in each of the distinct strata defined by the patterns of missingness, does not coincide with the distribution of the same components of Y_i in the target population. Therefore, the sample means, variances, and covariances based on either the "completers" or the available data are biased estimates of the corresponding parameters in the target population. As a result, GLS no longer provides valid estimates of β without making correct assumptions about the joint distribution of the longitudinal responses. On the other hand, ML estimation of β is valid when data are MAR provided the multivariate normal distribution has been correctly specified. This requires correct specification of not only the model for the mean response, but also the model for the covariance among the responses. In a sense, ML estimation allows the missing values to be validly "predicted" or "imputed" using the observed data and a correct model for the joint distribution of the responses.

To summarize, we have distinguished between two types of missing data mechanisms that are referred to as *missing completely at random* (MCAR) and *missing at random* (MAR). The MAR assumption is far less restrictive than MCAR. The distinction between these two mechanisms determines the appropriateness of maximum likelihood estimation under the assumption of a multivariate normal distribution for the responses and GLS without requiring assumptions about the shape of the distribution. The general properties of GLS described in the previous section require that either the data are complete or that any missing data are MCAR. If data are MAR, GLS based only on the means, variances, and covariances of the available data can yield biased estimates of β. In contrast, ML estimation yields valid estimates of β when data are MCAR or MAR, but for the latter mechanism, at the cost of requiring that the joint distribution of the responses is correctly specified. A more detailed discussion of missing data mechanisms, with concrete examples, and the implications of different types of missing data mechanisms for analysis, is presented in Chapter 14.

4.4 STATISTICAL INFERENCE

Next we consider how to make inferences about β. In particular, we consider the construction of confidence intervals and tests of hypotheses. To construct confidence intervals and tests of hypotheses about β, we can make direct use of the ML estimate $\widehat{\beta}$, and its estimated covariance matrix

$$\widehat{\mathrm{Cov}}(\widehat{\beta}) = \left\{ \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} X_i \right) \right\}^{-1},$$

where Σ_i in (4.5) is replaced by $\widehat{\Sigma}_i$, the ML estimate of Σ_i. For example, for any single component of β, say β_k, a natural method for constructing 95% confidence limits is

by taking $\widehat{\beta}_k$ plus or minus 1.96 times the standard error of $\widehat{\beta}_k$. Note that different confidence limits (e.g., 90%) can be obtained by choosing appropriate multiples of the standard error, based on the standard normal distribution. The standard error of $\widehat{\beta}_k$ is simply the square-root of the diagonal element of $\widehat{\mathrm{Cov}}(\widehat{\beta})$ corresponding to $\widehat{\beta}_k$,

$$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}.$$

Similarly, a test of the null hypothesis, H_0: $\beta_k = 0$ versus H_A: $\beta_k \neq 0$, can be based on the following Wald statistic:

$$Z = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}},$$

where $\widehat{\mathrm{Var}}(\widehat{\beta}_k)$ denotes the diagonal element of $\widehat{\mathrm{Cov}}(\widehat{\beta})$ corresponding to $\widehat{\beta}_k$. This test statistic can be compared with a standard normal distribution.

More generally, it may be of interest to construct confidence intervals and tests of hypotheses about certain linear combinations of the components of β. Let L denote a vector or matrix of *known* weights and suppose that it is of interest to test H_0: $L\beta = 0$. The linear combination of the components of β, $L\beta$, represents a contrast of scientific interest. For example, suppose that $\beta = (\beta_1, \beta_2, \beta_3)'$ and let $L = (0, 0, 1)$, then H_0: $L\beta = 0$ is equivalent to H_0: $\beta_3 = 0$. Alternatively, if $L = (0, 1, -1)$, then H_0: $L\beta = 0$ is equivalent to H_0: $\beta_2 - \beta_3 = 0$ or H_0: $\beta_2 = \beta_3$. A natural estimate of $L\beta$ is given by $L\widehat{\beta}$. Moreover, it can be shown that the sampling distribution of $L\widehat{\beta}$ is multivariate normal with mean, $L\beta$, and with covariance matrix, $L\mathrm{Cov}(\widehat{\beta})L'$.

Note that in the two examples considered earlier, L is a single, 1×3 row vector, $L = (0, 0, 1)$ or $L = (0, 1, -1)$. If L is a single row vector then $L\mathrm{Cov}(\widehat{\beta})L'$ is a single value (or scalar) and its square-root provides an estimate of the standard error for $L\widehat{\beta}$. Thus an approximate 95% confidence interval for $L\beta$ is given by

$$L\widehat{\beta} \pm 1.96\sqrt{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'}.$$

Similarly, in order to test H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$, we can use the Wald statistic,

$$Z = \frac{L\widehat{\beta}}{\sqrt{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'}},$$

and compare this test statistic to a standard normal distribution. Also, recall that if Z is a standard normal random variable, then Z^2 has a χ^2 distribution with 1 degree of freedom (df), denoted χ^2_1. Thus an identical test of H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$, uses the statistic

$$W^2 = (L\widehat{\beta})\{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'\}^{-1}(L\widehat{\beta}),$$

and compares W^2 to a χ^2 distribution with 1 degree of freedom. This latter observation helps to motivate how the Wald test readily generalizes when L has more

than one row, thereby allowing simultaneous testing of a single multivariate hypothesis. For example, suppose that $\beta = (\beta_1, \beta_2, \beta_3)'$ and it is of interest to test the equality of the three regression parameters. The null hypothesis can be expressed as H_0: $\beta_1 = \beta_2 = \beta_3$. Letting

$$L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

this null hypothesis can also be expressed as H_0: $L\beta = 0$, since if

$$\begin{aligned} L\beta &= \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \end{pmatrix} = 0, \end{aligned}$$

then

$$\begin{pmatrix} \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix},$$

or, equivalently, $\beta_1 = \beta_2 = \beta_3$. In general, suppose that L has r rows (e.g., representing r contrasts of scientific interest), then a simultaneous test of H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$ is given by

$$W^2 = (L\widehat{\beta})'\{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'\}^{-1}(L\widehat{\beta}),$$

which has a χ^2 distribution with r df. The latter test is often referred to as a multivariate Wald test.

One alternative to the Wald test is the *likelihood ratio test*. The likelihood ratio test of H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$ is obtained by comparing the maximized log-likelihoods for two models, one model that incorporates the constraint that $L\beta = 0$ (e.g., $\beta_3 = 0$ or $\beta_2 = \beta_3$), the other model unconstrained (i.e., without the constraint, $L\beta = 0$). The latter is referred to as the "full" model and the former is referred to as the "reduced" model. Note that these two models are *nested*, in the sense that the "reduced" model is a special case of the "full" model. That is, when the reduced model is *nested* within the full model it is a particular version of the full model, so that when the reduced model holds the full model must necessarily hold.

The likelihood ratio test for two nested models can be constructed by comparing their respective maximized log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, for the full and reduced models, respectively. The former is at least as large as the latter. The larger the difference between $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, the stronger the evidence that the reduced model is inadequate. A formal statistical test is obtained by taking twice the difference in the respective maximized log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing the statistic to a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models. This test is called the *likelihood ratio test*.

This use of the likelihood can also provide confidence limits for β or $L\beta$. Rather than calculating confidence limits for β (or $L\beta$) as the maximum likelihood estimate, $\widehat{\beta}$, plus or minus an appropriate multiple of the standard errors, likelihood-based confidence intervals can be constructed. The basic idea behind likelihood-based confidence intervals is to consider all values of β (or $L\beta$) that are consistent with the data at hand. More formally, for a single component of β, say β_k, we can define a *profile log-likelihood*, $l_p(\beta_k)$, obtained by maximizing the log-likelihood over the remaining parameters while holding β_k at some fixed value. A likelihood-based confidence interval for β_k is obtained by considering values of β_k that are reasonably consistent with the data. Specifically, an approximate 95% likelihood-based confidence interval is given by the set of all values of β_k satisfying

$$2 \times \left\{ l_p(\widehat{\beta}_k) - l_p(\beta_k) \right\} \leq 3.84,$$

where the critical value on the right-hand side of the equation is obtained from a chi-squared distribution with 1 degree of freedom. More generally, confidence intervals for $L\beta$ can be obtained by inverting the corresponding test of H_0: $L\beta = 0$ in a similar way.

Although the construction of likelihood ratio tests and likelihood-based confidence intervals is more involved (e.g., requiring an additional fit of the model under the null hypothesis) than the corresponding Wald-based tests and confidence intervals, the likelihood-based tests and confidence intervals often have superior properties. This is especially the case when the response variable is discrete. For example, in logistic regression with binary data, likelihood ratio tests have better properties than the corresponding Wald tests. Thus, when in doubt, we recommend the use of likelihood-based tests and confidence intervals. However, for ease of presentation, many of the results presented in later chapters rely on Wald-based tests and confidences intervals; likelihood-based tests and confidence intervals are presented only in cases where the discrepancies might change the substantive conclusions of the analysis.

Finally, we note that likelihood ratio tests can also be used for hypotheses about the covariance parameters. However, there are some potential problems with the standard use of the likelihood ratio test for comparing nested models for the covariance; we will return to this topic in Chapter 7. In general, we do not recommend testing hypotheses about the covariance parameters using Wald tests (i.e., based on the ratio of the parameter estimate to its standard error). In particular, the sampling distribution of the Wald test statistic for a variance parameter does not have an approximate normal distribution when the sample size is relatively small and the population variance is close to zero. Due to the fact that the variance has a lower bound of zero, very large samples are required to justify the normal approximation for the sampling distribution of the Wald test statistic when the variance is close to zero.

Comment on Denominator Degrees of Freedom

So far, in our discussion of confidence intervals and tests of hypotheses about β we have relied on the large sample properties of the sampling distribution of the ML estimate of β. That is, we have used the standard normal and chi-squared distributions instead of t and F distributions. It can be argued that the use of the standard normal and chi-squared distributions is more "liberal" (or "anti-conservative") than the corresponding t and F distributions because there is an implicit assumption of infinite denominator degrees of freedom. By "liberal", we mean that nominal p-values may be too small and confidence intervals may be too narrow.

With large denominator degrees of freedom (e.g., due to a large sample size) estimation of θ or Σ_i does not introduce any additional uncertainty. However, with small sample sizes there is some uncertainty attached to the estimation of θ that needs to be acknowledged in our inferences about β. Ordinarily, this additional source of uncertainty is recognized by use of the t and F distributions instead of the standard normal and chi-squared distributions.

A practical difficulty with the use of the t and F distributions in this setting is that the denominator degrees of freedom associated with tests and confidence intervals for components of β is not easy to determine except in certain special cases where the data are balanced and the model for the mean has a relatively simple form. To circumvent this difficulty, various approximations for the denominator degrees of freedom have been proposed. One well-known method is the Satterthwaite approximation, a somewhat tedious and computationally demanding procedure. If Satterthwaite's (1946) method is used to obtain approximate denominator degrees of freedom, say $\widehat{v}$, then an approximate 95% confidence interval for $L\beta$ is given by

$$L\widehat{\beta} \pm t_{\widehat{v},0.025}\sqrt{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'},$$

where $t_{\widehat{v},0.025}$ is the upper 2.5% cutoff from a t distribution with $\widehat{v}$ degrees of freedom (i.e., for the t distribution with $\widehat{v}$ degrees of freedom, 95% of the area lies between $-t_{\widehat{v},0.025}$ and $t_{\widehat{v},0.025}$). Similarly, to test H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$, we can use the Wald statistic

$$\frac{L\widehat{\beta}}{\sqrt{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'}}$$

and compare this test statistic to a t distribution with $\widehat{v}$ degrees of freedom. The Satterthwaite approximation can also be applied to multivariate Wald statistics, with the chi-squared distribution replaced by the F distribution (when the multivariate Wald statistic has been divided by the number of rows of the matrix L or the numerator degrees of freedom).

Recently, Kenward and Roger (1997) proposed an alternative approximation that adjusts the test statistics and provides approximate denominator degrees of freedom. Although the Satterthwaite approximation and the approximation proposed by Kenward and Roger (1997) are implemented as options in some statistical software packages (e.g., PROC MIXED in SAS), it must be emphasized that the small sample properties of these approximations in regression models for longitudinal data have not been extensively studied.

In summary, the use of the standard normal and chi-squared distributions is valid when Σ_i (or θ) is known, or when Σ_i has been estimated with a large number of degrees of freedom. Recall that there is not much practical difference between the use of the standard normal and t distributions once the degrees of freedom of the latter exceed 100. With small sample sizes, there is some uncertainty in the estimation of θ that should be accounted for and the use of the t and F distributions, with degrees of freedom approximated by the methods of Satterthwaite (1944) or Kenward and Roger (1997), is preferred. Fortunately, in many applications in the health sciences the numbers of subjects is reasonably large relative to the number of measurement occasions. As a result, the unknown denominator degrees of freedom, especially for components of β that represent time trends and their interactions with covariates (e.g., group $\times$ time interactions), will be sufficiently large that the standard normal and chi-squared distributions are reasonable approximations to the corresponding t and F distributions. For the remainder of the book, we construct confidence intervals and tests of hypotheses about β using the standard normal and chi-squared distributions; we use approximations for the denominator degrees of freedom only in cases where it might change the substantive conclusions of the analysis.

4.5 RESTRICTED MAXIMUM LIKELIHOOD (REML) ESTIMATION

We conclude this chapter with a discussion of a variant on ML estimation, known as *restricted maximum likelihood* (REML) estimation. Recall that the ML estimates of β and θ (or Σ_i) were obtained by maximizing the following log-likelihood function:

$$l = -\tfrac{K}{2}\log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\,\Sigma_i^{-1}\,(y_i - X_i\beta)\right\}.$$

Although the ML estimates of β and $\Sigma_i(\theta)$ have desirable large sample (or asymptotic) properties, the ML estimate of Σ_i has well-known bias in finite samples. For example, the diagonal elements of Σ_i are underestimated.

To illustrate the problem, consider the case where data arise from a series of cross-sectional studies that are repeated at n different occasions. Here we can assume that the observations are independent of one another, and for ease of exposition we also assume that the variance is constant, say σ^2. As noted earlier, the ML estimates of β and σ^2 are obtaining by maximizing

$$-\frac{K}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - X_{ij}'\beta)^2/\sigma^2.$$

The ML estimator of β is

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{ij} X_{ij}') \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{ij} y_{ij});$$

while the ML estimator of σ^2 is

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{n} \left(Y_{ij} - X_{ij}' \widehat{\beta} \right)^2 / K,$$

where $K = n \times N$. Furthermore, it can be shown that

$$E(\widehat{\sigma}^2) = \left(\frac{K - p}{K} \right) \sigma^2,$$

where p is the dimension of β. As a result, the ML estimate of σ^2 is biased in small samples and underestimates σ^2. An unbiased estimator is obtained by using $K - p$ (or the residual degrees of freedom) as the denominator instead of K,

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{n} \left(Y_{ij} - X_{ij}' \widehat{\beta} \right)^2 / (K - p).$$

This estimator for σ^2 is also known as the REML estimator. Note that the bias of the ML estimate of σ^2 is a decreasing function of the total number of observations, K.

In effect, the bias arises because the ML estimate has not taken into account the fact that β is also estimated from the data. In the estimator of σ^2 we have replaced β by $\widehat{\beta}$ but have failed to acknowledge in some sense that β was estimated from the data. If there are problems of bias with the ML estimate of σ^2 with independent observations, then it should not come as a great surprise that similar problems arise in the estimation of Σ_i (or θ) with correlated data.

The theory of restricted (or residual) maximum likelihood (REML) estimation was developed to address this problem. The main idea behind REML estimation is to separate that part of the data used for estimation of Σ_i from that used for estimation of β. Estimation of Σ_i is then based only on the relevant part of the data. Thus, in effect, the fundamental idea in REML estimation of Σ_i is to eliminate β from the likelihood so that it is defined only in terms of Σ_i. This can be achieved in a number of ways. One possible way to obtain the restricted likelihood is to transform the data to a set of linear combinations of observations that have a distribution that does not depend on β. For example, the residuals after estimating β by ordinary least squares (OLS) can be used as the data for estimating Σ_i (or θ). The likelihood for these residuals depends only on θ, and not on β. Thus, rather than maximizing the log-likelihood

$$-\frac{1}{2} \sum_{i=1}^{N} \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^{N} \left(y_i - X_i \widehat{\beta} \right)' \Sigma_i^{-1} \left(y_i - X_i \widehat{\beta} \right), \tag{4.7}$$

REML maximizes the following slightly modified log-likelihood (formed from the residuals)

$$-\frac{1}{2} \sum_{i=1}^{N} \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^{N} \left(y_i - X_i \widehat{\beta} \right)' \Sigma_i^{-1} \left(y_i - X_i \widehat{\beta} \right) - \frac{1}{2} \log \left| \sum_{i=1}^{N} X_i' \Sigma_i^{-1} X_i \right|. \tag{4.8}$$

When the residual likelihood given by (4.8) is maximized, we obtain an estimate of θ (or $\Sigma_i(\theta)$) that has made a correction for the fact that β has also been estimated. Of note, the additional term in the REML log-likelihood involves a determinant term,

$$-\frac{1}{2} \log \left| \sum_{i=1}^{N} X_i' \Sigma_i^{-1} X_i \right| = \frac{1}{2} \log \left| \left(\sum_{i=1}^{N} X_i' \Sigma_i^{-1} X_i \right)^{-1} \right| = \log \left| \operatorname{Cov}(\widehat{\beta}) \right|^{\frac{1}{2}},$$

that can be expressed as the covariance of $\widehat{\beta}$. As a result, the REML likelihood multiplies that usual ML likelihood by a factor that is the square-root of the *generalized variance* of $\widehat{\beta}$, a single number summary of the variation in the estimate of β. This makes a correction or adjustments that is analogous to the correction to the denominator in $\hat{\sigma}^2$.

We recommend the use of the REML estimator for Σ_i. In general, the REML estimator will be less seriously biased than the ML estimator for Σ_i. It should be notes that the difference between ML and REML estimation becomes less important when the sample size, N, is substantially larger than p, the dimension of β. Finally, when REML estimation is used to estimate Σ, β is estimated by the usual generalized least squares (GLS) estimator

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} X_i \right) \right\}^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} Y_i \right),$$

where $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$ is the REML estimate of Σ_i.

On a final note, while the REML log-likelihood can be used to compare nested models for the covariance (e.g., in terms of likelihood ratio tests comparing nested models for the covariance), it should not be used to compare nested regression models for the mean. The extra determinant term in the REML log-likelihood depends upon the regression model specification. As a result, the REML likelihoods for two nested models for the mean response are based on quite different transformations of the data (to obtain linear combinations of Y_i whose distributions do not depend on β). In short, the REML likelihoods for two nested models for the mean are based on

two entirely different sets of transformed responses, making comparisons between the models meaningless. Instead, the standard ML log-likelihood should be used for constructing likelihood ratio tests that compare nested regression models for the mean.

In conclusion, we recommend the use of REML for estimation of Σ_i (with β estimated using the GLS estimator that substitutes the REML estimate, $\widehat{\Sigma}_i$, for Σ_i). The REML log-likelihood should also be used to comparing nested models for the covariance. However, the construction of likelihood ratio tests comparing nested models for the mean should always be based on the ML, not the REML, log-likelihood.

4.6 FURTHER READING

Many textbooks on statistical theory and methods include a discussion of the methods of least squares and maximum likelihood estimation. Weisberg (1985) provides a useful introduction to the method of least squares in the context of regression; Chapter 4 of Cox and Wermuth (1996) presents a concise but remarkably lucid description of least squares, generalized least squares, and maximum likelihood estimation.

Bibliographic Notes

A discussion of the properties of generalized least squares (GLS) estimators can be found in, for example, Amemiya (1985) and Newey and McFadden (1994). Kakwani (1967) and Kackar and Harville (1981) discuss the unbiasedness properties of GLS estimators when the assumption of normally distributed errors is replaced by the weaker assumption that the distribution of the errors is symmetric.

The use of REML, as an alternative to maximum likelihood, for covariance parameter estimation was originally proposed by Patterson and Thompson (1971). Special cases of REML estimation had previously been considered by Anderson and Bancroft (1952), Russell and Bradley (1958), and Thompson (1962) in the context of balanced ANOVA models. Harville (1974) presented a Bayesian interpretation of REML.

5

Modelling the Mean: Analyzing Response Profiles

5.1 INTRODUCTION

In this chapter we present a method for analyzing longitudinal data that imposes minimal structure or restrictions on the mean response over time and on the covariance among the repeated measures. The method focuses on analyzing response profiles and can be applied to longitudinal data when the design is balanced, with the timing of the repeated measures common to all individuals in the study. Although we focus on study designs where all subjects are measured at the same set of n occasions (i.e., *balanced* longitudinal designs), as we will show, the analysis of response profiles can also handle incompleteness due to missing data (i.e., incomplete longitudinal studies with balanced designs).

Methods for analyzing response profiles are appealing when there is a single categorical covariate (perhaps denoting different treatment or exposure groups) and when no specific *a priori* pattern for the differences in the response profiles between groups can be specified. When repeated measures are obtained at the same sequence of occasions, the data can be summarized by the estimated mean response at each occasion, stratified by levels of the group factor. At any given level of the group factor, the sequence of means over time is referred to as the mean *response profile*.

For example, consider the blood lead level data from the TLC trial. The mean response profiles for the two groups randomized to succimer and placebo are presented in Figure 5.1. This plot is produced by simply calculating the arithmetic average of the responses at each occasion, within each treatment group, and joining adjacent means with a series of line segments. In settings where data on some subjects are missing, such a plot can still be made but it is obtained from the estimated means

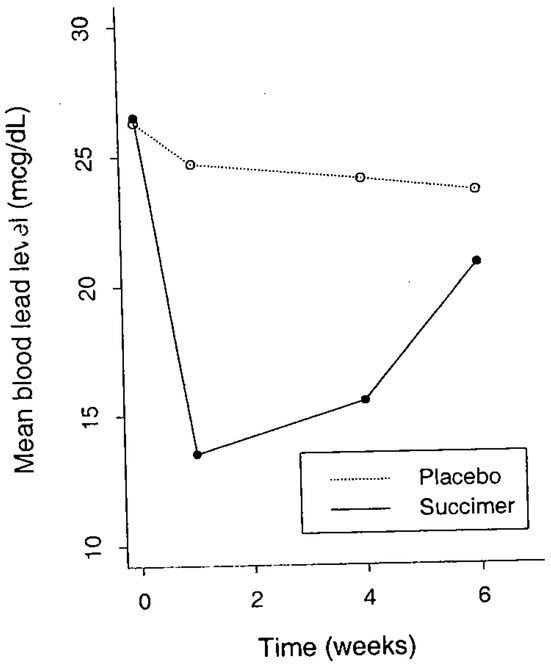

Fig. 5.1 Mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups.

for each occasion, stratified by group. We will show how to estimate these mean response profile curves in Section 5.3.

The main goal in the analysis of response profiles is to characterize the patterns of *change* in the mean response over time in the groups and to determine whether the shapes of the mean response profiles differ among the groups. For example, in the TLC trial, the major question of scientific interest is concerned with whether *changes* in the mean blood lead levels are the same for the succimer and placebo groups. In Sections 5.2 and 5.3 we emphasize how questions about whether the patterns of change are the same in all groups translate into hypotheses about the interaction between the group factor and time.

Methods for analyzing response profiles can be extended in a straightforward way to handle the case where there is more than a single group factor and when there are baseline covariates that need to be adjusted for. However, for ease of exposition, we focus on the case where there is only a single group factor. For example, in an observational study the groups might be defined by characteristics of the study subjects, such as age, gender, or exposure level. Alternatively, groups might be defined by random assignment to different treatments or interventions. The distinction

between observational studies and randomized trials is important and, as we will see later, has ramifications for the analysis of response profiles.

A characteristic feature of longitudinal studies is the presence of a baseline measurement. In the TLC trial, the objective is to compare the patterns of change in blood lead levels from baseline over time across the treatment groups. The baseline measurement is an outcome like those measured subsequently, but is unique in that, being pre-randomization, it can be assumed not to depend on treatment group. Indeed, this is apparent in the plot of the mean response profiles in Figure 5.1. This is a common feature of longitudinal studies which involve randomization after baseline. The baseline response may play a special role in other settings as well. For example, sometimes the baseline response is range restricted, as when only subjects with values greater than or less than a threshold are included in the study. With observational studies of growth or decline, groups may be known to differ at baseline, or comparison groups may be selected by matching so that baseline means are comparable.

Thus the question naturally arises as to how to handle the baseline measurement in the assessment of change. This is important, since it will affect how we construct hypothesis tests, and how they should be interpreted. In addition, how we handle the baseline response in the analysis will have an impact on efficiency and the power of tests of hypotheses. In Section 5.6, we describe two ways of adjusting for the baseline value in a simple setting and discuss their relative merits under different longitudinal study designs. In Section 5.7, we compare and contrast a number of alternative strategies for handling the baseline response in more general settings and make recommendations about the preferred strategies in different situations. Many of our readers may find the level of detail in Section 5.7 somewhat daunting. We note that Section 5.7 can be omitted at first reading without loss of continuity. However, we encourage all of our readers to eventually tackle the material in Section 5.7 since appropriate adjustment for baseline is an important aspect of the analysis of longitudinal change.

5.2 HYPOTHESES CONCERNING RESPONSE PROFILES

In our discussion of the analysis of response profiles, we focus initially on the two-group design, but generalizations to more than two groups are straightforward. Given a sequence of n repeated measures on a number of distinct groups of individuals, three main questions concerning the response profiles can be posed:

1. Are the mean response profiles similar in the groups, in the sense that the mean response profiles are parallel?
   This is a question that concerns the *group* $\times$ *time interaction effect*. A graphical representation of the null hypothesis of parallel mean response profiles is displayed in Figure 5.2(a).

2. Assuming that the population mean response profiles are parallel, are the means constant over time, in the sense that the mean response profiles are flat?
   This is a question that concerns the *time effect*. A graphical representation of

the null hypothesis that the mean response profiles are fiat is displayed in Figure 5.2(b).

3. Assuming that the population mean response profiles are parallel, are they also at the same level in the sense that the mean response profiles for the groups coincide?
   This is a question that concerns the *group effect*. A graphical representation of the null hypothesis that the mean response profiles are at the same level is displayed in Figure 5.2(c).

In longitudinal studies, it is the first question that is of main scientific interest. The hypothesis of parallel response profiles corresponds to the hypothesis that the patterns of change in the mean response over time are the same across groups. This comparison of change in the response over time is the *raison d'être* of a longitudinal study. In contrast, as we will see later, the second and third questions may not have any scientific relevance. That is, even when the response profiles can be assumed to be parallel, any interest in the second and third questions is secondary and depends upon the longitudinal study design.

Note that the second and third questions have implicitly made an assumption about the answer to the first. There is a very good reason for doing so. Except in very rare circumstances, it is not meaningful to ask the second and third questions if the mean response profiles are not parallel. Indeed, this is consistent with the general principle that *main effects* (e.g., group or time effects) are ordinarily not of interest when there is an interaction among them. That is, when there is a group × time interaction, the mean response profiles in the groups are different (non-parallel profiles); consequently, their shape can be described only with reference to a specific group and their level can be described only with reference to a specific time.

The appropriate scientific hypotheses in any particular study must be derived from the relevant scientific issues in that investigation. Here, it becomes important to distinguish between longitudinal data arising from a randomized trial and from an observational study. In the former case, when study participants have been randomized to treatment groups and the baseline values of the response has been obtained prior to any study interventions, the mean response at occasion 1 is independent of treatment assignment. That is, by design, the group means are equal at baseline (occasion 1). In contrast, in an observational study, there is no *a priori* reason to assume the groups have the same mean response at baseline unless the groups were selected by matching on baseline response.

Consider a randomized longitudinal clinical trial comparing treatments where the measurement at the first occasion is a baseline response, obtained prior to any study interventions. For example, in the TLC trial, the blood lead levels at baseline were obtained prior to receiving placebo or succimer. In that case, the only question of scientific interest is the first because it addresses whether the patterns of change in the mean response over time are the same in all groups. For example, in the TLC trial, the test of the group × time interaction assesses whether changes in the mean blood levels are the same for the succimer and placebo groups. In a randomized trial, the second question is usually of less importance because it does not involve a direct

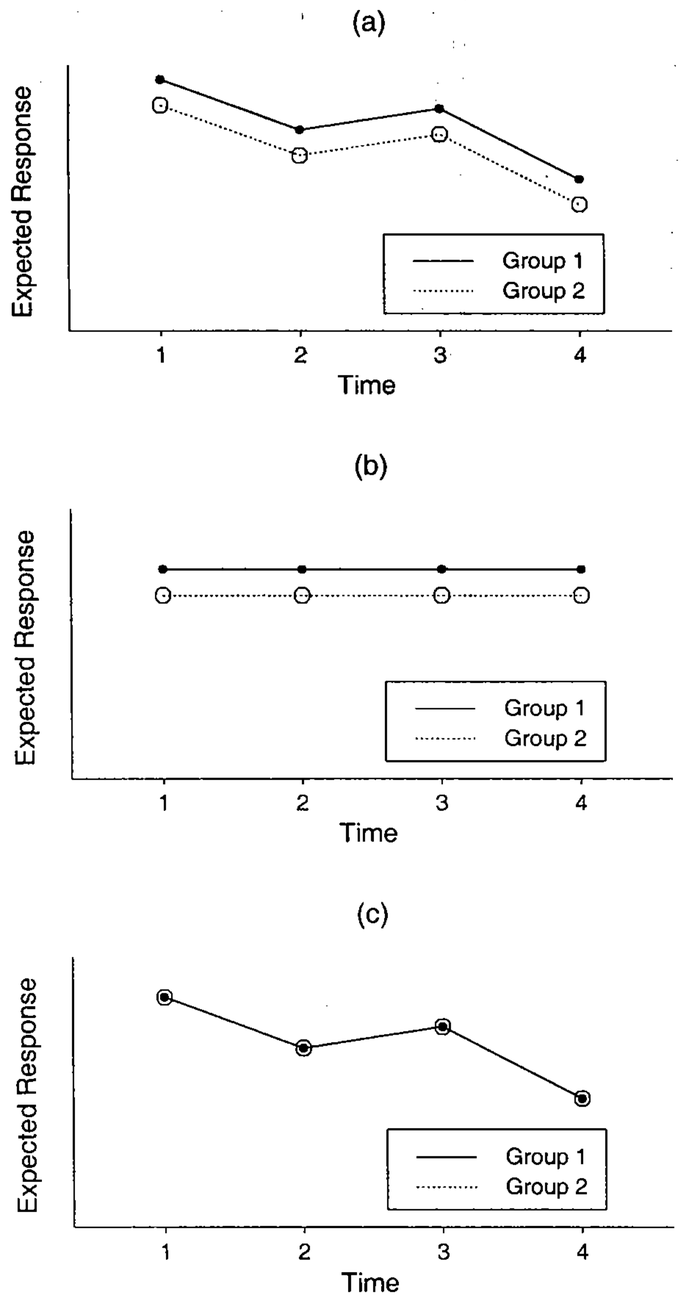

Fig. 5.2 Graphical representation of the null hypotheses of (a) no group × time interaction effect, (b) no time effect, and (c) no group effect.

comparison of groups. The second question concerns the time effect, where the focus is on the comparison of the mean response at each occasion averaged over the groups. Hypotheses concerning the main effect of time translate into questions concerning whether the overall (i.e., averaged over groups) mean response has changed from baseline. Finally, in a randomized trial, where the baseline response has been obtained before any study interventions, there is no interest in the third question. The third question concerns the group effect. However, in this setting, the absence of a group $\times$ time interaction implies that there is no group effect. That is, if the groups have the same pattern of change over time and, by design, do not differ at baseline, their mean response profiles must necessarily coincide. As a result, the test of group effect is subsumed within the test of group $\times$ time interaction.

In an observational study, where the group factor might represent different exposures or inherent characteristics of the individuals, the first question is usually of primary interest. It addresses the fundamental question of whether patterns of change over time in the mean response vary by group. In contrast to a randomized trial, however, the second and third questions may also be of substantive interest. For example, in a longitudinal study of growth or aging, there may be interest in the pattern of change in the mean response over time, even when the pattern of change is the same in all groups. This concerns the main effect of time and is addressed by the second question. Ordinarily, when there is interest in the time effect, the preferred way to describe the trend in the mean response is with a relatively simple parametric curve; methods for fitting parametric or semi-parametric curves to the mean response are described in Chapter 6. Finally, in an observational study, there may be interest in group comparisons of the mean response averaged over time. This concerns the group effect and is addressed by the third question. However, in the absence of any group $\times$ time interaction, it must be recognized that the test for the main effect of group represents a comparison of the groups in terms of their baseline (occasion 1) response. That is, if the groups have the same pattern of change over time, any group differences in the overall (i.e., averaged over occasions) mean response must reflect existing baseline differences among the groups.

To highlight the main features of the analysis of response profiles, consider the following example from a two-group study comparing a novel *treatment* and a *control*. We assume that the two groups have repeated measurements at the same set of n occasions. The analysis of response profiles is based on comparing the mean response profiles in the two groups. In a somewhat relaxed notation, let $\mu(T) = \{\mu_1(T), ..., \mu_n(T)\}'$ denote the mean response profile for the treatment group and $\mu(C) = \{\mu_1(C), ..., \mu_n(C)\}'$ denote the mean response profile for the control group. The population means in the two groups at each occasion are given in Table 5.1.

In this hypothetical study we are primarily interested in testing scientific hypotheses that compare the novel treatment and the control in terms of *changes* in the mean response over time. This can be determined by considering the null hypothesis of no group $\times$ time interaction, that is, the null hypothesis that the mean response profiles are parallel. If the mean response profiles are parallel, the difference in the means between the two groups is constant over time. As a result, in terms of Δ_j, the null

Table 5.1 Mean response profile over time in the treatment and control groups.

	Measurement Occasion			
Group	1	2	...	n
Treatment	$\mu_1(T)$	$\mu_2(T)$	...	$\mu_n(T)$
Control	$\mu_1(C)$	$\mu_2(C)$	...	$\mu_n(C)$
Difference	Δ_1	Δ_2	...	Δ_n

Note: $\Delta_j = \mu_j(T) - \mu_j(C)$.

hypothesis is

$$H_{01}: \Delta_1 = \Delta_2 = \cdots = \Delta_n,$$

where $\Delta_j = \mu_j(T) - \mu_j(C)$. If the null hypothesis is rejected, the two groups have non-parallel mean response profiles and the patterns of change over time differ in the two groups. Note that the number of constraints on the mean responses under this null hypothesis is $n - 1$. As a result, the test of this null hypothesis has $n - 1$ degrees of freedom. In Section 5.3 we describe how the constraints can be expressed in terms of specific contrasts of the means.

The previous illustration has focused on the special case of $G = 2$ groups; however, the main ideas can be generalized in a straightforward way when there are more than two groups. When there are G groups with repeated measurements at the same set of n occasions, we let $\mu(g) = \{\mu_1(g), ..., \mu_n(g)\}'$ denote the mean response profile for the g^{th} group ($g = 1, ..., G$). The population means in the G groups at each occasion are given in Table 5.2. With $G > 2$, we can compare groups in a number of different ways. However, with G groups, there are only $G - 1$ non-redundant comparisons. We define $\Delta_j(g) = \mu_j(g) - \mu_j(G)$, (for $j = 1, ..., n$; $g = 1, ..., G - 1$). That is, $\Delta_j(g)$ is a contrast or comparison of the mean response at the j^{th} occasion for the g^{th} group (for $g = 1, ..., G - 1$) with the mean response at the j^{th} occasion in group G. Then, the null hypothesis that the mean response profiles are parallel is

$$H_{01}: \Delta_1(g) = \Delta_2(g) = \cdots = \Delta_n(g); \quad \text{for } g = 1, ..., G - 1.$$

With $G \geq 2$, the test of the null hypothesis of no group $\times$ time interaction effect has $(G - 1) \times (n - 1)$ degrees of freedom.

So far, our discussion of the analysis of response profiles has focused on an omnibus test of the group $\times$ time interaction. However, unless the test of the group $\times$ time interaction has only a single degree of freedom, this test does not help in discerning in what manner the patterns of change over time differ across groups. For the latter, we must consider estimates (and their standard errors) of relevant contrasts of the

Table 5.2 Mean response profile over time in G groups.

	Measurement Occasion			
Group	1	2	...	n
1	$\mu_1(1)$	$\mu_2(1)$	...	$\mu_n(1)$
2	$\mu_1(2)$	$\mu_2(2)$	...	$\mu_n(2)$
⋮	⋮	⋮		⋮
g	$\mu_1(g)$	$\mu_2(g)$	...	$\mu_n(g)$
⋮	⋮	⋮		⋮
G	$\mu_1(G)$	$\mu_2(G)$	...	$\mu_n(G)$

means. In the next section, we describe a general linear model formulation of the analysis of response profiles. As we will show, the analysis of response profiles can be formulated in such a way that certain regression parameters have interpretations that bear directly on the questions of main scientific interest.

5.3 GENERAL LINEAR MODEL FORMULATION

Before we illustrate the main ideas with a numerical example, we consider how the analysis of response profiles can be implemented in the general linear model

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

for appropriate choices of X_i. We also describe how the main hypothesis of no group × time interaction effect can be expressed in terms of β. Let n be the number of repeated measures and N the number of subjects. To express the model for the longitudinal design with G groups and n occasions of measurement, we require $G \times n$ parameters for the G mean response profiles.

For example, with two groups measured at three occasions, there are $2 \times 3 = 6$ mean parameters (see Table 5.2). For the first group, let the design matrix X_i be the following 3×6 matrix:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix};$$

while for the second group, let the design matrix be

$$X_i = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then, in terms of the model

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_6)'$ is a 6×1 vector of regression coefficients,

$$\mu(1) = \begin{pmatrix} \mu_1(1) \\ \mu_2(1) \\ \mu_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix};$$

similarly,

$$\mu(2) = \begin{pmatrix} \mu_1(2) \\ \mu_2(2) \\ \mu_3(2) \end{pmatrix} = \begin{pmatrix} \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix}.$$

As a result, hypotheses about the mean response profiles in the two groups that were previously expressed in terms of $\mu(1) = \{\mu_1(1), \mu_2(1), \mu_3(1)\}'$ and $\mu(2) = \{\mu_1(2), \mu_2(2), \mu_3(2)\}'$ can easily be re-expressed in terms of hypotheses about the components of β. Specifically, the hypothesis of no group × time interaction effect can be expressed as

$$H_{01}: (\beta_1 - \beta_4) = (\beta_2 - \beta_5) = (\beta_3 - \beta_6).$$

In this parameterization, hypotheses about the group × time interaction cannot be expressed in terms of certain components of β being zero; instead, these hypotheses can be expressed in terms of $L\beta = 0$, for particular choices of vectors or matrices L. For example, the null hypothesis of no group × time interaction effect,

$$H_{01}: (\beta_1 - \beta_4) = (\beta_2 - \beta_5) = (\beta_3 - \beta_6),$$

can be expressed as

$$H_{01}: L\beta = 0,$$

where

$$L = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix}.$$

An attractive feature of the general linear model formulation,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

is that it can handle settings where the data for some subjects are missing. For example, suppose that the i^{th} subject belongs to the first group and is missing the response at the third occasion. The appropriate design matrix for that subject is the following 2×6 matrix, obtained by removing the last row of the full data design matrix for subjects from the first group:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

For more general patterns of missingness, the appropriate design matrix for the i^{th} subject is simply obtained by removing rows of the full data design matrix corresponding to the missing responses. This allows the analysis of response profiles to be based on all available observations of the subjects.

Note that the general linear model for two groups measured at three occasions,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

could also have been expressed in terms of the following two design matrices:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix},$$

for the first group, and

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

for the second group. In that case,

$$\mu(2) = \begin{pmatrix} \mu_1(2) \\ \mu_2(2) \\ \mu_3(2) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + \beta_3 \end{pmatrix};$$

and

$$\mu(1) = \begin{pmatrix} \mu_1(1) \\ \mu_2(1) \\ \mu_3(1) \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_4 \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5) \\ (\beta_1 + \beta_4) + (\beta_3 + \beta_6) \end{pmatrix}.$$

The choice of "reference group" (here the second group) is arbitrary and we have used the convention adopted by many of the procedures in SAS; a more detailed discussion of the "reference group" parameterization is given at the end of this section. With this choice of design matrices for the two groups, the interpretation of the regression coefficients β has changed. Re-expressing hypotheses about the mean response profiles for the two groups in terms of hypotheses about the components of β, the hypothesis of no group $\times$ time interaction is

$$H_{01}: \beta_5 = \beta_6 = 0.$$

Although both the alternative parameterizations considered thus far allow testing of hypotheses about the response profiles, the second parameterization is more convenient since the hypothesis of no group $\times$ time interaction is represented by the vanishing (or setting to zero) of certain components of β. Also, the second parameterization, often called the "reference group" parameterization, is the one that is commonly adopted by many statistical software packages (e.g., PROC MIXED in SAS).

As indicated earlier, when the hypothesis of parallel profiles cannot be rejected, hypotheses concerning the main effects of time and/or group may be of secondary interest, although their relevance depends upon the design of the study. Hypotheses concerning the main effects of time and group can similarly be represented by the vanishing (or setting to zero) of certain components of β. For example, with two groups measured at three occasions and assuming parallel profiles ($\beta_5 = \beta_6 = 0$), the hypothesis of no time effect is

$$H_{02}: \beta_2 = \beta_3 = 0;$$

the hypothesis of no group effect is

$$H_{03}: \beta_4 = 0.$$

For the more general case with G groups measured at n occasions, the number of constraints under H_{02} is $n - 1$ and is the same regardless of the number of groups; the test of H_{02} has $n - 1$ degrees of freedom. Similarly, the number of constraints under H_{03} is $G - 1$ and is the same regardless of the number of occasions; the test of H_{03} has $G - 1$ degrees of freedom. Both of these hypotheses can be tested by considering a model with only main effects of group and time. That is, valid tests of the main effects of group and time are obtained from the reduced model that excludes the group $\times$ time interaction.

Finally, given that the analysis of response profiles can be expressed in terms of the linear regression model

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_p)'$ is a $p \times 1$ vector of regression coefficients (with $p = G \times n$), maximum likelihood estimation of β, and the construction of tests of the group $\times$ time interaction (and the main effects of time and group), are possible once the covariance of Y_i has been specified. In the analysis of response profiles, the covariance of Y_i

is usually assumed to be unstructured with no constraints on the $\frac{n(n+1)}{2}$ covariance parameters other than the requirement that they yield a symmetric matrix that is positive-definite (the condition that the covariance matrix is positive-definite ensures that while the repeated measures can be highly correlated, there must be no redundancy in the sense that one of the repeated measures can be expressed as a linear combination of the others; the condition also ensures that no linear combination of the responses can have a negative variance). Given REML (or ML) estimates of β, and their standard errors (and the estimated covariance of $\widehat{\beta}$), tests of the group × time interaction (and the main effects of time and group), can be constructed using multivariate Wald tests. Alternatively, likelihood ratio tests can be constructed but require that the model be fit to the data with and without the constraints under the null hypothesis (i.e., fitting the "reduced" and "full" models, respectively). Before illustrating the analysis of response profiles, we present a brief review of the "reference group" parameterization.

REVIEW: REFERENCE GROUP PARAMETERIZATION

Consider a group factor with G levels. To represent this factor in a linear model we can define a set of "dummy" or indicator variables,

$$Z_{ig} = \begin{cases} 1 & \text{if } i^{th} \text{ subject belongs to group } g; \\ 0 & \text{otherwise.} \end{cases}$$

Letting $X_i = (Z_{i1}, \ldots, Z_{iG})$, the mean response in the G groups, denoted by $\mu_i(1), ..., \mu_i(G)$, can be expressed in terms of the following linear model:

$$E(Y_i|X_i) = \mu_i = X_i\beta.$$

In this parameterization,

$$\begin{pmatrix} \mu_i(1) \\ \mu_i(2) \\ \vdots \\ \mu_i(G-1) \\ \mu_i(G) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{G-1} \\ \beta_G \end{pmatrix}.$$

If we wish to include an "intercept", say β_1, by setting the first column of X_i to 1 (for all $i = 1, ..., N$), then there is redundancy in X_i if all G indicator variables, $Z_{i1}, \ldots, Z_{iG}$, are also included in the design vector, X_i. To avoid this overspecification, one of the indicator variables must be excluded from X_i. Arbitrarily, we can drop Z_{iG}. Then, with $X_i = (1, Z_{i1}, \ldots, Z_{iG-1})$, the mean response in the G groups can be expressed in terms of the following linear model:

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_G)'$. In this parameterization,

$$\begin{pmatrix} \mu_i(1) \\ \mu_i(2) \\ \vdots \\ \mu_i(G-1) \\ \mu_i(G) \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_2 \\ \beta_1 + \beta_3 \\ \vdots \\ \beta_1 + \beta_G \\ \beta_1 \end{pmatrix}.$$

Because the "intercept" term, β_1, is also the mean of group G, and all of the remaining components of β represent deviations from the mean of group G, this parameterization is often referred to as the "reference group" parameterization. Here, the last level of the group factor (i.e., group G) is the reference group and it is no coincidence that this is the same group whose indicator variable was excluded from X_i. Other choices of reference group can be obtained by excluding the relevant indicator variable for the group in question from X_i.

5.4 CASE STUDY

Next, we illustrate the main ideas by conducting an analysis of response profiles of the blood lead data of the 100 children from the succimer and placebo groups of the Treatment of Lead-Exposed Children (TLC) Trial.

Treatment of Lead-Exposed Children Trial

Recall that the TLC trial was a placebo-controlled, randomized trial of an orally administered chelating agent, succimer, in children with confirmed blood lead levels of 20–44 μg/dL. The children in the trial were aged 12–33 months and lived in deteriorating inner city housing. The following analysis is based on data on blood lead levels at baseline (or week 0), week 1, week 4, and week 6 during the first treatment period. The mean response profiles for the two groups were displayed in Figure 5.1.

In Table 5.3 the REML estimates of the components of the unstructured covariance matrix are displayed. Note the discernible increase in the variability in blood lead levels from pre- to post-randomization. This increase in variability from baseline is probably due to two factors. First, within each treatment group, there may be natural heterogeneity in the individual response trajectories over time. Second, the trial had an inclusion criterion that blood lead levels at baseline were in the range of 20–44 μg/dL; this may partially account for the smaller variance at baseline.

In Table 5.4 the results of the analysis of response profiles are presented. The main interest is in the test of the group × time interaction. The test of the group × time interaction is based on a multivariate Wald test. The test provides a simultaneous test of H_0: $L\beta = 0$ versus H_A: $L\beta \neq 0$, for a suitable choice of L, and the test statistic can be constructed as

Table 5.3 Estimated covariance matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Covariance Matrix			
25.2	19.1	19.7	22.2
19.1	44.3	35.5	29.7
19.7	35.5	47.4	30.6
22.2	29.7	30.6	58.7

Table 5.4 Wald tests of fixed effects based on an analysis of response profiles of the blood lead level data at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Variable	DF	Chi-Squared	*P*-Value
Group	1	25.43	<0.0001
Week	3	184.48	<0.0001
Group × Week	3	107.79	<0.0001

$$W^2 = (L\widehat{\beta})'\{L\widehat{\mathrm{Cov}}(\widehat{\beta})L'\}^{-1}(L\widehat{\beta}),$$

and compared to a χ^2 distribution with degrees of freedom equal to the number of rows of L. The corresponding likelihood ratio test could be constructed and would require the comparison of the maximized ML log-likelihood for two models, one model that incorporates the constraint that $L\beta = 0$ (i.e., the model without group × time interaction), the other model unconstrained (i.e., the model with group × time interaction).

In the TLC trial the question of main scientific interest concerns the comparison of the two treatment groups in terms of their patterns of change from baseline in the mean blood lead levels. This question translates directly into a test of the group × time interaction. From Table 5.4 the test of the group × time interaction yields a Wald statistic of 107.79 with 3 degrees of freedom (the corresponding likelihood ratio test yields $G^2 = 74.2$). When compared with the reference chi-squared distribution with 3 degrees of freedom, there is strong evidence to reject the null hypothesis and conclude that the patterns of change from baseline are not the same in the two groups. Given the pattern of observed responses (see Figure 5.1), this result is expected.

Table 5.5 Estimated regression coefficients and standard errors based on an analysis of response profiles of the blood lead level data at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.272	0.710	36.99
Group	S		0.268	1.005	0.27
Week		1	−1.612	0.792	−2.04
Week		4	−2.202	0.815	−2.70
Week		6	−2.626	0.889	−2.96
Group × Week	S	1	−11.406	1.120	−10.18
Group × Week	S	4	−8.824	1.153	−7.66
Group × Week	S	6	−3.152	1.257	−2.51

The tests of main effects of group and time in Table 5.4 are not meaningful, for two quite different reasons. First, the TLC data are from a randomized trial and the test of the main effect of time is not of subject-matter interest while the test of the main effect of group is subsumed within the test of group × time interaction. Second, in general, the tests of main effects of time and group in Table 5.4 are not meaningful in the presence of a significant group × time interaction. This underscores our earlier advice that tests of main effects should only be considered when the assumption of parallel response profiles is tenable. When the profiles are parallel, and there is scientific interest in the main effects of time and/or group, the tests of the main effects of time and group require that the model be re-fit to the data, excluding the group × time interaction. The resulting Wald tests for the main effects from this reduced model have the desired interpretations.

So far, our analysis of response profiles has provided an omnibus test of the group × time interaction. However, unless the test of the group × time interaction has only a single degree of freedom, this test does not indicate how the two groups differ. For the latter, we must consider the REML estimates of β and their standard errors presented in Table 5.5; alternative single degree of freedom tests for group × time interaction will be discussed in Section 5.5. For ease of interpretation, the baseline (week 0) is chosen as the reference level for time and the placebo group is chosen as the reference level for treatment group. From an examination of the three single-degree-of-freedom contrasts for the group × time interaction, the results indicate that children treated with succimer have a discernibly greater decrease in mean blood lead levels from baseline at all occasions when compared to the children treated with placebo. For example, when compared to the placebo group, the succimer group has an additional 3.152 μg/dL (with SE = 1.257) decrease in mean blood lead levels

from baseline to week 6. Of note, there are even larger differences between the two treatment groups earlier in the trial. For example, when compared to the placebo group, the succimer group has an additional 11.406 μg/dL decrease in mean blood lead levels from baseline to week 1. The apparent rebound in blood lead levels after week 1 in the succimer group is thought to be due to lead that is stored in the bones being mobilized, resulting in a new equilibrium in blood lead levels in the children treated with succimer.

We remind the reader that our earlier warning about testing main effects in the presence of interactions applies also to the results in Table 5.5. For example, the test of the main effect of group in Table 5.5 ($Z = 0.27$) does not compare the average (over occasions) response in the succimer and placebo groups; instead, it compares the mean response at baseline (here the reference level for time) in the two treatment groups. The lack of equivalence between the tests for the main effect of group in Tables 5.4 and 5.5 is a direct consequence of the reference group parameterization adopted here (and commonly used by many statistical software packages, for example, PROC MIXED in SAS).

Finally, because the TLC data are from a randomized trial, the mean response at baseline is independent of treatment assignment (as was confirmed by the non-significant test of the main effect of group in Table 5.5). Because of the random assignment to treatment groups, this suggests that the analysis of response profiles could be simplified by fitting a model that omits the main effect of group, thereby forcing the two groups to have the same mean response at baseline. In Sections 5.6 and 5.7 we consider the merits of such an adjustment and compare and contrast alternative strategies for handling the baseline response in different settings. In these sections, we highlight how the analysis of response profiles is a flexible method that can easily be adapted to account for the design of the study and to address questions that are scientifically relevant to any particular study.

5.5 ONE-DEGREE-OF-FREEDOM TESTS FOR GROUP BY TIME INTERACTION

As we saw in the previous section, the test for group $\times$ time interaction is quite general. It posits no specific pattern for the difference in the response profiles between groups. This lack of specificity becomes a problem in studies with a large number of occasions of measurement because the general test for group $\times$ time interaction, with $(G-1) \times (n-1)$ degrees of freedom, becomes less sensitive to an interaction with a specific pattern as n increases. Even with as few as three or four occasions of measurement, the general test for interaction will not be as sensitive to specific departures from parallelism as the more focused tests we discuss in this section.

In the typical randomized trial of interventions, subjects are randomized to the intervention groups at baseline and the investigator seeks to determine whether the pattern of response after randomization differs between groups. Randomization implies that the mean at baseline is independent of treatment group, that is, by design, the

groups have the same mean response at baseline. In that setting, analysts frequently specify a single contrast believed to best represent the direction in which the pattern of response will differ most markedly. For example, if we assume the first parameterization described in Section 5.3 with two groups and wish to test for equality of the difference between the average response at occasions 2 through n and the baseline value in the two groups, we can choose the contrast

$$L = (-L_1, L_1),$$

where

$$L_1 = \left(-1, \frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right).$$

Here, L_1 computes the mean response from occasions 2 through n and subtracts the mean response at baseline for a single group. The latter can be thought of as the average change over the interval for a single group. Thus L is a group contrast of this average change in the two groups.

A variant of this approach, known as *Area Under the Curve Minus Baseline*, or sometimes simply AUC, corresponds to a calculation of the area under the trapezoidal curve created by connecting the responses plotted at the respective time points and subtracting $y_1 \times (t_n - t_1)$, the area of the rectangle of height y_1 and width $t_n - t_1$. For illustrative purposes, the AUC of the profile of blood lead levels for a single subject in the TLC trial is shown in Figure 5.3. For this participant, as for most participants in the TLC trial, the AUC is negative because the responses after intervention begins are smaller than the baseline value. The AUC (minus baseline) can be constructed by subtracting the baseline mean, μ_1, from each of the means, μ_1 through μ_n, and calculating the area under the trapezoid constructed by connecting these differences. To test for the equality of the AUC in two groups, one employs the contrast

$$L = (-L_2, L_2),$$

where

$$L_2 = \frac{1}{2} \times (t_1 + t_2 - 2\,t_n,\ t_3 - t_1, \ldots, t_{j+1} - t_{j-1}, \ldots, t_n - t_{n-1})$$

and $\frac{1}{2} \times (t_{j+1} - t_{j-1})$ is the value of the contrast vector for time points other than 1 (baseline) or n (the last occasion). These contrast weights are not intuitively obvious, but can be derived from the formula for the area of a trapezoid. Although the curve presented in Figure 5.3 suggests that L is applied to the individual observations, we must emphasize that the contrast weights are applied to the estimated means, not the individual observations. As a result, the AUC can be estimated in setting where some subjects have missing response data.

A third popular method for constructing a single-degree-of-freedom test corresponds to a test of the hypothesis that the trend over time is the same in the several treatment groups. Because this method is a special case of growth curve analysis, to be discussed in depth in the next chapter, and because the expected pattern of response to chelation therapy in the TLC trial would not predict a linear trend in blood lead levels during the treatment period in the group receiving succimer therapy, we defer a discussion of this approach to Chapter 6.

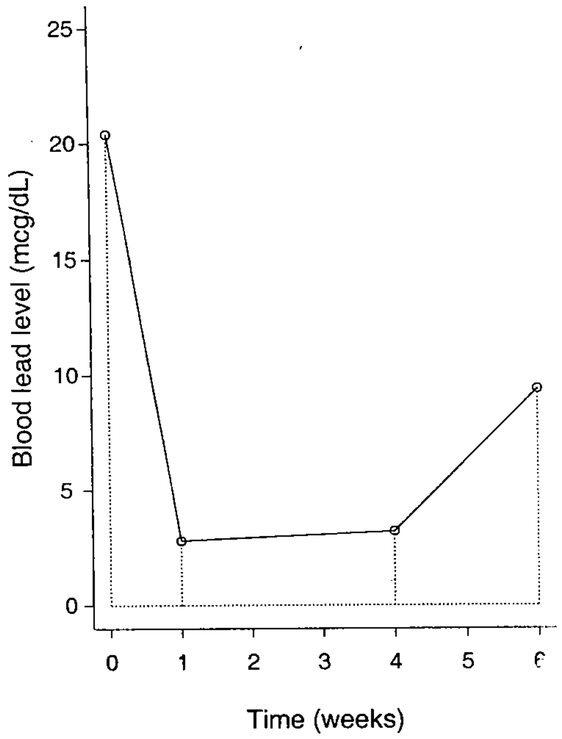

Fig. 5.3 Area under the curve, calculated using the trapezoidal rule, for the profile of blood lead levels for a single subject in the TLC trial.

Application to the Treatment of Lead-Exposed Children Trial

Since the TLC trial measured blood lead levels at four time points during the first treatment period, the vector representing the contrast based on the mean response at times 2 through n minus baseline is given by

$$L = (-L_1, L_1) = \left(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -1, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

For the data displayed in Figure 5.3, the change in mean response relative to baseline is -5.0. Because blood lead levels declined for this subject, as for most participants in the TLC trial, the mean response relative to baseline is negative. From the descriptive statistics in Table 5.6 we can easily determine that the average value of the mean response minus baseline is -9.90 in the succimer group and -2.17 in the placebo group. Thus, if we assume the first parameterization described in Section 5.3, then $L\widehat{\beta} = 7.73$ and the value of the Wald test statistic is $Z = 8.21$ (or $W^2 = 67.4$, with

Table 5.6 Mean blood lead levels (and standard deviation) at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.5	15.5	20.8
	(5.0)	(7.7)	(7.8)	(9.2)
Placebo	26.3	24.7	24.1	23.6
	(5.0)	(5.5)	(5.8)	(5.6)

one degree of freedom), indicating a highly significant difference in the response pattern between treatment groups.

Similarly, because the time points in the TLC trial were 0, 1, 4, and 6 weeks, the contrast for comparing the AUC (minus baseline) in the two treatment groups is given by

$$L = (-L_2, L_2) = (5.5, -2, -2.5, -1, -5.5, 2, 2.5, 1).$$

The area under the curve for the single subject shown in Figure 5.3 is -89.2. From Table 5.6, the estimated mean AUC is -59.20 in the succimer group and -11.40 in the placebo group. Thus, if we assume the first parameterization described in Section 5.3, then $L\widehat{\beta} = 47.8$, yielding a Wald statistic of $Z = 8.97$ (or $W^2 = 80.5$, with one degree of freedom), again highly statistically significant. Thus both methods of analysis provide a clear signal that the response profile differs in the two treatment groups.

Because the TLC trial data provide unequivocal evidence of an effect of succimer on blood lead level, the added sensitivity to treatment effects achieved by the greater specificity of a one-degree-of-freedom test is not important in this application. In many applications, however, the one-degree-of-freedom test will be statistically significant when the overall test for group × time interaction is not. For valid application of conventional significance levels, however, the form of the contrast must be specified prior to data analysis. Otherwise, one would be at risk of seeking the best contrast and testing its significance as if it had been chosen in advance. To guard against this criticism, the protocols for randomized trials usually specify the form of the contrast. This requirement highlights a hazard of one-degree-of-freedom tests. The added sensitivity comes at the price of reduced generality. If the difference between treatment groups takes a form quite different from the pattern anticipated by the contrast, one can fail to obtain a statistically significant result for a one-degree-of-freedom test even when the overall test for group × time interaction is statistically significant. Thus one-degree-of-freedom tests should be employed only when there is sufficient prior information to specify the contrast with confidence.

5.6 ADJUSTMENT FOR BASELINE RESPONSE

When the data are complete, each of the one-degree-of-freedom tests described in the previous section can be constructed by calculating a univariate summary statistic for each study participant and performing a test for equality of means of these summary statistics in the G groups. With complete data, group comparisons of these summary statistics are equivalent to applying the corresponding contrast weights to the mean responses. This is because the difference in the means is the mean of the differences when each subject is measured at every occasion. Moreover, for each of the two tests described in detail, mean response minus baseline and AUC minus baseline, the summary statistic corresponds to subtracting the baseline value from a summary of the responses on occasions 2 through n. For example, for the test for equality of mean response minus baseline, the summary statistic for the i^{th} participant is given by

$$\frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1} - Y_{i1}. \tag{5.1}$$

With this representation in mind, some analysts have suggested an alternative approach analogous to analysis of covariance (ANCOVA), in which a summary of the response at times 2 through n becomes the dependent variable and the baseline value enters the analysis as a covariate. When the response variable is the mean at occasions 2 through n and we wish to test for the equality of the mean in two treatment groups, we can write the corresponding univariate model as

$$Y_i^* = \beta_1 + \beta_2 Y_{i1} + \beta_3 \, \mathtt{trt}_i + e_i^*, \tag{5.2}$$

where

$$Y_i^* = \frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1}$$

is the mean response at occasions 2 through n for the i^{th} subject, $\mathtt{trt_i}$ is an indicator variable distinguishing the two treatment groups, and e_i^* is the error term in the univariate model. This model assumes that the data are complete and it cannot be fit with missing data; we defer a discussion of more general approaches for handling baseline response to Section 5.7.

An analysis based on either (5.1) or (5.2) will be especially appealing in settings where initial changes from baseline are expected to persist throughout the duration of follow up. For example, in a trial where the impact of the intervention on changes in the mean response at the start of follow up is expected to be similar to that toward the end of follow up; this pattern for the mean response profiles is illustrated in Figure 5.4. Tests based on (5.1) or (5.2) correspond to a comparison between groups of the mean responses on occasions 2 through n, with adjustment for baseline, and have $G-1$ degrees of freedom irrespective of the number of occasions of measurement.

This raises a question about whether one should incorporate the baseline value through the contrast given by (5.1) or through the analysis of covariance model given by (5.2) in a specific application. The answer depends critically on whether the data arose from an observational study or a randomized trial. If the study is an

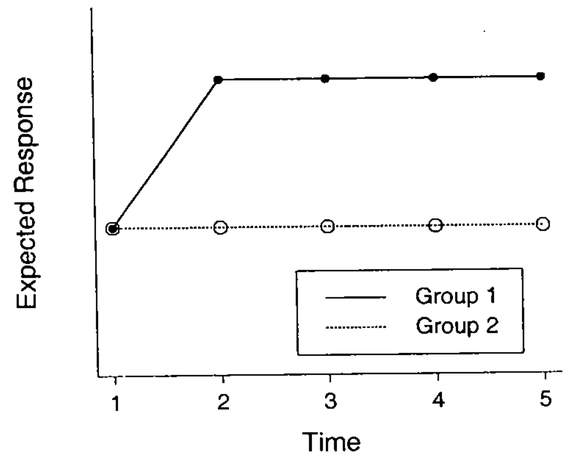

Fig. 5.4 Graphical representation of changes in the mean response from baseline (in Group 1) that persist throughout the duration of follow up.

observational one, for example, a longitudinal study of the determinants of rate of decline of pulmonary function in adults, it is usually not advisable to employ the analysis of covariance approach because the baseline value may be associated with other variables whose effects are to be studied, raising problems of confounding in an analysis intended to describe how the pattern of response over time is influenced by the characteristics of study participants. For example, individuals who are smokers as adults may have smoked during adolescence. If smoking affected the attained pulmonary function level for young adults, then smoking will likely be associated with pulmonary function level later in adult life, even if cigarette smoking does not influence the rate of decline of pulmonary function with age. Thus, adjustment for baseline pulmonary function level using (5.2) could introduce an association between smoking status and rate of decline of pulmonary function, even if the unadjusted rates of decline are nearly equivalent in the various smoking groups.

When participants have been randomized to the several treatment groups and the baseline value has been obtained before any study interventions, adjustment for baseline through analysis of covariance is of interest. In that setting, the mean response at time 1 is independent of treatment assignment. One can then show the one-degree-of-freedom test for equality of response profiles based on a contrast and the corresponding test based on analysis of covariance represent alternative tests of the same null hypothesis and that the test based on the analysis of covariance approach will always be more efficient. That is, the analysis of covariance approach yields estimates of treatment effects with smaller standard errors than those obtained by calculating contrasts.

For example, the greater efficiency of the analysis of covariance can be highlighted by examining the relative efficiency of (5.1) to (5.2) in simple settings. The relative efficiency is defined as the ratio of the variance of the estimator based on (5.2) to the variance of the estimator based on (5.1); a relative efficiency less than one implies that the estimator based on (5.2) has smaller variance. When the covariance among the repeated measures is assumed to have a compound symmetry pattern, with common variance σ^2 and common correlation ρ, the relative efficiency is given by

$$\frac{1}{n}\{1 + (n-1)\rho\}. \tag{5.3}$$

The derivation of (5.3) is not important. What this simple expression indicates is that the two methods of adjustment for baseline response are equally efficient only when $\rho = 1$. When $\rho = 0$, the analysis based on (5.1) is only $\frac{1}{n}$ times as efficient as the analysis of covariance. The greater efficiency of the analysis of covariance depends on both the number of repeated measures and the strength of the correlation among them. For example, when $n = 5$ and $\rho = 0.4$ the analysis of covariance is approximately twice as efficient as subtracting the baseline response.

In general, the analysis of longitudinal data from a randomized trial is the only setting where we recommend adjustment for baseline through analysis of covariance. In that setting, in contrast to observational studies, adjustment leads to meaningful tests of hypothesis of scientific interest. Moreover, the tests based on the analysis of covariance approach will be more powerful. The notion of adjustment for baseline can also be applied more generally in the analysis of response profiles; in Section 5.7 we compare and contrast a number of alternative strategies for handling the baseline response in more general settings and make recommendations about the preferred strategies in different situations.

We conclude this section by noting that adjustment for baseline in the analysis of longitudinal change is a topic that has generated heated debate among analysts. When longitudinal data arise from an observational study, the two methods of adjusting for baseline described in this section can yield discernibly different and, apparently conflicting, results. This conundrum is also known as *Lord's paradox* (named after Frederic Lord, who eloquently brought the issue to light) and has led many researchers astray over the years. The paradox lies in the interpretation of the two types of analyses and is resolved by noting that these two alternative methods of adjusting for baseline answer qualitatively different scientific questions when the data arise from an observational study. This can be illustrated in the simplest setting where there are two groups or sub-populations (e.g., males and females) measured at two occasions. The overall goal of such a study is to compare the changes in response for the two groups. The analysis that subtracts baseline response, thereby creating a simple change score, addresses the question of whether the two groups differ in terms of their mean change over time. In contrast, adjustment for baseline using analysis of covariance addresses the question of whether an individual belonging to one group is expected to change more (or less) than an individual belonging to the other group, *given that they have the same baseline response*. The latter question is a conditional one and, depending on the study design, may address a different scientific question than the former.

For example, in an observational study examining gender difference in weight gain of infants between the ages of 12 and 24 months, a measure of body weight might be obtained at 12 months (baseline) and at 24 months. The analysis of the simple change score addresses the question of whether boys and girls differ in terms of their changes in mean body weight over the 12 months of follow up. At baseline, boys are on average $1\frac{1}{2}$ pounds heavier than girls, but there is no evidence of a gender effect on the 12 month changes in body weight, with boys and girls both gaining approximately $5\frac{1}{4}$ pounds. In contrast, the analysis of covariance of the same data reveals a discernible gender effect, with boys showing more weight gain than girls. Thus, even though the unadjusted (or *unconditional*) increases in body weight are approximately the same for this age cohort of boys and girls, the analysis of covariance is directed at the *conditional* question of whether boys are expected to gain more weight than girls, *given that they have the same initial weight at 12 months*. That is, if we compare boys and girls within sub-populations with the same initial weight at 12 months, are their average weights at 24 months the same. When the conditional question is posed in this way we would expect boys to gain more weight than girls. The reasoning is as follows: If a boy and girl have the same initial weight at 12 months then there are two possibilities: (i) the girl is initially overweight and is expected to gain less weight over the 12 months, or (ii) the boy is initially underweight and is expected to gain more.

A more thorough discussion of this issue is beyond the scope of this book, but we advise readers to employ the analysis of covariance approach in longitudinal settings only if the approach and its implications are fully understood.

In summary, the choice between the two methods of adjusting for baseline discussed in this section should be made on substantive grounds. That is, the design of the longitudinal study and the research question of interest should guide the choice of analytic method. The analysis that subtracts baseline response is appropriate when the primary goal of the study is to compare distinct populations in terms of their average change over time. The analysis addresses the question: "Do the populations differ in terms of their average change?" and is appropriate when the data have arisen from either an observational study or a randomized trial. On the other hand, analysis of covariance will, in general, be appropriate only in cases where individuals have been assigned to groups at random (e.g., a randomized trial) or where the population distributions of the baseline responses can reasonably be assumed to be equal (even though the sample means of the baseline responses may differ across groups). In cases where the population distributions of the baseline responses are equal, it is then meaningful to ask the question: "Is the expected change the same in all groups, when we compare individuals having the same baseline response?". Furthermore, the analysis of covariance will provide a more powerful test of group differences. The latter has often been touted as the main reason why analysis of covariance should be the preferred method of adjusting for baseline. This faulty rationale, however, has blinded many researchers to the potential difficulties in interpreting the results of analysis of covariance when the assumption of equal population distributions of baseline response is not tenable. In conclusion, it is the study design and the scientific question of interest,

and not issues of statistical precision and power, that should primarily determine the choice of analytic methods for adjusting for baseline response.

5.7 ALTERNATIVE METHODS OF ADJUSTING FOR BASELINE RESPONSE*

One feature of longitudinal studies that sets them apart from repeated measures and related designs is the presence of a baseline measurement. In a randomized longitudinal trial comparing treatments, the measurement at the first occasion is usually a baseline response obtained prior to any study interventions. For example, in the TLC trial, the blood lead levels at baseline were obtained prior to receiving placebo or succimer. In that case, due to randomization, we can assume that the treatment group means are equal at baseline. Thus the question naturally arises as to how to handle the baseline measurement in the assessment of whether patterns of change in the mean response over time are the same in the groups.

In the previous section we considered two methods of adjustment for baseline in a relatively simple setting. The notion of adjustment for baseline can also be applied more generally in the analysis of response profiles. In this section† we compare and contrast a number of alternative strategies for handling the baseline response and make recommendations about the preferred strategies in different situations.

We consider four ways of handling the baseline value:

1. We can retain it as part of the outcome vector and make no assumptions about group differences in the mean response at baseline.
2. We can retain it as part of the outcome vector and assume the group means are equal at baseline, as might be appropriate in a randomized trial.
3. We can subtract the baseline response from all of the remaining post-baseline responses, and analyze the differences from baseline.
4. We can use the baseline value as a covariate in the analysis of the post-baseline responses.

We now consider the appropriateness and merits of each of these four strategies and illustrate their application to the blood lead level data from the TLC trial.

The first two strategies retain the baseline measurement as part of the outcome vector, but differ in terms of assumptions about the mean response at baseline. The first strategy corresponds to a standard analysis of response profiles without incorporating any constraints on the group means at baseline. This was the method of analysis highlighted in Sections 5.2–5.4. The second strategy corresponds to an analysis of response profiles where the group means at baseline are constrained to be equal. In a

†*Note*: Readers may find the level of detail in this section challenging; this section can be omitted at first reading without loss of continuity.

Table 5.7 Estimated regression coefficients and standard errors based on an analysis of response profiles of the blood lead level data assuming equal mean blood lead levels at baseline in the succimer and placebo groups.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.406	0.500	52.83
Week		1	−1.645	0.782	−2.10
Week		4	−2.231	0.807	−2.76
Week		6	−2.642	0.887	−2.98
Group × Week	S	1	−11.341	1.093	−10.38
Group × Week	S	4	−8.765	1.131	−7.75
Group × Week	S	6	−3.120	1.251	−2.49

randomized trial, where treatment assignment is random, both strategies yield valid estimates of treatment group comparisons, but the second strategy is in general more powerful. Thus, in randomized trials, or in observational studies where there is good reason to assume the groups have the same mean response at baseline (e.g., due to matching on baseline response), the second strategy for handling baseline is preferred and should be routinely used. In contrast, in observational studies where there is no *a priori* reason to assume the groups have the same mean response at baseline, the second strategy in not appropriate and only the first strategy should be used.

In the analyses of the blood lead level data from the TLC trial presented in Section 5.4, the first strategy was employed in the results presented in Tables 5.4 and 5.5. The second strategy can be implemented by excluding the treatment group main effect from the model for the response profiles. This model is unusual in that it contains an interaction between group and time, but no main effect of group. This model appears to contradict the conventional wisdom that interactions should not be included in a regression model without their main effects. However, this is an important exception to the rule. Because baseline (week 0) was chosen as the reference level for time, the exclusion of the group main effect forces the two groups to have the same mean response at baseline. The results of such an analysis are presented in Table 5.7 and are qualitatively similar to those in Table 5.5. Note the absence of the main effect of group, which has been set to zero. Also, the omnibus test of the group × time interaction from this model yields a Wald statistic of 111.96 with 3 degrees of freedom. In contrast, the analysis of response profiles without any adjustment for baseline (strategy 1) yielded a Wald statistic of 107.79, with 3 degrees of freedom (see Table 5.4). The difference between these two statistics reflects the increased power of the second method for handling baseline response.

The third and fourth strategies do not retain the baseline response as part of the outcome vector. Instead, they focus on raw and adjusted changes from baseline and restrict the outcome vector to measurements obtained post-baseline. The third strategy is to subtract the baseline response from the remaining post-baseline responses, and analyze the differences from baseline. We refer to these differences from baseline as "raw change scores". With responses at n occasions, we can define the $(n-1) \times 1$ vector of raw change scores,

$$D_i = (Y_{i2} - Y_{i1}, Y_{i3} - Y_{i1}, \ldots, Y_{in} - Y_{i1})',$$

and conduct an analysis of response profiles with D_i as the outcome vector. Because the outcome is a change score (the change from baseline), this approach alters the interpretation of the tests for all three effects in the analysis of response profiles. The test for group × time interaction becomes a test for parallel profiles for the changes from baseline in the mean response on occasions 2 through n; the test for group effect becomes a test that the changes from baseline at occasion 2 are the same across groups (assuming that occasion 2 is chosen as the reference level for time). Thus, to address the question of whether patterns of change over time are the same in all groups, the test of interest under the third strategy must be modified. It is now a joint test which combines the main effect of group and the group × time interaction. The test has the same $(G-1) \times (n-1)$ degrees of freedom because it combines a $(G-1)$ degrees of freedom test for group with a $(G-1) \times (n-2)$ degrees of freedom test for group × time interaction. This is in contrast to the conventional analysis of response profiles (with baseline response included as part of the response vector), where only the group × time interaction addresses important questions concerning group comparisons of the patterns of changes in the mean response and the group main effect is not of scientific interest.

Of note, this joint test of the main effect of group and the group × time interaction is formally equivalent to the test of the group × time interaction (with $(G-1) \times (n-1)$ degrees of freedom) under the first strategy. Moreover, for the purposes of analyzing changes in the mean response and how these changes differ among groups, the first and third strategies are completely equivalent, that is, the first and third strategies for handling baseline produce identical tests and estimates of effects. Thus it is clear that the third strategy offers no efficiency gain.

Using the blood lead level data from the TLC trial, the results of the analysis of change scores are presented in Table 5.8. At first glance, the regression parameter estimates in Table 5.8 do not appear to agree with those in Table 5.5. However, it can easily be shown that the 6 parameter estimates in Table 5.8 are simple linear combinations of the estimates for the time and group × time interaction effects reported in Table 5.5. For example, the group effect, −11.406, in Table 5.8 is identical to the estimate of the group × time interaction in Table 5.5 that represents the group contrast of the changes from baseline (week 0) to week 1. Similarly, the estimated group contrast of the changes from baseline to week 4, −8.824, from Table 5.5 can also be obtained from the estimates in Table 5.8 (−11.406 + 2.582 = −8.824). It can be shown that all estimates of change from baseline, and the group comparisons of these changes, reported in Tables 5.5 and 5.8 are identical.

Table 5.8 Estimated regression coefficients and standard errors based on an analysis of response profiles of the changes from baseline in blood lead levels at week 1, week 4, and week 6 for the children from the TLC trial.

Variable	Group	Week	Estimate	SE	Z
Intercept			−1.612	0.792	−2.04
Group	S		−11.406	1.120	−10.18
Week		4	−0.590	0.643	−0.92
Week		6	−1.014	0.934	−1.09
Group × Week	S	4	2.582	0.909	2.84
Group × Week	S	6	8.254	1.321	6.25

The analysis of change scores reported in Table 5.8 produced a Wald statistic of 107.79, with 3 degrees of freedom, for jointly testing the effect of group and the group × time interaction. As expected, this agrees with the omnibus test of the group × time interaction reported in Table 5.4 from the analysis of response profiles without any adjustment for baseline (strategy 1).

Because the first and third strategies yield identical analyses of changes in the mean response, and how these changes differ among groups, in principle either method can be used. However, from a practical standpoint, we recommend the first strategy over the third for two main reasons. First, the analysis of change scores has implications for the interpretation of the hypothesis tests that are more consequential. For example, while the test of primary interest in the conventional analysis of response profiles is usually the test for group × time interaction (with $(G-1) \times (n-1)$ degrees of freedom), the test of interest in the analysis of change score is a $(G-1) \times (n-1)$ degrees of freedom test that incorporates the main effect of group (with $(G-1)$ degrees of freedom) and the $(G-1) \times (n-2)$ degrees of freedom group × time interaction in this model. The analysis of change scores requires the construction of joint tests of main effects and interactions; these tests are not routinely produced as standard output from statistical software for analyzing response profiles. Second, when there are subjects with missing baseline response, all of their data are excluded from the analysis of change scores; in contrast, the first strategy incorporates all available data in the analysis.

The fourth strategy is to analyze the post-baseline responses and make an adjustment for the baseline response by including it as a covariate. If we have responses at n occasions, the analysis of response profiles is based on the $(n-1) \times 1$ vector,

$$Y_i = (Y_{i2}, Y_{i3}, \ldots, Y_{in})',$$

and the baseline response, Y_{i1}, is regarded as a covariate. This type of analysis corresponds to an analysis of covariance (ANCOVA), albeit one where the outcome

is a $(n-1) \times 1$ vector of responses. This strategy for handling baseline is appropriate when analyzing data from randomized trials. Note that, because of randomization, hypotheses of equality of the *conditional* means of the response at occasions 2 through n, given the baseline response, imply hypotheses of equality of the *unconditional* means of the response at occasions 2 through n. This strategy for handling baseline is appropriate also for observational studies where there is good reason to assume the groups have the same mean response at baseline. It should not be used, however, in observational studies where there is no *a priori* reason to assume the groups have the same mean response at baseline.

Interestingly, the fourth strategy for handling baseline can be implemented by conducting the analysis of response profiles (with Y_{i1} included as a covariate) on either the post-baseline responses or the post-baseline change scores. That is, the estimates of all effects of interest are identical whether the analysis is based on the $(n-1) \times 1$ vector of post-baseline response, $Y_i = (Y_{i2}, Y_{i3}, \ldots, Y_{in})'$, or the $(n-1) \times 1$ vector of post-baseline differences, $D_i = (Y_{i2} - Y_{i1}, Y_{i3} - Y_{i1}, \ldots, Y_{in} - Y_{i1})'$. The intuition for why these two analyses are identical is as follows: Because the two outcomes differ by Y_{i1}, and both analyses estimate effects that are adjusted for Y_{i1} by holding the baseline value fixed, they produce the same regression coefficients for all effects of interest. The two analyses differ only in terms of the estimated slope for Y_{i1}. (The estimated slope from the analysis based on Y_i is simply one unit larger than the estimated slope from the analysis based on D_i.) Because of this equivalence, we can regard the fourth strategy as an analysis of the "adjusted change scores" (i.e., D_i adjusted for Y_{i1}) in contrast to an analysis of the raw or unadjusted change scores (strategy 3).

Because the outcome is an adjusted change score, the fourth strategy for handling baseline also alters the interpretation of the tests for all three effects in the analysis of response profiles. The test for group × time interaction becomes a test for parallel profiles for the adjusted changes from baseline in the mean response on occasions 2 through n; the test for group effect becomes a test that the adjusted changes from baseline in the mean response at occasion 2 are the same across groups (assuming that occasion 2 is chosen as the reference level for time). Thus, similar to the third strategy, the test of interest is a joint test of the main effect of group and the group × time interaction.

Using the blood lead level data from the TLC trial, the results of the analysis of adjusted change scores are presented in Table 5.9. Although the results of this analysis are qualitatively similar to the results from the analysis of raw change scores (strategy 3), note that the estimated main effect of group (and the intercept) in Table 5.9 is slightly different from the corresponding estimate in Table 5.8.

The analysis of adjusted change scores reported in Table 5.9 produced a Wald statistic of 111.13, with 3 degrees of freedom, for jointly testing the effect of group and the group × time interaction. This is larger than the corresponding statistic under the third strategy, reflecting the greater efficiency of the analysis of adjusted change scores. The greater efficiency of analysis of covariance (adjusted change score analysis) over simple contrasts (raw change score analysis) was highlighted in Section 5.6.

Table 5.9 Estimated regression coefficients and standard errors based on an analysis of response profiles of the adjusted changes from baseline in blood lead levels at week 1, week 4, and week 6 for the children from the TLC trial.

Variable	Group	Week	Estimate	SE	Z
Intercept			−1.638	0.777	−2.11
Baseline† $(Y_{i1} - 26.406)$			−0.196	0.094	−2.08
Group	S		−11.354	1.099	−10.34
Week		4	−0.590	0.643	−0.92
Week		6	−1.014	0.934	−1.09
Group × Week	S	4	2.582	0.909	2.84
Group × Week	S	6	8.254	1.321	6.25

†Centering baseline response on its overall mean (26.406) gives the intercept a meaningful interpretation.

Observe that the Wald statistic of 111.13 produced by the analysis of adjusted change scores is quite similar to that obtained under the second strategy for adjusting for baseline. Thus the question naturally arises as to which approach is preferred: strategy 2 or strategy 4. One could argue that strategy 2 is preferred over strategy 4 for exactly the same reasons given for preferring strategy 1 over strategy 3. That is, the analysis of adjusted change scores requires the construction of joint tests of main effects and interactions and these tests are not routinely produced as standard output from statistical software for analyzing response profiles. Second, when there are subjects with missing baseline response, all of their data are excluded from the analysis of adjusted change scores; in contrast, the second strategy incorporates all available data in the analysis. Finally, there is a third reason why strategy 2 might be preferred over strategy 4. There is an implicit assumption in the adjusted change score analysis that the regression slope relating Y_{ij} to Y_{i1} (for $j = 2, \ldots n$) is the same at all $n-1$ post-baseline occasions. This implies a strong assumption about the covariances among $Y_{i1}, \ldots, Y_{in}$. Specifically, it constrains

$$\text{Cov}(Y_{i1}, Y_{i2}) = \text{Cov}(Y_{i1}, Y_{i3}) = \cdots = \text{Cov}(Y_{i1}, Y_{in}).$$

As a result of these constraints, there is the potential for misspecification of the model for the covariance and misleading inferences about change over time. In contrast, the second strategy imposes no such structure on the covariance matrix. Finally, we note that if the assumption of homogeneous regression slopes (relating Y_{ij} to Y_{i1}, for $j = 2, \ldots, n$) is relaxed and $n-1$ separate regression slopes are estimated, then strategy 4 and strategy 2 are completely equivalent and produce identical tests and estimates of effects. Thus, the second strategy for handling baseline can be seen to enjoy all of the efficiency gains that have been highlighted in Section 5.6 for

ANCOVA (adjusted change score analysis). However, on practical grounds, for the reasons outlined above, it can be argued that strategy 2 is preferred over strategy 4.

To summarize, there are many ways to handle baseline response in the analysis of longitudinal data. In this section we have reviewed four strategies. We have seen that the two methods that retain the baseline value as part of the outcome are completely equivalent to corresponding strategies that restrict the outcome vector to measurements obtained post-baseline. However, on practical grounds, it can be argued that the first and second strategies are preferable. The first and second strategies differ in terms of efficiency. The second strategy enjoys all the efficiency gains that have been highlighted in Section 5.6 for ANCOVA. But the choice between the first and second strategies should be guided by the study design. In randomized trials, or in observational studies where there is good reas n to assume the groups have the same mean response at baseline, the second strategy is in general more powerful and should be routinely used. In contrast, in observational studies where there is no *a priori* reason to assume the groups have the same mean response at baseline, the second strategy in not appropriate and the first should be used.

5.8 STRENGTHS AND WEAKNESSES OF ANALYZING RESPONSE PROFILES

The analysis of response profiles is a conceptually straightforward way to analyze data from a longitudinal study when the design is balanced, with the timing of the repeated measures common to all individuals in the study, and when all the covariates are discrete (e.g., representing different treatments, interventions, or characteristics of the study subjects). The main feature of the analysis of response profiles is that it allows arbitrary patterns in the mean response over time and arbitrary patterns in the covariance of the responses. As a result, this method for longitudinal analysis has a certain robustness since the potential risks of bias due to misspecification of the models for the mean and covariance are minimal. Although the analysis of response profiles requires that the data arise from a balanced design, it can be applied when the data are incomplete due to missing response data.

The method for analyzing response profiles described in this chapter is related to a more traditional approach known in the statistical literature as "profile analysis". However, we make a distinction between the method presented in this chapter and a traditional profile analysis. In a traditional profile analysis, the three hypotheses concerning response profiles described at the beginning of Section 5.2 are placed on an equal footing and there is an overwhelming emphasis on hypothesis testing rather than estimation of effects. As we have seen, however, tests of hypotheses concerning main effects of time and/or group often have no direct bearing on questions of scientific interest in a longitudinal study. This is especially the case for longitudinal data arising from randomized trials. Consequently, the routine use of traditional profile analysis for longitudinal data coerces the analyst to test certain hypotheses that do not necessarily translate into meaningful scientific questions about longitudinal change

in the response. In addition, because profile analysis is often implemented within a multivariate analysis of variance (MANOVA) that requires a complete response vector on each individual, it does not permit subjects with missing responses. The resulting analysis is very inefficient because it is based only on data from the so-called complete cases; it can also produce biased estimates of change in the mean response over time when the so-called "completers" are not a random sample from the target population. Finally, traditional profile analysis lacks flexibility in handling the baseline response and requires that it be part of the response vector. In contrast, the method for analyzing response profiles presented in this chapter can be readily adapted to address specific questions that are well grounded in the science, can be applied when the data are incomplete due to missing response data, and permits alternative approaches for making adjustments for the baseline response.

Although it was not considered here, the analysis of response profiles can be extended in a straightforward way to handle the case where individuals can be grouped according to more than a single factor. For example, if there are two covariates that are discrete (e.g., treatment group and gender), the analysis will include tests of the 3-way and 2-way interactions among these two factors and time (in addition to their main effects). The general linear model can also be used to provide estimated means for summary measure analyses that are based on linear combinations of the mean response vector, for instance, "area under the curve" analysis, when the data are incomplete.

The analysis of response profiles does have a number of potential drawbacks that make it either unappealing or unsuitable for analyzing data from many longitudinal studies. First, the requirement that the longitudinal design be balanced implies that the method cannot be applied when the vectors of repeated measures are obtained at different sequences of time except by "moving" an observation to the nearest planned measurement time. As a result, the method is not well suited to handle mistimed measurements, a common problem in many longitudinal studies. Note, however, that the general method for analyzing response profiles can handle unbalanced patterns of observations due to missing response data. Second, the analysis of response profiles ignores the time ordering of the repeated measures in a longitudinal study. Indeed, the analysis of response profiles could be applied when each individual has a vector of multivariate outcomes that are distinct and non-commensurate (i.e., measures of more than one outcome) rather than repeated measures of a single outcome. Because the analysis of longitudinal response profiles allows for an arbitrary pattern in the mean responses, and does not impose any time trends, the results of the analysis provide only a very broad or general statement about group differences in patterns of change over time. Ordinarily, a significant group $\times$ time interaction effect that has more than a single degree of freedom will require additional analysis to provide a more informative description of how the groups differ in their patterns of change in the mean response. Third, because the analysis of response profiles produces an overall or omnibus test of effects, it may have low power to detect group differences in specific trends in the mean response over time (e.g., linear trends in the mean response). Single-degree-of-freedom tests of specific time trends are more powerful. Finally, in the analysis of response profiles, the number of estimated parameters

($G \times n$ mean parameters and $\frac{n(n+1)}{2}$ covariance parameters) grows rapidly with the number of measurement occasions. For example, with two groups measured at three occasions, the number of parameters is 12. However, with two groups measured at 10 occasions, the number of parameters is 75. Consequently, this method is more appealing when the total number of subjects, N, is relatively large in comparison to the number of measurement occasions, n.

5.9 COMPUTING: ANALYZING RESPONSE PROFILES USING PROC MIXED IN SAS

The MIXED procedure in SAS is a very general and versatile procedure for fitting linear models to longitudinal and clustered data. No attempt is made here to give a comprehensive review of the main features of PROC MIXED. Instead, we present illustrative source code for an analysis of response profiles in general terms and then describe the most salient parts of the command syntax. Many of the later chapters will include a description of additional commands and features of PROC MIXED as they are needed. Although these concluding sections in each chapter will not provide a training manual for the use of PROC MIXED, they should provide a firm basis for understanding the command syntax required for analyzing longitudinal data using PROC MIXED in SAS.

Before discussing the command syntax for PROX MIXED, we note that the procedure requires each repeated measurement in a longitudinal data set to be a separate "record". For example, in the TLC trial, the data are recorded as follows:

(ID	Group	Baseline	Week 1	Week 4	Week 6)
001	P	30.8	26.9	25.8	23.8
002	S	26.5	14.8	19.5	21.0
003	S	25.8	23.0	19.1	23.2
004	P	24.7	24.5	22.0	22.5
005	S	20.4	2.8	3.2	9.4
006	S	20.4	5.4	4.5	11.9
⋮	⋮	⋮	⋮	⋮	⋮
100	P	31.1	31.2	29.2	30.1

with a single "record" of the 4 repeated measurements for each child in the study. When the data set is in this form it is said to be in a *multivariate* mode or *wide* format.

Table 5.10 Illustrative commands in SAS for transforming a data set with a single record for each individual to a data set with multiple records corresponding to each measurement occasion.

```
DATA lead;
   INFILE 'tlc.dat';
   INPUT id group $ y1 y2 y3 y4;
   y=y1; time=0; OUTPUT;
   y=y2; time=1; OUTPUT;
   y=y3; time=4; OUTPUT;
   y=y4; time=6; OUTPUT;
   DROP y1-y4;
```

Prior to analysis, these data must be converted to a data set with 4 records for each child, one for each measurement occasion. In the latter form, the data set is said to be in a *univariate* mode or *long* format. This can be accomplished using the illustrative SAS commands in Table 5.10, which produce the following data set:

(ID	Group	Time	Y)
001	P	0	30.8
001	P	1	26.9
001	P	4	25.8
001	P	6	23.8
002	S	0	26.5
002	S	1	14.8
002	S	4	19.5
002	S	6	21.0
⋮	⋮	⋮	⋮
100	P	0	31.1
100	P	1	31.2
100	P	4	29.2
100	P	6	30.1

Table 5.11 Illustrative commands for an analysis of response profiles using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS  id  group  time;
    MODEL  y=group  time  group*time / S  CHISQ;
    REPEATED  time / TYPE=UN  SUBJECT=id  R  RCORR;
```

To conduct an analysis of response profiles with data from two or more treatment groups measured repeatedly over time, we can use the illustrative SAS commands given in Table 5.11. This model assumes that the covariance matrix is unstructured. Alternative assumption about the covariance can be considered, and this may be advantageous when the number of measurement occasions is relatively large in comparison to the number of subjects. Choosing a model for the covariance matrix is a topic that will be discussed in Chapter 7. Next, we present a brief description of each of the command statements in Table 5.11.

PROC MIXED <options>;

The PROC MIXED statement calls the procedure MIXED in SAS. It can also include an option for the choice of method of estimation. By default, PROC MIXED uses REML estimation; ML estimation can be invoked by including the option METHOD=ML.

CLASS variables;

The CLASS statement is used to define all variables that are to be regarded as categorical or factors. By default, this statement will create indicator variables for each factor using a reference group coding, with the last level (where "last" here refers to the level with the largest alpha-numeric value) regarded as the reference group. Of note, this default indicator variable coding is not the most natural for a categorical variable denoting the occasions of measurement; for the latter, the "first" level of the factor (e.g., the baseline measurement occasions) is usually the natural reference group.

The default coding can be changed with the inclusion of the ORDER= option in the PROC MIXED statement. For example, the ORDER=DATA option forces the levels of all variables included in the CLASS statement to be sorted by their order of appearance in the input data set. Therefore, by previously sorting the data set in descending order of a categorical variable denoting the occasions of measurement, a more natural reference group coding is obtained for that variable (with the lowest, rather than the highest, level of the categorical variable for time used as the reference). However, one unappealing consequence of

circumventing the default coding of time in this way is that the estimates of the covariance matrix are printed in reverse (or descending) order of time. To avoid potential confusion when extracting the estimates of the covariance matrix, it is advisable to re-run the analysis without the ORDER=DATA option.

MODEL dependent = <fixed effects> / <options>;

The MODEL statement specifies the response variable and the fixed effects. The fixed effects can include both discrete (defined in the CLASS statement) and quantitative (excluded from the CLASS statement) covariates.

The covariates included in the MODEL statement determine the design matrix X_i. Of note, by default, PROC MIXED includes a column of 1's in X_i for the intercept. The option NOINT requests that no intercept be included in the model.

Various options that can be included on the MODEL statement modify how test statistics are computed and the type of output produced. The option DDFM=SATTERTH requests Satterthwaite's approximation for the denominator degrees of freedom for tests of the fixed effects. Alternatively. the option CHISQ requests that multivariate Wald tests be computed and compared to the reference chi-squared distribution. The option S (or SOLUTION) requests that the estimates of the fixed effects, and their standard errors, be displayed.

REPEATED <repeated effect> / SUBJECT = effect <options>;

The REPEATED statement is primarily used to distinguish which observations are correlated and which can be regarded as independent of one another. This is achieved with the SUBJECT option which is used to denote a variable that distinguishes clusters of correlated responses. By including a variable in the SUBJECT option (e.g., a subject identifier), pairs of observations with the same value of that variable are regarded as correlated while pairs of observations with distinct values are regarded as independent.

The REPEATED statement also includes options for specifying assumptions about the nature of the covariance among the errors. This is achieved with the TYPE=<pattern> option (e.g., TYPE=UN specifies an unstructured covariance matrix). A full listing and description of all the possible covariance patterns can be found in the SAS documentation. There are also various options that modify the type of output that is produced. The option R and RCORR print the covariance and correlation matrices, respectively.

Finally, a variable denoted the "repeated effect" can also be included on the REPEATED statement and this identifies "units within a cluster". In the context of longitudinal data, the "repeated effect" identifies the measurement occasions. While it is not always necessary to include this variable, failure to do so may have unforseen consequences when there are vectors of repeated measures of different length and/or when the vector of responses are not in the same order for all subjects. In Table 5.11. the REPEATED statement identifies "time" as the repeated effect. To avoid any potential problems, it is recommended that this variable be included in the REPEATED statement to ensure that the covariance is structured and estimated appropriately.

5.10 FURTHER READING

A useful review of traditional profile analysis, targeted at applied researchers, can be found in Chapter 3 (Section 3.4, pp. 48–52) of Hand and Taylor (1987).

A more detailed discussion of the subtle issues surrounding adjustment for baseline response in the analysis of change can be found in the articles by Lord (1967), Laird (1983), and Fitzmaurice (2001), and in Chapter 7 (Section 7.3, pp. 489–496) of Bock (1975).

Bibliographic Notes

One of the earliest descriptions of traditional profile analysis appeared in an article by Greenhouse and Geisser (1959). Profile analysis is also discussed in detail in Chapter 6 of Johnson and Wichern (2002), Chapters 4 and 5 of Morrison (1990), and Chapters 5 and 6 of Rencher (2002).

Problems

5.1 In the National Cooperative Gallstone Study (NCGS), one of the major interests was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones (Schoenfield *et al.*, 1981; Wei and Lachin, 1984). In this study, patients were randomly assigned to high-dose (750 mg per day), low-dose (375 mg per day), or placebo. We focus on a subset of data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups.

In the NCGS it was suggested that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result, serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20, and 24 months of follow-up. Many cholesterol measurements are missing because of missed visits, laboratory specimens were lost or inadequate, or patient follow-up was terminated.

The NCGS serum cholesterol data are stored in an external file: `cholesterol.dat`

Each row of the data set contains the following seven variables:

Group ID Y_1 Y_2 Y_3 Y_4 Y_5

Note: The categorical variable Group is coded 1 = High-Dose, 2 = Plecebo.

5.1.1 Read the data from the external file and keep it in a "multivariate" or "wide" format.

5.1.2 Calculate the sample means, standard deviations, and variances of the serum cholesterol levels at each occasion for each treatment group.

5.1.3 On a single graph, construct a time plot that displays the mean serum cholesterol versus time (in months) for the two treatment group. Describe the general characteristics of the time trends for the two groups.

5.1.4 Next, read the data from the external file and put the data in a "univariate" or "long" format, with 4 "records" per subject.

5.1.5 Assuming an unstructured covariance matrix, conduct an analysis of response profiles. Determine whether the patterns of change over time differ in the two treatment groups.

5.1.6 Display the estimated 5×5 covariance and correlation matrices for the five repeated measurements of serum cholesterol.

5.1.7 With baseline (month 0) and the placebo group (group 2) as the *reference group*, write out the regression model for mean serum cholesterol that corresponds to the analysis of response profiles in Problem 5.1.5.

5.1.8 Let L denote a matrix of known weights and β the vector of linear regression parameters from the model assumed in Problem 5.1.7. The null hypothesis that the patterns of change over time do not differ in the two treatment groups can be expressed as H_0: $L\beta = 0$. Describe an appropriate weight matrix L for this null hypothesis.

5.1.9 Show how the *estimated* regression coefficients from an analysis of response profiles can be used to construct the time-specific means in the two groups. Compare these estimated means with the sample means obtained in Problem 5.1.2.

5.1.10 With baseline (month 0) and the placebo group (group 1) as the *reference group*, provide an interpretation for each of the estimated regression coefficients in terms of the effect of the treatments on the patterns of change in mean serum cholesterol.

6

Modelling the Mean: Parametric Curves

6.1 INTRODUCTION

In the previous chapter we described an approach to modelling longitudinal data that effectively imposed no structure on the underlying mean response trend over time. This approach has some appeal when all subjects are measured at the same set of occasions and the number of measurement occasions is relatively small (e.g., not more than 4 or 5). But as the number of occasions increases and/or when the repeated measures are irregularly timed, analyzing response profiles becomes much less appealing. Even in cases where the number of repeated measures is relatively small, there are two obvious drawbacks of the analysis of response profiles that limit its usefulness for the analysis of longitudinal data. The first is that a statistical test of the null hypothesis of no group × time interaction is an omnibus or global test and provides only a broad assessment of whether the mean response profiles are the same in the different groups. If the null hypothesis is rejected, this does not indicate the specific ways in which the mean response profiles differ. As a result, additional analyses are invariably required. Second, by completely ignoring the time-ordering of the repeated measurements, the analysis of response profiles fails to recognize that they can be considered as observations of some continuous, underlying response process over time. The mean response over time can very often be described by relatively simple parametric (e.g., linear or quadratic) or semi-parametric (e.g., piecewise linear) curves. From a purely substantive point of view, it is unlikely that the pattern of change in the mean response over the duration of a longitudinal study will be so complicated that its description requires as many parameters as there are measurement occasions. The analysis of response profiles uses a saturated model for the mean response over

time, and thereby produces a perfect fit to the observed mean response profile. At first glance, this might seem like a desirable feature of any analytic approach; namely, that it fits the observed mean responses well. (In fact, not just well, but perfectly!) However, in doing so, the method fails to describe the most salient aspects of the changes in the mean response over time in terms of some pattern that can be given a substantive or theoretical interpretation. In summary, in the analysis of response profiles there is no reduction in complexity.

In contrast, the fitting of parametric or semi-parametric curves to longitudinal data can be justified on both substantive and statistical grounds. Substantively, in many longitudinal studies the true underlying mean response process is likely to change over time in a relatively smooth, monotonically increasing or decreasing pattern, at least for the duration of the study. As a result, simple parametric or semi-parametric curves can be used to describe how the mean response changes over time. From a statistical perspective, the fitting of parsimonious models for the mean response will result in statistical tests of covariate effects (e.g., treatment × time interactions) that have greater power than in an analysis of response profiles. The reason for the greater power is that the tests of covariate effects focus only on a relatively narrow range of alternative hypotheses. In contrast, the test statistics in the analysis of response profiles disperse their power over a much broader, but in many cases less substantively plausible or relevant, range of alternative hypotheses. For example, when trends in the mean response over time are assumed to be linear, and a linear trend actually provides a reasonable approximation to the true underlying shape of the mean response profile, the resulting tests of time trends and covariate effects will have greater power than the global tests in an analysis of response profiles. Note, however, that the tests based on parametric curves will only be more powerful at detecting changes in the mean response that exhibit a linear trend. They will not be more powerful, however, if the underlying shape of the mean response over time is U-shaped, rather than linear. Finally, simple parametric curves provide a parsimonious description of changes in the mean response over time in terms of a relatively small number of parameters. The results can be communicated easily to investigators and empirical researchers. In the following two sections, we describe two broad approaches for describing patterns of change in the mean response over time: polynomial trends and linear splines.

6.2 POLYNOMIAL TRENDS IN TIME

One widely adopted approach for analyzing longitudinal data is to describe the patterns of change in the mean response over time in terms of simple polynomial trends, for example, linear or quadratic trends. In this approach the means are modelled as an explicit function of time. This approach can handle highly unbalanced designs in a relatively seamless way. For example, mistimed measurements are easily incorporated in the model for the mean response.

LINEAR TRENDS OVER TIME

The simplest possible curve for describing changes in the mean response over time is a straight line. In this model, the slope for time has direct interpretation in terms of a constant change in the mean response for a single-unit change in time. Consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed in Section 5.2. If the mean response changes in an approximately linear fashion over the duration of the study we can adopt the following linear trend model

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \texttt{Time}_{ij} + \beta_3 \, \texttt{Group}_i + \beta_4 \, \texttt{Time}_{ij} \times \texttt{Group}_i, \qquad (6.1)$$

where $\texttt{Group}_i = 1$ if the i^{th} individual was assigned to the novel treatment, and $\texttt{Group}_i = 0$ otherwise; and $\texttt{Time}_{ij}$ denotes the measurement time for the j^{th} measurement on the i^{th} individual. Note that $\texttt{Group}_i$ requires only a single index i, since individuals do not change treatment groups over the course of the study. Also, by using two indices for $\texttt{Time}_{ij}$ we are implicitly allowing for the fact that there may potentially be mistimed measurements (in the latter case, $\texttt{Time}_{ij} \neq \texttt{Time}_{i'j}$, where i and i' denote two different subjects).

In the linear model given by (6.1), the model for the mean for subjects assigned to the control group is

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \texttt{Time}_{ij},$$

while for subjects assigned to the treatment group

$$E(Y_{ij}) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) \, \texttt{Time}_{ij}.$$

Thus each group's mean response is assumed to change linearly over time. This model with linear trends for two groups is depicted graphically in Figure 6.1, where the two groups have different intercepts and slopes. Here β_1 is the intercept in the control group (the "reference" group), while $(\beta_1 + \beta_3)$ is the intercept in the treatment group. The intercepts for each of the two groups have interpretation in terms of the mean response when $\texttt{Time}_{ij} = 0$; more generally, β_1 has interpretation as the mean response when all of the covariates are set to zero. Unless some care is taken with how the covariates are scaled (e.g., by centering all quantitative covariates prior to inclusion in the model), β_1 is not always readily interpretable and may represent an extrapolation beyond the data at hand. There can also be good reason, beyond issues of parameter interpretation, for centering the variable that denotes the time of measurement; this issue will be discussed later. Finally, the slope, or constant rate of change in the mean response per unit change in time, is β_2 in the control group, while the corresponding slope in the treatment group is $(\beta_2 + \beta_4)$. Ordinarily, in a longitudinal study the question of primary interest concerns a comparison of the changes in the mean response over time; this can be translated into a comparison of the slopes. Thus, if $\beta_4 = 0$, then the two groups do not differ in terms of changes in the mean response over time.

The model with linear trend over time is the simplest parametric "curve" that can be used to describe changes in the mean response over time. This model can easily incorporate both discrete (e.g., treatment or exposure group) and quantitative (e.g.,

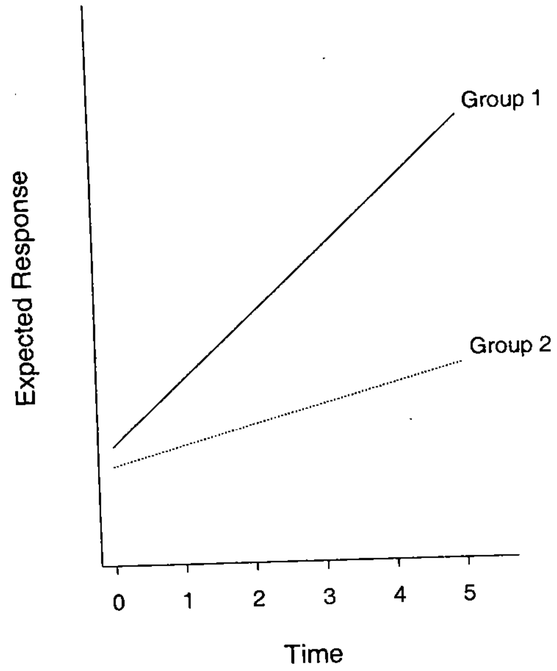

Fig. 6.1 Graphical representation of model with linear trends for two groups.

dose) covariates. Hypotheses about the dependence of changes in the mean response over time on covariates can be expressed in terms of hypotheses about whether the slope varies as a function of the covariates, that is, in terms of interactions between the covariates and the linear trend in time.

QUADRATIC TRENDS OVER TIME

When changes in the mean response over time are not linear, higher-order polynomial trends can be considered. For example, if the means are monotonically increasing or decreasing over the course of the study, but in a curvilinear way, a model with quadratic trends can be considered. In a quadratic trend model changes in the mean response are no longer constant (as in the linear trend model) throughout the duration of the study. Instead, the rate of change in the mean response depends on time, that is, the rate of change in the mean response depends upon whether the focus is on changes that occur early or later in the study. As a result, the rate of change must be expressed in terms of two parameters.

Consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier. Assuming that the changes in the mean response can be approximated by quadratic trends, the following model can be adopted:

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2 \, \texttt{Time}_{ij} + \beta_3 \, \texttt{Time}_{ij}^2 + \beta_4 \, \texttt{Group}_i \\ &\quad + \beta_5 \, \texttt{Time}_{ij} \times \texttt{Group}_i + \beta_6 \, \texttt{Time}_{ij}^2 \times \texttt{Group}_i. \end{aligned} \tag{6.2}$$

In this model, the mean response over time for subjects in the control group is given by

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \texttt{Time}_{ij} + \beta_3 \, \texttt{Time}_{ij}^2;$$

while the corresponding mean response over time in the novel treatment group is given by

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) \, \texttt{Time}_{ij} + (\beta_3 + \beta_6) \, \texttt{Time}_{ij}^2.$$

This model with quadratic trends for two groups is depicted graphically in Figure 6.2, where the two groups have different intercepts (or mean response at time 0) and non-constant rates of change over time that differ between the two groups.

Note that in the quadratic trends model the mean response changes at a different rate, depending upon $\texttt{Time}_{ij}$. For example, the rate of change in the control group is given by $\beta_2 + 2\beta_3 \, \texttt{Time}_{ij}$ (the derivation of this instantaneous rate of change requires some familiarity with calculus and is omitted). Thus, early in the study when $\texttt{Time}_{ij} = 1$, the rate of change in the mean response is $\beta_2 + 2\beta_3$; whereas later in the study, say $\texttt{Time}_{ij} = 4$, the rate of change in the mean response is $\beta_2 + 8\beta_3$. The rate of change is different at the two occasions and the magnitude and sign of the regression coefficients β_2 and β_3 determine whether the mean response is increasing or decreasing over time and how the rate of change depends on time. The regression coefficients, $(\beta_2 + \beta_5)$ and $(\beta_3 + \beta_6)$, have similar interpretations for the treatment group.

When fitting polynomial trend models one must take care to avoid extrapolation beyond the data at hand. While polynomial trend models can fit a flexible class of curves to the data, inferences beyond the measurement occasions should be avoided as these will be sensitive to the underling model assumptions. While a quadratic trend might be a reasonable approximation for the data, recall that a quadratic trend necessarily has a turning point where the trend changes (e.g., from an increasing trend over time to a decreasing trend, or vice versa). In the absence of a strong theoretical rationale for the model, extrapolation beyond the data can produce nonsensical results and should be avoided.

With polynomial trend models, there is a natural hierarchy of effects that has implications for testing hypotheses about linear, quadratic, and higher-order polynomial trends. That is, higher-order terms should be tested (and, if appropriate, removed from the model) before lower-order terms are assessed. Thus, in the quadratic model,

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \texttt{Time}_{ij} + \beta_3 \, \texttt{Time}_{ij}^2,$$

it is not meaningful or appropriate to test the coefficient for the linear trend, β_2, in a model that also includes a coefficient for the quadratic trend, β_3. Instead, a test for

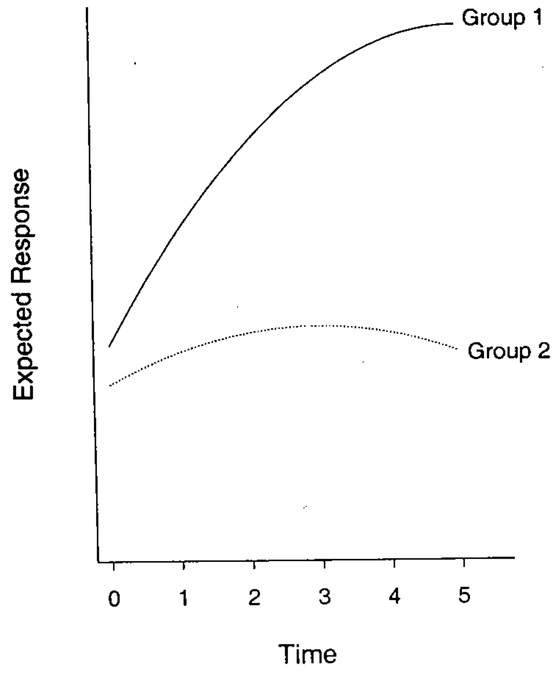

Fig. 6.2 Graphical representation of model with quadratic trends for two groups.

quadratic trend (versus linear trend) can be performed by testing the null hypothesis that $\beta_3 = 0$. If this null hypothesis cannot be rejected, it is then appropriate to remove the quadratic term from the model and consider the model with only linear trend. The test for linear trend is performed by testing the null hypothesis that $\beta_2 = 0$ in the model that only includes the linear term. This hierarchy is completely analogous to the testing of interactions, that is, tests of main effects (or attempts to interpret the main effects) are not meaningful in the presence of interaction. By the same token, tests of lower-order polynomials in time (e.g., linear trend) are not meaningful in the presence of higher-order polynomials in time (e.g., quadratic trend).

Finally, we return to the issue of "centering" variables. Although "centering" $\texttt{Time}_j$ at zero leads to a simple interpretation of the intercept when $\texttt{Time}_j$ represents time since baseline, to avoid potential problems of collinearity in the quadratic (or in any higher-order polynomial) trend model, it is advisable to "center" $\texttt{Time}_j$ on its mean value. That is, prior to the analysis, replace $\texttt{Time}_j$ by its deviation from the mean of $\texttt{Time}_1, \texttt{Time}_2, ..., \texttt{Time}_n$. To highlight the impact of this centering, consider the following example where $\texttt{Time}_j \in \{0, 1, 2, ..., 10\}$. The correlation between $\texttt{Time}_j$ and $\texttt{Time}_j^2$ is 0.963. When two covariates are so highly correlated, computational problems associated with collinearity may arise in the estimation of β. Centering $\texttt{Time}_j$ *before* quadratic (or any higher-order polynomial) terms are included in the model helps to alleviate problems associated with collinearity. For example, when $\texttt{Time}_j \in \{0, 1, 2, ..., 10\}$ and we create a "centered" variable, say $\texttt{CTime}_j = (\texttt{Time}_j - 5.0)$, where 5.0 is the mean of $\{0, 1, 2, ..., 10\}$, then the correlation between $\texttt{CTime}_j$ and $\texttt{CTime}_j^2$ is zero, thereby avoiding any potential problems associated with collinearity. With balanced longitudinal data (requiring only a single index j for $\texttt{Time}_j$), it is natural to center $\texttt{Time}_j$ on the mean of $\texttt{Time}_1, \texttt{Time}_2, ..., \texttt{Time}_n$. However, with unbalanced longitudinal data, it is important to center $\texttt{Time}_{ij}$ at some common value for all individuals. By centering at a common value, the regression intercept is interpretable as the mean response at that common value for time. In general, centering of $\texttt{Time}_{ij}$ at individual-specific values (e.g., the mean of the n_i times of measurement for the i^{th} individual) should be avoided as these may vary considerably from one individual to another, thereby making the interpretation of the regression intercept meaningless.

6.3 LINEAR SPLINES

Simple parametric curves can provide a parsimonious description of longitudinal trends in the mean response. The simplest case is the linear trend model that characterizes change in the mean response over time in terms of a single slope parameter representing a constant rate of change. By introducing higher-order polynomials in time, various kinds of non-linearities in the longitudinal trends can also be accommodated. However, as the degree of the polynomial increases, the interpretation of the regression coefficients becomes more difficult. As a result, the use of polynomials in time is most appealing when any non-linearity can be approximated by quadratic trends.

In some applications, the longitudinal trends in the mean response cannot be characterized by first and second degree polynomials in time (i.e., linear or quadratic trends). In addition, there are other applications where non-linear trends in the mean response cannot be well approximated by polynomials in time of any order. This will most often occur when the mean response increases (or decreases) rapidly for some duration and then more slowly thereafter (or vice versa) When this type of pattern of change arises, it can often by handled by using linear spline models.

If the simplest possible curve is a straight line, then one way to extend the curve is to have a sequence of joined or connected line segments that produces a piecewise linear pattern. Linear spline models provide a very useful and flexible way to accommodate many of the non-linear trends that cannot be approximated by simple polynomials in time. The basic idea behind linear spline models is remarkably simple: divide the time axis into a series of segments and consider a model for the trend over time that is comprised of piecewise linear trends, having different slopes within each segment but joined or tied together at fixed times. The locations where the lines meet or are tied together are known as the "knots". This allows the mean response to increase or decrease as time proceeds, depending on the sign and magnitude of

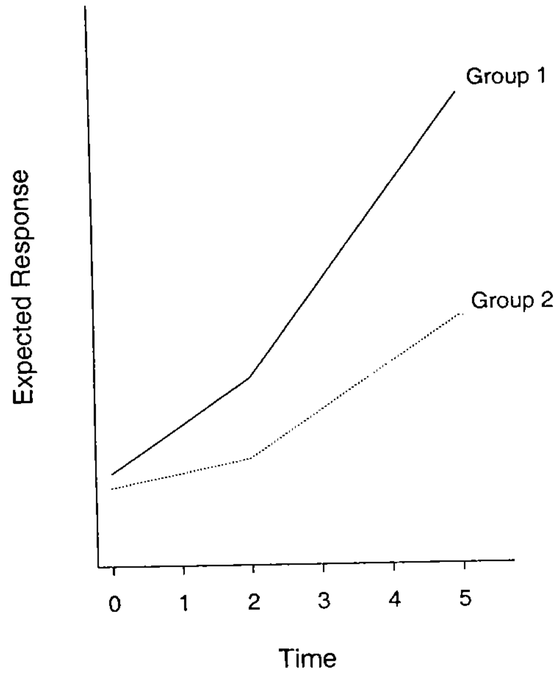

Fig. 6.3 Graphical representation of model with linear splines for two groups, with common knot at Time = 2.

the regression slopes for the line segments. The resulting piecewise linear curve is called a spline. Figure 6.3 provides an illustration of a linear spline model for two groups with common knot at time 2. Note that the slopes of the two lines, before and after time 2, are different, with a greater increase in the mean response in the second time segment, and a more attenuated increase in the mean response in the first time segment. This spline model is sometimes referred to as a piecewise linear or "broken-stick" model.

The simplest possible spline model has only one knot and can be parameterized in a number of different ways. Returning to the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier, if the mean response changes over time in a piecewise linear way, we can fit the following linear spline model with knot at $\mathtt{t}^*$

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2\,\mathtt{Time_{ij}} + \beta_3\,(\mathtt{Time_{ij}} - \mathtt{t}^*)_+ + \beta_4\,\mathtt{Group_i} \\ &\quad + \beta_5\,\mathtt{Time_{ij}} \times \mathtt{Group_i} + \beta_6\,(\mathtt{Time_{ij}} - \mathtt{t}^*)_+ \times \mathtt{Group_i}, \end{aligned} \tag{6.3}$$

where $(x)_+$ is defined as a function that equals x when x is positive and is equal to zero otherwise. Thus, $(\mathtt{Time_{ij}} - \mathtt{t}^*)_+$ is equal to $(\mathtt{Time_{ij}} - \mathtt{t}^*)$ when $\mathtt{Time_{ij}} > \mathtt{t}^*$

and is equal to zero when $\mathtt{Time_{ij}} \leq \mathtt{t}^*$. In the model given by (6.3), the means for subjects in the control group are

$$E(Y_{ij}) = \beta_1 + \beta_2\,\mathtt{Time_{ij}} + \beta_3\,(\mathtt{Time_{ij}} - \mathtt{t}^*)_+.$$

When expressed in terms of the mean response prior to and after $\mathtt{t}^*$,

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2\,\mathtt{Time_{ij}}, && \mathtt{Time_{ij}} \leq \mathtt{t}^*; \\ E(Y_{ij}) &= (\beta_1 - \beta_3 t^*) + (\beta_2 + \beta_3)\,\mathtt{Time_{ij}}, && \mathtt{Time_{ij}} > \mathtt{t}^*. \end{aligned}$$

Thus, in the control group, the slope prior to $\mathtt{t}^*$ is β_2 and following $\mathtt{t}^*$ is $(\beta_2 + \beta_3)$. Similarly, the means for subjects in the treatment group are given by

$$E(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\,\mathtt{Time_{ij}} + (\beta_3 + \beta_6)\,(\mathtt{Time_{ij}} - \mathtt{t}^*)_+.$$

When expressed in terms of the mean response prior to and after $\mathtt{t}^*$,

$$\begin{aligned} E(Y_{ij}) &= (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\,\mathtt{Time_{ij}}, && \mathtt{Time_{ij}} \leq \mathtt{t}^*; \\ E(Y_{ij}) &= \{(\beta_1 + \beta_4) - (\beta_3 + \beta_6)\,t^*\} \\ &\quad + (\beta_2 + \beta_3 + \beta_5 + \beta_6)\,\mathtt{Time_{ij}}, && \mathtt{Time_{ij}} > \mathtt{t}^*. \end{aligned}$$

Then, in terms of group comparisons, the null hypothesis of no group differences in patterns of change over time can be expressed as H_0: $\beta_5 = \beta_6 = 0$. Comparisons of the groups before and after $\mathtt{t}^*$ are also possible. For example, the null hypothesis of no group differences in patterns of change prior to $\mathtt{t}^*$ can be expressed as H_0: $\beta_5 = 0$.

The simple spline model considered so far can be extended to include more than one knot or more than two joined line segments. More generally, a spline model with K knots or break-points will produce $K + 1$ line segments and there will be $K + 1$ corresponding slopes. Thus, in principle, it is possible to accommodate quite complex non-linear patterns for changes in the mean response by including a sufficient number of variables, $(\mathtt{Time_{ij}} - \mathtt{t}^*_k)_+$, with knots located at $\mathtt{t}^*_k$ (for $k = 1, ..., K$). However, in practice, the data from many longitudinal studies can be well-approximated by simple piecewise linear models with at most one or two knots that are located at judiciously chosen time points.

Finally, our discussion thus far has avoided the thorny problem of the choice of location(s) for the knot(s). There is an extensive body of research in statistics on automated choices for the knot location, where the location is effectively determined by the data at hand. Ideally, the choice of knot location should also incorporate subject-matter considerations. For example, in studies of growth, certain ages are associated with growth spurts. Similarly, measures of hormonal response are known to change quite dramatically with the onset of puberty and menopause. In other settings, there may be a body of evidence that the response profile changes in a discernible way at certain time points. An example of the latter is in studies of HIV-infected patients. In early studies of the treatment of HIV-infected patients with AZT, it was increasingly

recognized that CD4 counts, a measure of the body's immune response, increased sharply over a 4–6 week period following treatment with AZT, but then levelled off. In summary, the choice of knot location is a mixture of art and science. When it is available, subject-matter knowledge should be brought to bear on the empirical evidence for the most appropriate choice of knot location.

6.4 GENERAL LINEAR MODEL FORMULATION

Next we demonstrate how both polynomial trend and spline models can be expressed in terms of the general linear model

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

for appropriate choices of X_i. Let n_i be the number of repeated measures on the i^{th} individual ($i = 1, ..., N$). To illustrate how the polynomial trend model can be expressed in terms of the general linear model, consider the hypothetical two-group study comparing a novel *treatment* and a *control* discussed earlier. Let us assume that the mean response changes over time in a quadratic trend. Then, the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & 0 & 0 & 0 \\ 1 & t_{i2} & t_{i2}^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 0 & 0 & 0 \end{pmatrix};$$

while for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 & 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 & 1 & t_{in_i} & t_{in_i}^2 \end{pmatrix},$$

where t_{ij} denotes the time of the j^{th} measurement on the i^{th} individual. Then, in terms of the general linear model

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_6)'$ is a 6×1 vector of regression coefficients, the mean responses in the control group are

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_2 t_{i1} + \beta_3 t_{i1}^2 \\ \beta_1 + \beta_2 t_{i2} + \beta_3 t_{i2}^2 \\ \vdots \\ \beta_1 + \beta_2 t_{in_i} + \beta_3 t_{in_i}^2 \end{pmatrix};$$

while the mean responses in the treatment group are

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix} = \begin{pmatrix} (\beta_1 + \beta_4) + (\beta_2 + \beta_5)t_{i1} + (\beta_3 + \beta_6)t_{i1}^2 \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5)t_{i2} + (\beta_3 + \beta_6)t_{i2}^2 \\ \vdots \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5)t_{in_i} + (\beta_3 + \beta_6)t_{in_i}^2 \end{pmatrix}.$$

For the spline model, let us assume that the mean response changes over time in a piecewise linear way, with knot at $t^* = 4$. Then the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & (t_{i1} - 4)_+ & 0 & 0 & 0 \\ 1 & t_{i2} & (t_{i2} - 4)_+ & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i} - 4)_+ & 0 & 0 & 0 \end{pmatrix};$$

while for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & (t_{i1} - 4)_+ & 1 & t_{i1} & (t_{i1} - 4)_+ \\ 1 & t_{i2} & (t_{i2} - 4)_+ & 1 & t_{i2} & (t_{i2} - 4)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & (t_{in_i} - 4)_+ & 1 & t_{in_i} & (t_{in_i} - 4)_+ \end{pmatrix}.$$

The spline model can then be expressed in terms of the general linear model,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

where $\beta = (\beta_1, ..., \beta_6)'$ is a 6×1 vector of regression coefficients.

Given that both the polynomial trend and spline models can be expressed in terms of the general linear regression model,

$$E(Y_i|X_i) = \mu_i = X_i\beta,$$

restricted maximum likelihood estimation of β, and the construction of confidence intervals and tests of hypotheses, are possible once the covariance of Y_i has been specified. Unlike the analysis of response profiles, where the covariance of Y_i is assumed to be unstructured with no constraints on the covariance parameters other than the requirement that they yield a symmetric matrix (and one that is positive-definite), more parsimonious models for the covariance can be adopted. Indeed, the use of parametric curves for the mean response is most appealing in settings where the longitudinal data are inherently unbalanced over time. As a result, an unstructured covariance matrix may not be well-defined, let alone estimated, when, in principle, each individual can have a unique sequence of measurement times. However, the discussion of models for the covariance is postponed until Chapter 7; here, we simply

assume that some appropriate model for the covariance has been adopted. Given models for both the mean and covariance, REML estimates of β, and their standard errors (based on the estimated covariance of $\widehat{\beta}$), can be obtained using the method of estimation described in Chapter 4.

6.5 CASE STUDIES

Next, we illustrate the main ideas by considering polynomial trend models for data on lung function (FEV_1) from a longitudinal epidemiologic study of current and former smokers aged 36 and older. The application of spline models is illustrated using the blood lead data on the 100 children from the treatment and placebo groups of the Treatment of Lead-Exposed Children (TLC) Trial.

The Vlagtwedde-Vlaardingen Study

In an epidemiologic study conducted in two different areas in The Netherlands, the rural area of Vlagtwedde in the north-east and the urban, industrial area of Vlaardingen in the south-west, residents were followed over time to obtain information on the prevalence of and risk factors for chronic obstructive lung diseases (van der Lende *et al.*, 1981; Rijcken *et al.*, 1987). Here we focus on a sub-sample of men and women from the rural area of Vlagtwedde. The sample, initially aged 15–44, participated in follow-up surveys approximately every 3 years for up to 21 years. At each survey, information on respiratory symptoms and smoking status was collected by questionnaire and spirometry was performed. Pulmonary function was determined by spirometry and a measure of forced expiratory volume (FEV_1) was obtained every three years for the first 15 years of the study, and also at year 19.

In this study, FEV_1 was not recorded for every subject at each of the planned measurement occasions. That is, the data are unbalanced due to incompleteness. The number of repeated measurements of FEV_1 on each subject varied from 1 to 7. For the purpose of this illustration we focus on a subset of the data on 133 residents aged 36 or older at their entry into the study and whose smoking status did not change over the 19 years of follow-up. Each study participant was either a current or former smoker. Current smoking was defined as smoking at least one cigarette per day. The trends in the mean FEV_1 over time, for current and former smokers, are displayed in Figure 6.4. The goal of our analysis is to describe changes in lung function over the 19 years of follow-up with parametric curves and to determine whether the time trends differ for current and former smokers. We summarize differences in mean change between current and former smokers, assuming change does not depend strongly on either age or gender (neither variable was available in the data set).

First we consider a linear trend in the mean response over time, with intercepts and slopes that differ for the two smoking exposure groups. For all of the analyses reported here, we assume an unstructured covariance matrix. Based on the REML estimates of the regression coefficients in Table 6.1, the mean response for participants

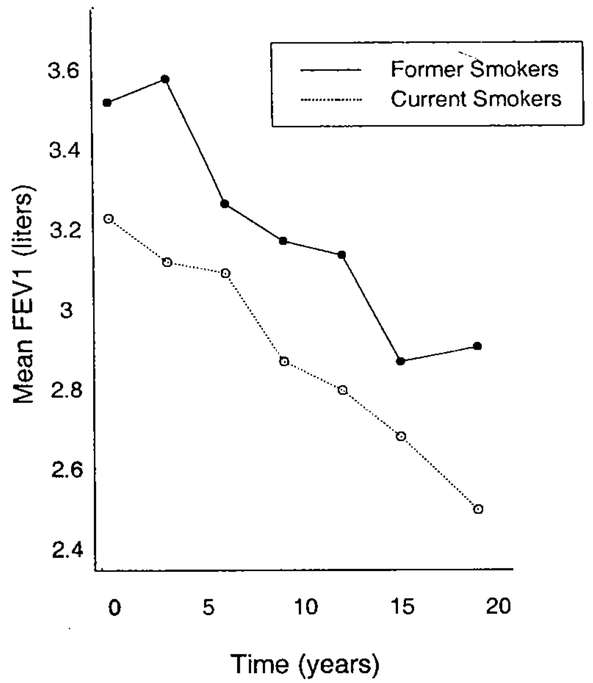

Fig. 6.4 Mean FEV_1 at baseline (year 0), year 3, year 6, year 9, year 12, year 15, and year 19 in the current and former smoking exposure groups.

who are former smokers is estimated to be

$$E(Y_{ij}) = 3.507 - 0.033\ \texttt{Time}_{ij},$$

while for participants who are current smokers,

$$\begin{aligned} E(Y_{ij}) &= (3.507 - 0.262) - (0.033 + 0.005)\ \texttt{Time}_{ij} \\ &= 3.245 - 0.038\ \texttt{Time}_{ij}. \end{aligned}$$

Thus, it would appear that both groups have a significant decline in mean FEV_1 over time, but there is no discernible difference between the two smoking exposure groups in the constant rate of change, since the $\texttt{Smoke}_i \times \texttt{Time}_{ij}$ interaction (i.e., the comparison of the two slopes) is not significant, with $Z = -1.42$, $p > 0.15$.

The adequacy of the linear trend model can be assessed by including higher-order polynomial trends. For example, we can consider a model that allows quadratic trends for changes in FEV_1 over time. Recall that the linear trend model is nested within the quadratic trend model. If the linear trend model is adequate for these data, the difference in maximized log-likelihoods (or the likelihood ratio test statistic) should

Table 6 1 Estimated regression coefficients and standard errors based on a model with linear trends for the FEV_1 data from the Vlagtwedde-Vlaardingen study.

Variable	Smoking Group	Estimate	SE	Z
Intercept		3.5073	0.1004	34.94
$\texttt{Smoke}_i$	Current	−0.2617	0.1151	−2.27
$\texttt{Time}_{ij}$		−0.0332	0.0031	−10.84
$\texttt{Smoke}_i \times \texttt{Time}_{ij}$	Current	−0.0050	0.0035	−1.42

Table 6.2 Comparison of the maximized (ML) log-likelihoods for the model with linear and quadratic trends for the FEV_1 data from the Vlagtwedde-Vlaardingen study.

Model	−2 (ML) Log-Likelihood
Quadratic Trend Model	237.2
Linear Trend Model	238.5
$-2 \times$ Log-Likelihood Ratio: $G^2 = 1.3,\ 2\ \text{df},\quad (p > 0.50)$	

not be large. The maximized log-likelihoods for the models with linear and quadratic trends are presented in Table 6.2. The likelihood ratio test statistic can be compared to a chi-squared distribution with 2 degrees of freedom (or 6, the number of parameters in the quadratic trend model, minus 4, the number of parameters in the linear trend model). Note that the likelihood ratio test is based on the ML, not REML, log-likelihood because we are comparing two different models for the mean; as a result, both models were re-fit using ML estimation. For both models, the polynomial trends over time are allowed to differ for the two smoking exposure groups. The likelihood ratio test, comparing the quadratic and linear trend models, produces $G^2 = 1.3$, with 2 degrees of freedom ($p > 0.50$). Thus, when compared to the quadratic trend model, the linear trend model appears to be adequate for these data. Finally, for illustrative purposes, we can make a comparison with a cubic trend model. This produces a likelihood ratio test statistic, $G^2 = 4.4$, with 4 degrees of freedom ($p > 0.35$), indicating again that the linear trend model is adequate for these data.

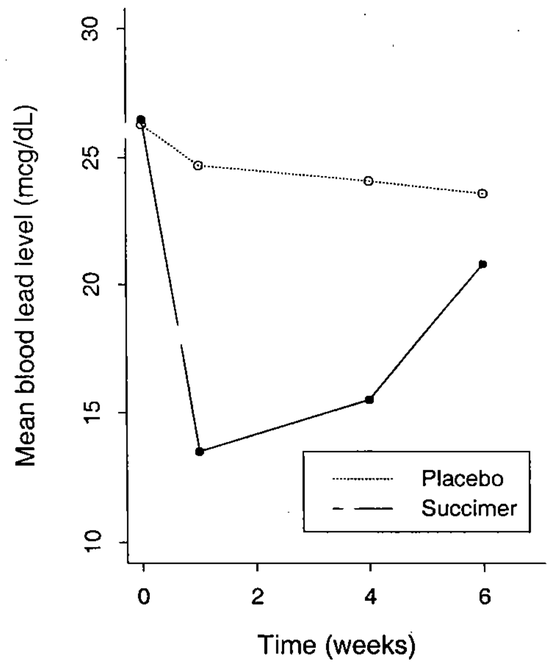

Fig. 6.5 Mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups.

Treatment of Lead-Exposed Children Trial

Recall that the TLC trial was a placebo-controlled, randomized trial of a chelating agent, succimer, in children with confirmed blood lead levels of 20–44 μg/dL. The children in the trial were aged 12–33 months and lived in deteriorating inner city housing. The following analyses are based on data on blood lead levels at baseline (or week 0), week 1, week 4, and week 6. Note that, from the plot of the means in Figure 6.5, it would appear that only the mean blood lead levels in the placebo group can be described by a linear trend; the mean in the succimer group decreases from baseline to week 1, but then increases thereafter.

Given that there are non-linearities in the trends over time, higher-order polynomial models (e.g., a quadratic trend model) could be fit to the data. However, to illustrate the application of spline models, we accommodate the non-linearity with a piecewise linear model with common knot at week 1,

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2\,\texttt{Week}_{ij} + \beta_3\,(\texttt{Week}_{ij} - 1)_+ + \beta_4\,\texttt{Group}_i \times \texttt{Week}_{ij} \\ &\quad + \beta_5\,\texttt{Group}_i \times (\texttt{Week}_{ij} - 1)_+, \end{aligned}$$

where $\texttt{Group}_i = 1$ if assigned to succimer, and $\texttt{Group}_i = 0$ otherwise. Because of the randomization of children to the two treatment groups, the model does not contain

Table 6.3 Estimated regression coefficients and standard errors based on a piecewise linear model, with common knot at week 1, for the blood lead level data from the TLC trial.

Variable	Group	Estimate	SE	Z
Intercept		26.3422	0.4991	52.78
$\texttt{Week}_{ij}$		−1.6296	0.7818	−2.08
$(\texttt{Week}_{ij} - 1)_+$		1.4305	0.8777	1.63
$\texttt{Group}_i \times \texttt{Week}_{ij}$	S	−11.2500	1.0924	−10.30
$\texttt{Group}_i \times (\texttt{Week}_{ij} - 1)_+$	S	12.5822	1.2278	10.25

a main effect of Group and we assume a common mean blood lead level at baseline. In this piecewise linear model, the means for subjects in the placebo group are given by

$$E(Y_{ij}) = \beta_1 + \beta_2 \texttt{Week}_{ij} + \beta_3 (\texttt{Week}_{ij} - 1)_+,$$

while in the succimer group the means are given by

$$E(Y_{ij}) = \beta_1 + (\beta_2 + \beta_4) \texttt{Week}_{ij} + (\beta_3 + \beta_5) (\texttt{Week}_{ij} - 1)_+.$$

The REML estimates of the regression parameters from the piecewise linear model are given in Table 6.3. When expressed in terms of the mean response prior to and after week 1, the estimated means in the placebo group are

$$\widehat{\mu}_{ij} = \widehat{\beta}_1 + \widehat{\beta}_2 \texttt{Week}_{ij}, \qquad \texttt{Week}_{ij} \leq 1;$$

$$\widehat{\mu}_{ij} = (\widehat{\beta}_1 - \widehat{\beta}_3) + (\widehat{\beta}_2 + \widehat{\beta}_3) \texttt{Week}_{ij}, \qquad \texttt{Week}_{ij} > 1.$$

Thus, in the placebo group, the slope prior to week 1 is $\widehat{\beta}_2 = -1.63$ and following week 1 is $(\widehat{\beta}_2 + \widehat{\beta}_3) = -1.63 + 1.43 = -0.20$. Similarly, when expressed in terms of the mean response prior to and after week 1, the estimated means for subjects in the succimer group are given by

$$\widehat{\mu}_{ij} = \widehat{\beta}_1 + (\widehat{\beta}_2 + \widehat{\beta}_4) \texttt{Week}_{ij}, \qquad \texttt{Week}_{ij} \leq 1;$$

$$\widehat{\mu}_{ij} = (\widehat{\beta}_1 - (\widehat{\beta}_3 + \widehat{\beta}_5) + (\widehat{\beta}_2 + \widehat{\beta}_3 + \widehat{\beta}_4 + \widehat{\beta}_5) \texttt{Week}_{ij}, \qquad \texttt{Week}_{ij} > 1.$$

The estimates of the mean blood lead levels for the placebo and succimer groups are presented in Table 6.4. The estimated means from the piecewise linear model appear to adequately fit the observed mean response profiles for the two treatment groups.

Table 6.4 Estimated mean blood lead levels for the placebo and succimer groups from the piecewise linear model, with common knot at week 1; the observed means are in parentheses.

Group	Week 0	Week 1	Week 4	Week 6
Succimer	26.3	13.5	16.7	19.1
	(26.5)	(13.5)	(15.5)	(20.8)
Placebo	26.3	24.7	24.1	23.7
	(26.3)	(24.7)	(24.1)	(23.6)

Note that the model with linear trends (and common intercept) is nested within the piecewise linear model (since the former can be obtained by setting $\beta_3 = \beta_5 = 0$ in the latter). When these two models are compared in terms of their maximized log-likelihoods (see Table 6.5), the likelihood ratio test statistic is $G^2 = 121.8$ and can be compared to a chi-squared distribution with 2 degrees of freedom (or 5, the number of parameters in the linear spline model, minus 3, the number of parameters in the linear trend model). Note that the likelihood ratio test is based on the ML, not REML, log-likelihood because we are comparing two different models for the mean; as a result, both models were re-fit using ML. The magnitude of the likelihood ratio test statistic (with $p < 0.0001$) indicates that the piecewise linear model significantly improves the overall fit to the mean response over time when compared to a linear trend model. This simply confirms what was already obvious from the plot of the means in Figure 6.5. Although the piecewise linear and quadratic trend models (with common intercept for the two treatment groups) are not nested, they both have the same number of parameters and therefore their respective log-likelihoods can be directly compared (see Table 6.5). From a comparison of the maximized log-likelihoods it is apparent that the piecewise linear model fits these data discernibly better than the quadratic trend model (−2 ML log-likelihood = 2436.2 for the piecewise linear model versus −2 ML log-likelihood = 2551.7 for the quadratic trend model).

Finally, it is quite instructive to examine the estimated unstructured covariance matrix for the linear trend model, a model that does not fit these data well. In Table 6.6 the REML estimates of the unstructured covariance matrix are presented. Note that the estimated variances at weeks 1 and 4 are approximately three to four times greater than at baseline. Moreover, the estimated covariance matrix is discernibly different from that obtained in the analysis of response profiles in Section 5.4 that placed no structure on the means. Indirectly, the inflation of the variance at weeks 1 and 4 is an indication that the lack of fit to the means at these two occasions is being attributed to error variability and hence the inflation of the variance at these two occasions. This highlights an important issue that will be discussed in greater detail in

Table 6.5 Comparison of the maximized (ML) log-likelihoods for the models with linear and quadratic trends, and piecewise linear trend with common knot at week 1, for the blood lead level data from the TLC trial.

Model	−2 (ML) Log-Likelihood
Piecewise Linear (Spline) Model	2436.2
Quadratic Trend Model	2511.7
Linear Trend Model	2558.0

Table 6.6 Estimated unstructured covariance matrix for the linear trend model for the blood lead levels at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Covariance Matrix			
25.5	13.8	16.1	21.4
13.8	111.2	81.2	38.4
16.1	81.2	78.3	36.8
21.4	38.4	36.8	59.4

Chapter 7, namely, that there is an interdependence between the mean response and the covariance that has important implications for how these two aspects of longitudinal data are jointly modelled.

In conclusion, one of the main aspects of summarizing trends in the mean response over time via parametric curves is that trends over time and their relation to covariates can be expressed as a function of a small number of parameters. That is, covariate effects on changes in the mean response over time can be captured in one or two regression parameters, leading to more powerful tests when the models are appropriate for the data at hand. Also, the parametric curves define the mean of Y_{ij}, $E(Y_{ij}|X_{ij})$, as an explicit function of the times of measurement, $\texttt{Time}_{ij}$. As a result, there is no reason to require all individuals to have the same set of measurement times, nor even the same number of measurements. In our examples we have used only data sets where subjects are measured at the same set of occasions, but this is because we will require models for the covariance when subjects are measured at arbitrary points in time. Hence, examples of this will be given in the next chapter.

Table 6.7 Illustrative commands for a linear trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS  id  group  t;
   MODEL  y=group  time  group*time / S  CHISQ;
   REPEATED  t / TYPE=UN  SUBJECT=id  R  RCORR;
```

6.6 COMPUTING: FITTING PARAMETRIC CURVES USING PROC MIXED IN SAS

To fit a linear trend model to data from two or more groups measured repeatedly over time, we can use the illustrative SAS commands given in Table 6.7. Note that this model assumes that the covariance matrix is unstructured. In principle, alternative assumption about the covariance can be considered. Indeed, when the data are unbalanced over time, it will be necessary to consider parametric models for the covariance; this topic will be discussed in greater detail in Chapter 7.

Note that the CLASS statement includes a variable `t`. This variable is an additional copy of the variable `time`. The difference is that while `t` is declared as a categorical variable on the CLASS statement, `time` is not and is treated as a quantitative covariate in the MODEL statement. The reason for having two versions, `time` and `t`, one quantitative and the other categorical, is that it is good practice to include, wherever possible, a REPEATED effect. This ensures that the covariance is estimated correctly when the design is balanced but incomplete due to missingness or when the study is balanced and complete but the repeated measures are not in the same order for each subject in the data set (e.g., this might arise when the data set has previously been sorted on another variable). With unbalanced data, it will very often not be possible to include a REPEATED effect; instead, the covariance model will need to be defined explicitly in terms of the times of measurement. A further discussion of this point is postponed until Chapter 7.

Next, we present illustrative commands for fitting a quadratic trend model in Table 6.8. The MODEL statement now includes both `time` and `timesqr`, the latter is simply an additional variable that is the square of `time` (i.e., $\texttt{time}^2$). Note that the MODEL statement includes both main effects of `time` and `timesqr`, and their interactions with `group`.

Finally, we present illustrative commands for fitting spline models. In Table 6.9 we present commands in SAS for fitting a model with a single knot at $\texttt{time} = 4$. The MODEL statement includes `time` and `time_4`, where `time_4` is a derived variable for $(\texttt{time} - 4)_+$. The latter variable can easily be computed in SAS as

$$\texttt{time\_4} = \text{MAX}(\texttt{time} - 4, 0);$$

Table 6.8 Illustrative commands for a quadratic trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time timesqr group*time group*timesqr / S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Table 6.9 Illustrative commands for a spline model, with knot at $time = 4$, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time time_4 group*time group*time_4 / S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

6.7 FURTHER READING

A concise and clear discussion of how to describe patterns of change over time using polynomial trends can be found in Section 3.5 of the book by Hand and Taylor (1987). A general discussion of splines and piecewise linear regression can be found in Chapter 11 (Section 11.5) of Neter *et al.* (1996).

Bibliographic Notes

The use of simple parametric curves to describe changes in the mean response over time has its origins in growth curve analysis. Methods for estimation and testing of growth curves were developed by Wishart (1938), Box (1950), and Rao (1958). Potthoff and Roy (1964 proposed an extension of the repeated measures analysis by MANOVA for growth curves; alternative formulations were developed by Rao (1965), Khatri (1966), and Grizzle and Allen (1969).

An excellent discussion of spline models can be found in Chapter 3 of Ruppert *et al.* (2003), and the references therein.

Problems

6.1 In a study of weight gain (Box, 1950), investigators randomly assigned 30 rats to three treatment groups: treatment 1 was a control (no additive); treatments 2 and 3 consisted of two different additives (thiouracil and thyroxin respectively) to the rats drinking water. Weight was measured at baseline (week 0) and at weeks 1, 2, 3, and 4.

The raw data are stored in an external file: `rat.dat`

Each row of the data set contains the following seven variables:

ID Group Y_1 Y_2 Y_3 Y_4 Y_5

Note: The variable Group is coded 1 = control, 2 = thiouracil, and 3 = thyroxin.

6.1.1 On a single graph, construct a time plot that displays the mean weight versus time (in weeks) for the three groups. Describe the general characteristics of the time trends for the three groups.

6.1.2 Read the data from the external file and put the data in a "univariate" or "long" format, with 5 "records" per subject.

6.1.3 Assume that the rate of increase in each group is approximately constant throughout the duration of the study. Assuming an unstructured covariance matrix, construct a test of whether the rate of increase differs in the groups.

6.1.4 On a single graph, construct a time plot that displays the *estimated* mean weight versus time (in weeks) for the three treatment groups from the results generated from Problem 6.1.3.

6.1.5 Based on the results from Problem 6.1.3, what is the estimated rate of increase in mean weight in the control group (group 1)? What is the estimated rate of increase in mean weight in the thiouracil group (group 2)? What is the estimated rate of increase in mean weight in the thyroxin group (group 3)?

6.1.6 The study investigators conjectured that there would be an increase in weight, but that the rate of increase would level-off towards the end of the study. They also conjectured that this pattern of change may differ in the three treatment groups. Assuming an unstructured covariance matrix, construct a test of this hypothesis.

6.1.7 Compare and contrast the results from Problems 6.1.3 and 6.1.6. Does a model with only a linear trend in time adequately account for the pattern of change in the three treatments groups? Provide results that support your conclusion.

6.1.8 Given the results of all the previous analyses, what conclusions can be drawn about the effect of the additives on the patterns of change in weight?

7

Modelling the Covariance

7.1 INTRODUCTION

Since one of the defining features of longitudinal data is that they are correlated, we must consider approaches for appropriately modelling the covariance or time dependence among the repeated measures obtained on the same individuals. When an appropriate model for the covariance has been adopted, correct standard errors are obtained and valid inferences about the regression parameters can be made. Accounting for the covariance among repeated measures usually increases efficiency or the precision with which the regression parameters can be estimated, that is, the positive correlation among the repeated measures reduces the variability of the estimate of change over time within individuals. Thus, in a longitudinal study, the positive correlation among repeated measures can be used to advantage in the study of change over time. In addition, when there are missing data correct modelling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In general, failure to take account of the covariance among the repeated measures will result in incorrect estimates of the sampling variability and can lead to quite misleading scientific inferences.

Longitudinal data present us with two aspects of the data that require modelling: the mean response over time and the covariance among repeated measures on the same individuals. Although these two aspects of the data can, in a certain sense, be modelled separately, they are also interrelated. That is, the choice of models for the mean response and the covariance are interdependent. This interdependence arises because the vector of residuals (observed responses minus fitted responses) depends upon the specification of the model for the mean. Put more formally, the covariance

between any pair of residuals, say $\{Y_{ij} - \mu_{ij}(\beta)\}$ and $\{Y_{ik} - \mu_{ik}(\beta)\}$, depends on the model for the mean (i.e., depends on β). A model for the covariance must be chosen on the basis of some model for the mean response; it represents an attempt to account for the covariance among the residuals that results from a specific model for the mean. A different choice of model for the mean, or moreover, any misspecification of the model for the mean, can potentially result in a different choice of model for the covariance. As a result of this interdependence between the models for the mean and covariance, we will need to develop an overall modelling strategy that takes this interdependence into account.

7.2 IMPLICATIONS OF CORRELATION AMONG LONGITUDINAL DATA

Before considering approaches for modelling the covariance or correlation among repeated measures, it is worth stepping back and considering some of the implications of the correlation among longitudinal data. First, it should be kept in mind that longitudinal data are not only correlated, but for the most part, they are also positively correlated. Moreover, the positive correlation among repeated measures can be used to advantage in the study of change over time. That is, we can capitalize on the positive correlation among longitudinal data when the main focus of the analysis is on change in the mean response over time.

Consider a simple longitudinal study design where it is of interest to measure change in a health outcome "before" and "after" receiving some health intervention. With only two repeated measures of the outcome, the statistical analysis of these data will focus on the difference score, say $Y_{i2} - Y_{i1}$, for each individual. Note that the variability of the difference score is given by

$$\begin{aligned} \text{Var}(Y_{i2} - Y_{i1}) &= \text{Var}(Y_{i1}) + \text{Var}(Y_{i2}) - 2\,\text{Cov}(Y_{i1}, Y_{i2}) \\ &= \sigma_1^2 + \sigma_2^2 - 2\,\sigma_{12} \\ &= \sigma_1^2 + \sigma_2^2 - 2\,\rho_{12}\sigma_1\sigma_2, \end{aligned}$$

where ρ_{12} is the correlation among the pair of responses, Y_{i1} and Y_{i2}. On the other hand, suppose that an alternative study design is adopted to assess the impact of the health intervention. Rather than using a longitudinal design, a cross-sectional design is adopted where study participants are randomly assigned to two groups, a group that receives the intervention and a control group that does not. Then the variance of the difference between the responses of any two individuals, when one individual is randomly selected from the intervention group and the other from the control group, is given by

$$\begin{aligned} \text{Var}(Y_{i2} - Y_{i1}) &= \text{Var}(Y_{i1}) + \text{Var}(Y_{i2}) \\ &= \sigma_1^2 + \sigma_2^2, \end{aligned}$$

(where Y_{i1} and Y_{i2} now denote the responses from two different individuals from the control and intervention groups, respectively).

Thus, provided the correlation among repeated measures is positive, the variability of the within-individual differences is always smaller than the variability of the between-individual differences. If in this simple illustration we further assume that the variance of the response is constant (over time in the longitudinal study design, and across groups in the cross-sectional study design), with $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the variance of the within-individual differences is simply $2\sigma^2(1 - \rho)$, while the variability of the between-individual differences is $2\sigma^2$. The ratio of these two variances provides an index of the precision of within-individual differences when compared to between-individual differences. Their ratio (within-individual variance / between-individual variance) can be expressed as $(1 - \rho)$. Thus, when the correlation is relatively large and positive, the variability of the within-individuals differences (or within-individual changes) can be substantially smaller than that for the corresponding between-individual differences. It is in this sense that a longitudinal study can provide a more precise (i.e., less variable) estimate of change in the mean response than a cross-sectional study with the same number and pattern of observations.

Finally, it must be emphasized that failure to adequately account for the correlation among repeated measures can result in misleading inferences. For instance, if it is assumed that the repeated measures are uncorrelated, when in fact there is strong positive correlation, the nominal standard errors (resulting from the naive assumption of independent or uncorrelated repeated measures) will be incorrect. Specifically, for contrasts that estimate change in the mean response over time, the nominal standard errors will be too large. In this case, one fails to get the full benefit of longitudinal data. With incorrect standard errors, test statistics and p-values will also be incorrect and thus can lead to incorrect inferences about patterns of change and their relation to covariates. In addition, when there is missingness that is MAR, but not MCAR, likelihood-based estimation of the regression parameters, β, requires that the entire joint distribution of the vector of responses be correctly specified. As a result, the model for the covariance must be correctly specified to ensure that valid estimates of β are obtained. In general, when there are missing data, greater care must be exercised when modelling the covariance among the responses (see Chapter 14).

In summary, the positive correlation among repeated measures is an inescapable feature of longitudinal data that must be accounted for in the analysis in order to make appropriate inferences. Although the correlation, or more generally, the covariance among the repeated responses, is not usually of intrinsic interest, it cannot simply be ignored. Moreover, the positive correlation among longitudinal data enables us to estimate changes in the mean response, and their relation to covariates, with far greater precision than would be possible if the data were uncorrelated. Recognizing that the covariance is an important aspect of the data that must be properly accounted for to complete the specification of any regression model for longitudinal data, there are three broad approaches to modelling the covariance that can be distinguished. The first is to allow any arbitrary pattern of covariance among the repeated measures; this approach results in an "unstructured" covariance and is the topic of Section 7.3. Alternatively, structure can be placed on the covariance matrix and there are two main strategies for doing so. The first modelling approach borrows ideas from the time series literature and assumes that the variances and covariances are not arbitrary but

follow distinctive patterns. As a result, we refer to these models as covariance pattern models and they are the topic of Section 7.4. Finally, in a somewhat less direct way, structure can be imposed on the covariance through the introduction of *random effects* in the model for the mean response. That is, by assuming that the mean response depends on a combination of population parameters, β (also known as fixed effects), and individual-specific random effects, one induces a very distinctive structure on the covariance matrix. Because of the important role of the random effects structure in modelling the covariance in longitudinal data, discussion of these models will be the topic of Chapter 8.

7.3 UNSTRUCTURED COVARIANCE

When the number of measurement occasions is relatively small and all individuals are measured at the same set of occasions, it may be reasonable to allow the covariance matrix to be arbitrary, with all of its elements unconstrained. The only formal requirement is that the covariance matrix be symmetric and positive-definite (recall that the latter condition ensures that while the repeated measures can be highly correlated, there must be no redundancy, that is, none of the repeated measures can be expressed as a linear combination of the others). When no explicit structure is assumed for the covariance among the repeated measures (other than the homogeneity of covariance across different individuals, $\text{Cov}(Y_i) = \Sigma_i = \Sigma$), the resulting covariance is referred to as an "unstructured" covariance. The chief advantage of an "unstructured" covariance is that no assumptions are made about the variances and covariances. The absence of restrictions on the variances is especially important since our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. For example, the variability of baseline measurements is often discernibly different from the variability of post-baseline measurements.

With n measurement occasions, the "unstructured" covariance matrix has $\frac{n\times(n+1)}{2}$ parameters: the n variances at each occasion and the $n \times (n-1)/2$ pairwise covariances (or correlations),

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

Herein lies one of the potential drawbacks of assuming an unstructured covariance: the number of covariance parameters to be estimated grows rapidly with the number of measurement occasions. For example, when there are three occasions ($n = 3$), the number of covariance parameters is 6 (3 variances and 3 pairwise covariances). However, when $n = 5$, the number of covariance parameters has grown to 15, while when $n = 10$ the number of covariance parameters is 55 (and may be fast approaching

the number of subjects enrolled in some longitudinal studies!). When the number of covariance parameters that need to be estimated is large, relative to the sample size, estimation is likely to be very unstable. Thus the use of an unstructured covariance will be appealing only in cases where the number of subjects, N, is large relative to the number of covariance parameters, $\frac{n\times(n+1)}{2}$.

Setting aside the issue of the potentially large number of covariance parameters that may need to be estimated, the use of an unstructured covariance matrix is problematic when there are mistimed measurements or, more generally, measurement made at grossly irregular intervals. Even the most carefully designed longitudinal study will frequently suffer from deviations from the measurement protocol, resulting in measurements made at arbitrary, irregularly timed intervals. When this problem arises, as it frequently does in studies in the health sciences, the resulting mistimed repeated measurements cannot be accommodated in an unstructured covariance. Thus, when the longitudinal data are inherently unbalanced and/or when the sample size is not sufficiently large to estimate an unstructured covariance, it is usually desirable to impose some structure on the covariance matrix.

7.4 COVARIANCE PATTERN MODELS

When attempting to impose some structure on the covariance, a subtle balance needs to be struck. If too little structure is imposed there may be too many parameters to estimate with the limited amount of data at hand. This was one of the main drawbacks of the unstructured covariance considered in the previous section; by imposing no structure on the covariance, the number of parameters to be estimated grows rapidly with the number of measurement occasions. In a certain sense, any given data set contains but a fixed amount of longitudinal information. If too little structure is imposed on the covariance, there will be too many covariance parameters to be estimated from the limited amount of data available and this will adversely affect the precision with which the main parameters of interest, β, can be estimated. As a result, imposing too little structure on the covariance can result in weaker inferences concerning β. When structure is imposed on the covariance, it is possible to improve the precision with which β can be estimated. However, if too much structure is imposed, there is a potential risk of model misspecification that could ultimately result in misleading inferences concerning β. Once again, this is the classic tradeoff between bias and precision. In modelling the covariance, a balance must be struck between these two competing forces.

Structure can be built into the covariance by adopting a covariance pattern model. Covariance pattern models for longitudinal data have their basis in models for serial correlation that were originally developed for time series data. While time series data have a structure that is somewhat different than longitudinal data, being composed of a small number of replications or individuals (in some cases only a single replication) and a large number of repeated measures, they share a common characteristic: the repeated measures are positively correlated and measures taken closer together in

time are expected to be more highly correlated than measures further apart in time. Because there are few, if any, replications, much of the statistical literature on the analysis of time series data has focused on parametric models that can describe the covariance structure among the repeated measures with only a few parameters. Many of the models for time series data result in relatively parsimonious models for the covariance that can also be adopted for longitudinal data. Here we describe some of the most widely used covariance pattern models for longitudinal data. Many of these covariance pattern models are available as options in standard statistical software packages for analyzing longitudinal data (e.g., PROC MIXED in SAS).

Compound Symmetry

Historically, one of the first covariance pattern models used for the analysis of repeated measures data was compound symmetry. With a compound symmetry covariance it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$ for all j and k. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix},$$

with the constraint that $\rho \geq 0$.

The compound symmetry covariance has a randomization justification in certain repeated measures designs (e.g., split-plot designs). In an experiment where the within-subject factor is randomly allocated to subjects, randomization arguments can be made to show that the constant variance and constant correlation conditions hold. (See Chapter 16 for a more detailed discussion of the randomization argument.) However, the randomization argument is simply not justifiable in the longitudinal data setting since measurement occasions cannot be randomly allocated to subjects.

As mentioned in Chapter 3, the compound symmetry covariance does have a theoretical justification when the mean response is thought to depend on a combination of population parameters, β, and a single individual-specific random effect. When the model for the longitudinal responses is expressed as

$$Y_{ij} = X_{ij}'\beta + b_i + e_{ij},$$

where b_i is a random effect and e_{ij} is a within-individual measurement error, this induces marginally (or averaged over the random effect) a compound symmetry structure on the covariance matrix (with the constraint that $\rho \geq 0$). A more detailed discussion of random effects structures for the covariance will be given in Chapter 8.

The compound symmetry covariance is very parsimonious, with only two parameters regardless of the number of measurement occasions. However, it does make the rather strong assumption that the correlation between any pair of measurements

is the same regardless of the time interval between the measurements. This latter aspect of the compound symmetry covariance, the constraint on the correlation among repeated measurements, is somewhat unappealing for most longitudinal data, where the correlations are expected to decay with increasing separation in time. Also, the assumption of constant variance across time is unrealistic in many settings. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. In many settings the assumption of constant variance is the one that is not valid with longitudinal data.

Toeplitz

The Toeplitz covariance pattern makes the assumption that any pair of responses that are equally separated in time have the same correlation. When the covariance has a Toeplitz form it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho_k$ for all j and k. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

Because a Toeplitz covariance assumes that the correlation among responses at adjacent measurement occasions is constant, ρ_1, this structure is only appropriate when the measurements are made at equal (or approximately equal) intervals of time. Note that the Toeplitz covariance has n parameters (1 variance parameter, and $n-1$ correlation parameters). A special case of the Toeplitz covariance is the (first-order) autoregressive covariance.

Autoregressive

In the autoregressive model for the covariance it is assumed that the variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho^k$ for all j and k, and $\rho \geq 0$. That is,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

The autoregressive covariance is very parsimonious and has only two parameters, regardless of the number of measurement occasions. Because the autoregressive

covariance has a Toeplitz form, this structure is only appropriate when the measurements are made at equal (or approximately equal) intervals of time. Note that the correlations decline over time as the separation between pairs of repeated measures increases. However, as mentioned in Section 2.5, in many settings the correlations among repeated measures on the same individual rarely decay that quickly over time.

The autoregressive covariance has a theoretical justification when the errors, e_{ij}, are thought of as arising from the following first-order autoregressive process:

$$e_{ij} = \rho\, e_{ij-1} + w_{ij},$$

where $w_{ij} \sim N\left(0, \sigma^2\left[1-\rho^2\right]\right)$ and the process is initiated by an error, say e_{i0}, where $e_{i0} \sim N\left(0, \sigma^2\right)$. The autoregressive process is said to be "first-order" because there is only dependence on the previous error; dependence on the two previous errors would yield a "second-order" autoregressive process. Thus, the autoregressive covariance can be thought of as resulting from a process where the error term at the j^{th} occasion is a deterministic function of the error at the previous occasion, $\rho\, e_{i,j-1}$ (i.e., the recent past predicts the present), plus an additional (and independent) source of random error, w_{ij}. For such a process, it can be shown that

$$\text{Var}\left(e_{ij}\right) = \sigma^2,$$

and

$$\text{Cov}\left(e_{ij}, e_{ik}\right) = \sigma^2 \rho^{|j-k|}.$$

Finally, the compound symmetry, Toeplitz, and autoregressive covariances assume that the variances are constant across time. This assumption can easily be relaxed and it is possible to consider versions of these three covariance pattern models with heterogeneous variances, $\text{Var}(Y_{ij}) = \sigma_j^2$. Thus a heterogeneous (variances) autoregressive covariance pattern model is given by

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \cdots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \cdots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \cdots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \cdots & \sigma_n^2 \end{pmatrix},$$

and has $n+1$ parameters (n variance parameters and 1 correlation parameter).

Banded

The banded covariance patterns make the assumption that the correlation is zero beyond some specified interval. For example, a banded covariance pattern with a band size of 3 assumes that $\text{Corr}(Y_{ij}, Y_{ij+k}) = 0$ for $k \geq 3$. It is possible to apply a banded pattern to any of the covariance pattern models considered so far. Thus, a

banded Toeplitz covariance pattern with a band size of 2 is given by

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ 0 & \rho_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

where $\rho_2 = \rho_3 = \cdots = \rho_{n-1} = 0$.

Banding makes a very strong assumption about how quickly the correlation decays to zero with increasing separation between the repeated measurements. In our experience with longitudinal studies in the health sciences, it is rare for the correlation to decay to zero, even in studies where there is a lengthy period of follow-up.

Exponential

When the measurement occasions are not equally spaced over time, the formulation of the autoregressive covariance model can be generalized as follows: Let $\{t_{i1}, \ldots, t_{in}\}$ denote the observation times for the i^{th} individual and assume that the variance is constant across all measurement occasions, say σ^2, and

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|},$$

for $\rho \geq 0$. That is, the correlation between any pair of repeated measures decreases exponentially with the time separations between them. This structure is referred to as an "exponential" covariance model because it can be re-expressed as

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij}-t_{ik}|} \\ &= \sigma^2 \exp\left(-\theta\,|t_{ij} - t_{ik}|\right), \end{aligned}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$. Also, note that the exponential covariance model is invariant under linear transformation of the time scale. If we replace t_{ij} by $(a + bt_{ij})$ (e.g., if we replace time measured in "weeks" by time measured in "days"), the same form for the covariance matrix holds.

A distinctive feature of the exponential model is that it assumes that the correlation is one if measurements are made repeatedly at the same occasion (or replicate measurements on an individual can be obtained at the same occasion), and that the correlation decreases rapidly to zero as the time separation between measurements increases. This first aspect of the exponential covariance model corresponds to an assumption that the responses are measured without error; an unrealistic assumption in most longitudinal studies in the health sciences. The latter feature, correlations among repeated measurements that decay to zero, is rarely observed in longitudinal studies.

Hybrid Models

Finally, by combining the autoregressive and the compound symmetry models, it is possible to overcome many of the unappealing aspects of each of these models for longitudinal data. Consider a model for the covariance where

$$\mathrm{Cov}(Y_i) = \Sigma_1 + \Sigma_2,$$

where

$$\Sigma_1 = \sigma_1^2 \begin{pmatrix} 1 & \rho_1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & \rho_1 & 1 & \dots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \dots & 1 \end{pmatrix},$$

and

$$\Sigma_2 = \sigma_2^2 \begin{pmatrix} 1 & \rho_2^{|t_{i1}-t_{i2}|} & \rho_2^{|t_{i1}-t_{i3}|} & \dots & \rho_2^{|t_{i1}-t_{in}|} \\ \rho_2^{|t_{i2}-t_{i1}|} & 1 & \rho_2^{|t_{i2}-t_{i3}|} & \dots & \rho_2^{|t_{i2}-t_{in}|} \\ \rho_2^{|t_{i3}-t_{i1}|} & \rho_2^{|t_{i3}-t_{i2}|} & 1 & \dots & \rho_2^{|t_{i3}-t_{in}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2^{|t_{in}-t_{i1}|} & \rho_2^{|t_{in}-t_{i2}|} & \rho_2^{|t_{in}-t_{i3}|} & \dots & 1 \end{pmatrix}.$$

In this model,

$$\mathrm{Var}(Y_{ij}) = \sigma_1^2 + \sigma_2^2,$$

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2,$$

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

This implies that the correlation between replicate measurements on an individual obtained at the same occasion is

$$\frac{\rho_1\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

which is less than one when $\rho_1 < 1$. Furthermore, as the time separation increases, the correlation no longer decays to zero but has a minimum of

$$\frac{\rho_1\sigma_1^2}{\sigma_1^2 + \sigma_2^2},$$

which is greater than zero provided $\rho_1 > 0$. As noted, the compound symmetry model is also a random effects model, so that Σ_1 can be written as

$$\Sigma_1 = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_e^2 \end{pmatrix},$$

so that $\sigma_1^2 = \sigma_b^2 + \sigma_e^2$, and $\rho_1 = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$. Thus we can think of the total variance, $\mathrm{Var}(Y_{ij})$, as the sum of the autoregressive variance, σ_2^2, subject-to-subject variability, σ_b^2, and measurement error variability, σ_e^2.

7.5 CHOICE AMONG COVARIANCE PATTERN MODELS

As mentioned at the beginning of this chapter, the choices of models for the covariance and for the mean are interdependent. As a result, it is important to follow a modelling strategy that will result in a sensible choice of models for both aspects of the data. Since model selection criteria for the mean response depend upon the correct specification of the model for the covariance (e.g., confidence intervals and tests of hypotheses concerning components of β depend critically upon the correct model for the covariance), the first step is to choose a suitable model for the covariance.

It must be recognized that any model for the covariance depends on the assumed model for the mean. A model for the covariance tries to account for the covariance among the residuals, say $\{Y_{ij} - \mu_{ij}(\beta)\}$ and $\{Y_{ik} - \mu_{ik}(\beta)\}$, that result from a specific model for the mean. Therefore, the choice of model for the covariance should be based on a "maximal" model for the mean that minimizes any potential misspecification of the model for the mean. Recall that any misspecification of the model for the mean can result in a certain amount of spurious covariance among the residuals, and can induce spurious dependence of the covariance on the covariates.

In longitudinal studies with balanced designs and a very small number of discrete covariates that can be classified as between-subject factors (e.g., treatment assignments, exposure levels, or some characteristic of the subjects), the choice of maximal model is relatively straightforward since it is possible to choose as the maximal model one that includes the main effects of time (regarded as a within-subject factor) and all other main effects, in addition to their two- and higher-way interactions. For example, with n measurement occasions and a single grouping factor with G levels (e.g., treatment versus control), it is possible to fit a saturated model for the mean response with separate parameters for the $G \times n$ means. This corresponds to a model with main effects for both the grouping factor and time, in addition to their interaction. This strategy of fitting saturated models for the mean response will be appropriate for longitudinal studies with balanced designs and where the number of qualitatively different levels of the covariates is relatively small. A saturated model for the mean

response allows an arbitrary pattern for the mean response profile at every different level of the covariates and thereby minimizes any potential concerns about the impact of misspecification of the model for the mean.

However, in longitudinal studies where there are many covariates (some of which may be quantitative, rather than discrete), the choice of a maximal model is somewhat more difficult. In this case it is not realistic to consider a saturated model for the mean response; instead a maximal model should be in a certain sense the most elaborate or complex model for the mean response that we would consider from a subject-matter point of view. Such a model may need to distinguish treatment covariates (e.g., treatment groups in experiments) or quasi-treatment covariates (e.g., exposure groups in observational studies) that are the main focus of the study from other covariates that are regarded as potential confounders or effect modifiers. The maximal model will ordinarily include the main effects of the treatment or quasi-treatment covariates and their interactions with time, since the latter effects characterize how changes in the mean response depend on these covariates. The choice of whether to include additional interactions, and so on, must be made on subject-matter grounds. In summary, when there are many potential covariates that can be included in the model for the mean, it is not straightforward to give a simple prescription for choosing the maximal model. The choice of maximal model, it must be recognized, cannot be made through any automatic procedure but must, rather, reflect substantive subject-matter considerations. The maximal model for the mean is a model that excludes certain higher-order interactions among the potential covariates and usually is more complex than any of the sequence of models for the mean response under consideration from a subject-matter point of view. In a sense, the reader should envisage a model that, in its degree of complexity, goes a step beyond any model for which empirical researchers in the field would care to provide a specific rationale. Once a maximal model has been chosen, the residual variation and covariation can then be used to select an appropriate model for the covariance.

Given a maximal model for the mean, a sequence of covariance pattern models can be fit to the data at hand. The choice among models can be made by comparing the maximized likelihoods for each of the covariance pattern models. That is, when any pair of models is nested, a likelihood ratio test statistic can be constructed that compares the "full" and "reduced" models. Recall that two covariance models are said to be nested when the "reduced" model is a special case of the "full" model, so that when the reduced model holds the full model must necessarily hold. For example, the compound symmetry model is nested within the Toeplitz model since if the compound symmetry model holds, then the Toeplitz model must necessarily hold, with $\rho_1 = \rho_2 = \cdots = \rho_{n-1}$. The likelihood ratio test for two nested covariance model can be constructed by comparing the maximized REML log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, for the full and reduced models, respectively. The use of REML, as an alternative to ML, is preferred because it reduces the well-known finite sample bias in the estimation of the covariance. The likelihood ratio test is obtained by taking twice the difference in the respective maximized REML log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing the statistic to percentiles from a chi-squared distribution with degrees of freedom equal to the difference between the number of covariance parameters in the full and reduced models.

In general, the likelihood ratio test provides a valid method for comparing nested models for the covariance. However, in certain cases, the likelihood ratio test may not be valid depending upon the nature of the null hypothesis that is being tested. In particular, when the likelihood ratio test is testing a null hypothesis that is "on the boundary of the parameter space", the usual conditions required for classical likelihood theory no longer apply. What is meant by testing a null hypothesis that is "on the boundary of the parameter space"? This rather technical point is best illustrated by considering variances. Recall that variances cannot be negative, they must be positive. As a result, variances are considered to be bounded from 0 to ∞. Thus a likelihood ratio test of the null hypothesis that a variance is zero is testing a null hypothesis that is "on the boundary of the parameter space" for a variance. One consequence is that the usual null distribution for the likelihood ratio test is no longer valid. That is, the null distribution for the likelihood ratio test is no longer a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models; instead, the null distribution is a mixture of chi-squared distributions. To illustrate the problem, consider the following model where the mean depends on a combination of population parameters, β, and a single individual-specific random effect,

$$Y_{ij} = X'_{ij}\beta + b_i + e_{ij},$$

where b_i is a random effect and e_{ij} is a within-individual measurement error, with variances σ_b^2 and σ^2, respectively. Marginally, this model induces a compound symmetry covariance, subject to the constraint that

$$\rho = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\tau_b^2}{\sigma_b^2 + \sigma^2} \geq 0.$$

Note that a test of the null hypothesis, H_0: $\sigma_b^2 = 0$ versus H_A: $\sigma_b^2 > 0$, is equivalent to a test of the null hypothesis, H_0: $\rho = 0$ versus H_A: $\rho > 0$, (subject to the constraint that ρ is non-negative). Under the null hypothesis the repeated measures are assumed to be uncorrelated; under the alternative hypothesis, they are assumed to be positively correlated. In both instances, the null hypothesis is testing on the boundary of the parameter space (i.e., testing that a variance is zero or testing that a non-negative correlation is zero). As a result, the null distribution of the likelihood ratio test is not a chi-squared distribution with 1 degree of freedom. Instead, it is an equally weighted mixture of chi-squared distributions with 0 and 1 degrees of freedom (a chi-square distribution with 0 degrees of freedom has all of its mass or probability at zero). Some intuition for why it is a mixture of chi-squared distributions with 0 and 1 degrees of freedom can be obtained by considering the fit of the model to the data under the null hypothesis. When H_0: $\rho = 0$ is true, the fit of the model to the data is equally likely to show some evidence of positive or negative correlation among the responses due to sampling variability. When there is evidence of positive correlation $\widehat{\rho}$ will

be positive, but when there is evidence of negative correlation $\widehat{\rho}$ will be zero (since under the alternative hypothesis, H_A: $\rho > 0$, ρ is constrained to be non-negative). When H_0: $\rho = 0$ is true, there is a 50:50 chance that $\widehat{\rho} > 0$ (or $\widehat{\rho} = 0$). As a result, $\widehat{\rho}$ only makes contributions to the likelihood ratio test statistic approximately half of the time, when $\widehat{\rho}$ is positive. The distribution of the likelihood ratio test statistic can be thought of as chi-squared with 1 degree of freedom half of the time, when $\widehat{\rho}$ is positive (and chi-squared with 0 degrees of freedom, when $\widehat{\rho}$ is zero, the other half of the time).

A more detailed discussion of the null distribution of the likelihood ratio test under non-standard conditions is beyond the scope of this book. However, the reader should be aware that the comparison of models for the covariance can sometimes be a non-standard problem. In general, when testing a null hypothesis that is on the boundary of the parameter space, the usual null distribution for the likelihood ratio test is no longer valid. If this problem is simply ignored, and the standard null distribution is naively used, the resulting p-value for the likelihood ratio test will be overestimated (i.e., a p-value that is too large will be obtained). Consequently, failure to account for this problem can lead to the selection of a model for the covariance that is too parsimonious. That is, there is a danger that the model for the covariance is too simple and ignores some inherent structure in the covariance. Because it is not straightforward to determine the correct null distribution for the likelihood ratio test in these non-standard settings, we recommend the use of $\alpha = 0.1$, instead of $\alpha = 0.05$, when judging the statistical significance of the likelihood ratio test. Use of the $\alpha = 0.1$ level is a somewhat *ad hoc* solution but protects against selection of a model for the covariance that is too parsimonious. Alternatively, for cases where the null distribution is a known 50:50 mixture of chi-squared distributions, the critical values given in Table C.1 in Appendix C can be used (see Section 8.5 for additional discussion of this topic).

Often it is of interest to compare non-nested models for the covariance. To compare non-nested models, an alternative approach is the Akaike Information Criterion (AIC). According to the AIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned} \text{AIC} &= -2(\text{maximized log-likelihood}) + 2(\text{number of parameters}) \\ &= -2(\widehat{l} - c), \end{aligned}$$

where $\widehat{l}$ is the maximized REML log-likelihood and c is the number of covariance parameters. Note that AIC can similarly be defined as selecting the model that minimizes

$$\text{AIC} = -\widehat{l} + c,$$

or the model that maximizes

$$\text{AIC} = \widehat{l} - c.$$

Although AIC can be defined in a number of different ways, the basic underlying idea behind AIC is to strike a balance between two competing objectives: the covariance

model must be sufficiently complex to provide a good fit to the data, but at the same time a premium is attached to a parsimonious model. This is achieved by extracting a penalty for the estimation of each additional covariance parameter. With these definitions of AIC, it can be used to compare models with the same fixed effects (i.e., the same model for the mean), but different models for the covariance. Note that expanding the definition of c to include the number of fixed effects parameters, β, would not alter the selection of the model for the covariance provided the model for the mean is held constant.

We note that AIC is but one of a variety of different "information criteria" that have been proposed. Another criterion is the Bayesian Information Criterion (BIC). According to the BIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned} \text{BIC} &= -2(\text{maximized log-likelihood}) + \log N^*(\text{number of parameters}) \\ &= -2(\widehat{l} - \log\sqrt{N^*}\, c), \end{aligned}$$

where N^* is the number of subjects. The BIC is sometimes defined where N^* is the number of "effective subjects", N in the case of ML estimation and $N - p$ in the case of REML estimation (where p is the dimension of β). The main idea underlying BIC requires some understanding of the Bayesian approach to model selection where the objective is to choose the model that has the highest posterior probability (or largest Bayes factor). While this is a legitimate model selection criterion, it must be emphasized that BIC only approximates this Bayesian criterion; furthermore, the BIC extracts a very large penalty for the estimation of each additional covariance parameter. In general, we do not recommend the use of BIC for covariance model selection as it entails a high risk of selecting a model that is too simple or parsimonious for the data at hand.

Finally, as mentioned earlier, inferences about β depend upon the correct specification of the model for the covariance. Recall that confidence intervals and tests of hypotheses concerning components of β rely on standard errors that are obtained by substituting the REML estimate of Σ_i in the expression for $\text{Cov}(\widehat{\beta})$ (see Eq. (4.5) in Section 4.2). Any misspecification of the model for the covariance has negligible impact on the estimates of the regression coefficients, that is, the regression parameter estimates are unbiased even when the covariance has been misspecified. However, misspecification of the covariance results in incorrect standard errors and this can lead to potentially misleading inferences concerning β (e.g., due to confidence intervals that are too narrow or wide and p-values that are too small or large). Fortunately, in many cases, valid standard errors for $\widehat{\beta}$ can be obtained when there is concern about misspecification of the covariance. In particular, valid standard errors for $\widehat{\beta}$ can be based on the so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; these standard errors are robust to any misspecification of the covariance. Although the "sandwich" estimator is more widely used in the marginal models for discrete longitudinal data that are the focus of Chapter 11, we note that the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ can be applied also in the linear models for longitudinal continuous data described in Part II. The "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ will be discussed in greater detail in Chapter 11.

Table 7.1 Estimated unstructured covariance matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	9.668	10.175	8.974	9.812	9.407
4	10.175	12.550	11.091	12.580	11.928
6	8.974	11.091	10.642	11.686	11.101
8	9.812	12.580	11.686	13.990	13.121
12	9.407	11.928	11.101	13.121	13.944

7.6 CASE STUDY

Next, we illustrate the main ideas by considering covariance pattern models for data from a trial examining the effectiveness of two different exercise therapy regimens.

Exercise Therapy Trial

In this study, subjects were assigned to one of two weightlifting programs to increase muscle strength. In the first program, hereafter referred to as treatment 1, the number of repetitions of the exercises was increased as subjects became stronger. In the second program, hereafter referred to as treatment 2, the number of repetitions was held constant but the amount of weight was increased as subjects became stronger. Measurements of muscle strength were taken at baseline and on days 2, 4, 6, 8, 10, and 12. However, to illustrate some of the main differences among the covariance models considered earlier, we focus only on measures of strength obtained at baseline (or day 0) and on days 4, 6, 8, and 12.

Before considering models for the covariance, it is necessary to choose a maximal model for the mean response. Here, with a balanced design on time and only two groups, we chose the maximal model to be the saturated model for the mean, with a total of 10 parameters for the response profiles for the two treatment groups.

First, we consider an unstructured covariance matrix, with all 15 of its elements unconstrained. The estimated covariance and correlation matrices are displayed in Tables 7.1 and 7.2, respectively. Note that the variance is larger by the end of the study when compared to the variance at baseline; this is a characteristic pattern observed in many longitudinal studies. Furthermore, from examination of Table 7.2, the correlations decrease as the time separation between the repeated measures increases.

Despite the apparent increase in the variance over time, we consider an autoregressive model for the covariance. This model is very parsimonious, with only two

Table 7.2 Estimated unstructured correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9237	0.8847	0.8437	0.8102
4	0.9237	1.0000	0.9597	0.9494	0.9017
6	0.8847	0.9597	1.0000	0.9577	0.9113
8	0.8437	0.9494	0.9577	1.0000	0.9394
12	0.8102	0.9017	0.9113	0.9394	1.0000

parameters, one describing the variance, σ^2, the other the correlation, ρ. When a first-order autoregressive model is fit to the data it results in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.94$. The resulting estimated pairwise correlations among the five repeated measurements are given in Table 7.3. This model was fit primarily for illustrative purposes; the model is not very appropriate for these data as they are unequally spaced over time (i.e., there is a four day interval between the first two repeated measures and the last two repeated measures, but all other adjacent repeated measurements were taken two days apart). In order to account for the unequal time interval, an exponential model for the covariance was considered, where

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|},$$

for $t_{i1} = 0, t_{i2} = 4, t_{i3} = 6, t_{i4} = 8$, and $t_{i5} = 12$ for all subjects. This resulted in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.98$. The resulting estimated pairwise correlations among the 5 repeated measurements are given in Table 7.4. Of note, the declines in the estimated correlations in Tables 7.3 and 7.4 are too fast when compared to the corresponding declines in Table 7.2.

Next we consider the choice among these covariance pattern models. The maximized REML log-likelihood and AIC for each of the covariance pattern models are displayed in Table 7.5. Note that there is a hierarchy among the models. The autoregressive and exponential models are both nested within the unstructured covariance. That is, if either the autoregressive or exponential model holds, then the unstructured covariance must necessarily hold. Comparisons of the autoregressive and exponential models with the unstructured covariance can be made using (REML) likelihood ratio tests. However, the autoregressive and exponential models are not nested models; indeed, both models have the same number of parameters. As a result, any comparison between these two models can be made directly in terms of their maximized log-likelihoods, since any penalty extracted by information criteria will be the same in both cases (e.g., with AIC a penalty of 4 is extracted for the estimation of the two

Table 7.3 Estimated autoregressive correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9402	0.8839	0.8311	0.7813
4	0.9402	1.0000	0.9402	0.8839	0.8311
6	0.8839	0.9402	1.0000	0.9402	0.8839
8	0.8311	0.8839	0.9402	1.0000	0.9402
12	0.7813	0.8311	0.8839	0.9402	1.0000

Table 7.4 Estimated exponential correlation matrix for the strength data at baseline (day 0), day 4, day 6, day 8, and day 12 from the exercise therapy trial.

Day	0	4	6	8	12
0	1.0000	0.9169	0.8780	0.8408	0.7709
4	0.9169	1.0000	0.9576	0.9169	0.8408
6	0.8780	0.9576	1.0000	0.9576	0.8780
8	0.8408	0.9169	0.9576	1.0000	0.9169
12	0.7709	0.8408	0.8780	0.9169	1.0000

covariance parameters). The likelihood ratio test, comparing the autoregressive and unstructured covariance, yields

$$G^2 = 621.1 - 597.3 = 23.8,$$

and can be compared to a chi-squared distribution with 13 (or 15 − 2) degrees of freedom. On the basis of the likelihood ratio test there is evidence that the autoregressive model does not provide an adequate fit to the covariance, when compared to the unstructured covariance ($p < 0.05$). On the other hand, the likelihood ratio test, comparing the exponential and unstructured covariance, yields

$$G^2 = 618.5 - 597.3 = 21.2,$$

Table 7.5 Comparison of the maximized (REML) log-likelihoods and AIC for the covariance pattern models for the strength data from the exercise therapy trial.

Covariance Pattern Model	−2 (REML) Log-Likelihood	AIC
Unstructured	597.3	627.3
Autoregressive	621.1	625.1
Exponential	618.5	622.5

and when compared to a chi-squared distribution with 13 degrees of freedom, $p > 0.05$. Thus the exponential covariance provides an adequate fit to the data. Also, in terms of AIC, the exponential model minimizes this criterion.

7.7 DISCUSSION: STRENGTHS AND WEAKNESSES OF COVARIANCE PATTERN MODELS

The defining feature of covariance pattern models is that they attempt to account for all the potential sources of variability that have an impact on the covariance among repeated measures on the same individual. That is, they do not distinguish between-subject and within-subject sources of variability. Covariance pattern models characterize the covariance among longitudinal data with a relatively small number of parameters. Many of the models (e.g., autoregressive, Toeplitz, and banded) are only appropriate when the repeated measurements are obtained at equal intervals and cannot handle irregularly timed measurements. Although there is a large selection of models for the correlations, the choice of models for the variances is somewhat limited. Covariance pattern models either make the strong assumption that the variances are constant over time, or relax this assumption entirely and allow the variances to depend arbitrarily on time.

For the most part, covariance pattern models are appropriate for balanced longitudinal designs, and many require that the repeated measurements are obtained at equal intervals. Although these models can handle imbalance due to missing data at any of the fixed occasions, they are not well suited for modelling data from inherently unbalanced longitudinal designs. With inherently unbalanced designs, many of the covariance pattern models are not well defined. In an attempt to overcome the latter limitation, a few covariance pattern models have been developed that allow for irregularly timed measurements (e.g., the exponential covariance pattern model). In these models the correlation is assumed to depend upon the time separation between pairs of repeated measurements. However, a potential problem with these models is that they assume the correlation decays rapidly with increasing time separation and that the correlation between two measurements taken at the same occasion is one. As

Table 7.6 Covariance pattern modelling options using PROC MIXED in SAS.

TYPE =	<pattern>	Specifies the covariance pattern
	UN	Unstructured
	CS	Compound symmetry
	AR(1)	First-order autoregressive
	TOEP	Toeplitz
	UN(n)	Banded unstructured, with n bands
	CSH	Heterogeneous compound symmetry
	ARH(1)	Heterogeneous first-order autoregressive

mentioned earlier, in our experience with longitudinal studies in the health sciences, the correlation among repeated measures rarely exhibits either of these two characteristics. Furthermore, although these covariance pattern models allow the correlation to depend on the time separation between repeated measurements, they do not allow the variances to depend on time. As a result, they make the strong and often unrealistic assumption that the variance remains constant over time.

In conclusion, covariance pattern models are appropriate for balanced longitudinal designs and many models require that the repeated measurements are obtained at equal intervals. In general, we do not recommend the use of covariance pattern models that make the strong assumption that the variances are constant over time. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. Because the assumption of constant variance is the one that is not valid in many setting, we recommend that covariance pattern models with heterogeneous variances, allowing the variances to depend arbitrarily on time, should generally be adopted.

7.8 COMPUTING: FITTING COVARIANCE PATTERN MODELS USING PROC MIXED IN SAS

In the following we assume that there is a single group factor and the maximal model is the saturated model for the mean. Different patterns can be fit to the covariance matrix among the residuals, denoted R in PROC MIXED in SAS, by using the TYPE= option on the REPEATED statement. Table 7.6 provides a summary of some of the commonly used covariance pattern models; a full description of all of the options can be found in the SAS documentation.

For example, to fit an autoregressive model for the covariance we can use the illustrative SAS commands given in Table 7.7. The options R and RCORR on the

Table 7.7 Illustrative commands for an autoregressive model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS  id  group  time;
   MODEL  y=group  time  group*time / S  CHISQ;
   REPEATED  time / TYPE=AR(1)  SUBJECT=id  R  RCORR;
```

Table 7.8 Illustrative commands for an exponential model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS  id  group  time;
   MODEL  y=group  time  group*time / S  CHISQ;
   REPEATED  time / TYPE=SP(EXP)(ctime)  SUBJECT=id  R RCORR;
```

REPEATED statement request that the estimated covariance matrix (R) and the corresponding correlation matrix be displayed as part of the output. By default, the covariance and correlation matrices are displayed for the first subject and will have row and column dimensions corresponding to the number of repeated measures obtained on the first subject. When the vector of responses on the first subject is incomplete it may be preferable to display the covariance and correlation matrices for a subject with complete responses. The options R=1,5,7 and RCORR=1,5,7 request that the estimated covariance and correlation matrices be displayed for the first, fifth, and seventh subjects.

To fit covariance pattern models to inherently unbalanced data requires the use of the "spatial" covariance pattern options in PROC MIXED. These are covariance pattern models developed for spatial data that are defined in terms of "distances" in two-dimensional space. However, these options can also be used where "distance" (or time separation) is defined along the single dimension of time. For example, to fit an exponential covariance pattern model, the following option is used:

TYPE = SP(EXP)(list)

where `list` is the name of the variable used to construct "distances" or time separation between repeated measurements. Table 7.8 contains illustrative commands for fitting an exponential covariance pattern model. Note that the variable `ctime` is

simply an additional copy of `time` that is treated as a continuous covariate for the purpose of constructing the time separation between repeated measurements.

Finally, when the EMPIRICAL option is included on the PROC MIXED statement standard errors for $\widehat{\beta}$ are based on the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. As mentioned earlier, these standard errors are robust to any misspecification of the model for the covariance. The "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ will be discussed in Chapter 11.

7.9 FURTHER READING

Additional discussion of covariance pattern models can be found in Chapter 6, Section 6.2, of Brown and Prescott (1999) and in the tutorial by Littell *et al.* (2000).

Bibliographic Notes

Jennrich and Schluchter (1986) describe covariance pattern models for longitudinal data. For a more recent and comprehensive overview of this topic, see the review article by Zimmerman and Nunez-Anton (2001), and the references therein. Finally, Pourahmadi (1999) presents a flexible approach for parametric modelling of the covariance structure.

Altham (1984) discusses the advantages, in terms of increased precision of estimation of the parameters of interest, that can result from fitting a parsimonious model to complex data. Altham's (1984) general discussion of this issue has great relevance for the modelling of the covariance in longitudinal data.

The large-sample distribution theory for testing a null hypothesis that is "on the boundary of the parameter space" (e.g., testing that a variance is zero) is discussed in Miller (1977), Self and Liang (1987), Stram and Lee (1994, 1995), Silvapulle and Silvapulle (1995), Silvapulle (1996), and Verbeke and Molenberghs (2003).

Problems

7.1 In a study of dental growth, measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14 (Potthoff and Roy, 1964).

The raw data are stored in an external file: `dental.dat`

Each row of the data set contains the following six variables:

ID Gender Y_1 Y_2 Y_3 Y_4

Note: The categorical (character) variable Gender is coded F = Female, M = Male. The 3rd measure (at age 12) on subject ID = 20 is a potential outlier.

7.1.1 On a single graph, construct a time plot that displays the mean distance (mm) versus age (in years) for boys and girls. Describe the time trends for boys and girls.

7.1.2 Read the data from the external file and put the data in a "univariate" or "long" format, with 4 "records" per subject.

7.1.3 For the "maximal" model, assume a saturated model for the mean response. Fit the following models for the covariance:

(a) unstructured covariance
(b) compound symmetry
(c) heterogeneous compound symmetry
(d) autoregressive
(e) heterogeneous autoregressive

Choose a model for the covariance that adequately fits the data.

7.1.4 Given the choice of model for the covariance from Problem 7.1.3, treat age (or time) as a categorical variable and fit a model which includes the effects of age, gender, and their interactions. Determine whether the pattern of change over time is different for boys and girls.

7.1.5 Show how the *estimated* regression coefficients from Problem 7.1.4 can be used to estimate the means in the two groups at ages 8 and 14.

7.1.6 Given the choice of model for the covariance from Problem 7.1.3, treat age as a continuous variable and fit a model which includes the effects of a linear trend in age, gender, and their interaction. Compare and contrast the results with those obtained in Problem 7.1.4.

7.1.7 On a single graph, construct a time plot that displays the *estimated* mean distance (mm) versus age (in years) for boys and girls from the results generated from Problem 7.1.6.

7.1.8 Show how the regression coefficients from Problem 7.1.6 can be used to estimate the means in the two groups at ages 8 and 14.

7.1.9 Does a model with only a linear trend in age adequately account for the pattern of change in the two groups?

7.1.10 The 3rd measure (at age 12) on subject ID = 20 is a potential outlier. Repeat the analyses in Problems 7.1.3, 7.1.4, 7.1.6 and 7.1.9 excluding the 3rd measure on subject ID = 20. Do the substantive conclusions change?

7.1.11 Given the results of all the previous analyses, what conclusions can be drawn about gender differences in patterns of dental growth?

8

Linear Mixed Effects Models

8.1 INTRODUCTION

In Chapters 5 and 6 we introduced models for longitudinal data where changes in the mean response, and their relation to covariates, can be expressed as

$$E(Y_i) = X_i\beta,$$

and where the primary goal is to make inferences about the population regression parameters, β. In Chapter 7 we described how the specification of this regression model for longitudinal data can be completed by making additional assumptions about the structure of $\text{Cov}(Y_i) = \Sigma_i$. In this chapter we consider an alternative, but closely related, approach for analyzing longitudinal data using linear mixed effects models. The underlying premise of linear mixed effects models is that some subset of the regression parameters vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population. That is, individuals in the population are assumed to have their own subject-specific mean response trajectories over time and a subset of the regression parameters are now regarded as being random. The distinctive feature of linear mixed effects models is that the mean response is modelled as a combination of population characteristics, β, that are assumed to be shared by all individuals, and subject-specific effects that are unique to a particular individual. The former are referred to as *fixed effects*, while the latter are referred to as *random effects*. The term *mixed* is used in this context to denote that the model contains both fixed and random effects.

Although the linear mixed effects model assumes that the longitudinal responses depend on a combination of population (or fixed effects) parameters, β, and subject-

specific effects, it nonetheless leads to a model for the marginal mean response (averaged over the distribution of the random effects) that can be expressed in the familiar form

$$E(Y_i) = X_i\beta.$$

However, the introduction of random effects induces covariance among the responses and $\text{Cov}(Y_i) = \Sigma_i$ has a distinctive random effects structure. With the inclusion of random effects, the covariances among the repeated measures can be expressed as functions of time. Unlike the covariance pattern models considered in Chapter 7, which do not distinguish the different sources of variability that have an impact on the covariance, linear mixed effects models explicitly distinguish between-subject and within-subject sources of variability. Moreover, the induced random effects covariance structure can often be described with relatively few parameters, regardless of the number and timing of the measurement occasions.

Because linear mixed effects models explicitly distinguish between fixed and random effects, they allow the analysis of between-subject and within-subject sources of variation in the longitudinal responses. In addition, it is not only possible to estimate parameters that describe how the mean response changes in the population of interest, but it is also possible to predict how individual response trajectories change over time. For example, linear mixed effects models can be used to obtain predictions of individual growth trajectories over time. The latter will be of interest when the focus of inference is on the individual rather than the population of individuals. For example, in the physician–patient context, these predictions can be used to identify those patients who do not respond well to their assigned treatment in a clinical trial.

One very appealing aspect of linear mixed effects models is their flexibility in accommodating any degree of imbalance in longitudinal data, coupled with their ability to account for the covariance among the repeated measures in a relatively parsimonious way. That is, with linear mixed effects models we do not require the same number of observations on each subject nor that the measurements be taken at the same set of measurement occasions. As a result, these models are particularly well suited for analyzing inherently unbalanced longitudinal data. While the regression models for the mean response described in Chapter 6 can also handle unbalanced longitudinal data, the class of covariance pattern models suitable for unbalanced data is very limited.

Example: Random Intercept Model

Recall that in earlier chapters we encountered the simplest possible case of a linear mixed effects model: the linear model with a randomly varying subject effect. In this model, each subject is assumed to have an underlying level of response that persists over time. This is incorporated in the linear mixed effects model by regarding this subject effect as random, yielding the following model

$$Y_{ij} = X_{ij}'\beta + b_i + e_{ij}, \tag{8.1}$$

where b_i is the random subject effect and the e_{ij} are regarded as measurement or sampling errors. Let us examine this simple model more closely. In this model, the response for the i^{th} subject at the j^{th} occasion is assumed to differ from the population mean, $X_{ij}'\beta$, by a subject effect, b_i, and a within-subject measurement error, e_{ij}. Both the subject effect and the measurement error are assumed to be random, with mean zero, and with variances, $\text{Var}(b_i) = \sigma_b^2$ and $\text{Var}(e_{ij}) = \sigma^2$, respectively. In addition, it is assumed that b_i and e_{ij} are independent of one another. Note that this model describes the mean response trajectory over time for any individual,

$$E(Y_{ij}|b_i) = X_{ij}'\beta + b_i,$$

in addition to the mean response profile in the population,

$$E(Y_{ij}) = X_{ij}'\beta,$$

where the averaging is over all individuals in the population. We refer to the former as the *conditional* mean of Y_{ij}, given the subject-specific effect, and the latter as the *marginal* mean of Y_{ij} (averaged over the distribution of the subject-specific effects, b_i). There is potential for confusion in our use of this terminology, however, since in both cases the mean response is conditional also upon the covariates, X_{ij}.

Next, consider the interpretation of the parameters in the model given by (8.1). The regression parameters β describe patterns of change in the mean response over time (and their relation to covariates) in the population of interest, while b_i describes how the trend over time for the i^{th} individual deviates from the population average. That is, b_i represents an individual's deviation from the population mean intercept, after the effects of the covariates have been accounted for. Thus, when combined with the fixed effects, b_i describes the mean response trajectory over time for any individual. This interpretation is often obscured by the use of vector and matrix notation, but is apparent if we express the model given by (8.1) as

$$\begin{aligned} Y_{ij} &= X_{ij}'\beta + b_i + e_{ij} \\ &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + e_{ij} \\ &= \beta_1 + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + e_{ij} \\ &= (\beta_1 + b_i) + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij}, \end{aligned}$$

where $X_{ij1} = 1$ for all i and j, and β_1 is then the fixed effect intercept term in the model. When expressed in this way, it can be seen that the intercept for the i^{th} individual is $\beta_1 + b_i$ and varies randomly from one individual to another. Because the mean of the random effect b_i is assumed to be zero, b_i represents the deviation of the i^{th} individual's intercept $(\beta_1 + b_i)$ from the population intercept, β_1.

For this simple example of a linear mixed effects model the fundamental ideas can be best understood by considering the graphical representation of the model equations. Figure 8.1 displays how the marginal mean response over time in the population changes linearly with time (denoted by the solid line), but also indicates how the conditional mean responses for two specific individuals, say subjects A and

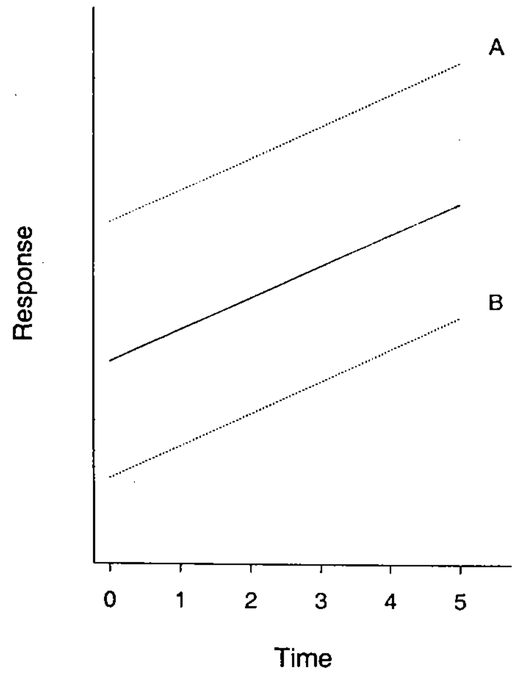

Fig. 8.1 Graphical representation of the marginal and conditional mean responses over time.

B, deviate from the population trend (denoted by the broken lines). In this simple illustration, individual A responds "higher" than the population average and thus has a positive b_i. On the other hand, individual B responds "lower" than the population average and has a negative b_i. Note that the mixed effects model with randomly varying intercepts does not posit that the repeated measures for individual A or B fall perfectly along these subject-specific response trajectories (represented by the broken lines in Figure 8.1). The inclusion of the measurement errors, e_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories; this is illustrated in Figure 8.2.

Next, consider the marginal covariance among the repeated measurements on the same individual. When averaged over the individual-specific effects, the marginal mean of Y_{ij} is given by

$$E(Y_{ij}) = \mu_{ij} = X'_{ij}\beta.$$

The marginal covariance among the Y_{ij} is defined in terms of deviations of Y_{ij} from the marginal mean, μ_{ij}. For example, in Figure 8.2 these deviations are positive at all measurement occasions for individual A and negative at all measurement occasions for individual B, indicating a strong positive correlation (marginally) among the responses

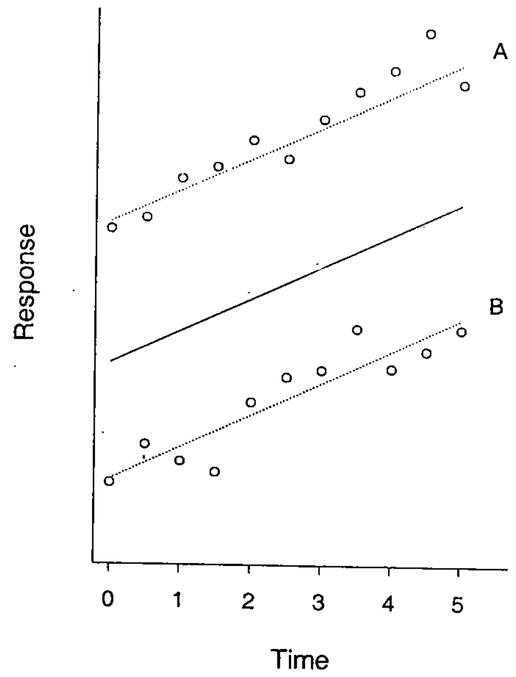

Fig. 8.2 Graphical representation of the marginal and conditional mean responses over time, plus measurement errors.

over time. For the model with randomly varying intercepts, the marginal variance of each response is given by

$$\begin{aligned}
\text{Var}\,(Y_{ij}) &= \text{Var}\,(X'_{ij}\beta + b_i + e_{ij}) \\
&= \text{Var}\,(b_i + e_{ij}) \\
&= \text{Var}\,(b_i) + \text{Var}\,(e_{ij}) \\
&= \sigma_b^2 + \sigma^2.
\end{aligned}$$

Similarly, the marginal covariance between any pair of responses, Y_{ij} and Y_{ik}, is given by

$$\begin{aligned}
\text{Cov}\,(Y_{ij}, Y_{ik}) &= \text{Cov}\,(X'_{ij}\beta + b_i + e_{ij},\ X'_{ik}\beta + b_i + e_{ik}) \\
&= \text{Cov}\,(b_i + e_{ij},\ b_i + e_{ik}) \\
&= \text{Cov}\,(b_i,\ b_i) \\
&= \text{Var}\,(b_i) \\
&= \sigma_b^2.
\end{aligned}$$

Thus the marginal covariance matrix of the repeated measurements has the following compound symmetry pattern

$$\mathrm{Cov}(Y_i) = \begin{pmatrix} \sigma_b^2+\sigma^2 & \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & .. & \sigma_b^2+\sigma^2 \end{pmatrix}.$$

This is the only covariance model that arises in both the patterned (see Section 7.4) and random effects families.

Given that the covariance between any pair of repeated measurements is σ_b^2, the correlation is

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2+\sigma^2}.$$

This simple expression for the correlation emphasizes an important aspect of mixed effects models: the introduction of a random subject effect, b_i, can be seen to induce correlation among the repeated measurements. Although the randomly varying intercepts model is the simplest example of a linear mixed effects model, and the resulting covariance structure is not usually appropriate for longitudinal data, the basic ideas can be generalized to provide a very versatile model for analyzing longitudinal data.

8.2 LINEAR MIXED EFFECTS MODELS

In this section we consider generalizations of (8.1) by allowing additional regression coefficients to vary randomly. We also highlight some of the appealing aspects of the linear mixed effects model alluded to earlier. The underlying premise of the model is that some subset of the regression coefficients vary randomly from one individual to another. In the simplest case considered above, we assumed that the intercept varied randomly. The introduction of this single random effect induces covariance among the repeated measures, albeit with a somewhat restricted form. By allowing a subset of the regression coefficients to vary randomly, a very flexible, and yet quite parsimonious, class of random effects covariance structures becomes available.

To fix ideas, consider the following example of a linear mixed effects model with intercepts and slopes that vary randomly among individuals. That is, for the i^{th} subject at the j^{th} measurement occasion,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij}, \quad j = 1, ..., n_i.$$

In this model, each subject varies not only in their baseline level of response (when $t_{i1} = 0$), but also in terms of changes in their responses over time. This can be best understood by considering the graphical representation of the model equations. Figure 8.3 displays how the marginal mean response in the population changes linearly

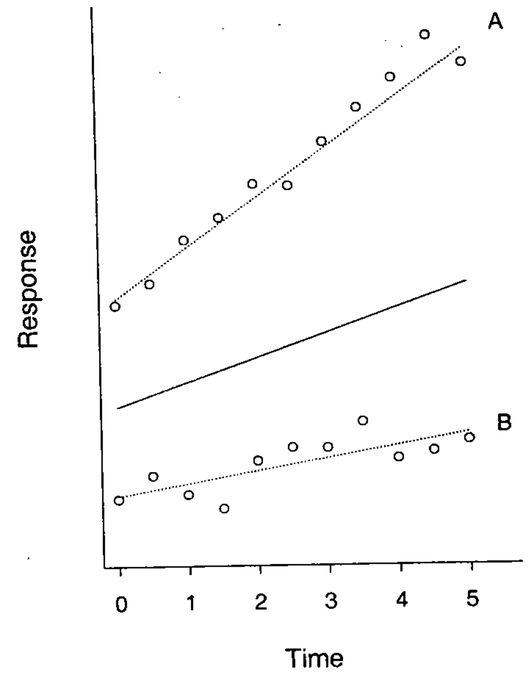

Fig. 8.3 Graphical representation of the marginal and conditional mean responses over time, plus measurement errors.

with time (denoted by the solid line), but also indicates how the conditional mean responses for two specific individuals, say subjects A and B, deviate from the population trend (denoted by the broken lines). In this simple illustration, individual A has a "higher" baseline level of response ($\beta_1 + b_{1i}$) than the population average (β_1) and thus has a positive b_{1i}. On the other hand, individual B has a "lower" baseline level of response than the population average and thus has a negative b_{1i}. In addition, individual A has a steeper rate of increase over time ($\beta_2 + b_{2i}$) than the population average (β_2) and thus has a positive b_{2i}. Individual B has a less steep rate of increase over time than the population average and thus has a negative b_{2i}. Finally, the inclusion of the measurement errors, e_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories. In this illustration, there are randomly varying intercepts and slopes. However, the linear mixed effects model can be generalized to incorporate additional randomly varying regression coefficients and to allow the means of the random effects to depend on covariates.

In the following we assume there are N individuals on whom we have collected n_i repeated observations, with the response variable Y_{ij} measured at time t_{ij}. Thus

the longitudinal data can be inherently unbalanced over time. In the most extreme case, each individual has a unique sequence of measurement occasions, $t_{i1}, ..., t_{in_i}$. Although no longitudinal study would ever be intentionally designed in this way, a change in the metameter for "time" may induce such a design. For example, a longitudinal design can be perfectly balanced when time is defined relative to the baseline measurement but become highly unbalanced if time is defined relative to some landmark event (e.g., puberty, menarche, or menopause).

Using vector and matrix notation, the linear mixed effects model can be expressed as

$$Y_i = X_i\beta + Z_i b_i + e_i, \tag{8.2}$$

where β is a $(p \times 1)$ vector of fixed effects, b_i is a $(q \times 1)$ vector of random effects, X_i is a $(n_i \times p)$ matrix of covariates, and Z_i is a $(n_i \times q)$ matrix of covariates, with $q \leq p$. Here, Z_i is a known design matrix linking the vector of random effects b_i to Y_i. In particular, the columns of Z_i are a subset of the columns of X_i. The reason for this restriction on the columns of Z_i will become evident in Section 8.4. In model (8.2) the particular subset of the regression parameters β that vary randomly is determined by the columns of X_i that comprise Z_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding column of X_i in Z_i, the design matrix for the random effects. The random effects, b_i, are assumed to have a multivariate normal distribution with mean zero and covariance matrix G. That is, $E(b_i) = 0$ and $\text{Cov}(b_i) = G$. In principle, any multivariate distribution for b_i could be assumed; in practice, b_i are assumed to have a multivariate normal distribution.

If, in model (8.2), the vector of random effects, b_i, has mean zero, the random effects then have interpretation in terms of how the subset of regression parameters for the i^{th} individual deviate from those in the population. As mentioned previously, the particular subset of the regression parameters, β, that are assumed to vary randomly is determined by the columns of X_i that comprise Z_i. For example, in a model with only randomly varying intercepts, Z_i is a $(n_i \times 1)$ vector composed of 1's (since $X_{ij1} = 1$ for all i and j). Later, we will consider the form of the design matrix Z_i for more general models.

An important distinction in the linear mixed effects model is that between the conditional and marginal means of Y_{ij}. The *conditional* or *subject-specific* mean of Y_i, given b_i, is

$$E(Y_i|b_i) = X_i\beta + Z_i b_i,$$

while the *marginal* or population-averaged mean of Y_i, when averaged over the distribution of the random effects b_i, is

$$\begin{aligned} E(Y_i) &= \mu_i \\ &= E\{E(Y_i|b_i)\} \\ &= E(X_i\beta + Z_i b_i) \\ &= X_i\beta + Z_i E(b_i) \\ &= X_i\beta, \end{aligned}$$

since $E(b_i) = 0$. Thus, in the linear mixed effects model, the vector of regression parameters β (the *fixed effects*), are assumed to be the same for all individuals and have population-averaged interpretations, for example, in terms of changes in the mean response, averaged over all individuals in the population. In contrast to β, the vector b_i (when combined with the corresponding fixed effects) is comprised of subject-specific regression coefficients. These are the *random effects* and, when combined with the fixed effects, they describe the mean response profile of any *individual*. That is, the mean response profile for the i^{th} individual is given by

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

Finally, the $(n_i \times 1)$ vector of errors, e_i, is assumed to be independent of b_i, and to also have a multivariate normal distribution with mean zero and covariance matrix R_i. Ordinarily, it is further assumed that R_i is the diagonal matrix, $\sigma^2 I_{n_i}$, where I_{n_i} denotes an $n_i \times n_i$ identity matrix. In that case, e_{ij} and e_{ik} are uncorrelated, with equal variance, and the e_{ij}'s can be thought of as sampling or measurement errors. In principle, we can allow correlation among the e_{ij}'s by assuming R_i has a covariance pattern of the kind considered in Section 7.4. However, doing so would raise two potential complications. First, the e_{ij}'s would no longer have a simple interpretation as measurement or sampling errors. This would alter the interpretation of the e_{ij}'s, and hence b_i, implying that the e_{ij}'s include a component of model misspecification at the individual level. Second, there can be subtle issues of model identification when R_i is assumed to have a non-diagonal covariance pattern since it may not be possible to estimate both G and R_i from the data at hand. For example, it is not possible to estimate both G and an unstructured R_i. Throughout the remainder of this chapter we assume that the e_{ij}'s are pure measurement or sampling errors and that $R_i = \sigma^2 I_{n_i}$.

Although we have assumed multivariate normality for both the random effects, b_i, and the measurement errors, e_i, these distributional assumptions are not required for the model development. The form of the conditional and marginal means only requires that the measurement errors are independent of the random effects and that both have mean zero, $E(b_i) = 0$ and $E(e_i) = 0$. The multivariate normal assumption is required in subsequent sections where we consider estimation, testing, and prediction of random effects.

To clarify the vector and matrix notation introduced so far, consider the following linear mixed effects model with intercepts and slopes that vary randomly among individuals (see Figure 8.3). For the i^{th} subject at the j^{th} measurement occasion, assume that

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij}, \quad j = 1, ..., n_i.$$

Using vector and matrix notation, this model can be expressed as

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

where

$$X_i = Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}.$$

Here, $q = p = 2$ and Z_i is composed of the two columns of X_i. This model posits that individuals vary not only in their baseline level of response (when $t_{i1} = 0$), but also in terms of their changes in the mean response over time. The effects of covariates (e.g., due to treatments, exposures, or background characteristics of the individuals) can be incorporated by allowing the means of the intercepts and slopes to depend upon these covariates (e.g., by allowing them to vary across the different treatment groups or levels of exposure).

For example, consider the hypothetical two-group study comparing a *treatment* and a *control* group discussed in Section 5.2. If the mean response changes in an approximately linear fashion over time, but with the means of the intercepts and slopes depending on group, the following linear mixed effects model can be adopted:

$$Y_{ij} = \beta_1 + \beta_2\, t_{ij} + \beta_3\, \texttt{Group}_\texttt{i} + \beta_4\, t_{ij} \times \texttt{Group}_\texttt{i} + b_{1i} + b_{2i}\, t_{ij} + e_{ij},$$

where $\texttt{Group}_\texttt{i} = 1$ if the i^{th} individual was assigned to the treatment, and $\texttt{Group}_\texttt{i} = 0$ otherwise. In this model, the design matrix X_i has the following form for the control group

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ 1 & t_{i2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,n_i} & 0 & 0 \end{pmatrix};$$

whereas for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & 1 & t_{i1} \\ 1 & t_{i2} & 1 & t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,n_i} & 1 & t_{i,n_i} \end{pmatrix}.$$

Note that the design matrix Z_i has the same form for both the treatment and control groups,

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{i,n_i} \end{pmatrix}.$$

Next, consider the covariance among the components of Y_i in this linear mixed effects model with randomly varying intercepts and slopes. Let $\text{Var}\,(b_{1i}) = g_{11}$, $\text{Var}\,(b_{2i}) = g_{22}$, and $\text{Cov}\,(b_{1i}, b_{2i}) = g_{12}$. These are the three unique elements of the

(2×2) covariance matrix $G = \text{Cov}(b_i)$. If we also assume that $R_i = \text{Cov}(e_i) = \sigma^2 I_{n_i}$, then it can be shown that

$$\begin{aligned} \text{Var}\,(Y_{ij}) &= \text{Var}\,\left(X'_{ij}\beta + Z'_{ij}b_i + e_{ij}\right) \\ &= \text{Var}\,\left(Z'_{ij}b_i + e_{ij}\right) \\ &= \text{Var}\,(b_{1i} + b_{2i}\, t_{ij} + e_{ij}) \\ &= \text{Var}(b_{1i}) + 2\, t_{ij}\, \text{Cov}(b_{1i}, b_{2i}) + t_{ij}^2\, \text{Var}(b_{2i}) + \text{Var}(e_{ij}) \\ &= g_{11} + 2\, t_{ij}\, g_{12} + t_{ij}^2\, g_{22} + \sigma^2. \end{aligned}$$

Similarly, it can be shown that

$$\text{Cov}\,(Y_{ij}, Y_{ik}) = g_{11} + (t_{ij} + t_{ik})\; g_{12} + t_{ij}\, t_{ik}\, g_{22}.$$

Thus, in this model for longitudinal data the covariance matrix, $\text{Cov}(Y_i)$, can be expressed as a function of time, t_{ij}. In Section 8.3, we consider the form of the induced random effects covariance structure in the more general case.

Finally, an important issue in the linear mixed effects model concerns the "centering" of the times of measurement. In Chapter 6 we emphasized that "centering" can avoid problems of collinearity when the model for the mean includes linear, quadratic (and possibly higher-order polynomial) time trends. In the linear mixed effects model "centering" has implications for the proper interpretation of both the mean response and the variance of the random effects. In the illustration above, if t_{ij} represents time since baseline then $\beta_1 + b_{1i}$ represents the subject-specific mean response at baseline (in the control group) and $\text{Var}(b_{1i}) = g_{11}$ is the between-subject variation in the mean response at baseline. On the other hand, if t_{ij} is an individual's age at the j^{th} measurement occasion, then $\beta_1 + b_{1i}$ does not have a useful interpretation since it represents the subject-specific mean response at age zero; similarly, $\text{Var}(b_{1i})$ does not have a useful interpretation. In that case there are two obvious choices for centering: (i) center the times of measurement for all subjects at some common fixed age within the age range of the study participants (i.e., $t_{ij} - a$, for some fixed value a), or (ii) center at the mean age of each subject, when averaged over the subject's period of follow-up (i.e., $t_{ij} - \bar{a}_i$, where $\bar{a}_i$ is the average age, over the period of follow-up, for the i^{th} subject). The first option is preferable because $\beta_1 + b_{1i}$ represents the subject-specific mean response at the common age a and g_{11} is the between-subject variation in the mean response at that age. The second option should be avoided because $\beta_1 + b_{1i}$ then represents the subject-specific mean response at a specific subject's mean age over the period of follow-up. Since the mean age may vary considerably from one subject to another, g_{11} will be inflated and will not have a meaningful interpretation. In summary, with unbalanced longitudinal data, mean centering of the times of measurement should be avoided. Instead, we recommend that times of measurement should be centered at some common value of time (or age) in the center of the range of values for all individuals. By centering at a common value, the intercept is interpretable as the mean response at that common value for time (or age) and $\text{Var}(b_{1i})$ also has a meaningful interpretation.

8.3 RANDOM EFFECTS COVARIANCE STRUCTURE

Next, we consider the form of the induced random effects covariance structure for longitudinal data in the more general case. In the linear mixed effects model

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

$R_i = \text{Cov}\,(e_i)$ describes the covariance among the longitudinal observations when focusing on the conditional mean response profile of a *specific* individual. That is, it is the covariance of the i^{th} individual's deviations from her mean response profile,

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

For example, in Figures 8.2 and 8.3 these deviations are positive and negative, and vary randomly about zero, for individuals A and B. As mentioned previously, it is usually assumed that R_i is a diagonal matrix, $\sigma^2 I_{n_i}$, where I_{n_i} denotes an $n_i \times n_i$ identity matrix. The latter is often referred to as a "conditional independence assumption", that is, given the random effects b_i, the measurement errors are independently distributed with a common variance σ^2.

In the linear mixed effects model we can distinguish the conditional mean of Y_i, given b_i,

$$E\,(Y_i|b_i) = X_i\beta + Z_i b_i,$$

from the *marginal* or population-averaged mean of Y_i,

$$E\,(Y_i) = X_i\beta,$$

where averaging is over the distribution of the random effects, b_i. In a similar way, we can distinguish between conditional and marginal covariances. The conditional covariance of Y_i, given b_i, is

$$\text{Cov}\,(Y_i|b_i) = \text{Cov}\,(e_i) = R_i,$$

while the marginal covariance of Y_i, averaged over the distribution of b_i, is

$$\begin{aligned}\text{Cov}\,(Y_i) &= \text{Cov}\,(Z_i b_i) + \text{Cov}\,(e_i)\\ &= Z_i\text{Cov}(b_i)Z_i' + \text{Cov}(e_i)\\ &= Z_i G Z_i' + R_i.\end{aligned}$$

This latter expression for the marginal covariance may be somewhat daunting at first glance. Even when $R_i = \text{Cov}(e_i) = \sigma^2 I_{n_i}$, a diagonal matrix (with all pairwise correlations equal to zero),

$$\text{Cov}\,(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i}$$

is emphatically not a diagonal matrix. That is, $\text{Cov}\,(Y_i)$ will, in general, have non-zero off-diagonal elements, thereby accounting for the correlation among the repeated

observations on the same individuals in a longitudinal study. Thus the introduction of random effects, b_i, induces correlation among the components of Y_i. An additional property of the linear mixed effects model is that $\text{Cov}\,(Y_i)$ has been described in terms of a set of covariance parameters, some defining the matrix G and some defining the matrix R_i. That is, the linear mixed effects model allows for the explicit analysis of between-subject (G) and within-subject (R_i) sources of variation in the responses. Finally, the marginal covariance of Y_i is a function of the times of measurement. For example, in the model with randomly varying intercepts and slopes considered in Section 8.2, we saw that

$$\begin{aligned}\text{Var}\,(Y_{ij}) &= g_{11} + 2\,t_{ij}\,g_{12} + t_{ij}^2\,g_{22} + \sigma^2,\\ \text{Cov}\,(Y_{ij}, Y_{ik}) &= g_{11} + (t_{ij} + t_{ik})\,g_{12} + t_{ij}\,t_{ik}\,g_{22},\end{aligned}$$

so that both $\text{Var}\,(Y_{ij})$ and $\text{Cov}\,(Y_{ij}, Y_{ik})$ depend on the measurement times.

The induced random effects covariance structure,

$$\text{Cov}\,(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i},$$

can be contrasted with the covariance pattern models described in Chapter 7 (see Section 7.4). Recall that a defining feature of the covariance pattern models is that they take into account all sources of variability that have an impact on the covariance, but do not distinguish between the different sources of variability. In contrast, linear mixed effects models explicitly distinguish between-subject and within-subject sources of variability. For linear models for longitudinal continuous data, both approaches yield the same model for the *marginal* or population-averaged mean of Y_i

$$E\,(Y_i) = X_i\beta,$$

and differ only in terms of the assumed model for the covariance. As we will see in Chapters 11–13, longitudinal models for discrete responses do not share this property; for discrete responses, different approaches for accounting for the covariance among the longitudinal responses can lead to models for the mean response having regression parameters with quite distinct interpretations.

The induced random effects covariance structure has certain features that are different from the covariance pattern models considered in Chapter 7. First, unlike many covariance pattern models, the random effects covariance structure does not require a balanced longitudinal design. Because the covariance is expressed as an explicit function of the times of measurement (when times of measurement, or functions of time, are included in Z_i), in principle, each individual can have a unique sequence of measurement times. This makes linear mixed effects models well suited for modelling data from inherently unbalanced longitudinal designs. In addition, the number of covariance parameters is the same regardless of the number and timing of the measurements. Finally, unlike many of the covariance pattern models that make strong assumptions about homogeneity of variance over time, the random effects covariance structure allows the variance and covariance to increase or decrease as a function of the times of measurement (e.g., in the random intercepts and slopes model, the variance is a quadratic function of the times of measurement).

8.4 TWO-STAGE RANDOM EFFECTS FORMULATION

The linear mixed effects model given by (8.2) can be motivated by a two-stage random effects formulation of the model. Indeed, some of the main ideas behind the mixed effects model are often better understood by considering the model as arising from a two-stage specification. For purely pedagogical purposes, we find the two-stage specification to be quite helpful; however, we must caution the reader that the two-stage formulation of the linear mixed effects model does introduce some unnecessary restrictions on the model.

Stage 1

As the term implies, a two-stage random effects model can be conceived in two separate stages. In the first stage subjects are assumed to have their own unique individual-specific mean response trajectories. That is, in stage 1 we posit that the repeated measures on each individual follow a regression model having the same set of covariates, but with separate or distinct regression coefficients for each individual. This is expressed more formally as

$$Y_i = Z_i\beta_i + e_i,$$

where the vector of errors, e_i, are assumed to have a normal distribution, with mean equal to zero and variance σ^2. That is, the e_i can be thought of as measurement or sampling errors, with $e_i \sim N(0, \sigma^2 I_{n_i})$. Note that the number of individual-specific regression coefficients is the same (i.e., the dimension of β_i is q), regardless of the number of longitudinal responses n_i. These individual-specific regression coefficients, β_i, can be interpreted as the i^{th} individual's "true" regression coefficients. Alternatively, $Z_i\beta_i$ can be thought of as the i^{th} individual's "true" underlying mean response trajectory. When viewed in this way, the longitudinal responses on the i^{th} individual are assumed to follow the individual-specific response trajectory given by $Z_i\beta_i$, but with the addition of measurement or sampling errors, e_i.

Note that the matrix Z_i specifies how an individual's mean response changes over time and/or how the mean response changes with other time-varying covariates (e.g., height). For example, it might be assumed that the mean response trajectory is linear, quadratic, or a spline function of time. Consider a model that assumes the individual-specific trajectories are linear in time. Then, the first-stage model can be written as

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}.$$

The essential idea underlying the first-stage model is to fit separate linear regression models to the data for each individual, but with the proviso that these regressions

involve the same set of covariates, Z_i. This is an important observation since it implies that, in principle (and given a sufficient number of repeated measures on each individual), it should be possible to estimate β_i (and σ^2) using data from only the i^{th} individual.

Recall that a particular feature of this first-stage formulation is that the matrix of covariates Z_i is restricted to contain only within-individual or time-varying covariates (with the exception of the column of 1's for the intercept). Time-invariant or between-individual covariates (e.g., gender, treatment group, exposure group) cannot be included in Z_i since their effects would simply be absorbed into the intercept term. Instead, between-individual covariates are introduced in the second stage of the model formulation.

Stage 2

In the second stage we make the assumption that the individual-specific effects, β_i, are random. Given that the β_i are random variables, they have some probability distribution, with a mean and covariance. The mean and covariance of the β_i are the population parameters that are modelled in the second stage. Specifically, variation in β_i from one individual to another is modelled as a function of a set of between-individual (or time-invariant) covariates (e.g., gender, treatments group). In particular, the mean of the β_i can be expressed as a linear function of a set of between-individual covariates, A_i,

$$E(\beta_i) = A_i\beta,$$

where A_i is a $q \times p$ matrix. The remaining residual between-individual variation in the β_i that cannot be explained by A_i is expressed as

$$\text{Cov}(\beta_i) = G.$$

Specification of a model for the mean and covariance of the β_i completes the second stage of the model formulation[1].

For example, consider the hypothetical two-group study comparing a *treatment* and a *control* discussed earlier. If we assume that individual-specific changes in the mean response over time are linear, the first stage model is given by

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}.$$

In the second stage, we can allow the mean of β_i (i.e., the mean intercept and slope) to depend upon group. For example, a model that allows both the mean intercept and

[1] Note that, in this section, in a slight abuse of notation, we are using β to denote a fixed parameter and β_i to denote a random variable.

slope to depend on group is given by

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_2\,\mathtt{Group_i} \\ E(\beta_{2i}) &= \beta_3 + \beta_4\,\mathtt{Group_i} \end{aligned}$$

where $\mathtt{Group_i} = 1$ if the i^{th} individual was assigned to the treatment, and $\mathtt{Group_i} = 0$ otherwise. In this model, β_1 is the mean intercept in the control group, while $\beta_1 + \beta_2$ is the mean intercept in the treatment group. That is, β_2 represents the treatment group difference in the mean intercept. When t_{ij} is the time since baseline, β_2 has a useful interpretation in terms of a treatment group difference in the mean response at baseline. Similarly, β_3 is the mean slope, or rate of change in the mean response over time, in the control group, while $\beta_3 + \beta_4$ is the mean slope in the treatment group. That is, β_4 has interpretation in terms of a treatment group difference in the mean slope or rate of change in the mean response over time. In this model, the design matrix A_i of between-individual covariates has the following form:

$$A_i = \begin{pmatrix} 1 & \mathtt{Group_i} & 0 & 0 \\ 0 & 0 & 1 & \mathtt{Group_i} \end{pmatrix}.$$

Thus, for the control group, the model for the mean is

$$E\begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix};$$

similarly, for the treatment group, the model for the mean is

$$E\begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_2 \\ \beta_3 + \beta_4 \end{pmatrix}.$$

It is also assumed that the remaining residual variation in β_i, that cannot be explained by the effect of group, is

$$\mathrm{Cov}(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \mathrm{Var}(\beta_{1i})$, $g_{22} = \mathrm{Var}(\beta_{2i})$, and $g_{12} = g_{21} = \mathrm{Cov}(\beta_{1i}, \beta_{2i})$. Thus, g_{11} is the variance of β_{1i}, after adjusting for the effect of treatment group, and so on.

The two components of the two-stage model can be combined to yield a linear mixed effects model for Y_i, albeit one that has some restrictions. To see how this can be achieved let us rewrite the subject-specific effects, β_i, as

$$\beta_i = A_i\beta + b_i,$$

where b_i has a multivariate normal distribution with mean zero and covariance matrix, G. Here the b_i yield the regression coefficients from an individual's *residual* trajectory over time, after the covariate effects have been accounted for. Put another way, the b_i represent the i^{th} individual's deviation from the population mean response. Next, by combining the two components of the two-stage model, we obtain

$$\begin{aligned} Y_i &= Z_i\beta_i + e_i \\ &= Z_i(A_i\beta + b_i) + e_i \\ &= (Z_iA_i)\beta + Z_ib_i + e_i \\ &= X_i\beta + Z_ib_i + e_i, \end{aligned}$$

where $X_i = Z_iA_i$. When averaged over the random effects, b_i,

$$E(Y_i) = (Z_iA_i)\beta = X_i\beta,$$

and

$$\mathrm{Cov}(Y_i) = Z_iGZ_i' + \sigma^2 I_{n_i}.$$

Note that in the two-stage formulation, Z_i appears in both the models for the marginal mean and covariance.

While this model is remarkably similar to the linear mixed effects model introduced in the previous section, there is one important difference. The two-stage model places a constraint on the choice of the design matrix for the fixed effects. That is, the two-stage formulation requires that the design matrix for the fixed effects has the special structure $X_i = Z_iA_i$, where A_i contains only between-subject (or time-invariant covariates and Z_i contains only within-subject (or time-varying) covariates. This form for the design matrix for the fixed effects implies that any time-varying covariates must be specified as random effects to ensure their inclusion in the model for the population mean response. This constraint is unnecessary and, in many settings, it can be somewhat inconvenient. In some applications this constraint forces us to consider rather more complex models than may be necessary. For example, in order to allow a sufficiently complex model for the mean response over time (specified in terms of $Z_iA_i\beta$), it may be necessary to include many covariates in Z_i. However, in the two-stage model formulation, this can only be achieved by also introducing an equally complex model for the covariance, since

$$\mathrm{Cov}(Y_i) = Z_iGZ_i' + \sigma^2 I_{n_i}.$$

An example of this arises in developing a model for $\mathrm{FEV_1}$ in children. Previous studies have shown that both age (as a linear spline) and log height are important covariates. Thus four subject-specific regression coefficients (an intercept, two coefficients for age, and one coefficient for log height) are needed to model the mean. But a 4×4 covariance matrix for G is very unwieldy and difficult to fit without very large samples.

Alternatively, in the two-stage formulation a very simple structure for the covariance imposes an often unrealistically simple structure on the mean response. The

most extreme example of this is the two-stage model, which induces a compound symmetry covariance. In that case, a compound symmetry covariance is obtained from a two-stage model with randomly varying intercepts,

$$Y_i = Z_i\beta_i + e_i,$$

where Z_i is a $(n_i \times 1)$ vector of 1's. While marginally (or averaged over the random effects) this model induces a simple compound symmetry covariance structure

$$\text{Cov}(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i} = g_{11} J_{n_i} + \sigma^2 I_{n_i},$$

where J_{n_i} denotes a $(n_i \times n_i)$ matrix of 1's, this model precludes any dependence of the mean response on time. That is,

$$E(Y_i) = (Z_i A_i)\beta$$

cannot depend on time since time, a witnin-subject covariate, has not been included in Z_i in the first stage. Thus, when formulated as a two-stage model, the randomly varying intercepts model excludes the most salient within-subject covariate in a longitudinal study (i.e., time), and thereby does not allow for estimation of changes in the mean response over time, the primary goal in a longitudinal analysis!

In summary, we view the two-stage formulation as be ıg most useful for motivating the main ideas and concepts underling linear mixed effects models. The inherent restrictions in the two-stage formulations can be circumvented by considering linear mixed effects models with an arbitrary design matrix, X_i, for the fixed effects, and by allowing the dimension of Z_i to be arbitrary. The only restrictions placed on X_i and Z_i is that Z_i is composed of a subset of the columns of X_i. The latter constraint ensures that $Z_i b_i$ can be interpreted as a zero mean between-subject residual trajectory (or, put another way, the discrepancy between the i^{th} individual's conditional mean response trajectory and the mean response trajectory in the population). Thus, in the linear mixed effects model

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

the restriction that the columns of Z_i are a subset of the columns of X_i allows us to partition the columns of X_i into a set of columns corresponding to the effects that are fixed and a complementary set of columns corresponding to the effects that are random. If we denote the former by $X_i^{(F)}$ and the latter by $X_i^{(R)}$, the model for Y_i can then be rewritten as

$$Y_i = X_i^{(F)}\beta^{(F)} + X_i^{(R)}\beta_i^{(R)} + e_i,$$

where β has been similarly partitioned into effects that are considered to be fixed, $\beta^{(F)}$, and effects that are considered to be random, $\beta_i^{(R)}$.

Finally, on an historical note, a version of the two-stage formulation was popularized by biostatisticians working at the U.S. National Institutes of Health (NIH). They proposed a method for analyzing repeated measures data where in the first stage

subject-specific regression coefficients are estimated using ordinary least-squares regression (based only on the observations for each subject). In the second stage, the estimated regression coefficients are then analyzed as summary measures using standard parametric (or nonparametric) methods. This method for analyzing repeated measures data became known as the "NIH method"[2] and is a variant of the summary measure analyses considered in Section 3.6.

8.5 CHOICE AMONG RANDOM EFFECTS COVARIANCE MODELS

Although the linear mixed effects model assumes that the longitudinal responses depend on a combination of population and subject-specific effects, when averaged over the distribution of the random effects,

$$E(Y_i) = X_i\beta,$$

and the covariance among the responses has the distinctive random effects structure,

$$\text{Cov}\,(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i}.$$

From the perspective of modelling the covariance, the random effects structure is appealing because the number of covariance parameters, $q \times (q+1)/2 + 1$, is the same regardless of the number and timing of the measurement occasions. In many applications, it will be sufficient to include only random intercepts and slopes for time (a total of $2 \times (2+1)/2 + 1 = 4$ covariance parameters), thereby allowing for heterogeneity in the variances and correlations that can be expressed as functions of time. In other applications, a more complex random effects structure may be required.

In choosing a model for the covariance, it will often be of interest to compare two nested models, one with q correlated random effects, the other with $q+1$ correlated random effects. The difference in the number of covariance parameters between these two models is $q+1$, since there is one additional variance and q additional covariances in the "full" model. As mentioned in Section 7.5, in general, the likelihood ratio test provides a valid method for comparing nested models for the covariance. However, in certain cases, the usual null distribution for the likelihood ratio test is no longer valid. In particular, the comparison of random effects models for the covariance is such a non-standard problem.

In general, when testing a null hypothesis that is on the boundary of the parameter space (e.g., the variance of a random effect equals zero), the usual null distribution for the likelihood ratio test is no longer a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models; instead, the null distribution is a mixture of chi-squared distributions. For example, when comparing two nested models, one with q correlated random

[2] It is difficult to attribute the popularization of the so-called "NIH method" to any single biostatistician at NIH. During their time at NIH, Sam Greenhouse, Max Halperin, and Jerry Cornfield introduced many biostatisticians to this technique.

effects, the other with $q + 1$ correlated random effects, the null distribution of the likelihood ratio test is a 50:50 mixture of chi-squared distributions with q and $q + 1$ degrees of freedom. In Section 7.5, we considered the special case where $q = 0$. A table of critical values, when the null distribution is a known 50:50 mixture of chi-squared distributions with q and $q + 1$ degrees of freedom, is provided in Table C.1 in Appendix C. The critical values in Table C.1 can be used for making inferences about the complexity of the random effects covariance structure. For example, when comparing a model with 2 correlated random effects (e.g., random intercepts and slopes) versus a model with 1 random effect (e.g., random intercepts only), the critical value for the 0.05 significance level can be found in the second row ($q = 1$) of Table C.1. This yields a critical value of 5.14, which is somewhat smaller than the critical value of 5.99 from a standard chi-squared distribution with 2 degrees of freedom.

Alternatively, and especially for more complex comparisons among nested random effects models for the covariance where the null distribution of the likelihood ratio test is not well understood (e.g., comparisons of nested models with q correlated random effects and $q + k$ correlated random effects, where $k > 1$), we recommend the use of $\alpha = 0.1$, instead of $\alpha = 0.05$, when judging the statistical significance of the likelihood ratio test. The latter procedure is somewhat *ad hoc* but will protect against selection of a model that is too parsimonious. In conclusion, for simple comparisons among nested random effects models, the likelihood ratio test statistic can be compared with the critical values in Table C.1. For more complex comparisons, we recommend the use of the $\alpha = 0.1$, instead of $\alpha = 0.05$, significance level.

8.6 PREDICTION OF RANDOM EFFECTS

In this section we provide a non-technical discussion on the prediction of random effects. A good grasp of the material in this section is all that is required for an understanding of the notion of predicting random effects. In Section 8.7 we present a more detailed and technical discussion of the same topic. Many of our readers may find the level of mathematical difficulty of the material in Section 8.7 too challenging. While we encourage all of our readers to tackle Section 8.7, we note that it can be omitted at first reading without loss of continuity.

In many applications where longitudinal data arise, inference is focused on the fixed effects, β. These regression parameters have interpretation in terms of changes in the mean response over time, and their relation to covariates. However, in some longitudinal studies, we may want to predict subject-specific response profiles. For example, in studies of growth it may be of interest to obtain subject-specific growth trajectories. In other types of longitudinal studies, it may be of interest to identify those individuals who showed the greatest increase or decrease in the response over time. Because the linear mixed effects model explicitly distinguishes between fixed and random effects, we can also estimate (or predict) individual-specific response trajectories over time. That is, it is possible to obtain predictions of the subject-specific effects, b_i, or of the subject-specific response trajectories, $X_i\beta + Z_ib_i$. Technically,

because the b_i are random variables, and not fixed population parameters, we customarily refer to "predicting" the random effects rather than "estimating" them.

In general, the problem of predicting a random variable can be shown to be that of predicting its conditional mean, given the available data. Thus, the best predictor of b_i is the conditional mean of b_i, given the vector of responses Y_i (and $\widehat{\beta}$),

$$E(b_i|Y_i) = GZ_i'\Sigma_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\Sigma_i = \text{Cov}(Y_i) = Z_iGZ_i' + R_i$. This is known as the "best linear unbiased predictor" (or BLUP). This predictor of b_i depends upon the unknown covariance among the Y_i. When the unknown covariance parameters are replaced by their REML (or ML) estimates, the resulting predictor

$$\widehat{b}_i = \widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}),$$

is referred to as the "empirical BLUP". Given the "empirical BLUP", $\widehat{b}_i$, we can also obtain the i^{th} subject's predicted response profile as follows:

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i.$$

Interestingly, the i^{th} subject's predicted response profile can also be expressed as a weighted average of the estimated population-averaged mean response profile, $X_i\widehat{\beta}$, and the i^{th} subject's observed response profile Y_i. Specifically,

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i = (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I_{n_i} - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i.$$

This expression helps to explain why it is often said that the empirical BLUP estimator "shrinks" the i^{th} subject's predicted response profile towards the population-averaged mean response profile.

The amount of "shrinkage" toward the population mean depends on the relative magnitude of R_i and Σ_i. Recall that R_i characterizes the within-subject variability, while Σ_i incorporates both within-subject and between-subject sources of variability. As a result, when R_i is relatively "large", and the within-subject variability is greater than the between-subject variability, more weight is assigned to $X_i\widehat{\beta}$, the estimated population-averaged mean response profile, than to the i^{th} individual's observed responses. On the other hand, when the between-subject variability is large relative to the within-subject variability, more weight is given to the i^{th} subject's observed responses, Y_i. Intuitively, this weighting scheme is quite sensible since greater weight should be given to the i^{th} individual's observed responses when any within-subject variability in the longitudinal responses (e.g., due to measurement error) is relatively small when compared to the natural heterogeneity in the individual-specific longitudinal response trajectories. On the other hand, less weight should be given to the i^{th} individual's observed responses when the within-subject variability is relatively large and the population is relatively homogeneous. Finally, the amount of "shrinkage" toward the population mean depends also upon n_i, the number of observation on the i^{th} subject. In general, there is more shrinkage toward the population mean curve when n_i is small. Intuitively, this is also quite sensible since less weight should be given to the i^{th} individual's observed responses when fewer data points are available.

8.7 PREDICTION AND SHRINKAGE*

In this section† we present a more detailed and technical discussion on prediction of random effects in the linear mixed effects model. In doing so, we provide some motivation for, and expressions that support, the main results outlined in Section 8.6.

Because the linear mixed effects model explicitly distinguishes between fixed and random effects, we can estimate both types of effects. As noted in the previous section, the prediction of random effects translates into the problem of predicting the conditional mean of b_i, given the vector of responses Y_i, $E(b_i|Y_i)$. Under the assumptions of the linear mixed effects model, Y_i and b_i have a joint multivariate normal distribution. Using well-known properties of the multivariate normal distribution, it can be shown that the conditional mean of b_i given Y_i (and $\widehat{\beta}$) is

$$E(b_i|Y_i) = GZ_i'\Sigma_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\Sigma_i = \mathrm{Cov}(Y_i) = Z_iGZ_i' + R_i$. This is known as the "best linear unbiased predictor" (or BLUP). From a practical standpoint, this predictor of b_i is unusable because it depends upon the unknown covariance parameters. When the unknown covariance parameters are replaced by their REML estimates, the resulting predictor

$$\widehat{b}_i = \widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}),$$

is referred to as the "empirical BLUP" or the "empirical Bayes" (EB) estimator (since $\widehat{b}_i$ can also be derived from a fully Bayesian formulation). In addition to obtaining a prediction of b_i, we can also obtain standard errors for the prediction based on the following expression

$$\mathrm{Cov}(\widehat{b}_i - b_i) = G - GZ_i'\Sigma_i^{-1}Z_iG + GZ_i'\Sigma_i^{-1}X_i\left(\sum_{i=1}^{N} X_i'\Sigma_i^{-1}X_i\right)^{-1} X_i'\Sigma_i^{-1}Z_iG.$$

Note that $\mathrm{Cov}(\widehat{b}_i - b_i)$ is used to assess the precision of the prediction of b_i, rather than $\mathrm{Cov}(\widehat{b}_i)$, because the latter would fail to recognize the random variation of b_i. Standard errors for the prediction are obtained by simply substituting $\widehat{\Sigma}_i$ and $\widehat{G}$, the REML estimates of the covariance parameters, in the previous expression for $\mathrm{Cov}(\widehat{b}_i - b_i)$.

Given the prediction of b_i, the i^{th} subject's predicted response profile is given by

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i.$$

†This section provides a more technical presentation of prediction of random effects and the notion of shrinkage; it can be omitted at first reading without loss of continuity.

This expression for the predicted response profile can be re-expressed as follows:

$$\begin{aligned}
\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\
&= X_i\widehat{\beta} + Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}) \\
&= (I_{n_i} - Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1}Y_i \\
&= (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I_{n_i} - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i,
\end{aligned}$$

since

$$\widehat{\Sigma}_i\widehat{\Sigma}_i^{-1} = I_{n_i} = (Z_i\widehat{G}Z_i' + \widehat{R}_i)\widehat{\Sigma}_i^{-1} = Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1} + \widehat{R}_i\widehat{\Sigma}_i^{-1}.$$

The latter expression for $\widehat{Y}_i$ shows how the empirical Bayes estimator "shrinks" the i^{th} subject's predicted response profile toward the population-averaged mean response profile. As noted in Section 8.6, the amount of "shrinkage" depends on the relative magnitude of R_i and Σ_i. When the within-subject variability, R_i, is large relative to the between-subject variability, more weight is assigned to $X_i\widehat{\beta}$ than to the i^{th} individual's observed responses. Conversely, when the between-subject variability is large relative to the within-subject variability, more weight is given to the i^{th} subject's observed responses, Y_i.

Similarly, it can be shown that the prediction of individual-specific regression coefficients is a weighted average of the REML estimate of the fixed effects and the corresponding OLS estimate based only on the individual's observations. Specifically, when the linear mixed effects model has a two-stage representation, with $X_i = Z_iA_i$ and $R_i = \sigma^2 I_{n_i}$, the "empirical BLUP" of $\beta_i = A_i\beta + b_i$ is a weighted average of $A_i\widehat{\beta}$ and $\widehat{\beta}_i^{\mathrm{OLS}}$, where $\widehat{\beta}$ is the usual REML estimate of β obtained from the available data on all subjects and $\widehat{\beta}_i^{\mathrm{OLS}}$ is the ordinary least squares estimate of β_i based only on the n_i observations for the i^{th} subject. More formally, the "empirical BLUP" of β_i can be expressed as

$$\widehat{\beta}_i = A_i\widehat{\beta} + \widehat{b}_i = W_i\widehat{\beta}_i^{\mathrm{OLS}} + (I_q - W_i)A_i\widehat{\beta},$$

where the "weight", W_i, is a ratio of the between-subject variability to the sum of the between- and within-subject variability,

$$W_i = G(G + \sigma^2(Z_i'Z_i)^{-1})^{-1},$$

and I_q denotes a $q \times q$ identity matrix. Although this expression for the "weight", W_i, appears somewhat daunting, note that when there is very little within-subject variability (and $\sigma^2 \approx 0$), $W_i \approx I_d$ and then $\widehat{\beta}_i \approx \widehat{\beta}_i^{\mathrm{OLS}}$. That is, when there is very little within-subject variability, we have almost perfect information about b_i from Y_i alone. On the other hand, when there is very little between-subject variability (and $G \approx 0$), $W_i \approx 0$ and then $\widehat{\beta}_i \approx A_i\widehat{\beta}$. Thus, when there is very little between-subject variability in the individual-specific trajectories, it is quite sensible to base our "estimate" or prediction of b_i on data from all of the individuals in the study. The expression for the weight W_i also highlights how the number of repeated measurements influences

the compromise between $\widehat{\beta}_i^{\text{OLS}}$ and $A_i\widehat{\beta}$. For example, consider the special case of the model with randomly varying intercepts, where Z_i is an $n_i \times 1$ vector of 1's. It can be shown that

$$W_i = G(G + \sigma^2(Z_i'Z_i)^{-1})^{-1} = \frac{n_i\, g_{11}}{n_i\, g_{11} + \sigma^2},$$

where $g_{11} = \text{Var}(b_{1i})$ is the variance of the random intercept. Thus, for fixed values of the within- and between-subject variability, the more observations that are available on the i^{th} individual the more the "empirical BLUP" of β_i relies on that individual's data to "estimate" the random effect.

8.8 CASE STUDIES

Next we illustrate the main ideas presented in this chapter by considering linear mixed effects models for analyzing data from three different studies. The first illustration uses lung function growth data in a sample of children and adolescents from the Six Cities Study of Air Pollution and Health. The second illustration uses data on body fat accretion from a prospective study of the development of obesity in a cohort of girls. The third illustration uses data on CD4 counts from a randomized clinical trial of AIDS patients with advanced immune suppression.

Six Cities Study of Air Pollution and Health

The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth (Dockery *et al.*, 1983). A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the United States: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian. The basic maneuver in simple spirometry is maximal inspiration (or breathing in) followed by forced exhalation as rapidly as possible into a closed container. Many different measures can be derived from the spirometric curve of volume exhaled versus time. One widely used measure is the total volume of air exhaled in the first second of the maneuver (FEV_1).

In this section we present an analysis of a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV_1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time. Note that, although girls with only a single observation do not directly provide information about longitudinal or intra-individual change over time, their observations at a single occasion do contribute to the analysis (e.g., these observations contribute information to the estimation of variances and between-subject effects). Data for four selected girls are presented in Table 8.1. Examination of Table 8.1 reveals that these data are inherently unbalanced over time, and the degree of imbalance is even more marked when the age of the child is used as the metameter for time. That is, in this data set children enter the study at different ages and also have different occasions of measurement. Figure 8.4 displays a time plot, with joined line segments, of $\log(\text{FEV}_1/\text{height})$ versus age for 50 randomly selected girls.

Table 8.1 Data on age, height, and FEV_1 for four girls selected from the Topeka data set.

Subject ID	Age	Height	Time	FEV_1
001	9.34	1.20	0.00	1.24
001	10.39	1.28	1.05	1.45
001	11.45	1.33	2.11	1.63
001	12.46	1.42	3.12	2.12
001	13.42	1.48	4.08	2.30
001	15.47	1.50	6.13	2.44
001	16.37	1.52	7.03	2.39
002	6.58	1.13	0.00	1.36
002	7.65	1.19	1.06	1.42
002	12.74	1.49	6.15	2.13
002	13.77	1.53	7.19	2.38
002	14.69	1.55	8.11	2.85
002	15.82	1.56	9.23	3.17
002	16.67	1.57	10.08	2.52
002	17.63	1.57	11.04	3.11
003	6.91	1.18	0.00	1.54
003	7.97	1.23	1.06	1.47
003	8.97	1.30	2.05	1.82
003	9.99	1.35	3.08	2.12
003	11.08	1.47	4.16	2.63
003	13.07	1.57	6.16	2.45
003	14.10	1.59	7.19	2.77
003	15.08	1.60	8.17	3.02
003	16.02	1.60	9.10	2.96
007	6.43	1.18	0.00	0.97
007	7.50	1.25	1.06	1.10
007	13.63	1.64	7.19	2.62
007	14.56	1.67	8.12	2.53
007	15.64	1.68	9.21	2.76
007	16.50	1.69	10.06	2.80
007	17.49	1.69	11.06	2.67

Note: Time represents time since entry to study.

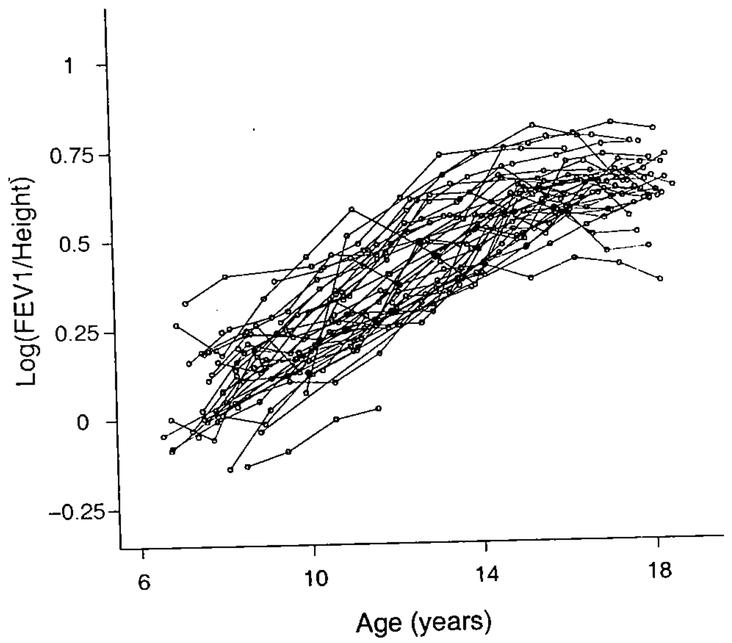

Fig. 8.4 Time plot, with joined line segments, of log(FEV_1/height) versus age in years for 50 randomly selected girls from the Topeka data set.

When age is used as the metameter for time, there are two sources of information about the relationship between FEV_1 and age. First, there is "cross-sectional" or between-subject information that arises because children enter the study at different ages. For example, there is information about how FEV_1 changes with age in the baseline (or time = 0) observations. Second, there is "longitudinal" or within-subject information that arises because children are measured repeatedly over time, yielding measurements of FEV_1 at different ages. Because there are two potentially conflicting sources of information about the relationship between FEV_1 and age, it is important to model these data in a way that allows for separate estimation of the "cross-sectional" and "longitudinal" effects of age of FEV_1. In doing so, it is then possible to test whether there are differences between the "cross-sectional" and "longitudinal" effects of age on FEV_1, and report separate effects where necessary or estimate a combined effect, based on both sources of information, if appropriate. Note that the same issues arise in examining the relationship between FEV_1 and height. A more detailed discussion of the main issues surrounding the analysis of longitudinal designs that provide both longitudinal and cross-sectional sources of information can be found in Chapter 15 (see Section 15.4).

The Six Cities Study was designed to characterize lung function growth between the ages of six and eighteen. The goal of the following analyses is to determine

Table 8.2 Estimated regression coefficients (fixed effects) and standard errors for the log(FEV_1) data from the Six Cities Study.

Variable	Estimate	SE	Z
Intercept	−0.2883	0.0387	−7.45
Age	0.0235	0.0014	16.86
Log(Height)	2.2372	0.0435	51.39
Initial Age	−0.0165	0.0075	−2.21
Initial Log(Height)	0.2182	0.1455	1.50

how changes in lung function (as determined by FEV_1) over time are related to the age and height of the child. Previous research has indicated that log(FEV_1) has an approximately linear relationship with age and log(height) in children and adolescents. To distinguish between the "cross-sectional" and "longitudinal" effects of age and log(height) on log(FEV_1), baseline and current values of these covariates were included in the model for the mean. Because these data are inherently unbalanced, accounting for the covariance among the repeated observation on the same child via a random effects structure is very appealing. Here we allow the intercept and slope for age to varying randomly from one child to another. Specifically, we consider the following model for log(FEV_1):

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 \mathtt{Age}_{ij} + \beta_3 \mathtt{log(Ht)}_{ij} + \beta_4 \mathtt{Age}_{i1} + \beta_5 \mathtt{log(Ht)}_{i1} + b_{1i} + b_{2i} \mathtt{Age}_{ij},$$

where Y_{ij} is the log(FEV_1) for the i^{th} child at the j^{th} occasion, and $\mathtt{Age}_{i1}$ and $\mathtt{log(Ht)}_{i1}$ are the initial or baseline age and log(height) for the i^{th} child. In this model, β_2 and β_3 are the "longitudinal" effects of age and log(height), respectively, while $(\beta_2 + \beta_4)$ and $(\beta_3 + \beta_5)$ are the corresponding "cross-sectional" effects. That is, β_4 and β_5 represent the differences between the longitudinal and cross-sectional effects of age and log(height), respectively. (See Section 15.4 for a more detailed discussion of models that allow for estimation of both longitudinal and cross-sectional effects.)

Preliminary analysis of the data revealed a measurement of FEV_1 that was clearly outlying. This observation was from a girl who had only a baseline measurement available. This observation was removed and all subsequent analyses are based on the data from 299 girls (with a total of 1993 measurements). The REML estimates of the fixed effects are displayed in Table 8.2. Based on the magnitude of the estimates of β_4 and β_5, relative to their standard errors, there is evidence of a significant difference between the longitudinal and cross-sectional effects of age, but not of log(height). From a subject-matter point of view, the magnitudes of the longitudinal and cross-sectional effects of log(height) are quite similar (2.24 versus 2.46), whereas the magnitudes of the longitudinal and cross-sectional effects of age are strikingly

different (0.024 versus 0.007). That is, the longitudinal and cross-sectional effects of age on changes in FEV_1 ($e^{0.024} \approx 1.025$ versus $e^{0.007} \approx 1.007$) are discernibly different. This may be due, in part, to the relatively small amount of variability in ages at baseline (relative to the variability in ages throughout the duration of the study), resulting in the cross-sectional effect of age being poorly estimated. Focusing on the longitudinal effects of age and log(height), there is clear evidence that changes in $\log(\text{FEV}_1)$ are related to both age and height.

Next we consider the interpretation of the fixed effects estimates. The model for the mean, averaged over the distribution of the subject-specific random effects, is given by

$$E(Y_{ij}) = \beta_1 + \beta_2\,\texttt{Age}_{ij} + \beta_3\,\texttt{log(Ht)}_{ij} + \beta_4\,\texttt{Age}_{i1} + \beta_5\,\texttt{log(Ht)}_{i1}.$$

Furthermore, this model can be re-expressed in terms of two models, a cross-sectional model and a longitudinal model. The former is given by

$$\begin{aligned} E(Y_{i1}) &= \beta_1 + \beta_2\,\texttt{Age}_{i1} + \beta_3\,\texttt{log(Ht)}_{i1} + \beta_4\,\texttt{Age}_{i1} + \beta_5\,\texttt{log(Ht)}_{i1} \\ &= \beta_1 + (\beta_2 + \beta_4)\,\texttt{Age}_{i1} + (\beta_3 + \beta_5)\,\texttt{log(Ht)}_{i1}; \end{aligned}$$

while the latter is given by

$$\begin{aligned} E(Y_{ij} - Y_{i1}) &= \beta_1 + \beta_2\,\texttt{Age}_{ij} + \beta_3\,\texttt{log(Ht)}_{ij} + \beta_4\,\texttt{Age}_{i1} + \beta_5\,\texttt{log(Ht)}_{i1} \\ &\quad - \{\beta_1 + \beta_2\,\texttt{Age}_{i1} + \beta_3\,\texttt{log(Ht)}_{i1} + \beta_4\,\texttt{Age}_{i1} + \beta_5\,\texttt{log(Ht)}_{i1}\} \\ &= \beta_2\,(\texttt{Age}_{ij} - \texttt{Age}_{i1}) + \beta_3\,\{\texttt{log(Ht)}_{ij} - \texttt{log(Ht)}_{i1}\}. \end{aligned}$$

The longitudinal effect of log(height), β_3, has interpretation in terms of the changes in mean $\log(\text{FEV}_1)$ for a single-unit increase in log(height), for any given change in age (e.g., during a two-year interval). Similarly, the longitudinal effect of age, β_2, has interpretation in terms of the changes in mean $\log(\text{FEV}_1)$ for a one-year increase in age, for any given change in log(height). The coefficient for log(height), 2.24, is not directly interpretable because a single-unit change in log(height) corresponds to an almost threefold (or $e^{1.0} \approx 2.7$) increase in height. Instead, it is probably more meaningful to consider the effect of a 10% increase in height. On the logarithmic scale, this corresponds to a 0.1 increase in log(height), since $e^{0.1} \approx 1.1$. Therefore, a 10% increase in height (corresponding to an approximate 0.1 increase in log(height)) is associated with a 0.224 increase in $\log(\text{FEV}_1)$. Note that a 0.224 increase in $\log(\text{FEV}_1)$ corresponds to a 25% increase in FEV_1 (since $e^{0.224} = 1.25$). On the other hand, the coefficient for age, 0.024, is more directly interpretable. The estimate of the longitudinal effect of age indicates that a single year increase in age is associated with a 0.024 increase in $\log(\text{FEV}_1)$ or an approximate 2.5% ($e^{0.024} \approx 1.025$) increase in FEV_1, for any fixed change in height.

Next consider the estimates of the variances and covariances of the random effects. When these estimates are compared to their standard errors (see Table 8.3), there is evidence to support their retention in the model. The marginal covariance among the repeated observations is a function of these variance and covariance parameters (and σ^2) and the ages of the child when the observations were obtained. The estimated correlations for annual measurements from ages 7 to 18 are displayed in Table 8.4 and these results indicate that there is strong positive correlation among measurements of $\log(\text{FEV}_1)$ that declines by a modest amount over the 11 years of follow-up. This pattern of correlation reinforces an observation that we made in earlier chapters of the book: the correlation among repeated measurements of many health outcomes rarely decays to zero, even when they are separated by many years.

Table 8.3 Estimated covariance of the random effects and standard errors (×100) for the $\log(\text{FEV}_1)$ data from the Six Cities Study.

Parameter	Estimate	SE	Z
$\text{Var}(b_{i1}) = g_{11}$	1.2207	0.1924	6.34
$\text{Cov}(b_{i1}, b_{i2}) = g_{12}$	−0.0435	0.0122	−3.55
$\text{Var}(b_{i2}) = g_{22}$	0.0050	0.0010	5.11
$\text{Var}(e_i) = \sigma^2$	0.3629	0.0133	27.21

Table 8.4 Estimated marginal correlations among repeated measures of $\log(\text{FEV}_1)$ between the ages of 7 and 18.

1.00	0.70	0.69	0.68	0.67	0.66	0.64	0.62	0.60	0.58	0.56	0.54
0.70	1.00	0.70	0.69	0.68	0.67	0.66	0.65	0.63	0.61	0.60	0.58
0.69	0.70	1.00	0.70	0.70	0.69	0.68	0.67	0.66	0.64	0.63	0.61
0.68	0.69	0.70	1.00	0.70	0.70	0.70	0.69	0.68	0.67	0.66	0.64
0.67	0.68	0.70	0.70	1.00	0.71	0.71	0.70	0.70	0.69	0.68	0.67
0.66	0.67	0.69	0.70	0.71	1.00	0.72	0.72	0.71	0.71	0.70	0.70
0.64	0.66	0.68	0.70	0.71	0.72	1.00	0.73	0.73	0.73	0.72	0.72
0.62	0.65	0.67	0.69	0.70	0.72	0.73	1.00	0.74	0.74	0.74	0.74
0.60	0.63	0.66	0.68	0.70	0.71	0.73	0.74	1.00	0.75	0.75	0.75
0.58	0.61	0.64	0.67	0.69	0.71	0.73	0.74	0.75	1.00	0.76	0.76
0.56	0.60	0.63	0.66	0.68	0.70	0.72	0.74	0.75	0.76	1.00	0.77
0.54	0.58	0.61	0.64	0.67	0.70	0.72	0.74	0.75	0.76	0.77	1.00

Table 8.5 Estimated regression coefficients (fixed effects) and standard errors for the log(FEV_1) data from the Six Cities Study.

Variable	Estimate	SE	Z
Intercept	−0.2846	0.0390	−7.30
Age	0.0233	0.0012	18.65
Log(Height)	2.2523	0.0461	48.82
Initial Age	−0.0163	0.0074	−2.19
Initial Log(Height)	0.1808	0.1455	1.24

Finally, we note that the correlation among repeated measurements has been accounted for by the introduction of random intercepts and slopes for age. Alternatively, we could have considered a random effects model with randomly varying slopes for log(height). By assuming that the slope for log(height) varies randomly for individuals this would also induce covariance among the repeated observations but with correlations that are a function not of age, but of the height of the child. For illustrative purposes we considered the following model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 \mathtt{Age}_{ij} + \beta_3 \mathtt{log(Ht)}_{ij} + \beta_4 \mathtt{Age}_{i1} + \beta_5 \mathtt{log(Ht)}_{i1} + b_{1i} + b_{2i} \mathtt{log(Ht)}_{ij}.$$

The REML estimates of the fixed effects are displayed in Table 8.5 and are qualitatively very similar to those presented in Table 8.2. The reader might then ask which model is more appropriate for the data at hand, a model with randomly varying slopes for age or randomly varying slopes for log(height)? Fortunately, since both models have the same number of covariance parameters, we can make that judgement based on a direct comparison of their maximized REML log-likelihoods. For the model with randomly varying slopes for age the maximized REML log-likelihood is 2283.9, while for the model with randomly varying slopes for log(height) the maximized REML log-likelihood is 2294.7. The comparison of the maximized log-likelihoods indicates that the model with randomly varying slopes for log(height) is to be preferred. For illustrative purposes we also considered a random effects model with randomly varying slopes for both age and log(height). By assuming that the slopes for age and log(height) vary randomly this would induce covariances among the repeated observations that are functions of both the age and height of the child. For the latter model the maximized REML log-likelihood is 2294.9 and does not lead to a discernible improvement in fit over the model with randomly varying slopes for log(height) only. Formally, the likelihood ratio test statistic, $G^2 = 0.4$, can be compared to the critical values in the third row ($q = 2$) of Table C.1 in Appendix C.

Study of Influence of Menarche on Changes in Body Fat Accretion

The second illustration uses longitudinal data from a prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study (Bandini *et al.*, 2002; Phillips *et al.*, 2003). The data represent a subset of the study materials and should not be used to draw substantive conclusions.

It is known that increases in body fatness in girls begin just before or around menarche. Although it has been presumed that the increase in body fatness levels off approximately four years after menarche, these changes in body fat accretion had not been studied in population-based samples. Naumova *et al.* (2001) examined changes in body fat before and after menarche. At the start of the study, all of the girls were pre-menarcheal and non-obese, as determined by a triceps skinfold thickness less than the 85th percentile. All girls were followed over time according to a schedule of annual measurements until four years after menarche. The final measurement was scheduled on the fourth anniversary of their reported date of menarche.

At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis. Percent body fat (%BF) was derived from three basic measurements of body weight (Wt. in kg), height (Ht. in cm) and bioelectric impedance resistance (R). Percent body fat is calculated using the equation:

$$\%\mathrm{BF} = \left(1 - \frac{\mathrm{TBW}}{0.73}\mathrm{Wt}\right) \times 100\%,$$

where total body water, $\mathrm{TBW} = (0.7\mathrm{Ht}^2/\mathrm{R}) - 0.32$.

In this section we present an analysis of the changes in percent body fat before and after menarche. For the purposes of these analyses, "time" is coded as time since menarche and can be positive or negative. Although the measurement protocol is the same for all girls and the study design is balanced if the timing of measurement is defined as the time since the baseline measurement, it is inherently unbalanced when the timing of measurements is defined as the time since a girl experienced menarche.

In this data set there are a total of 1049 individual percent body fat measurements, with an average of 6.4 measurements per subject. The numbers of measurements per subject pre- and post-menarche are approximately equal, with 497 measurements for the pre-menarcheal period (producing an average of 3.1 measurements per subject) and 552 measurements for the post-menarcheal period (producing an average of 3.5 measurements per subject). In this sample, the average age at menarche was 12.8 years.

Figure 8.5 shows a time plot of the individual response profiles (where time is relative to the individual age at menarche). This graph reveals some information about the greater variability of measurement times before menarche. However, it is difficult to discern whether the changes in percent body fat in the pre-menarcheal period are similar to the changes in the post-menarcheal period. In Figure 8.6, the trend in the mean response is assessed using a *lowess* smoothed curve. Recall that *lowess* is a nonparametric, robust regression method that traces the salient features of the mean response as a function of time while making only minimal assumptions about the form of the relationship. The *lowess* curve reveals that the mean response

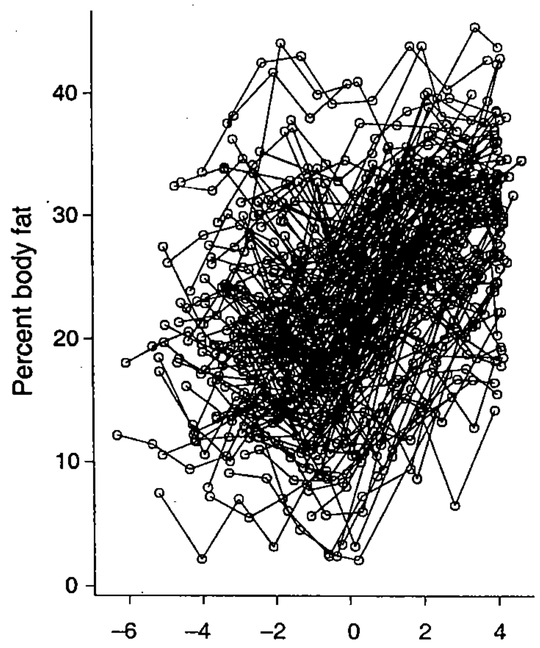

Fig. 8.5 Time plot of percent body fat against time, relative to age of menarche (in years).

remains relatively flat during the pre-menarcheal period and then rises sharply during the post-menarcheal period.

In the following analysis we consider the hypothesis that percent body fat accretion increases linearly with age, but with different slopes before and after menarche. Specifically, we assume that each girl has a piecewise linear spline growth curve with a knot at the time of menarche. That is, each girl's growth curve can be described with an intercept and two slopes, one slope for the changes in response before menarche, another slope for the changes in response after menarche. Note that unlike the piecewise linear splines considered in Section 6.3, the knot is not the same age for all subjects.

Let t_{ij} denote the time of the j^{th} measurement on the i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche). We fit the following linear mixed effects model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$. In this model, $(\beta_1 + b_{1i})$ is the intercept for the i^{th} subject and has interpretation as the true percent body fat

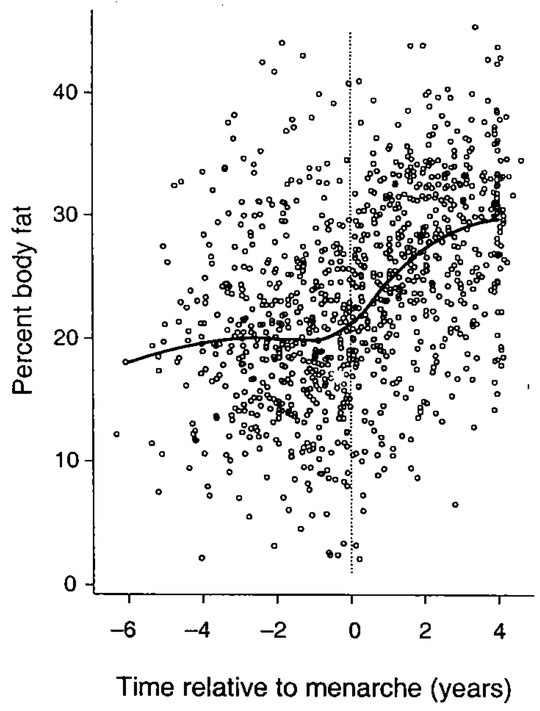

Fig. 8.6 Time plot of percent body fat against time, relative to age of menarche (in years), with *lowess* smoothed curve.

at menarche (when $t_{ij} = 0$). Of note, the actual percent body fat at menarche is not observed and cannot be directly estimated from the data at hand. As a result, we use the term "true" percent body fat at menarche to remind the reader that this is a parameter in the model. Similarly, $(\beta_2 + b_{2i})$ is the i^{th} subject's slope, or rate of change in percent body fat during the pre-menarcheal period. Finally, the i^{th} subject's slope during the post-menarcheal period is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$. Since the overall goal of the analysis is to assess whether the population slopes for fat accretion differ before and after menarche, this can be translated into the null hypothesis, H_0: $\beta_3 = 0$.

The REML estimates of the fixed effects and the variance components are displayed in Tables 8.6 and 8.7, respectively. Based on the magnitude of the estimate of β_3, relative to its standard error, there is a significant difference between the slopes before and after menarche. The estimate of the population mean pre-menarcheal slope is 0.42, which is statistically significant at the 0.05 level. This estimated slope is rather shallow and indicates that the annual rate of body fat accretion is less that 0.5%. Note that the estimated variance of b_{2i} is 1.63, indicating that there is substantial variability from girl to girl in rates of fat accretion and that many girls are losing body fat while

Table 8.6 Estimated regression coefficients (fixed effects) and standard errors for the percent body fat data.

Variable	Estimate	SE	Z
Intercept	21.3614	0.5646	37.84
Time	0.4171	0.1572	2.65
$(\text{Time})_+$	2.0471	0.2280	8.98

Table 8.7 Estimated covariance of the random effects and standard errors for the percent body fat data.

Parameter	Estimate	SE	Z
$\text{Var}(b_{1i}) = g_{11}$	45.9413	5.7393	8.00
$\text{Var}(b_{2i}) = g_{22}$	1.6311	0.4331	3.77
$\text{Var}(b_{3i}) = g_{33}$	2.7497	0.9675	2.85
$\text{Cov}(b_{1i}, b_{2i}) = g_{12}$	2.5263	1.2185	2.07
$\text{Cov}(b_{1i}, b_{3i}) = g_{13}$	−6.1096	1.8730	−3.26
$\text{Cov}(b_{2i}, b_{3i}) = g_{23}$	−1.7505	0.5980	−2.93
$\text{Var}(e_i) = \sigma^2$	9.4732	0.5443	17.40

others are gaining body fat during the pre-menarcheal period. For example, approximately 95% of girls have changes in percent body fat between −2.09% and 2.92% (i.e., $0.42 \pm 1.96 \times \sqrt{1.63}$). The estimate of the population mean post-menarcheal slope is 2.46 (with SE = 0.12), which is statistically significant at the 0.05 level. This indicates that the annual rate of body fat accretion is approximately 2.5%, almost six times higher than the corresponding rate in the pre-menarcheal period. The estimated variance of the individual slopes during the post-menarcheal period, $\text{Var}(b_{2i} + b_{3i})$, is 0.88 (or [1.63 + 2.75 −2 × 1.75]), indicating that there is less variability in the slopes after menarche. For example, approximately 95% of girls have changes in percent body fat between 0.62% and 4.30% (i.e., $2.46 \pm 1.96 \times \sqrt{0.88}$). In other words, more than 95% of girls are expected to have increases in body fat during the post-menarcheal period, while substantially fewer (approximately 63%) are expected to have increases in body fat during the pre-menarcheal period.

Table 8.8 Estimated marginal correlations (on the off-diagonals) among repeated measures of percent body fat between 4 years pre- and post-menarche, with estimated variances along the main diagonal.

61.3	0.82	0.78	0.71	0.61	0.60	0.57	0.52	0.47
0.82	54.9	0.81	0.76	0.70	0.68	0.64	0.60	0.54
0.78	0.81	51.8	0.80	0.76	0.74	0.71	0.66	0.60
0.71	0.76	0.80	52.0	0.81	0.79	0.76	0.71	0.64
0.61	0.70	0.76	0.81	55.4	0.81	0.78	0.73	0.66
0.60	0.68	0.74	0.79	0.81	49.1	0.79	0.76	0.70
0.57	0.64	0.71	0.76	0.78	0.79	44.6	0.77	0.74
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	0.76
0.47	0.54	0.60	0.64	0.66	0.70	0.74	0.76	40.8

The estimated marginal correlations among annual measurements of percent body fat, based on the estimated covariances among the random effects, are displayed in Table 8.8. These results indicate that there is strong positive correlation (approximately 0.8) between annual measurements of percent body fat. Although the strength of the correlation declines over time, it does not decay to zero even when measurements are taken eight years apart. In general, the variability of percent body fat is greater in the pre-menarcheal period.

The mixed effects model can be used to obtain estimates of each girl's growth trajectory over time. Figure 8.7 displays the estimated population mean growth curve and the predicted (empirical BLUP) growth curves for two girls, based on the fixed and random effects estimates reported in Tables 8.6 and 8.7. Note that the two girls selected for display in Figure 8.7 differ in the number of measurements that were obtained (with 6 and 10 measurements, respectively). A noticeable feature of the predicted growth curves is that there is more shrinkage toward the population mean curve when fewer data points are available. That is, the predicted growth curve for the girl with only 6 data points is pulled closer to the population mean curve (or further away from her own data points) while the predicted growth curve for the girl with 10 observation follows her data more closely. This feature becomes more apparent when the empirical BLUPs are compared to the ordinary least squares (OLS) estimates based only on the longitudinal observations from each girl (see Figure 8.8). Examination of Figure 8.8 reveals that the empirical BLUP for the girl with 10 observations is largely based on her longitudinal observations. On the other hand, the empirical BLUP for the girl with 6 observations "borrows strength" from the population mean curve. This is a characteristic feature of the empirical BLUPs that was noted in Sections 8.6 and 8.7. When there is less information available for estimating an individual's growth

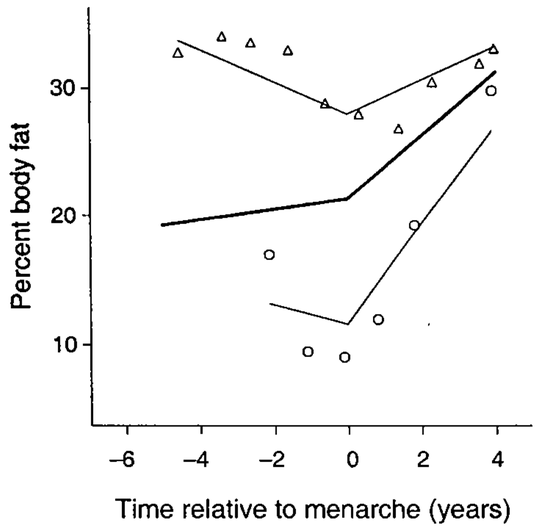

Fig. 8.7 Population average curve (thicker solid line) and empirical BLUPs for two randomly selected girls.

curve, there is a greater "borrowing of strength" from the data obtained on all girls in the study.

Finally, we can use these data to illustrate a hybrid random effects and covariance pattern model by fitting the following model to the percent body fat

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + \epsilon_{ij},$$

where $\epsilon_{ij} = U_i(t_{ij}) + e_{ij}$. The $U_i(t_{ij})$ are assumed to have a normal distribution, with zero mean, variance σ_u^2, and correlation

$$\text{Corr}\{U_i(t_{ij}), U_i(t_{ik})\} = \rho(|t_{ij} - t_{ik}|).$$

The $U_i(t_{ij})$ induce serial correlation among the responses, such that the correlation becomes weaker as the time separation increases. Two popular choices for $\rho(|t_{ij} - t_{ik}|)$ are the exponential correlation function,

$$\rho(|t_{ij} - t_{ik}|) = e^{-\alpha|t_{ij} - t_{ik}|},$$

and the Gaussian correlation function,

$$\rho(|t_{ij} - t_{ik}|) = e^{-\alpha|t_{ij} - t_{ik}|^2},$$

for some $\alpha > 0$. Finally, the e_{ij} are the usual sampling or measurement errors and these are assumed to be independent, with mean zero and variance σ^2.

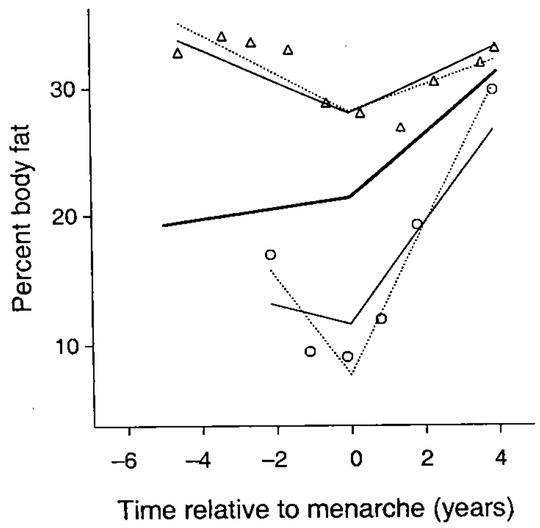

Fig. 8.8 Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

Table 8.9 Comparison of the maximized (REML) log-likelihoods and AIC for the mixed effects model and the hybrid models with exponential and Gaussian serial correlation.

Model	−2 (REML) Log-Likelihood	AIC
Mixed Effects	6062.4	6076.4
Hybrid: Exponential Serial Correlation	5999.9	6007.9
Hybrid: Gaussian Serial Correlation	5991.2	5999.2

We considered the goodness of fit of the hybrid model when the serial correlation function is exponential and Gaussian. Table 8.9 displays the maximized (REML) log-likelihood and AIC for the hybrid models with exponential and Gaussian serial correlation; the maximized (REML) log-lik lihood and AIC for the mixed effects model considered previously are also displayed. These results indicat. that the hybrid model with Gaussian serial correlation fits the data best, since it has the largest maximized log-likelihood (when compared to the hybrid model with exponential serial correlation) and the smallest AIC (when compared to the mixed effects model).

Table 8.10 Estimated regression coefficients (fixed effects) and standard errors for the hybrid model with Gaussian serial correlation.

Variable	Estimate	SE	Z
Intercept	21.2918	0.5400	39.43
Time	0.2168	0.1439	1.51
$(\text{Time})_+$	2.1655	0.2331	9.29

The REML estimates of the fixed effects from the hybrid model with Gaussian serial correlation are displayed in Table 8.10. The estimates of β are similar to those reported in Table 8.6. In particular, the estimate of β_3 is very similar and, when compared to its standard error, there is a significant difference between the slopes before and after menarche. On the other hand, the estimate of the population mean pre-menarcheal slope is 0.22, and is no longer statistically significant at the 0.05 level. Overall, the substantive conclusions are very similar in the two sets of analyses: there is at most a very weak pre-menarcheal slope, indicating that the annual rate of body rate accretion is very modest ($0.2 - 0.4\%$), while the annual rate of fat accretion during the post-menarcheal period is discernibly greater (approximately $2.4 - 2.5\%$) than the corresponding rate in the pre-menarcheal period. Of note, an attempt to fit an extended mixed effects model (with randomly varying intercepts and pre- and post-menarcheal slopes) by incorporation of a Gaussian serial correlation component failed to converge. This lack of convergence was taken as an indication that the observed data simply do not support the need for both randomly varying slopes and serially correlated residuals. As was mentioned in Section 8.2, there can be identifiability problems with the hybrid model unless the random effects structure is kept very simple (e.g., random intercepts only). That is, there may be insufficient information in the data at hand to support separate estimation of randomly varying slopes, serially correlated residuals, and measurement errors.

Randomized Study of Dual or Triple Combinations of HIV-1 Reverse Transcriptase Inhibitors

The final illustration uses data from a randomized, double-blind, study of AIDS patients with advanced immune suppression (CD4 counts of ≤ 50 cells/mm^3) (Henry *et al.*, 1998). Patients in AIDS Clinical Trial Group (ACTG) Study 193A were randomized to dual or triple combinations of HIV-1 reverse transcriptase inhibitors. Specifically, patients were randomized to one of four daily regimens containing 600 mg of zidovudine: zidovudine alternating monthly with 400 mg didanosine; zidovudine plus 2.25 mg of zalcitabine; zidovudine plus 400 mg of didanosine; or zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine (triple therapy). For the analyses presented here, we focus on the comparison of the first three treatment regimens (dual therapy) with the fourth (triple therapy).

Table 8.11 Data on log CD4 counts for four randomly selected subjects from ACTG study 193A.

Subject ID	Group	Time	log(CD4 + 1)
56	0	0.0	1.7047
56	0	8.1	1.7918
56	0	16.1	0.6932
56	0	25.4	1.0986
56	0	33.4	0.6932
56	0	39.1	0.6932
544	1	0.0	3.3844
544	1	7.6	3.2189
544	1	15.9	2.1972
544	1	31.9	1.6094
736	0	0.0	3.7495
736	0	8.9	3.4965
736	0	18.9	3.1780
736	0	30.9	2.7726
986	1	0.0	4.4659
986	1	17.4	3.3322
986	1	30.9	3.5553
986	1	39.6	3.3673

Note: Group = 1 if randomized to triple therapy, Group = 0 if randomized to dual therapy.

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to mistimed measurements and missing data that resulted from skipped visits and dropout (see data for four randomly selected subjects presented in Table 8.11). The number of measurements of CD4 counts during the first 40 weeks of follow-up varied from 1 to 9, with a median of 4. The goal of our analyses is to compare the dual and triple therapy groups in terms of short-term changes in CD4 counts from baseline to week 40 (approximately 10 months of follow-up). The analyses are based on log transformed CD4 counts, log(CD4 counts +1), available on 1309 patients.

In Figure 8.9, the trend in the mean response in the dual and triple therapy groups is assessed using *lowess* smoothed curves. The curves reveal a modest decline in the mean response during the first 16 weeks for the dual therapy group, followed by a steeper decline from week 16 to week 40. In contrast, for the triple therapy group, the mean response increases during the first 16 weeks and declines thereafter. The rate of decline from week 16 to week 40 appears to be similar for the two groups. A note of caution: Because there is a substantial amount of missing data the plot of the

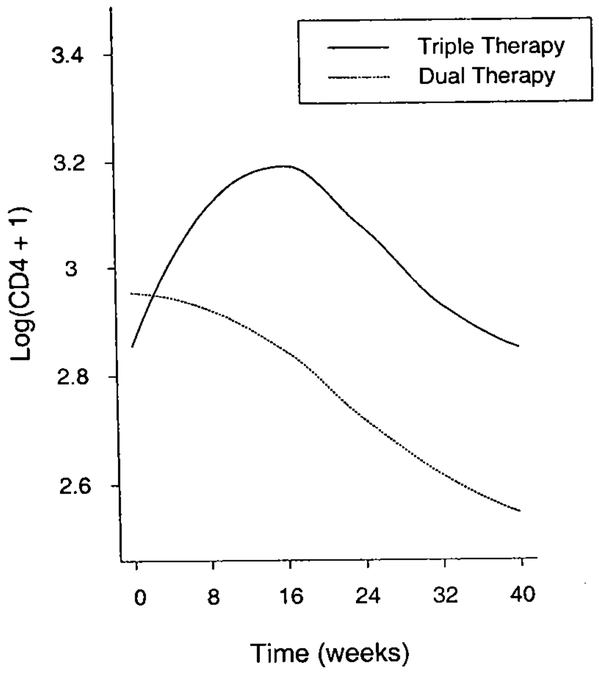

Fig. 8.9 Lowess smoothed curves of log(CD4 + 1) against time (in weeks), for subjects in the dual and triple therapy groups in ACTG study 193A.

mean response over time can be potentially misleading, unless the data are missing completely at random (MCAR). When data are missing, and especially when there is attrition of subjects whose responses are different from those who remain in the study, a plot of the mean response over time can be deceptive; the observed changes in the mean response may reflect the pattern of missingness or the attrition, and not within-individual change. (See Chapter 14 for a more detailed discussion of this issue.)

Next we consider a model for the mean response that allows the rates of change before and after week 16 to differ within and between groups. Specifically, we assume that each patient has a piecewise linear spline with a knot at week 16. That is, each patient's response trajectory can be described with an intercept and two slopes—one slope for the changes in response before week 16, another slope for the changes in response after week 16. The average slopes for changes in response before and after week 16 are allowed to vary by group. Because this is a randomized study, the mean response at baseline is assumed to be the same in the two groups.

Table 8.12 Estimated regression coefficients (fixed effects) and standard errors for the log CD4 counts.

Variable	Estimate	SE	Z
Intercept	2.9415	0.0256	114.81
t_{ij}	−0.0073	0.0020	−3.70
$(t_{ij} - 16)_+$	−0.0120	0.0032	−3.79
$\texttt{Group}_\texttt{i} \times t_{ij}$	0.0269	0.0039	6.98
$\texttt{Group}_\texttt{i} \times (t_{ij} - 16)_+$	−0.0277	0.0062	−4.47

Letting t_{ij} denote the time since baseline for the j^{th} measurement on the i^{th} subject (with $t_{ij} = 0$ at baseline), we consider the following linear mixed effects model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \texttt{Group}_\texttt{i} \times t_{ij} + \beta_5 \texttt{Group}_\texttt{i} \times (t_{ij} - 16)_+ + b_{1i} + b_{2i}\, t_{ij} + b_{3i}\, (t_{ij} - 16)_+,$$

where $\texttt{Group}_\texttt{i} = 1$ if the i^{th} subject is randomized to triple therapy, and $\texttt{Group}_\texttt{i} = 0$ otherwise; $(t_{ij} - 16)_+ = t_{ij} - 16$ if $t_{ij} > 16$ and $(t_{ij} - 16)_+ = 0$ if $t_{ij} \leq 16$. In this model, $(\beta_1 + b_{1i})$ is the intercept for the i^{th} subject and has interpretation as the true log CD4 count at baseline (when $t_{ij} = 0$). Similarly, $(\beta_2 + b_{2i})$ is the i^{th} subject's slope, or rate of change in log CD4 counts from baseline to week 16, if randomized to dual therapy; $(\beta_2 + \beta_4 + b_{2i})$ is the i^{th} subject's slope if randomized to triple therapy. Finally, the i^{th} subject's slope from week 16 to week 40 is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$ if randomized to dual therapy and $\{(\beta_2 + \beta_3 + \beta_4 + \beta_5) + (b_{2i} + b_{3i})\}$ if randomized to triple therapy. The null hypothesis of no treatment group differences in the changes in log CD4 counts can be expressed as H_0: $\beta_4 = \beta_5 = 0$.

The REML estimates of the fixed effects are displayed in Table 8.12. A test of H_0: $\beta_4 = \beta_5 = 0$ yields a Wald test, $W^2 = 59.21$, with 2 degrees of freedom ($p < 0.0001$); the corresponding likelihood ratio test yields $G^2 = 57.99$, with 2 degree of freedom ($p < 0.0001$). Based on the magnitude of the estimate of β_4, relative to its standard error, there is a significant group difference in the rates of change from baseline to week 16. In the dual therapy group, there is a significant decrease in the mean of the log CD4 counts from baseline to week 16. The estimated change during the first 16 weeks is -0.12, or 16×-0.0073. On the untransformed scale, this corresponds to an approximate 10% decrease in CD4 counts (since $e^{-0.12} = 0.89$). In contrast, in the triple therapy group, there is a significant increase in the mean response. The estimated change during the first 16 weeks in the triple therapy group is 0.31, or $16 \times (-0.0073 + 0.0269)$; the estimated slope for the triple therapy group,

Table 8.13 Estimated covariance (×1000) of the random effects and standard errors for the log CD4 counts.

Parameter	Estimate	SE	Z
$\text{Var}(b_{1i}) = g_{11}$	585.742	34.754	16.85
$\text{Var}(b_{2i}) = g_{22}$	0.923	0.160	5.76
$\text{Var}(b_{3i}) = g_{33}$	1.240	0.395	3.14
$\text{Cov}(b_{1i}, b_{2i}) = g_{12}$	7.254	1.805	4.02
$\text{Cov}(b_{1i}, b_{3i}) = g_{13}$	−12.348	2.730	−4.52
$\text{Cov}(b_{2i}, b_{3i}) = g_{23}$	−0.919	0.236	−3.90
$\text{Var}(e_i) = \sigma^2$	306.163	10.074	30.39

0.0196, has a standard error of 0.0033. On the untransformed scale, this corresponds to an approximate 35% increase in CD4 counts (since $e^{0.31} = 1.36$).

The lowess curves in Figure 8.9 suggest that the rate of decline from week 16 to week 40 is similar for the two groups. The null hypothesis of no treatment group differences in the rates of change in log CD4 counts from week 16 to week 40 can be expressed as H_0: $\beta_4 + \beta_5 = 0$ (or H_0: $\beta_4 = -\beta_5$). The estimates of β_4 and β_5 in Table 8.12 appear to support the null hypothesis since they are of similar magnitude but opposite signs. A test of the null hypothesis, H_0: $\beta_4 + \beta_5 = 0$, yields a Wald test, $W^2 = 0.07$, with 1 degree of freedom ($p > 0.75$); the corresponding likelihood ratio test yields $G^2 = 0.07$, with 1 degree of freedom ($p > 0.75$).

The estimated variances of the random effects in Table 8.13 indicate that there is substantial variability from patient to patient in baseline CD4 counts and the rates of change in CD4 counts. For example, although many patients randomized to triple therapy show increases in CD4 counts during the first 16 weeks, some patients have declining CD4 counts. Specifically, approximately 95% of patients randomized to triple therapy are expected to have changes in log CD4 counts from baseline to week 16 between −0.64 and 1.27 (or $16 \times [0.0196 \pm 1.96 \times \sqrt{0.000923}]$). That is, approximately 26% of patients are expected to have decreases in CD4 counts during the first 16 weeks of triple therapy. There is also a substantial component of variability due to measurement error.

In a clinical trial, it is often of interest to predict the direction and magnitude of the treatment effect for patients with specific covariate values. In the physician–patient context, for example, these predictions can be used to identify those patients who do not respond well to their assigned therapy. When there is interest in subject-specific predictions, we must consider the relative magnitudes of the between-subject and

Table 8.14 Estimated regression coefficients (fixed effects) and standard errors for the revised model for the log CD4 counts.

Variable	Estimate	SE	Z
Intercept	2.6457	0.1280	20.67
t_{ij}	−0.0072	0.0019	−3.71
$(t_{ij} - 16)_+$	−0.0124	0.0029	−4.33
$\texttt{Group}_\texttt{i} \times \{t_{ij} - (t_{ij} - 16)_+\}$	0.0263	0.0034	7.68
$\texttt{Age}_\texttt{i}$	0.0100	0.0030	3.31
$\texttt{Gender}_\texttt{i}$	−0.0927	0.0754	−1.23

within-subject variability. When the within-subject or measurement error variability is relatively large, the observed response profile for a subject is unreliable and a better prediction can be obtained by "borrowing strength" from the data on all of the subjects. Next, we consider the prediction of patients' response trajectories from the following linear mixed effects model that also includes gender and baseline age:

$$\begin{aligned} E(Y_{ij}|b_i) &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \texttt{Group}_\texttt{i} \times t_{ij} - \beta_4 \texttt{Group}_\texttt{i} \times (t_{ij} - 16)_+ \\ &\quad + \beta_5 \texttt{Age}_\texttt{i} + \beta_6 \texttt{Gender}_\texttt{i} + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+ \\ &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \texttt{Group}_\texttt{i} \times \{t_{ij} - (t_{ij} - 16)_+\} \\ &\quad + \beta_5 \texttt{Age}_\texttt{i} + \beta_6 \texttt{Gender}_\texttt{i} + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+, \end{aligned}$$

where $\texttt{Age}_\texttt{i}$ is the baseline age (in years) of the patient, $\texttt{Gender}_\texttt{i} = 1$ if the i^{th} patient is male, and $\texttt{Gender}_\texttt{i} = 0$ otherwise. In this model, the mean rate of change from baseline to week 16 can differ in the two groups (with slopes of β_2 and $\beta_2 + \beta_4$ respectively), but the mean rate of change from week 16 to week 40 is assumed to be the same (with slope of $\beta_2 + \beta_3$).

The REML estimates of the fixed effects are displayed in Table 8.14 and the substantive conclusions about the treatment group comparisons are similar to those obtained from Table 8.12. Controlling for gender and age at baseline, there is a 10% decrease in CD4 counts (since $e^{16 \times -0.0017} = e^{-0.12} = 0.89$) in the dual therapy group. In contrast, in the triple therapy group, there is a 35% increase in CD4 counts (since $e^{16 \times (-0.0017 + 0.0263)} = e^{0.31} = 1.36$). In both treatment groups, there is a significant decline in the mean response from week 16 to week 40, corresponding to an approximate 40% decrease in CD4 counts (since $e^{24 \times (-0.0072 - 0.0124)} = e^{-0.47} = 0.63$).

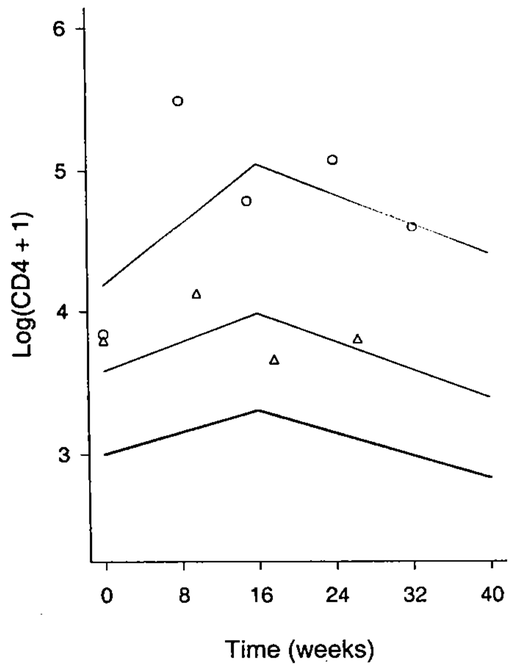

Fig. 8.10 Population average curve (thicker solid line) and empirical BLUPs for two male patients, aged 45, with similar baseline CD4 counts, and treated with triple therapy.

The inclusion of the random effects in the model allows each patient's response trajectory to be described with an intercept and two slopes, one slope for the changes in response before week 16, another slope for the changes in response after week 16. Based on the REML estimates of the fixed effects and variance components, the predicted (or BLUP) trajectory for each patient can be obtained. Figure 8.10 displays the estimated population mean curve and the predicted curves for two male patients, aged 45, and with similar baseline CD4 counts, who were randomized to triple therapy.

In general, the empirical BLUPs, or the predictions of summary measures (e.g., predicted area under the curve for a patient), can be used to identify those patients who have or have not responded well to their assigned therapy. In the physician–patient context, these predictions may be far more relevant than knowledge of the population mean curve. The appealing feature of the linear mixed effects model analysis is that it allows inferences about both the population trends and individual-specific trajectories.

Table 8.15 Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS  id  group;
   MODEL  y=group  time  group*time / S  CHISQ;
   RANDOM  INTERCEPT  time / TYPE=UN  SUBJECT=id  G  V;
```

8.9 COMPUTING: FITTING LINEAR MIXED EFFECTS MODELS USING PROC MIXED IN SAS

To fit linear mixed effects models we need to make use of the RANDOM statement in PROC MIXED. The RANDOM statement is used to define all effects that are considered to be random. Specifically, the RANDOM statement is used to define the covariates in the design matrix, Z_i, for the random effects, b_i. Ordinarily, these will be a subset of the covariates included on the MODEL statement. (Recall that the covariates in the design matrix, X_i, for the fixed effects appear in the MODEL statement.) While the MODEL statement is used to define the design matrix for the fixed effects and the RANDOM statement is used to define the design matrix for the random effects, note that an intercept is included by default in the former but not the latter. That is, unlike the MODEL statement, PROC MIXED does not include an intercept in the RANDOM statement by default. However, you can specify INTERCEPT (or INT) as a random effect on the RANDOM statement. The RANDOM statement is also used to specify the structure of the covariance matrix for the random effects, G. The structure of G is specified using the TYPE=option. The random effects can be assumed to be correlated (TYPE=UN) or uncorrelated (TYPE=VC); ordinarily, covariance pattern models are not used to account for the covariance among the random effects. However, to ensure that the unstructured covariance matrix for the random effects is constrained to be positive-definite, the TYPE=FAO(q) option can be used (where q is the number of random effects). The latter option can be useful when the TYPE=UN option yields an estimated G matrix that is not positive-definite.

For example, to fit a model with randomly varying intercepts and slopes to data from two or more groups measured repeatedly over time, we can use the illustrative SAS commands given in Table 8.15. Note that the SUBJECT option on the RANDOM statement is used in the same manner as on the REPEATED statement and denotes a variable that distinguishes clusters of correlated responses. By including a variable in the SUBJECT option (e.g., a subject identifier), pairs of observations with distinct values of that variable are regarded as independent. Pairs of observations with the same values of that variable share common values of the random effects.

Table 8.16 Illustrative commands for obtaining the estimated BLUPs and the predicted responses from a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
ODS OUTPUT SOLUTIONR=bluptable;

PROC MIXED;
   CLASS id group;
   MODEL y=group time group*time / S CHISQ OUTPRED=yhat;
   RANDOM INTERCEPT time / TYPE=UN SUBJECT=id SOLUTION G V;

PROC PRINT DATA=yhat;
   VAR id group time y Pred;

PROC PRINT DATA=bluptable;
```

Various options can be included on the RANDOM statement. The option G requests that the estimates of the variances and covariances of the random effects be displayed. The option GCORR requests that the estimates of the correlations among the random effects be displayed. The option V requests that the estimates of the marginal covariance matrix, averaged over the distribution of the random effects, be displayed for the first subject. That is, the option V produces estimates of $\Sigma_i = \text{Cov}(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i}$. Finally, when there is interest in predicting the random effects, the SOLUTION (or S) option can be used to request that the estimated BLUPs for the random effects, $\widehat{b_i}$, be displayed (in addition to standard errors for predictions based on the expression for $\text{Var}(\widehat{b_i} - b_i)$ given in Section 8.7). Alternatively, the predicted values of the response, $\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b_i}$, can be requested by using the OUTPRED (or OUTP) option on the MODEL statement. This option specifies a SAS data-set than contains the predicted values of Y_i, denoted by the variable name `Pred`, and some related quantities. For example, to obtain the estimated BLUPs for the random effects and predicted values of the response, $\widehat{Y}_i$, we can use the illustrative SAS commands given in Table 8.16. The OUTPRED option specifies an output SAS data-set containing the predicted values, $\widehat{Y}_i$, whereas the SOLUTION option on the RANDOM statement requests that the estimated BLUPs be produced as part of the standard output from PROC MIXED. Inclusion of the Output Delivery System (ODS) statement creates a SAS data-set containing the estimated BLUPs. Predicted values of the outcome, at occasions other than those actually observed, can also be obtained by including "pseudo-observations" in the data set that have missing values for the outcome variable and the desired values of the covariates.

Table 8.17 Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, within-subject errors with an exponential covariance, and independent measurement errors using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group;
   MODEL y=group time group*time / S CHISQ;
   REPEATED / TYPE=SP(EXP)(time) LOCAL SUBJECT=id;
   RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G V;
```

The alert reader may have noticed that the residual error variance, σ^2, has not been included on the RANDOM statement. Instead, it is included in an implicit REPEATED statement. Recall that the repeated statement is used to specify assumptions about the nature of the covariance among the errors. When the REPEATED statement is not included in PROC MIXED, it is assumed, by default, that the covariance among the errors, $R_i = \sigma^2 I_{n_i}$. To fit hybrid models that include both random effects and correlated errors, it is necessary to include both the RANDOM statement and the REPEATED statement. For example, to fit a hybrid model with (i) randomly varying intercepts and slopes, (ii) within-subject errors with an exponential covariance structure, and (iii) independent measurement or sampling errors, we can use the illustrative SAS commands given in Table 8.17. On the REPEATED statement we use the option TYPE=SP(EXP)(time) to specify an exponential covariance structure for the within-subject errors that depends on time. This command exploits the spatial covariance structures option built into PROC MIXED. Finally, the option LOCAL requests that a diagonal matrix, $\sigma^2 I_{n_i}$, be added to the exponential covariance structure for R_i.

8.10 FURTHER READING

Useful reviews of the linear mixed effects models, targeted at applied researchers, can be found in the articles by Feldman (1988), Gibbons *et al.* (1988), Naumova *et al.* (2001), and Chapters 3 and 4 of Singer and Willett (2003). A comprehensive, but more mathematically challenging discussion of linear mixed effects models can be found in Chapter 3 of Verbeke and Molenberghs (2000) and in the review article by Cnaan *et al.* (1997).

An excellent, non-technical, discussion of the notion of shrinkage can be found in Efron and Morris (1977); also, see the discussion of prediction of random effects in Naumova *et al.* (2001).

Finally, a tutorial description of fitting linear mixed effects models using PROC MIXED in SAS can by found in Singer (1998); also see Chapters 6–7 of Littell *et al.* (1996) and Chapter 8 of Verbeke and Molenberghs (2000).

Bibliographic Notes

Harville (1977) introduced a general class of linear mixed effects models suitable for the analysis of repeated measures and growth curves; also, see Hartley and Rao (1967). Laird and Ware (1982), Jennrich and Schluchter (1986), Laird *et al.* (1987), Lindstrom and Bates (1988), Diggle (1988), Chi and Reinsel (1989), and others, drew upon this family to propose a general class of models for longitudinal data. Ware (1985) provides a general overview of the application of linear mixed effects models to repeated measures and longitudinal data; also, see Chapter 3 of Davidian and Giltinan (1995) for a concise review of linear mixed effects models for repeated measures data.

The notion of shrinkage was first introduced in a seminal paper by Stein (1955). Best linear unbiased prediction (BLUP) is discussed in Henderson (1963); see Robinson (1991) for an interesting review of the prediction of random effects.

Problems

8.1 In a study of exercise therapies, 37 patients were assigned to one of two weightlifting programs (Freund *at al.*, 1988). In the first program (treatment 1), the number of repetitions was increased as subjects became stronger. In the second program (treatment 2), the number of repetitions was fixed but the amount of weight was increased as subjects became stronger. Measures of strength were taken at baseline (day 0), and on days 2, 4, 6, 8, 10, and 12.

The raw data are stored in an external file: `exercise.dat`

Each row of the data set contains the following nine variables:

ID Treatment Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7

Note: The categorical variable Treatment is coded 1 = Program 1 (increase number of repetitions), 2 = Program 2 (increase amount of weight).

8.1.1 On a single graph, construct a time plot that displays the mean strength versus time (in days) for the two treatment groups. Describe the general characteristics of the time trends for the two exercise programs.

8.1.2 Read the data from the external file and put the data in a "univariate" or "long" format, with 7 "records" per patient.

8.1.3 Fit a model with randomly varying intercepts and slopes, and allow the mean values of the intercept and slope to depend on treatment group (i.e., include main effect of treatment, a linear time trend, and a treatment by linear time trend interaction as fixed effects).

(a) What is the estimated variance of the random intercepts?

(b) What is the estimated variance of the random slopes?

(c) What is the estimated correlation between the random intercepts and slopes?

(d) Give an interpretation to the magnitude of the estimated variance of the random intercepts. For example, "approximately 95% of subjects have baseline measures of strength between a and b" (calculate the limits of the interval between a and b).

(e) Give an interpretation to the magnitude of the estimated variance of the random slopes.

8.1.4 Is a model with only randomly varying intercepts defensible? Explain?

8.1.5 What are the mean intercept and slope in the two exercise programs.

8.1.6 Based on the previous analysis, interpret the effect of treatment on changes in strength. Does your analysis suggest a difference between the two groups?

8.1.7 What is the estimate of $\mathrm{Var}(Y_{i1}|b_i)$? What is the estimate of $\mathrm{Var}(Y_{i1})$? Explain the difference.

8.1.8 Obtain the predicted (empirical BLUP) intercept and slope for each subject.

8.1.9 Using any standard linear regression procedure, obtain the ordinary least squares (OLS) estimates of the intercept and slope from the regression of strength on time (in days) for subject 24 (ID=24). That is, restrict the analysis to data on subject 24 only and estimate that subject's intercept and slope.

8.1.10 For subject 24 (ID=24), compare the predicted intercepts and slopes obtained in Problems 8.1.8 and 8.1.9. How and why might these differ?

8.2 AIDS Clinical Trial Group (ACTG) study 193A was a randomized, double-blind, study of AIDS patients with advanced immune suppression (CD4 counts of ≤ 50 cells/mm^3) (Henry *et al.*, 1998). Patients were randomized to one of four daily regimens containing 600 mg of zidovudine:

(1) zidovudine alternating monthly with 400 mg didanosine;

(2) zidovudine plus 2.25 mg of zalcitabine;

(3) zidovudine plus 400 mg of didanosine;

(4) zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine.

In the analyses reported in Section 8.8, the first three treatment groups were combined and compared to the fourth.

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to mistimed measurements and missing data that resulted from skipped visits and dropout. The number of measurements of CD4 counts during the first 48 weeks of follow-up varied from 1 to 9, with a median of 4. CD4 count refers to the number of T-lymphocyte cells in the body; these cells are directly affected by the HIV virus. A normal CD4 count is approximately 800 to 1000; a CD4 count below 200 is one of the diagnostic criteria for AIDS established by the Centers for Disease Control and Prevention (CDC).

The raw data are stored in an external file: `cd4.dat`

Each row of the data set contains the following four variables:

ID Group Week Log(CD4 + 1)

Note: The categorical variable `Group` is coded 1 = zidovudine alternating monthly with 400 mg didanosine, 2 = zidovudine plus 2.25 mg of zalcitabine, 3 = zidovudine plus 400 mg of didanosine, and 4 = zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine. The variable `Week` represents time since baseline (in weeks).

8.2.1 On a single graph, construct a smoothed time plot that displays the mean log CD4 counts versus time (in weeks) for the four treatment groups. Describe the general characteristics of the time trends for the four groups.

8.2.2 Fit a model where each patient's response trajectory is represented by a randomly varying piecewise linear spline with a knot at week 16. That is, fit a model with random intercepts and two randomly varying slopes, one slope for the changes in log CD4 counts before week 16, another slope for the changes in response after week 16. Allow the average slopes for changes in response before and after week 16 to vary by group, but assume the mean response at baseline is the same in the two groups.

8.2.3 Is a model with only randomly varying intercepts defensible? Explain?

8.2.4 Construct a 6-degrees-of-freedom test of the null hypothesis of no treatment group differences in the changes in log CD4 counts.

8.2.5 Based on the previous analysis, interpret the effects of treatment on changes in log CD4 counts from baseline to week 16.

8.2.6 Based on the previous analysis, interpret the effects of treatment on changes in log CD4 counts from week 16 to week 40.

8.2.7 Using the estimates of the fixed effects from the previous analysis, construct a time plot that displays the *estimated* mean log CD4 counts versus time (in weeks) for the four treatment groups. Does the plot suggest that one treatment regimen is superior to the others in terms of short-term (40 weeks) changes in CD4 counts?

9

Residual Analyses and Diagnostics

9.1 INTRODUCTION

The analysis of longitudinal data is not complete without an examination of the residuals. Residuals can be used to assess the adequacy of the fitted model and can also indicate the presence of outliers. Methods for residual analyses are well developed for standard regression settings with independent observations on a univariate response. In principle, many of the same properties of residual analysis can be extended to the longitudinal setting.

9.2 RESIDUALS

With longitudinal data we can define a vector of residuals for each individual,

$$r_i = Y_i - X_i\widehat{\beta}. \tag{9.1}$$

The vector of residuals has mean zero and provides an estimate of the vector of errors,

$$e_i = Y_i - X_i\beta.$$

The residuals defined in (9.1) can be used to check for any systematic departures from the model for the mean response; they can also form the basis of an assessment of the adequacy of the model for the covariance. For example, a scatter-plot of the residuals

$$r_{ij} = Y_{ij} - X'_{ij}\widehat{\beta}$$

against the predicted mean response

$$\widehat{\mu}_{ij} = X'_{ij}\widehat{\beta}$$

can be examined for the appearance of any systematic trend. The fitting of a smooth curve (e.g., a *lowess* curve; see Section 3.3) to the scatter-plot can often help in judging whether curvature is present. In a correctly specified model, the scatter-plot should display no systematic pattern, with a more or less random scatter around a constant mean of zero. Similarly, scatter-plots of the residuals against selected covariates from the model for the mean can be examined for any systematic trends. Such a trend may indicate the omission of a quadratic term or the need for transformation of the covariate.

For most practical purposes, graphical displays of the residuals can be used to detect discrepancies in the model for the mean response or the presence of outlying observations that require further investigation. However, there are two properties of the residuals from an analysis of longitudinal data that must be kept in mind. First, the components of the vector of residuals,

$$r_i = Y_i - X_i\widehat{\beta},$$

are correlated and do not necessarily have constant variance. Recall that the mean of the residuals is zero, mimicking the mean of the vector of errors,

$$e_i = Y_i - X_i\beta.$$

In contrast, the covariance of the residuals is not identical to the covariance of the errors. However, for all practical purposes we can approximate the covariance of the residuals by

$$\text{Cov}(r_i) \approx \text{Cov}(e_i) = \Sigma_i.$$

Because the residuals have approximate covariance matrix, Σ_i, this has important implications for the examination of plots of the residuals. First, because the variance is not necessarily constant, the scatter-plot of the residuals against the predicted values, or against time, will not necessarily have a constant range. As a result, standard residual diagnostics for examining either the homogeneity of the residual variance or autocorrelation among the residuals should be avoided altogether. Second, although residuals from a univariate linear regression are uncorrelated with the covariates, the residuals from a regression analysis of longitudinal data may be correlated with the covariates. As a result, there may be an apparent systematic trend in the scatter-plot of the residuals against a selected covariate.

9.3 TRANSFORMED RESIDUALS

To circumvent some of the aforementioned problems with the use of residuals from longitudinal data based on (9.1), we can transform the residuals. There are many possible ways to transform the residuals. In general, it would be desirable to standardize

and, for lack of a better term, "de-correlate" the residuals so that they mimic residuals from a standard linear regression. That is, we would like to transform the residuals so that they have constant variance and zero correlation. This can be achieved using a simple and well-known method called the *Cholesky decomposition* (or *Cholesky factorization*).

Given an estimate of the approximate covariance matrix for the residuals, $\widehat{\Sigma}_i$, the Cholesky decomposition of $\widehat{\Sigma}_i$ can be used to create a lower triangular matrix, L_i, such that

$$\widehat{\Sigma}_i = L_i L'_i;$$

note that a lower triangular matrix is simply one with all zeros above the diagonal. We can then use the matrix L_i or, more specifically, L_i^{-1}, to take us from a set of correlated residuals with heterogeneous variances to a set of transformed residuals,

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\widehat{\beta}), \tag{9.2}$$

which are uncorrelated and have unit variance.

Interestingly, the transformation used in (9.2) leads to a set of transformed residuals with appealing interpretations in the longitudinal setting. For example, the first element of

$$r_i^* = L_i^{-1}(Y_i - X_i\widehat{\beta})$$

is the standardized residual for the first repeated measurement (often the baseline observation). In contrast, the second through last transformed residuals represent standardized deviations from the conditional mean of the response given all previous observations. For example, the k^{th} transformed residual is an estimate of

$$\frac{Y_{ik} - E(Y_{ik}|Y_{i1}, .. \; Y_{ik-1})}{\sqrt{\text{Var}(Y_{ik}|Y_{i1}, ..., Y_{ik-1})}}.$$

This property of the transformed residuals for longitudinal data is not obvious; it requires a deeper understanding of matrix algebra than is assumed throughout this book.

Given the set of transformed residuals, r_i^*, all of the usual residual diagnostics for standard linear regression can be applied. For example, we can construct a scatter-plot of the transformed residuals, r_{ij}^*, versus the transformed predicted values, $\widehat{\mu}_{ij}^*$, where

$$\widehat{\mu}_i^* = L_i^{-1}\,\widehat{\mu}_i = L_i^{-1}\,X_i\widehat{\beta}.$$

In a correctly specified model, this scatter-plot should display no systematic pattern, with a random scatter around a constant mean of zero and with a constant range for varying $\widehat{\mu}_{ij}^*$. Similarly, we can construct a scatter-plot of the transformed residuals versus selected transformed covariates. With longitudinal data, a scatter-plot of the transformed residuals versus transformed time (or age) can be particularly useful for assessing the adequacy of the model assumptions about patterns of change in the mean response over time. Finally, the transformed residuals also make it somewhat easier to identify skewness and potential outliers that require further investigation. A

normal quantile plot (or so-called quantile-quantile or Q-Q plot) of the transformed residuals can be used to assess the normal distribution assumption and to identify outliers. That is, on the basis of the ranks of the transformed residuals, we can plot the sample quantiles of the residuals against the quantiles expected if they have a normal distribution. If the residuals depart discernibly from a straight line then the assumption of normality may not be tenable. For example, skewness is usually indicated by a bow-shaped pattern in the normal quantile plot; outliers will appear as "stragglers", far from the ends of the line.

When statistical software is available that automates the production of residual diagnostics for standard linear regression, the following procedure can be used. Given the estimated covariance, $\widehat{\Sigma}_i$, L_i can be obtained from the Cholesky decomposition of $\widehat{\Sigma}_i$. Then, a transformed response vector and covariate matrix can be constructed as follows:

$$Y_i^* = L_i^{-1} Y_i; \quad X_i^* = L_i^{-1} X_i.$$

Finally, the generalized least squares (GLS) estimate of β, from the regression of Y_i on X_i (with estimated covariance matrix, $\widehat{\Sigma}_i$), can be re-estimated from the ordinary least squares (OLS) regression of Y_i^* on X_i^*. That is, any standard linear regression program can be used to model the dependence of Y_i^* on X_i^*, and all the built-in residual diagnostics from the resulting OLS regression can be examined to check the adequacy of the model. Thus, once a model has been selected for the mean and the covariance, standard regression diagnostics for independent observations (with homogeneous variance) can be applied by re-fitting a standard linear regression of Y_i^* on X_i^* and making use of available residual diagnostic tools.

As mentioned earlier, the transformed residuals are useful for detecting outlying *observations*. They can also be used to detect outlying *individuals*. For each individual we can calculate a summary measure of multivariate distance between their observed and fitted responses, based on the *Mahalanobis distance*,

$$d_i = r_i^{*\prime} r_i^*. \tag{9.3}$$

If the model is correctly specified, the distances given by (9.3) have an approximate chi-squared distribution with degrees of freedom (df) equal to the dimension of r_i^* (i.e, df = n_i, the number of repeated measurements on the i^{th} subject). Outlying individuals will have distances, d_i, that have small associated p-values. The p-values provide a common metric for comparing and detecting large values of d_i, corresponding to unusual or outlying individuals, when the number of repeated measurements varies across subjects. A word of caution concerning the interpretation of these p-values. Because the major focus is on the most extreme values of d_i, the distribution of these extremes (e.g., the distribution of the maximum d_i) is somewhat more complicated than a chi-squared distribution with n_i degrees of freedom. In principle, a Bonferroni correction to the p-values could be applied (e.g., multiplying the p-values by the sample size, N); however, the Bonferroni correction is known to be very conservative. In general, we recommend that the p-values be used as a common metric for comparing d_i when the number of repeated measurements varies across subjects, while recognizing that p-values less than 0.05 occur with predictable regularity (when the sample size is 200, the expected number is $200 \times 0.05 = 10$).

So far, much of the discussion of residual diagnostics has focused on the adequacy of the model for the mean response. As alluded to above, the adequacy of the variance assumption can be informally assessed by examining the scatter-plot of the transformed residuals versus the transformed predicted values and/or time. In a correctly specified model for the variance, the range of the transformed residuals should be approximately constant over (transformed) time and for varying $\widehat{\mu}_{ij}^*$. A more informative plot is obtaining by considering the scatter-plot of the absolute values of the transformed residuals, $|r_{ij}^*|$, versus (transformed) time and/or $\widehat{\mu}_{ij}^*$ If the assumed model for the variance is adequate, there should be no systematic trend. The fitting of a smooth curve to the scatter-plot (e.g., a lowess curve) can often help in judging whether any curvature is present. The fitted curve should display no systematic departures from a horizontal line centered at approximately 0.8; note that, if the transformed residuals are assumed to be normal, with mean zero and unit variance, then the mean of the absolute values of the residuals is 0.798. Finally, an informal check on the overall adequacy of the model for the covariance, both the models for the variances and correlations, is provided by a smoothed plot of the so-called *empirical semi-variogram*. The definition of the empirical semi-variogram, and a description of its use as a diagnostic tool, is given in the next section.

9.4 SEMI-VARIOGRAM

Historically, the semi-variogram has been widely used in spatial statistics to represent the covariance structure in geostatistical data. Unlike two-dimensional spatial data, the coordinates for longitudinal data are along a single dimension, namely, time. For longitudinal data the semi-variogram is defined as one-half the expected squared difference between residuals obtained on the same individual. The semi-variogram, denoted $\gamma(h_{ijk})$, is given by

$$\gamma(h_{ijk}) = \frac{1}{2} E(r_{ij} - r_{ik})^2, \tag{9.4}$$

where h_{ijk} is the time elapsed between the j^{th} and k^{th} repeated measurement on the i^{th} individual. Since the residuals have mean zero, the semi-variogram given by (9.4) can be expressed as

$$\begin{aligned} \gamma(h_{ijk}) &= \tfrac{1}{2} E(r_{ij} - r_{ik})^2 \\ &= \tfrac{1}{2} E(r_{ij}^2 + r_{ik}^2 - 2 r_{ij} r_{ik}) \\ &= \tfrac{1}{2} \mathrm{Var}(r_{ij}) + \tfrac{1}{2} \mathrm{Var}(r_{ik}) - \mathrm{Cov}(r_{ij}, r_{ik}). \end{aligned}$$

Although the semi-variogram can be used to suggest appropriate models for the covariance, here we simply use it as a diagnostic tool for assessing the adequacy of a selected model for the covariance. When the semi-variogram is applied to the transformed residuals, r_{ij}^*, it simplifies to

$$\gamma(h_{ijk}) = \frac{1}{2} \mathrm{Var}(r_{ij}^*) + \frac{1}{2} \mathrm{Var}(r_{ik}^*) - \mathrm{Cov}(r_{ij}^*, r_{ik}^*) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

Thus, in a correctly specified model for the covariance matrix, the plot of the semi-variogram for the transformed residuals versus the time elapsed between the corresponding observations should fluctuate randomly around a horizontal line centered at 1.

The empirical or sample semi-variogram, $\widehat{\gamma}(h)$, is simply defined as one-half the average squared difference between pairs of residuals on the same individual whose corresponding observations are h units apart (and where the average is taken over all pairs of observations for which $h_{ijk} = h$). With inherently unbalanced longitudinal data, where subjects are not all measured at the same set of occasions, the empirical semi-variogram can be estimated by fitting a smooth curve (e.g., a lowess curve) to the scatter-plot of the observed half squared differences between residuals obtained on the same individual and the corresponding time lags. In a correctly specified model for the covariance matrix, a smooth plot of the empirical semi-variogram for the transformed residuals should be centered at 1 and display no systematic curvature. However, the construction of a smooth plot of the empirical semi-variogram requires extra care. Because the empirical semi-variogram is based on the squared differences between pairs of residuals it can be very sensitive to outliers; furthermore, because each residual contributes to $n_i - 1$ squared differences between pairs of residuals on the same individual, a single outlier can have an inordinate influence at several different time lags.

Finally, recall that in the linear mixed effects model Σ_i has a characteristic random effects structure given by

$$\Sigma_i = Z_i G Z_i' + \sigma^2 I_{n_i}.$$

Transformed residuals from the linear mixed effects model can be obtained by taking the Cholesky decomposition of $\widehat{\Sigma}_i = Z_i \widehat{G} Z_i' + \widehat{\sigma}^2 I_{n_i}$. The adequacy of the random effects covariance structure can be assessed from the plot of the empirical semi-variogram for the transformed residuals. In addition, with linear mixed effects models we can obtain predictions of the random effects (empirical BLUPs) and examine their distribution for any evidence of extremes or outliers, perhaps representing individuals whose subject-specific response profiles are somewhat unusual. Because the empirical BLUPs are known to be heavily influenced by the normal distribution assumption for the random effects, we caution that histograms and normal quantile plots of the empirical BLUPs should not be used to assess the adequacy of the normal distribution assumption for the random effects. Furthermore, because the empirical BLUPs have been "shrunk" toward the population fixed effects, β, their distribution does not accurately represent the distribution of the random effects (e.g., due to "shrinkage" toward the population mean, empirical BLUPs have smaller variance).

9.5 CASE STUDY

In Section 8.8 we presented the results of analyses of body fat accretion from a prospective study of the development of obesity in a cohort of girls. In this section we examine the residuals from the fitted model to assess the overall adequacy of the model. The residual analyses are also used to detect individuals with unusual response profiles.

Table 9.1 Estimated regression coefficients (fixed effects) and standard errors for the piecewise linear model for the percent body fat data.

Variable	Estimate	SE	Z
Intercept	21.3614	0.5646	37.84
Time	0.4171	0.1572	2.65
$(\text{Time})_+$	2.0471	0.2280	8.98

Study of Influence of Menarche on Changes in Body Fat Accretion

Recall that the data are from a prospective longitudinal study examining changes in body fat before and after menarche in a cohort of 162 girls from the MIT Growth and Development Study (Bandini *et al.*, 2002; Phillips *et al.*, 2003). At the start of the study, all of the girls were pre-menarcheal and non-obese. They were followed over time according to a schedule of annual measurements until four years after menarche. At each examination, a measure of body fatness, percent body fat (%BF) was derived from three basic measurements of (1) body weight, (2) height, and (3) bioelectric impedance resistance.

In Section 8.8 we presented analyses of the changes in percent body fat before and after menarche. For these analyses "time" was coded as time since menarche and could be positive or negative. We considered the hypothesis that percent body fat increases linearly with age, but with different slopes before and after menarche. Specifically, we assumed that each girl had a piecewise linear spline growth curve with a knot at the time of menarche and fitted the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2\, t_{ij} + \beta_3\, (t_{ij})_+ + b_{1i} + b_{2i}\, t_{ij} + b_{3i}\, (t_{ij})_+,$$

where t_{ij} denotes the time of the j^{th} measurement on the i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche), $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$. In this model, each girl's growth curve can be described with an intercept and two slopes, one slope for the changes in response before menarche, another slope for the changes in response after menarche.

The REML estimates of the fixed effects are displayed in Table 9.1 (the REML estimates of the variance components for the random effects are displayed in Table 8.7 in Chapter 8). The main goal of the analysis was to assess whether the population slopes for fat accretion differ before and after menarche. Based on the magnitude of the estimate of β_3, relative to its standard error, it was concluded that there was a significant difference between the slopes before and after menarche. In particular, the

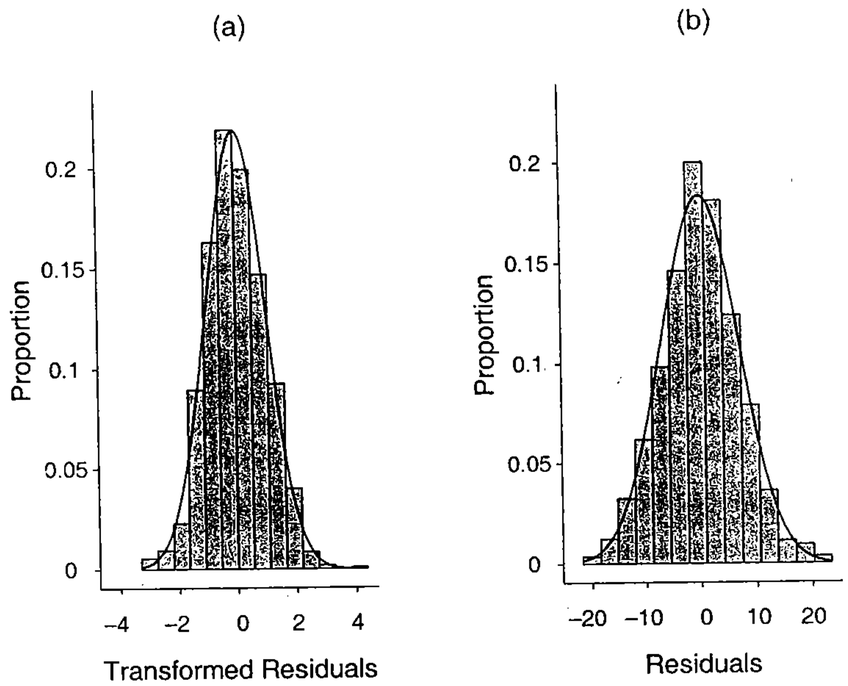

Fig. 9.1 Histogram, with normal density overlaid, of (a) the transformed residuals, and (b) the untransformed residuals, for the percent body fat data.

estimated pre-menarcheal slope is rather shallow (0.42) and indicates that the annual rate of body fat accretion is less that 0.5%. In contrast, the estimated post-menarcheal slope is 2.46 and indicates that the annual rate of body fat accretion is approximately 2.5%, almost six times higher than the corresponding rate in the pre-menarcheal period.

Next we use residual diagnostics to assess the adequacy of the fitted model. Based on the Cholesky decomposition of the estimated covariance matrix, $\widehat{\Sigma}_i$, we can calculate transformed residuals,

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\widehat{\Sigma}_i = L_i L_i'$. For illustrative purposes, we also examine the untransformed residuals and compare the diagnostic plots based on these two types of residuals.

Histograms of the transformed and untransformed residuals are presented in Figure 9.1 and do not indicate any discernible skewness. In addition, the normal quantile plots of the residuals do not display any systematic departures from a straight line (see Figure 9.2). The quantile plot of the transformed residuals does reveal one very extreme observation. This observation corresponds to a measurement on a girl, with subject ID = 128, obtained prior to menarche. Approximately two years prior to

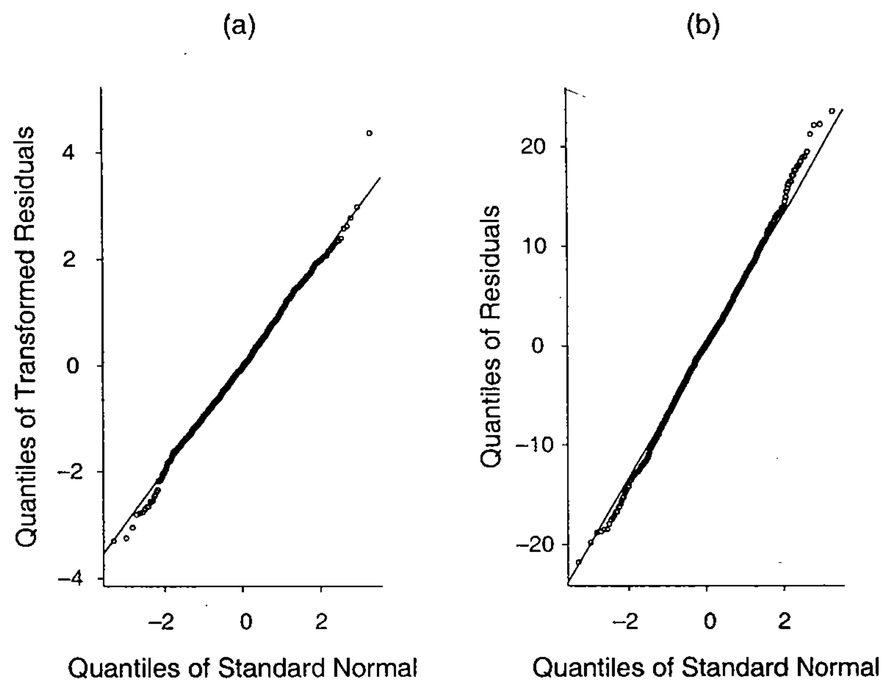

Fig. 9.2 Normal quantile plot of (a) the transformed residuals, and (b) the untransformed residuals, for the percent body fat data.

menarche, this girl's percent body fat increased to 44% (from 27% one year earlier); this is an extreme value, over 3 standard deviations above the mean for pre-menarcheal percent body fat. However, the observation is not a recording error and this girl had subsequent measurements of percent body fat of 40% and 41% at the next two occasions. Overall, the number of extreme residuals highlighted by Figure 9.2 is not more than what we would expect due to chance, given a total of 1049 observations. From an examination of the residuals in Figures 9.1 and 9.2, there is no evidence to suggest any discernible skewness and the normal assumption appears to be tenable.

Next we consider scatter-plots of the transformed and untransformed residuals versus the transformed and untransformed predicted values respectively. The scatter-plots of the residuals in Figure 9.3 display no obvious systematic pattern, with a random scatter around a constant mean of zero. However, when lowess smoothed curves are superimposed on the scatter-plots, they do reveal some apparent curvature. Focusing on the transformed residuals, there appears to be a quadratic trend, although the fall in the lowess curve at the largest values of the transformed predicted values should be cautiously interpreted as the fitted curve is based on few observations at the extremities and is therefore likely to be unreliable in that region.

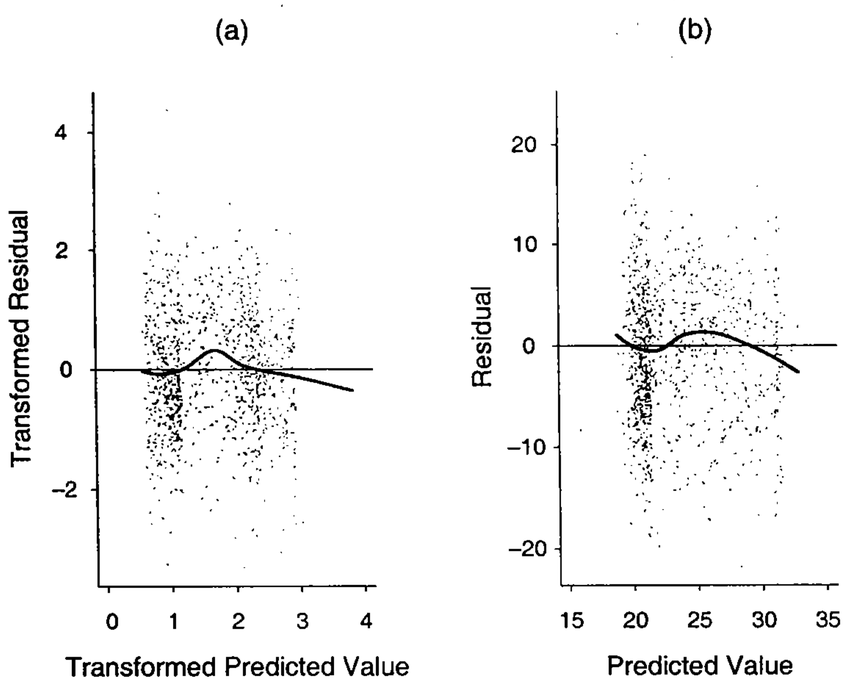

Fig. 9.3 Scatter-plot of (a) the transformed residuals versus transformed predicted values, and (b) the untransformed residuals versus predicted values, for the percent body fat data.

Because of the suggestion of curvature in Figure 9.3, we next examine scatter-plots of the (transformed) residuals versus (transformed) time (see Figure 9.4). These scatter-plots of the transformed and untransformed residuals suggest curvature at (untransformed) times corresponding to approximately 2 to 4 years post-menarche. The pattern is more apparent in the scatter-plot of the transformed residuals and can no longer be discounted due to sparseness of the observations at the extremities; this is the first pair of plots in which the transformed and untransformed data give a different impression. The curvature in the scatter-plots suggests that the model for the mean response might be improved by the inclusion of a quadratic trend in the post-menarcheal period.

Before considering a refinement to the model for the mean response, we illustrate how the transformed residuals can be used to identify unusual individuals. We can calculate the Mahalanobis distance,

$$d_i = r_i^{*\prime} r_i^*,$$

for each girl and then compare the values to reference chi-squared distributions with degrees of freedom (df) equal to the dimension of r_i^* (i.e, df = n_i, the number of repeated measurements obtained on each girl). For each girl, we calculated d_i and

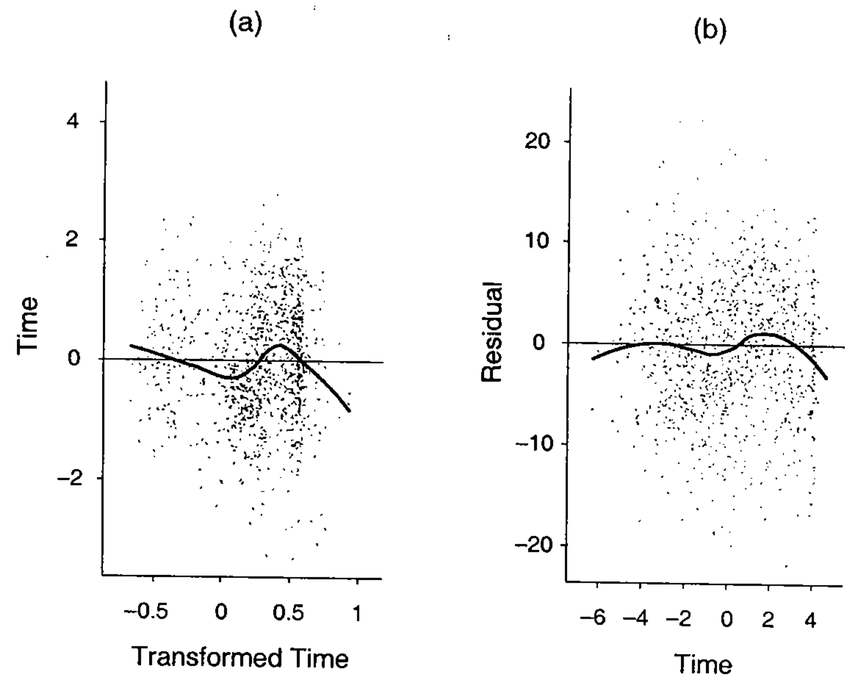

Fig. 9.4 Scatter-plot of (a) the transformed residuals versus transformed time, and (b) the untransformed residuals versus time, for the percent body fat data.

its associated p-value. There were 7 girls whose d_i yielded p-values less than 0.05 and 2 girls with p-values less than 0.01. Given that the sample is comprised of 162 girls, distances of these magnitudes are to be expected by chance alone. Nonetheless, it is useful to identify the 2 girls whose d_i yielded p-values less than 0.01. They correspond to girls with subject ID = 128 and subject ID = 79. Recall that the former was identified as having extreme observations prior to menarche, with percent body fat increasing from 27% to 44% in a one year interval. The other girl, with subject ID = 79, is unusual because she displayed a sudden decrease in percent body fat, from approximately 37% two years prior to menarche to 20% around the time of menarche, and then maintained that level of percent body fat during the 3 years post-menarche. This pattern of a drop in percent body fat prior to menarche, coupled with almost no gain post-menarche, is at odds with the general pattern of change in the population.

Next we consider a refinement to the model to allow for a quadratic trend in the post-menarcheal period. In particular, we assume that each girl has a piecewise linear-quadratic growth curve with a knot at the time of menarche and fit the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + \beta_4 (t_{ij})_+^2 + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+ + b_{4i} (t_{ij})_+^2,$$

Table 9.2 Estimated regression coefficients (fixed effects) and standard errors for the piecewise linear-quadratic model for the percent body fat data.

Variable	Estimate	SE	Z
Intercept	20.4201	0.5817	35.10
Time	−0.0155	0.1612	−0.10
$(\text{Time})_+$	4.8439	0.4055	11.94
$(\text{Time})_+^2$	−0.6469	0.0772	−8.38

where $(t_{ij})_+^2 = t_{ij}^2$ if $t_{ij} > 0$ and $(t_{ij})_+^2 = 0$ if $t_{ij} \leq 0$. In this model, each girl has a separate growth curve that can be described in terms of a linear trend for changes in response before menarche, and a quadratic trend for changes in response after menarche.

The REML estimates of the fixed effects, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$, are displayed in Table 9.2. These results suggest that there is significant non-linearity in the post-menarcheal trend. The estimate of β_4 indicates that increases in percent body fat are greatest around the time of menarche but level off at approximately 4 years following the onset of menarche. The results also suggest that there is no significant increase in percent body fat during the 3 to 4 years prior to menarche.

For this revised model, we consider scatter-plots of the (transformed) residuals versus (transformed) time (see Figure 9.5). The scatter-plots of the transformed and untransformed residuals do not reveal any obvious systematic trends. When lowess smoothed curves are superimposed on the scatter-plots, the curvature that was apparent in Figure 9.4(a) is no longer discernible in Figure 9.5(a). The apparent curvature in Figure 9.5(b), at approximately 5 to 6 years prior to menarche, can be discounted because the fitted curve is based on so few observations at this extremity and is therefore likely to be unreliable in that region. The inclusion of a quadratic trend in the post-menarcheal period has led to an improvement in fit as determined by both the Wald test for the quadratic trend ($Z = -8.38, p < 0.0001$) and the examination of residual diagnostics.

So far, we have considered residual diagnostics for assessing the goodness of fit of the assumed model for the mean response. We can also assess the adequacy of the model for the variance by constructing a scatter-plot of the absolute values of the transformed residuals, $|\widehat{r}_{ij}^*|$, versus the transformed predicted values and transformed time (see Figure 9.6). The scatter-plots in Figure 9.6 indicate that there is no obvious systematic trend. When lowess smoothed curves are superimposed on the scatter-plots there is no evidence of a discernible departure from a straight line centered at approximately 0.8. Recall that if the transformed residuals are normal, with mean zero and unit variance, then the mean of the absolute values of the transformed residuals

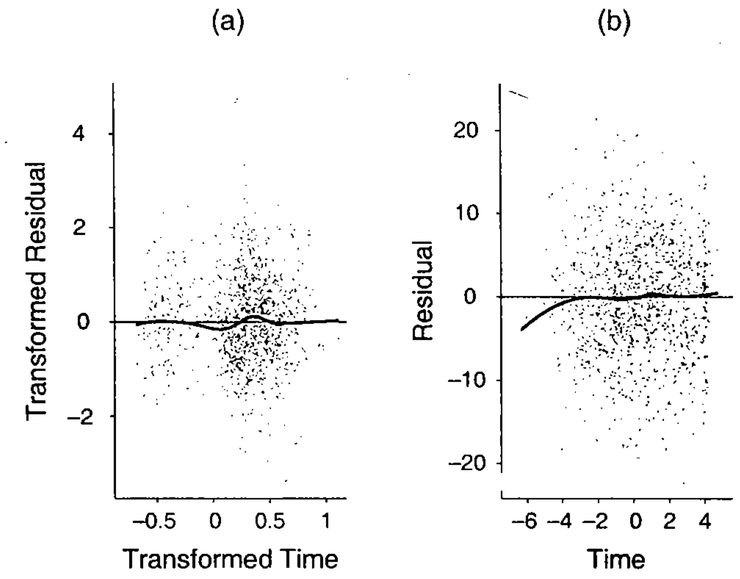

Fig. 9.5 Scatter-plot of (a) the transformed residuals versus transformed time, and (b) the untransformed residuals versus time, for the revised model for the percent body fat data.

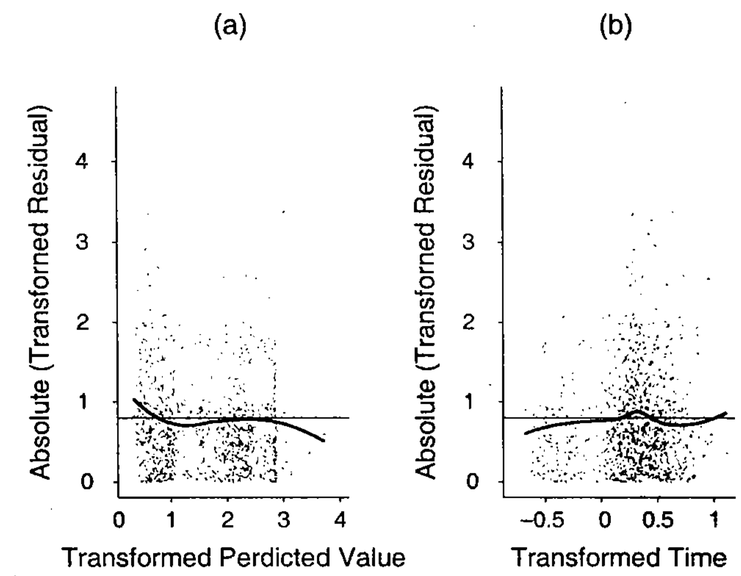

Fig. 9.6 Scatter-plot of the absolute value of the transformed residuals versus (a) transformed predicted values, and (b) transformed time, for the revised model for the percent body fat data.

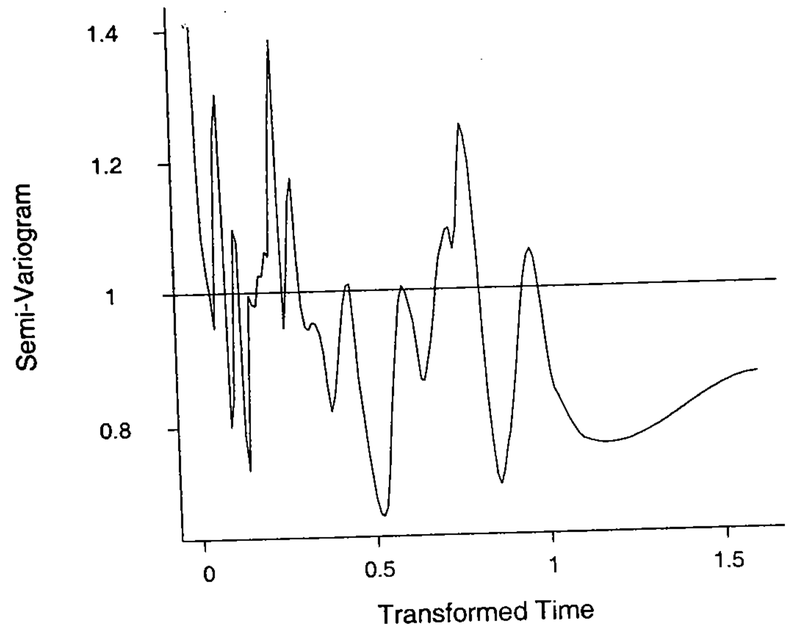

Fig. 9.7 Empirical semi-variogram, estimated by fitting a lowess smoothed curve, for transformed residuals obtained from the revised model for the percent body fat data.

is 0.798. Thus we conclude that the variability is approximately constant for varying (transformed) predicted values and times.

Finally, the adequacy of the overall model for the covariance matrix can be assessed by examining the empirical semi-variogram for the transformed residuals (see Figure 9.7). The empirical semi-variogram, estimated by fitting a lowess smoothed curve, appears to fluctuate randomly around the horizontal line centered around 1; it does not display any obvious systematic trend over time. This suggests that the assumed random effects structure for the covariance matrix is adequate for these data. For illustrative purposes, we consider the empirical semi-variogram for the transformed residuals from the following mixed effects model,

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2\, t_{ij} + \beta_3\, (t_{ij})_+ + \beta_4\, (t_{ij})_+^2 + b_{1i}.$$

This model retains the same set of fixed effects as before (i.e., the same model for the marginal means), but makes the strong assumption that the covariance matrix has a compound symmetry structure (i.e., a random intercepts only model). The adequacy of the compound symmetry model for the covariance matrix can be assessed by examining the empirical semi-variogram for the transformed residuals estimated from this model. Figure 9.8 displays the empirical semi-variogram, estimated by fitting a lowess smoothed curve. The empirical semi-variogram is no longer centered

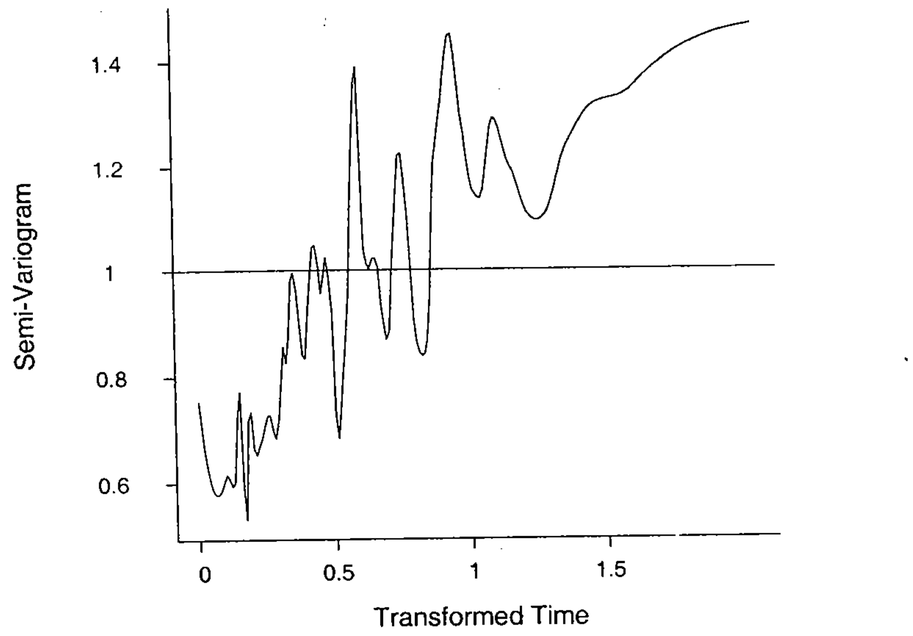

Fig. 9.8 Empirical semi-variogram, estimated by fitting a lowess smoothed curve, for transformed residuals obtained from the revised model for the fixed effects and with a compound symmetry assumption for the covariance matrix.

around 1 and displays an increasing trend with time. This indicates that the compound symmetry assumption of constant variance and constant correlation is not tenable for these data.

9.6 SUMMARY

In this chapter we have seen that many of the standard techniques for residual diagnostics and outlier detection in the univariate regression setting can be readily extended to longitudinal data. Simple residuals, based on the observed minus predicted responses at each occasion, are straightforward to calculate and can be produced by most statistical software packages for analyzing longitudinal data. Although these residuals have some shortcomings, they are probably adequate for most practical purposes. Systematic departures from model assumptions should be revealed by standard residual diagnostic plots of these simple residuals.

In Section 9.3 we discussed a particular transformation of the residuals for longitudinal data that makes them mimic residuals from a standard linear regression. The transformation is based on the Cholesky decomposition of the covariance matrix.

Given an estimate of the covariance among the residuals, the Cholesky decomposition of the covariance matrix can be implemented in many standard statistical software packages. The transformed residuals can also be produced as standard output from some statistical packages (e.g., using the "normalized" residuals option with the *lme* function in S-PLUS and the VCIRY option with PROC MIXED in SAS).

Residual diagnostics for assessing the adequacy of the model for the covariance among repeated measurements are less well developed. The adequacy of the variance assumption can be informally assessed by examining a scatter-plot of the absolute values of the transformed residuals versus the predicted values and/or time. A check on the adequacy of the overall model for the covariance is provided by a smoothed plot of the empirical semi-variogram

Finally, residual analysis is useful for detecting outlying observations. The detection of outlying observations is important because they can potentially have an inordinate influence on the analysis. Outliers require further investigation to ensure that they are not due to recording errors. When it has been established that the outliers are not the result of recording errors or other types of mistakes, then it can be useful to replicate the analysis with and without the outlying observations to determine their influence on substantive conclusions. However, we do not recommend the automatic exclusion or down-weighting of outliers. The residuals can also be used to identify outlying individuals who have unusual response profiles.

9.7 FURTHER READING

Methods for residual analysis in standard linear regression for independent observations are well developed; a comprehensive description of techniques for residual analysis can be found in Cook and Weisberg (1982), and in many standard textbooks on linear regression. A discussion of residual diagnostics and the use of the semi-variogram for longitudinal data can be found in the review article by Laird *et al.* (1992).

Bibliographic Notes

A discussion of the generalization of residual diagnostic to longitudinal data can be found in articles by Waternaux *et al.* (1989) and Waternaux and Ware (1991).

Historically, the semi-variogram has been widely used in spatial statistics to represent the covariance structure in geostatistical data. The use of the semi-variogram for longitudinal data is described in details in Chapter 10 (Section 10.4) of Verbeke and Molenberghs (2000) and Chapter 3 (Section 3.4) of Diggle *et al.* (2002); also, see Chapters 2 (Section 2.5) and 5 (Section 5.4) of Diggle (1990).

Problems

9.1 In Section 8.8 we presented the results of analyses of a subset of the pulmonary function data collected in the Six Cities Study of Air Pollution and Health (Dockery *et al.*, 1983). The data consist of measurements of FEV_1, height, and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. Specifically, we considered the following model for $\log(FEV_1)$:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 \mathtt{Age}_{ij} + \beta_3 \mathtt{log(Ht)}_{ij} + \beta_4 \mathtt{Age}_{i1} + \beta_5 \mathtt{log(Ht)}_{i1} + b_{1i} + b_{2i}\mathtt{Age}_{ij},$$

where Y_{ij} is the $\log(FEV_1)$ for the i^{th} child at the j^{th} visit, $\mathtt{Age}_{i1}$ and $\mathtt{log(Ht)}_{i1}$ are the initial or baseline age and log(height) for the i^{th} child. In this exercise, we examine the residuals from the fitted model to assess the overall adequacy of the model for the mean response.

The raw data are stored in an external file: `topeka.dat`

Each row of the data set contains the following six variables:

ID Height Age Initial Height Initial Age $\log(FEV_1)$

9.1.1 Calculate the untransformed residuals,

$$r_{ij} = Y_{ij} - \widehat{\mu}_{ij},$$

from the fitted model given above, where

$$\widehat{\mu}_{ij} = \widehat{\beta}_1 + \widehat{\beta}_2\, \mathtt{Age}_{ij} + \widehat{\beta}_3\, \mathtt{log(Ht)}_{ij} + \widehat{\beta}_4\, \mathtt{Age}_{i1} + \widehat{\beta}_5\, \mathtt{log(Ht)}_{i1}.$$

9.1.2 Construct a histogram of the residuals. Comment on the shape of the distribution of the residuals.

9.1.3 Construct a normal quantile plot (Q-Q plot) of the residuals. Does the plot display any systematic departures from a straight line? Does the plot suggest any potential outlying observations?

9.1.4 On a single graph, construct a scatter-plot of the residuals versus the predicted values and superimpose a lowess smoothed curve on the scatter-plot. Does the plot display any systematic pattern?

9.1.5 On a single graph, construct a scatter-plot of the residuals versus age and superimpose a lowess smoothed curve on the scatter-plot. Does the plot display any systematic pattern?

9.1.6 On a single graph, construct a scatter-plot of the residuals versus log(height) and superimpose a lowess smoothed curve on the scatter-plot. Does the plot display any systematic pattern?

9.1.7 On the basis of the residual diagnostics from Problems 9.1.2 through 9.1.6, comment on the overall adequacy of the model for the mean response. Can you suggest how the model for the mean response might be improved?

Part III

Generalized Linear Models for Longitudinal Data

10

Review of Generalized Linear Models

10.1 INTRODUCTION

In Part II we considered methods for analyzing longitudinal data when the response variable is continuous. In many biomedical applications the longitudinal response is not continuous, for example, the presence or absence of respiratory illness, or counts of the number of epileptic seizures in a four-week interval. When the longitudinal response is discrete (e.g., binary or a count) the methods that we have discussed in Part II are no longer appropriate.

In Part III we focus on methods for analyzing discrete longitudinal data. When the response is discrete, linear models are no longer appropriate for relating changes in the mean response to covariates. Instead, we consider extensions of *generalized linear models* for longitudinal data.

Generalized linear models provide a unified class of models for regression analysis of independent observations of a discrete or continuous response. A straightforward application of generalized linear models to longitudinal data is not appropriate due to the correlation (or lack of independence) among observations obtained from the same individual. Instead, we consider extensions of this broad class of models to handle longitudinal responses. There are many ways to extend generalized linear models to account for the correlation among longitudinal observations, we consider two general, but quite distinct, approaches in Chapters 11 and 12.

In Chapter 11 we present a unified methodology for analyzing longitudinal data when the response variable is discrete or continuous. It does not require distributional assumptions for the observations, only a regression model for the mean response. That is, we describe a general method for analyzing diverse types of longitudinal responses

that avoids making assumptions about the distribution of the vector of responses; the method relies solely on assumptions about the mean response. Recall that in previous chapters we noted that the multivariate normal assumption was not so crucial in longitudinal analysis of a continuous response, provided N, the number of subjects, is relatively large in comparison to n, the number of repeated measures and any missing data can be assumed to be MCAR. In Chapter 11, we will provide some rationale for why the distributional assumption for the vector of responses can be relaxed. In Chapter 12 we consider an alternative extension of generalized linear models that accounts for the correlation among longitudinal data via the introduction of random effects. These models extend in a natural way the conceptual approach represented by the linear mixed effects models discussed in Chapter 8. In Part III, we focus primarily on longitudinal analysis of a discrete response, although the general methodology described in these chapters can be applied equally to continuous responses.

A characteristic feature of generalized linear models is that a suitable non-linear transformation of the mean response is a linear function of the covariates. This non-linearity raises some additional issues concerning the interpretation of the regression coefficients in models for longitudinal data. That is, different approaches for accounting for the source of within-subject association among the longitudinal data can lead to models having regression coefficients with quite distinct interpretations. As a result, for the same data, there will be differences between the estimated regression coefficients obtained from the two distinct classes of models described in Chapters 11 and 12. In general, the choice among different classes of models for discrete longitudinal data must be made on subject-matter grounds.

One of the underlying themes that will be emphasized in Part III is that different models for discrete longitudinal data have somewhat different targets of inferences. Thus, to ensure that the regression model parameters bear directly on the question of scientific interest, somewhat greater care is needed in the choice of model for discrete longitudinal data.

10.2 SALIENT FEATURES OF GENERALIZED LINEAR MODELS

In this section we provide a non-technical summary of the most salient features of generalized linear models for a single, univariate response. In later chapters, we discuss how generalized linear models can be extended to handle longitudinal responses. A good grasp of the material in this section is all that is required for an understanding of the methodology for longitudinal data that will be described in Chapters 11 and 12. In Section 10.5 we present a detailed and somewhat more technical overview of generalized linear models. Many of our readers, in particular, those encountering this topic for the first time, may find the material in Section 10.5 challenging. While we encourage all of our readers to skim through Section 10.5, we note that it can be omitted without loss of continuity.

Generalized linear models provide a unified method for analyzing diverse types of univariate responses (e.g., continuous, binary, counts). Generalized linear models are actually a broad class or collection of regression models and they include as special

cases the standard linear regression and analysis of variance (ANOVA) models for a normally distributed continuous response, logistic regression models for a binary or dichotomous response, and log-linear or Poisson regression models for counts. Although generalized linear models encompass a much broader range of regression models, these three are among the most widely used regression models in biomedical research. In this chapter we focus primarily on generalized linear models for binary and count data since, with the exception of continuous responses, these two data types are by far the most commonly encountered in applications.

Notation

Throughout this chapter we assume that we have N *independent* observations of a response variable, Y. We let Y_i $(i = 1, \ldots, N)$ denote the response variable for the i^{th} subject. Associated with each response, Y_i, is a $p \times 1$ vector of covariates,

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad i = 1, ..., N;$$

where X_{ik} denotes the k^{th} covariate for the i^{th} subject. Typically, although not always, $X_{i1} = 1$ for all i, and then β_1 is the intercept term in the regression model. Generalized linear models extend the standard linear regression model in a number of important ways, while also retaining some of its distinctive features. In particular, a generalized linear model for Y_i has the following three-part specification:

1. a distributional assumption,
2. a systematic component, and
3. a link function.

We consider each of these three components in turn.

Distributional Assumption

Generalized linear models extend many of the basic concepts and ideas of standard linear regression analysis to settings where the response variable is discrete and can no longer be assumed to have a normal distribution. In particular, they extend the class of probability distributions for the response to include many of the distributions commonly used for modelling discrete responses. Generalized linear models assume that the response variable has a probability distribution belonging to the so-called *exponential family* of distributions. The exponential family includes many distributions that the reader may already have encountered. For example, the normal, Bernoulli, binomial, and Poisson distributions all belong to the exponential family. The first component of a generalized linear model, the distributional assumption, specifies the random component of the model. That is, it specifies a probabilistic mechanism by which the responses are assumed to be generated.

Table 10.1 Canonical link and variance functions for the normal, Bernoulli, and Poisson distributions.

Distribution	Variance Function, $v(\mu)$	Canonical Link
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1-\mu)$	Logit: $\log\left(\frac{\mu}{1-\mu}\right) = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

Because the normal, Bernoulli, binomial, and Poisson distributions are members of the same family, they share some common statistical properties. In particular, the variance of the response can be expressed in terms of the product of a single scale or dispersion parameter, ϕ, and a so-called *variance function*, denoted $v(\mu_i)$; the latter being a known function of the mean, μ_i. That is,

$$\text{Var}(Y_i) = \phi v(\mu_i),$$

where $\phi > 0$. The variance function, $v(\mu_i)$, describes how the variance of the response is functionally related to the mean of the response. The variance functions for the normal, Bernoulli, and Poisson distributions are summarized in Table 10.1. For many distributions for discrete data, ϕ is not a parameter that requires estimation but is a known constant (e.g., $\phi = 1$ for the Bernoulli and Poisson distributions); for other distributions ϕ is an unknown parameter (e.g., ϕ is the variance of the normal distribution).

For the Bernoulli and Poisson distributions, the variance depends on the mean. This dependence of the variance on the mean is a characteristic feature of most distributions for discrete responses. On the other hand, for the normal distribution, the variance does not depend on the mean, that is, $\text{Var}(Y_i) = \phi$ (and the variance function, $v(\mu_i) = 1$). This provides some rationale for why the assumption of homogeneity of variance (or common variance) is generally adopted in the standard linear regression model for normally distributed responses. In some applications, however, the homogeneity of variance assumption is too restrictive and the variance may depend upon covariates. In later sections we briefly mention how restrictive assumptions about the variance of Y_i can be relaxed.

Systematic Component

Generalized linear models not only share a common family of distributions, they also share a common regression formulation. An important aspect of the standard linear regression model that is retained in all generalized linear models is the linear regression component. This is the *systematic component* of a generalized linear

model and it specifies that the effects of the covariates, X_i, on the mean of Y_i can be expressed in terms of the following *linear predictor*, denoted by η_i,

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where, typically, $X_{i1} = 1$ for all i, and then β_1 is the intercept. The linear predictor is simply a linear combination of the unknown regression coefficients, $\beta = (\beta_1, \ldots, \beta_p)'$ and the covariates, X_i.

The key word here is *linear*. The term "linear" in generalized linear models means that η_i must be linear in the regression parameters. This implies that the mean response (or any transformation of the mean response) can be expressed as a simple weighted sum of the regression parameters, β. For example,

$$\eta_i = \beta_1 + \beta_2 X_i,$$

$$\eta_i = \beta_1 + \beta_2 \log(X_i),$$

and

$$\eta_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2,$$

are all cases where η_i is linear in the regression coefficients, even if it is non-linear in X_i. However,

$$\eta_i = \beta_1 + e^{\beta_2 X_i},$$

and

$$\eta_i = \beta_1 / (1 + \beta_2 e^{-\beta_3 X_i})$$

are examples where η_i is not linear in the regression parameters and the latter types of non-linearities are not included in the class of generalized linear models.

Thus, the linearity strictly applies to the regression parameters, β, but not necessarily to the covariates. As a result, the restriction to a linear structure for the covariates does not preclude relationships between the mean response and the covariates that are non-linear. This latter type of non-linearity is easily accommodated by taking appropriate transformations of the mean response (see below) and/or by transformation of the covariates (e.g., log(X)).

Link Function

The final way in which generalized linear models extend the standard linear regression model is by taking a suitable transformation of the mean response and relating the transformed mean response to the covariates. This is achieved by the introduction of a so-called *link function*. The link function applies a transformation to the mean and then links the covariates, via the linear predictor, to the transformed mean of the distribution of the responses,

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik} = X_i' \beta,$$

where the link function $g(\cdot)$ is some known function, for example, $\log(\mu_i)$. This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates.

Thus, while in the standard linear regression model the mean response is related directly to a linear combination of the covariates, in generalized linear models, it is some appropriate transformation of the mean response, for example, $\log(\mu_i)$, that is related to a linear combination of the covariates. The linearity applies to a transformation of the mean response, or, put in a somewhat different way, the effects of covariates are assumed to be additive on a suitably transformed scale for the mean response.

The use of non-linear link functions, for instance, $\log(\mu_i)$, ensures that the model produces predictions of the mean response that are within the allowable range. For example, when analyzing a binary response, μ_i has interpretation in terms of the probability of "success" (with $0 < \mu_i < 1$). If the mean response, here the probability of success, is related directly to a linear combination of the covariates, the model can yield predicted probabilities outside of the range from 0 to 1. The use of certain non-linear link functions ensures that this cannot happen.

We can distinguish two main types of link functions, *canonical* link functions and *non-canonical* link functions. The former are unique and can be derived for any selected distribution; the latter are somewhat arbitrary and bear no direct relation to the selected distribution. For example, the logit link function is the canonical link function associated with the Bernoulli and binomial distributions; the probit link function is a non-canonical link function for these distributions that is often adopted for the analysis of binary data from toxicological experiments. Although, in principal, any suitable link function can be used to relate the mean response to the covariates, the choice of a canonical link function produces many of the most widely used regression models. The canonical link functions for the normal, Bernoulli, and Poisson distributions are summarized in Table 10.1.

In summary, in generalized linear models, the distribution of the response is assumed to belong to a single family of distributions known as the exponential family. The exponential family includes the normal, Bernoulli, binomial, and Poisson distributions. A transformation of the mean response is then linearly related to the covariates, via an appropriate link function. Because generalized linear models make distributional assumptions about the response variable, the regression parameters can be estimated using the method of maximum likelihood. The maximum likelihood estimates of the regression coefficients, β, are simply those values of β that are most probable (or most "likely") for the data that have actually been observed. The method of maximum likelihood provides a very general technique for estimation and for inference, that is, for estimating β, constructing confidence intervals, testing hypotheses, and assessing the adequacy of models. All of these ideas will be elaborated in Section 10.3, where we focus on two special cases of generalized linear models: logistic regression for a binary response and log-linear regression for counts. Although generalized linear models provide a very broad and flexible collection of regression models for analyzing diverse types of responses, they do have one very important restriction: they assume that observations on the response variable are independent

of one another. In later chapters, we will discuss how this restriction to independent observation can be relaxed to accommodate the correlated nature of the responses arising from longitudinal studies.

10.3 ILLUSTRATIVE EXAMPLES

To clarify the main ideas presented in the previous sections, we consider in greater detail two special cases of generalized linear models: logistic regression for a binary response and log-linear regression for counts. We consider each of these models in terms of their three-part specification as a generalized linear model. We also emphasize the interpretation of the regression coefficients, β, in these models. The description of methods for extending generalized linear models for longitudinal responses presented in later chapters will assume a good working knowledge of these two important regression models. As a result, the reader is encouraged to master the material in this section before proceeding to Chapters 11 and 12. This section can be skimmed through for those with a strong background in logistic and log-linear regression models.

10.3.1 Logistic Regression for Binary Responses

Logistic regression is used widely to describe the relationship between a binary response variable (e.g., denoting "success" or "failure") and a set of covariates. In common with standard linear regression, the primary objective of logistic regression is to relate the mean of the response to a set of covariates. However, the response variable is binary rather than continuous and this has a number of consequences for modelling the mean. In this section we describe the main features of logistic regression and highlight its three-part specification as a generalized linear model. We also consider various aspects of interpretation of logistic regression coefficients. An example, using data of low-birth-weight infants, is used to illustrate the main ideas.

Let Y_i denote a binary response variable, whose two categories, for convenience, are often referred to as "success" or "failure". For example, Y_i might indicate the presence or absence of a disease. Denoting the two possible outcomes for Y_i by 1 (for "success") and 0 (for "failure"), the probability distribution of Y_i is Bernoulli, with $\Pr(Y_i = 1) = \mu_i$ (and, correspondingly, $\Pr(Y_i = 0) = 1 - \mu_i$). The primary goal of logistic regression is to describe the effects of changes in a set of covariates, X_i, on the mean μ_i. For ease of exposition, we will first consider the simple case where there is only a single covariate, X_i. Generalizations to more than one covariate will be considered later.

Since the analytic goal is to investigate the relationship between μ_i and X_i, and since linear regression plays such a dominant role in applications, it may at first seem natural to assume a linear model relating the mean of Y_i to X_I,

$$E(Y_i|X_i) = \mu_i = \beta_1 + \beta_2 X_i.$$

However, this linear model for the probabilities has one obvious difficulty. Expressing μ_i as a linear function violates the restriction that probabilities must lie within the range from 0 to 1. For sufficiently large or small values of X_i, this regression model will yield predicted probabilities outside of the range from 0 to 1. A further difficulty with the linear model for μ_i is that we often expect a non-linear relationship between μ_i and X_i. For example, a 0.2 unit increase in μ_i might be considered more "extreme" when $\mu_i = 0.1$ than when $\mu_i = 0.5$. In terms of ratios, the change from $\mu_i = 0.1$ to $\mu_i = 0.3$ represents a three-fold or 300% increase, whereas the change from $\mu_i = 0.5$ to $\mu_i = 0.7$ represents only a 40% increase. In a sense, the units of measurement for a probability (or proportion) are often not considered to be constant over the range from 0 to 1. The linear probability model simply does not take this into consideration when relating μ_i to X_i. Note also that the usual assumption of homogeneity of variance (or constant variance) in linear regression would be violated since the variance of a binary response explicitly depends upon the mean, with

$$\text{Var}(Y_i) = \mu_i (1 - \mu_i) .$$

To circumvent these difficulties with the linear probability model, a non-linear transformation can be applied to μ_i and the transformed probabilities are related linearly to X_i. When the logit or logistic function, $\log[\mu_i/(1 - \mu_i)]$, is adopted, the resulting model

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \text{logit}(\mu_i) = \beta_1 + \beta_2 X_i ,$$

is known as the *logistic regression* model. Recall that if μ_i is the probability of success, then $\mu_i/(1 - \mu_i)$ is known as the *odds* of success. For example, if the probability of success is 0.8 then the odds of success is 4 (or 0.8/0.2) to 1. That is, the probability of success is 4 times as large as the probability of failure. Thus the logistic regression model assumes a linear relationship between the log odds of success and X_i. For the reader unfamiliar with logistic regression, it is useful to bear in mind that the transformation of μ_i in logistic regression has the following property: as the probability of success, μ_i, increases, so too does the odds of success and the log odds of success; similarly, as the probability of success decreases, so too does the odds of success and the log odds of success.

Next consider the interpretation of the logistic regression coefficients, β_1 and β_2. For the special case where the predictor variable X_i is dichotomous, taking values of 0 and 1, the logistic regression slope, β_2, has a simple and very attractive interpretation in terms of the log odds ratio (comparing the log odds of success when $X_i = 1$ to the log odds of success when $X_i = 0$). That is,

$$\text{logit}(\mu_i|X_i = 1) - \text{logit}(\mu_i|X_i = 0) = (\beta_1 + \beta_2) - \beta_1 = \beta_2.$$

Thus $\exp(\beta_2)$ has interpretation as the odds ratio of the response for the two possible values of the covariate.

In simple linear regression the interpretation of the slope of the regression is in terms of changes in the mean of Y_i for a single-unit change in X_i. Similarly, for

arbitrary X_i, the logistic regression slope β_2 has interpretation as the change in the log odds (of success) for a unit change in X_i. Equivalently, a unit change in X_i increases or decreases the odds of success *multiplicatively* by a factor of $\exp(\beta_2)$. Also, recall that the intercept in simple linear regression has interpretation as the mean value of the response variable when X_i is equal to zero. Similarly, the logistic regression intercept, β_1 has interpretation as the log odds (of success) when $X_i = 0$; alternatively,

$$\frac{\exp(\beta_1)}{1 + \exp(\beta_1)}$$

is the probability of success when $X_i = 0$.

The logistic regression model can also be expressed in terms of the probability of success, μ_i,

$$\mu_i = \frac{\exp(\beta_1 + \beta_2 X_i)}{1 + \exp(\beta_1 + \beta_2 X_i)}.$$

While the latter expression may appear to be somewhat more complicated, this is simply an equivalent way of expressing the logistic regression model. That is, logistic regression describes how the log odds, $\log(\frac{\mu_i}{1-\mu_i})$, has a linear relationship with X_i which is equivalent to describing how μ_i has a sigmoidal or S-shaped relationship with increasing values of $\beta_2 X_i$. (See Figure 10.1 for a plot of μ versus X when $\beta_1 = 0.5$ and $\beta_2 = 0.9$.) Of note, the logistic transformation ensures that the predicted probabilities are restricted to the range from 0 and 1.

When viewed as a generalized linear model, logistic regression is simply the special case where the distribution of Y_i is assumed to be Bernoulli (a member of the exponential family) and a logit link function, the canonical link function, has been adopted. Because the Bernoulli distribution is a one-parameter exponential family distribution, the variance of Y_i can be expressed explicitly in terms of the mean, via the following variance function:

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i(1 - \mu_i),$$

and $\phi = 1$. For the Bernoulli distribution, the dispersion parameter is a fixed constant ($\phi = 1$).

When X_i is a discrete covariate with J distinct categories or levels (e.g., treatment groups), the binary responses for the N individuals can be grouped. Let m_j denote the number of individuals with the j^{th} covariate pattern and let Y_j denote the number of successes among the m_j individuals, for $j = 1, ..., J$. We may provisionally assume that all individuals within a group respond independently with constant probability of success, μ_j, depending only on group. Then Y_j, the number of successes in the j^{th} group, has a binomial distribution with probability of success, μ_j. The binomial distribution belongs to the exponential family and the probability of success, μ_j, can be related to group using a logit link function. For the binomial distribution, the mean or expected number of successes for the j^{th} covariate pattern is

$$E(Y_j) = m_j \mu_j.$$

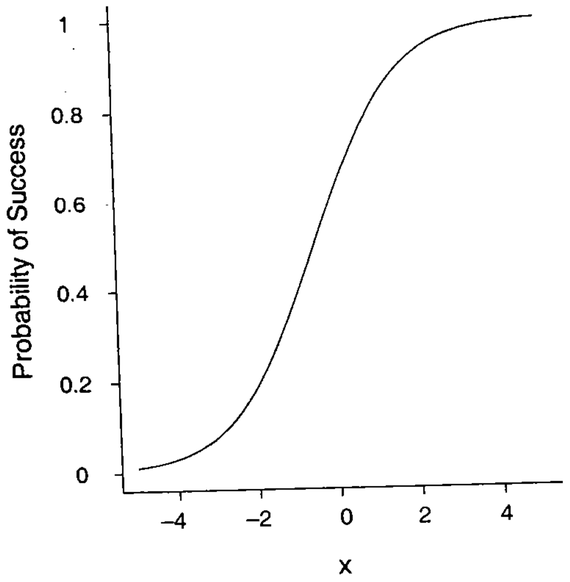

Fig. 10.1 Plot of logistic response function, with success probability, $\mu = \dfrac{e^{0.5+0.9X}}{1+e^{0.5+0.9X}}$.

There is a well-known relationship between the mean and variance of Y_j, with

$$\text{Var}(Y_j) = m_j\mu_j(1-\mu_j).$$

However, in many biomedical applications, counts of the number of successes have variability that far exceeds that predicted by the binomial distribution; this phenomenon is often referred to as *overdispersion* (although underdispersion can also arise, it is far less common). Overdispersion is a common indicator of failure of the binomial assumptions: independent observations with constant probability of success. That is, overdispersion can be represented either by a positive correlation between the responses or by variation in the response probabilities. To allow for overdispersion or extra-binomial variation, a scale factor ϕ (with $\phi \neq 1$) is often included in the specification of the binomial variance,

$$\text{Var}(Y_j) = \phi m_j\mu_j(1-\mu_j).$$

Failure to account for overdispersion has negligible impact of the estimated logistic regression coefficients. That is, the regression parameter estimates are consistent and there is usually little loss of efficiency. Neglecting overdispersion, however, results in the standard errors being underestimated and failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients (e.g., confidence intervals that are too narrow and p-values that are too

small). Also, model selection strategies based on likelihood ratio tests or on information criteria, such as the Akaike information criterion (AIC), will perform poorly. When overdispersion is ignored, a model with too many parameters is likely to be selected and thus can lead to overinterpretation of these parameters (e.g., unnecessary inclusion of interactions). Adjustments to the nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in the specification of the binomial variance,

$$\text{Var}(Y_j) = \phi m_j\mu_j(1-\mu_j),$$

or by basing standard errors on the so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; the latter will be discussed in greater detail in Chapter 11.

So far, we have only considered the simple case where there is a single predictor X_i. Next, we consider the case where X_i is a $p \times 1$ vector of covariates. The logistic regression model becomes

$$\log\{\mu_i/(1-\mu_i)\} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where $X_{i1} = 1$ for all $i = 1, \ldots, N$. The logistic regression coefficients in this model have the following interpretations. Each of the logistic regression slopes, β_k (for $k = 2, ..., p$), represents the change in the log odds (of success) for a unit change in X_{ik} given that all of the other predictor variables remain constant. This is completely analogous to the interpretation of the regression coefficients in multiple linear regression. Thus, holding the remaining predictors at some fixed set of values and not allowing them to vary with any changes in X_{ik}, a single-unit increase in X_{ik} is predicted to increase or decrease the log odds of success by an amount β_k. Equivalently, a single-unit increase in X_{ik} increases or decreases the odds of success *multiplicatively* by a factor of $\exp(\beta_k)$. The logistic regression intercept, β_1, now has interpretation as the log odds (of success) when all covariate values are set to zero. Alternatively,

$$\frac{\exp(\beta_1)}{1+\exp(\beta_1)}$$

is the probability of success when $X_{i2} = X_{i3} = \cdots = X_{ip} = 0$.

Finally, the logistic regression model for binary data can also be developed through the notion of an underlying latent variable distribution. Suppose that U_i is a latent (i.e., unobserved) continuous variable and that a positive response is observed only when U_i exceeds some threshold denoted by τ. The observed binary response can be thought of as a categorization of the unobservable latent variable, above and below the threshold τ. That is,

$$Y_i = 1 \quad \text{if} \quad U_i > \tau,$$

$$Y_i = 0 \quad \text{if} \quad U_i \leq \tau.$$

Suppose that the latent variable, U_i, has a standard logistic distribution. The standard logistic distribution (with mean zero and variance $\pi^2/3$) is a symmetric distribution and is very similar to the standard normal distribution, except that it has longer tails.

Then, using calculus, it can be shown that the relationship between the observable binary response variable, Y_i, and the unobservable, latent variable, U_i, is given by

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(U_i > \tau) \\ &= \int_{\tau}^{\infty} \frac{\exp(u)}{\{1+\exp(u)\}^2} du \\ &= \frac{\exp(-\tau)}{1+\exp(-\tau)}.\end{aligned}$$

Next, suppose that the following linear model for U_i holds:

$$U_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i = X_i'\beta + e_i,$$

where e_i (or $U_i - X_i'\beta$) is assumed to have a standard logistic distribution, with mean zero and variance $\pi^2/3$. Here, we regard the threshold of U_i as fixed and the location or mean of the distribution of U_i as changing with X_i. Without loss of generality, we can assume the threshold for categorizing U_i is $\tau = 0$, since any non-zero values for the threshold would simply be absorbed into the intercept term in the linear predictor, $X_i'\beta$. Then the relationship between Y_i and U_i results in a logistic regression model for $\Pr(Y_i = 1)$. That is,

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(U_i > 0) \\ &= \Pr(U_i - X_i'\beta > -X_i'\beta) \\ &= \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}.\end{aligned}$$

Thus, the linear model for U_i with standard logistic errors,

$$U_i = X_i'\beta + e_i,$$

implies the logistic regression model for Y_i,

$$\log\left\{\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right\} = X_i'\beta.$$

Similarly, if a probit link function is adopted instead of a logit link function, the linear model for U_i with standard normal errors, instead of standard logistic errors, implies a probit regression model for Y_i.

Although the logistic regression model can be derived from the notion of a latent variable distribution, assuming the existence of a latent variable is not a necessary requirement for the use of logistic regression models (indeed, in practice, the existence of a latent variable is usually not verifiable from the data). In later chapters, we use the notion of an underlying latent variable distribution to derive analogues of the between-subject and within-subject sources of variability in models for longitudinal binary responses.

Table 10.2 Estimated coefficients and standard errors for logistic regression of BPD on birth weight.

Variable	Estimate	SE	Z
Intercept	4.0343	0.6958	5.798
Birth Weight	−0.4229	0.0641	−6.599

Illustration

Next we consider an application of logistic regression to illustrate how the model can be used in practice. The data are from a study of low-birth-weight infants in a neonatal intensive care unit. In this example we are interested in the development of bronchopulmonary dysplasia (BPD), a chronic lung disease, in a sample of 223 infants weighing less than 1750 grams (Van Marter *et al.*, 1990).

Let Y_i be a binary response, with $Y_i = 1$ if the i^{th} infant develops BPD by day 28 of life and $Y_i = 0$ otherwise (where BPD is defined by both oxygen requirement and compatible chest radiograph). To examine whether there is an association between the risk of BPD and birth weight (in grams $\times 10^{-2}$), we consider the following logistic regression model:

$$\log\{\mu_i/(1-\mu_i)\} = \beta_1 + \beta_2\,\texttt{Weight}_i,$$

where $\mu_i = E(Y_i) = \Pr(Y_i = 1)$. For the 223 infants in the sample, the estimated logistic regression parameters (and standard errors), obtained using maximum likelihood, are displayed in Table 10.2.

The estimated logistic regression is

$$\log\{\widehat{\mu}_i/(1-\widehat{\mu}_i)\} = 4.0343 - 0.4229\,\texttt{Weight}_i.$$

When compared to its standard error, the ML estimate of β_2, the slope for birth weight, is significantly different from zero at the 0.05 level. The results from the logistic regression analysis indicate that the risk of BPD decreases with increasing birth weight. Specifically, the estimate of β_2 implies that, for every 100 gram increase in birth weight, the log odds of BPD decreases by 0.42. For example, the odds of BPD for an infant weighing 1200 grams (approximately 2.5 pounds) is

$$\exp(4.0343 - 12 \times 0.4229) = \exp(-1.0057) = 0.366.$$

Thus the predicted probability of BPD is $0.366/(1+0.366) \approx 0.27$. The estimated probability of BPD can be calculated at any birth weight and a plot of the estimated probability versus weight produces the characteristic sigmoidal or S-shaped curve displayed in Figure 10.2.

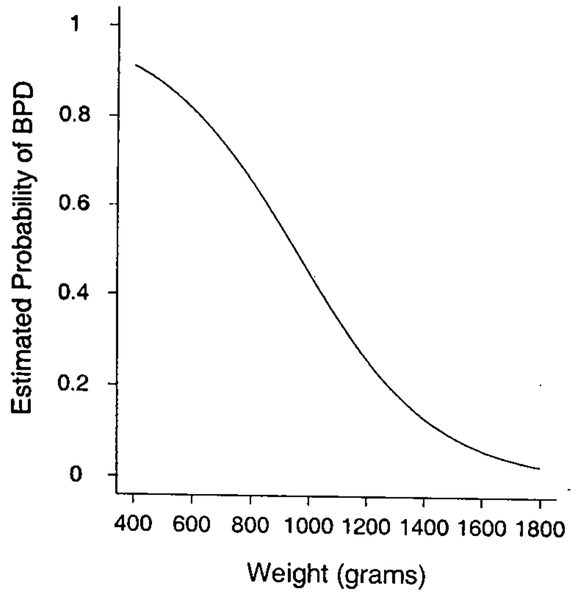

Fig. 10.2 Plot of estimated logistic response function of BPD on birth weight based on a sample of 223 infants with birth weight less than 1750 grams.

Next, suppose we include two additional covariates, gestational age (in weeks) and presence of toxemia (with 1 denoting the presence of toxemia and 0 its absence). That is, we consider the following logistic regression model:

$$\log\{\mu_i/(1-\mu_i)\} = \beta_1 + \beta_2\,\texttt{Weight}_i + \beta_3\,\texttt{Age}_i + \beta_4\,\texttt{Toxemia}_i.$$

For the 223 infants in the sample, the estimated logistic regression parameters (and standard errors) are displayed in Table 10.3.

The estimated coefficient for birth weight has now decreased, when gestational age and toxemia are included in the analysis. Nonetheless, the estimate of β_2 remains significantly different from zero at the 0.05 level. The estimated coefficient for gestational age has interpretation in terms of the change in the log odds of BPD for a 1 week increase in gestational age, while holding birth weight and toxemia constant. Specifically, a 1-week increase in gestational age is associated with a 0.39 decrease in the log odds of developing BPD. Finally, the estimated coefficient for toxemia has interpretation in terms of the log odds ratio, comparing mothers who were diagnosed with toxemia to mothers who were not, while adjusting for the effects of birth weight and gestational age. Specifically, the adjusted odds ratio is 0.26 (or $e^{-1.34}$) and indicates that infants of mother's diagnosed with toxemia have approximately a quarter the risk of developing BPD.

Table 10.3 Estimated coefficients and standard errors for logistic regression of BPD on birth weight, gestational age and toxemia.

Variable	Estimate	SE	Z
Intercept	13.9361	2.9826	4.672
Birth Weight	−0.2644	0.0812	−3.254
Gestational Age	−0.3885	0.1149	−3.382
Toxemia	−1.3437	0.6075	−2.212

10.3.2 Log-Linear Regression for Counts

Log-linear regression, often referred to as Poisson regression, is used widely for the analysis of counts of the number of times some event occurs in either time or space. For example, the response variable might be the count of the number of epileptic seizures a particular patient experiences in a four-week interval. Alternatively, the response might be a count of bacteria present in a fixed volume of bacterial suspension. In either case the response variable Y_i is a count and the objective of log-linear regression is to relate the mean or expected count to a set of covariates.

If the occurrences of some event are counted within an interval of time (or sometimes volume or area), then the *rate* at which the event occurs is usually of more direct interest than the corresponding count. That is, the count or absolute number of events is often not satisfactory because any comparisons depends almost entirely on the "time at risk" (or, in other contexts, the sizes of the groups or areas of regions) that generated the observations. For example, it would not be very meaningful to compare counts of the number of seizures in a four-week interval with counts of the number of seizures in a twelve-month interval since it seems reasonable to suppose that the number of seizures is directly proportional to the period at risk. When the "time at risk" is not the same for all observations, a rate provides a meaningful basis for direct comparison. In either case, the primary objective of log-linear regression is to relate the expected counts or rates to a set of covariates.

When the response is a count it is often reasonable to assume that Y_i has a Poisson distribution, although it is important to note that this is an assumption and it may not be valid. This is in contrast to the binary data case where the distribution of a binary response is always Bernoulli with mean μ_i. The Poisson distribution describes the probability that a specific number of events, say y_i, occurs,

$$\Pr(Y_i = y_i) = e^{-\mu_i}\mu_i^{y_i}/y_i!; \qquad y_i = 0, 1, 2, \ldots$$

where $y! = y \times (y-1) \times (y-2) \times \cdots \times 2 \times 1$. The Poisson distribution is completely determined by a single parameter, $\mu_i = E(Y_i) \geq 0$, the mean number of events. A distinctive property of the Poisson distribution is that the mean and variance of Y_i are identical,

$$E(Y_i) = \mu_i = \text{Var}(Y_i).$$

Note that μ_i is defined as the expected count or number of events. The expected rate is given by μ_i/T_i, where T_i is a relevant measure of the "time at risk" (e.g., T_i might be an interval of time, the person-years of observation, or the size of a group). In log-linear regression the goal is to describe the effects of a set of covariates, X_i, on the expected rate. Once again, for ease of exposition, we will first consider the simple case where there is only a single covariate, X_i. Generalizations to more than one covariate will be considered later.

Because a rate of occurrence of some event cannot be negative, a standard linear regression model relating μ_i/T_i directly to X_i is somewhat unappealing. That is, for sufficiently large or small values of X_i, a standard linear regression model could yield predicted rates that are negative. Instead, we can relate a transformation of the rate directly to X_i. When a logarithmic transformation is adopted, the resulting model

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 X_i$$

is known as the *log-linear regression* model. Recall that T_i is known and fully observed. As a result, the log-linear regression model can also be expressed as

$$\log(\mu_i) = \log(T_i) + \beta_1 + \beta_2 X_i,$$

since $\log(\mu_i/T_i) = \log(\mu_i) - \log(T_i)$. Note that, although $\log(T_i)$ appears on the right-hand side of the regression equation, it does not have a regression coefficient attached to it. That is, the regression parameter for $\log(T_i)$ is known to be equal to 1 and does not require estimation. We refer to $\log(T_i)$ as an *offset*. Thus the log-linear regression model assumes a linear relationship between the log rate of occurrence of some event and X_i.

When viewed as a generalized linear model, log-linear regression is simply the special case where the distribution of Y_i is assumed to be Poisson (a member of the exponential family) and a log link function, the canonical link function for the Poisson distribution, has been adopted. Because the Poisson distribution is a one-parameter exponential family distribution, the variance of Y_i can be expressed explicitly in terms of the mean, via the following variance function:

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i.$$

For the Poisson distribution, the dispersion parameter is a fixed constant ($\phi = 1$). However, in many biomedical applications, count data have variability that far exceeds that predicted by the Poisson distribution. Overdispersion is a common indicator of failure of the Poisson assumption when dealing with count data. The implications of overdispersion for count data are the same as for grouped binary data. As discussed earlier, neglecting overdispersion results in the standard errors being underestimated

and failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients (e.g., confidence intervals that are too narrow and p-values that are too small). Also, model selection strategies based on likelihood ratio tests or on information criteria (e.g., AIC) will perform poorly. Adjustments to the nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in the specification of the Poisson variance,

$$\text{Var}(Y_i) = \phi\, \mu_i,$$

or by basing standard errors on the so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; the "sandwich" estimator will be discussed in greater detail in Chapter 11.

Next, consider the interpretation of the log-linear regression coefficients, β_1 and β_2. For the special case where the predictor variable X_i is dichotomous, taking values of 0 and 1, the log-linear regression slope, β_2, has a simple and very attractive interpretation in terms of the log rate ratio (comparing the log expected rate when $X_i = 1$ to the log expected rate when $X_i = 0$). That is,

$$\log(\mu_i|X_i = 1) - \log(\mu_i|X_i = 0) = \{\log(T_i) + \beta_1 + \beta_2\} - \{\log(T_i) + \beta_1\} = \beta_2.$$

Thus $\exp(\beta_2)$ has interpretation as the rate ratio,

$$\frac{(\mu_i|X_i = 1)}{(\mu_i|X_i = 0)},$$

for the two possible values of the covariate.

For arbitrary X_i, the slope β_2 has interpretation as the change in the log expected rate for a single-unit change in X_i. Equivalently, a unit change in X_i increases or decreases (depending on the sign of β_2) the rate of occurrence of the event *multiplicatively* by a factor of $\exp(\beta_2)$. Thus, when exponentiated, the regression coefficients can be interpreted in terms of relative rates. This becomes more apparent if we express the log-linear regression model as

$$\mu_i = (\mu_i|X_i) = E(Y_i|X_i) = T_i \times e^{\beta_1} \times e^{\beta_2 X_i}.$$

From this expression it can be seen that a single-unit increase in X_i increases or decreases μ_i/T_i by a factor of e^{β_2}. That is,

$$(\mu_i|X_i + 1) = T_i \times e^{\beta_1} \times e^{\beta_2(X_i+1)} = T_i \times e^{\beta_1} \times e^{\beta_2 X_i} \times e^{\beta_2} = e^{\beta_2} \times (\mu_i|X_i).$$

On the other hand, the intercept, β_1, has interpretation as the log expected rate when $X_i = 0$; alternatively, $\exp(\beta_1)$ is the expected rate of occurrence of the event when $X_i = 0$.

So far, we have only considered the simple case where there is a single predictor X_i. Next, we consider the case where X_i is a $p \times 1$ vector of covariates. The log-linear regression model becomes

$$\log(\mu_i) = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where $X_{i1} = 1$ for all $i = 1, \ldots, N$. The log-linear regression coefficients in this model have the following interpretations. Each of the log-linear regression slopes, β_k (for $k = 2, ..., p$), has interpretation as the change in the log expected rate for a unit change in X_{ik} given that all of the other covariates remain constant. Thus, holding the remaining covariates at some fixed set of values and not allowing them to vary with any changes in X_{ik}, a single-unit increase in X_{ik} is predicted to increase or decrease the log expected rate by an amount β_k. Equivalently, a single-unit increase in X_{ik} increases or decreases the expected rate *multiplicatively* by a factor of $\exp(\beta_k)$. The log-linear regression intercept, β_1, now has interpretation as the log expected rate of occurrence of the event when all covariate values are set to zero. Alternatively, $\exp(\beta_1)$ is the expected rate when $X_{i2} = X_{i3} = \cdots = X_{ip} = 0$.

Illustration

Next we consider an application of log-linear regression to illustrate how the model can be used in practice. The data for this illustration arise from a prospective study of potential risk factors for coronary heart disease (CHD) (Rosenman *et al.*, 1975). The study observed 3154 men aged 40–50 for an average of 8 years and recorded the incidence of cases of CHD. The potential risk factors included smoking, blood pressure, and personality/behavior type. The data are summarized in Table 10.4.

Let Y_i denote the count of the number of cases of CHD and T_i denote the person-years of follow-up. Person-years of follow-up is calculated as the total duration of observed follow-up, from entry into the study until either disease detection or end of follow-up, for the individuals in each risk group. To examine whether the rates of CHD are related to the smoking exposure variable we consider the following log-linear regression model:

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 \, \mathtt{Smoke}_i,$$

where $\mu_i = E(Y_i)$ and $\mathtt{Smoke}_i$ is a quantitative measure of smoking exposure (0: Non-smoker, 10: 1–10 cigarette/day, 20: 11–20 cigarette/day, 30: 30+ cigarette/day). To adjust for differences in the total person-years of follow-up for each risk group, $\log(T_i)$ is included in the model for Y_i as an offset.

The ML estimate of the slope for smoking exposure, β_2, is 0.0318 (SE = 0.0056) and when compared to its standard error is significantly different from zero at the 0.05 level. This indicates that increases in the smoking exposure increases the log expected rate of CHD. That is, the expected rate of CHD for individuals who smoke one pack of cigarettes (or 20 cigarettes) per day is estimated to be approximately twice (or $e^{0.0316 \times 20} = 1.88$ times) as high as the rate of CHD for non-smokers.

Because risk factors for CHD are likely to be correlated, we consider the impact of smoking on the rates of CHD after adjusting for the potential confounding effects of blood pressure and personality type. High blood pressure and type A behavior pattern are known to be associated with high rates of CHD. Specifically, we consider the following log-linear regression model

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 \, \mathtt{Smoke}_i + \beta_3 \, \mathtt{BP}_i + \beta_4 \, \mathtt{Type}_i,$$

Table 10.4 Data on incidence of CHD and associated risk factors.

Person-Years	Smoking[a]	Blood Pressure[b]	Behavior[c]	CHD
5268.2	0	0	0	20
2542.0	10	0	0	16
1140.7	20	0	0	13
614.6	30	0	0	3
4451.1	0	0	1	41
2243.5	10	0	1	24
1153.6	20	0	1	27
925.0	30	0	1	17
1366.8	0	1	0	8
497.0	10	1	0	9
238.1	20	1	0	3
146.3	30	1	0	7
1251.9	0	1	1	29
640.0	10	1	1	21
374.5	20	1	1	7
338.2	30	1	1	12

[a] 0: Non-smoker, 10: 1–10 cigarette/day, 20: 11–20 cigarette/day, 30: 30+ cigarette/day;
[b] 0: < 140, 1: ≥ 140;
[c] 0: Type B Personality; 1: Type A Personality.

Source: From *Practical Biostatistical Methods*, 1st edition, by Steve Selvin. © 1995. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

where $\mathtt{BP}_i = 1$ if blood pressure ≥ 140 and 0 otherwise; $\mathtt{Type}_i = 1$ if Type A personality and $\mathtt{Type}_i = 0$ if Type B personality[1]. The estimated log-linear regression parameters (and standard errors) are displayed in Table 10.5.

The estimated coefficient for smoking, 0.027, has now decreased, when blood pressure and personality type have been controlled for in the analysis. Nonetheless,

[1]Type A personalities are characterized by impatience, competitiveness, aggressiveness, a sense of time urgency, and tenseness; Type B personalities are the opposite of Type A and exhibit traits such as being easy going, more relaxed about time, not competitive, and not easily angered or agitated.

Table 10.5 Estimated coefficients and standard errors for log-linear regression of expected rate of CHD on smoking, blood pressure and personality type.

Variable	Estimate	SE	Z
Intercept	−5.4202	0.1308	−42.69
Smoke	0.0273	0.0056	4.50
Personality Type	0.7526	0.1362	5.54
Blood Pressure	0.7534	0.1292	5.84

the estimate of β_2 remains significantly different from zero at the 0.05 level. The estimated coefficient for smoking has interpretation in terms of the change in the log expected rate of CHD, after adjusting for the effects of blood pressure and personality type. Specifically, the adjusted rate of CHD (controlling for blood pressure and behavior type) for individuals who smoke one pack of cigarettes (or 20 cigarettes) per day is estimated to be 1.7 (or $e^{0.027\times 20} = 1.704$) times higher than the rate of CHD for non-smokers.

There is also a very strong relationship between type A behavior pattern and CHD incidence. The rate ratio (comparing type A to type B behavior pattern) is 2.12 (or $e^{0.7526}$), indicating that the rate of CHD among type A individuals is approximately twice that among type B individuals. Moreover, this adjusted estimate of risk cannot be explained by the association of personality type with smoking and blood pressure since the latter two risk factors have been adjusted for in the analysis.

10.4 COMPUTING: FITTING GENERALIZED LINEAR MODELS USING PROC GENMOD IN SAS

To fit generalized linear models we can use the PROC GENMOD procedure in SAS. The GENMOD procedure fits generalized linear models using maximum likelihood estimation. It includes many of the commonly used exponential family distributions for the response variable and a wide variety of link functions for relating the mean response to the covariates. In addition, link functions that are not incorporated in the procedure can be specified through programming statements (the FWDLINK and INVLINK statements) used within the procedure. Finally, PROC GENMOD can be used to fit models to correlated responses using the generalized estimating equations approach. This latter aspect of the procedure will be describe in Chapter 11.

For example, to fit a logistic regression model to data from two groups (e.g., treatment or exposure groups), we can use the illustrative SAS commands given in

Table 10.6 Illustrative commands for logistic regression using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
    CLASS  group;
    MODEL  y=group / DIST=BINOMIAL  LINK=LOGIT;
```

Table 10.7 Illustrative commands for log-linear regression, with an offset, using PROC GENMOD in SAS.

```
PROC GENMOD;
    CLASS  group;
    MODEL  y=group / DIST=POISSON  LINK=LOG  OFFSET=logtime;
```

Table 10.6. Similarly, to fit a log-linear regression, with an offset, we can use the illustrative SAS commands given in Table 10.7. Next, we present a brief description of the most salient parts of the command syntax used in the two illustrations in Tables 10.6 and 10.7.

PROC GENMOD <options>;

This statement calls the procedure GENMOD in SAS. It can also include an option for specifying the level of the response variable that is modelled. By default, the lower response level is modelled. For a binary response, coded (0,1), it is the probability that $Y = 0$ that is modelled. Use of the DESCENDING option reverses the default ordering of the response levels, resulting in the highest response level being modelled (i.e., the probability that $Y = 1$ for binary data that are coded as 0 and 1).

CLASS variables;

The CLASS statement is used to identify all variables that are to be regarded as categorical or factors. By default, this statement will create indicator variables for each factor using a reference group coding, with the last level (where "last" here refers to the level with the largest alpha-numeric value) regarded as the reference group. Different sort orders for the CLASS variables can be requested by the ORDER=<option> on the PROC GENMOD statement.

MODEL response = <effects> / <options>;

MODEL events/trials = <effects> / <options>;

The MODEL statement specifies the response variable and the covariate effects. The second form of the MODEL statement, with the `events/trials` syntax, allows the response to be in the form of a ratio of two variables (e.g., counts of the number of successes and the number of trials) and is used for binomial response data. The covariate effects determine the linear predictor and can include both discrete (defined on the CLASS statement) and quantitative (excluded from the CLASS statement) covariates. By default, PROC GENMOD includes a column of 1's for the intercept in the model.

Two important options need to be included on the MODEL statement. The DIST=*keyword* specifies a built-in response variable distribution, from the exponential family, that is assumed for the model. The LINK=*keyword* specifies the choice of built-in link function relating the mean response to the linear predictor. If the LINK=*keyword* is omitted, the default link function is the canonical link function for the distribution specified on DIST=*keyword*. If both the LINK=<option> and the DIST=<option> are omitted, the default is a normal distribution with an identity link function.

A final option that is often required when modelling count data is an offset. The OFFSET=*variable* specifies a variable to be used as an offset. Note that this variable cannot be a CLASS variable and it should not be included as one of the covariates listed on the MODEL statement.

PROC GENMOD provides many options for handling the dispersion parameter, ϕ, in the exponential family distribution. Recall that for many discrete response distributions (e.g., Bernoulli, binomial, and Poisson), the dispersion parameter is a fixed constant ($\phi = 1$) and not a parameter to be estimated. As discussed earlier, in many biomedical applications, the data display more variability than is predicted by the variance–mean relationship for the assumed distribution of the response. Neglecting overdispersion (e.g., greater variability than that predicted by the binomial or Poisson distributions) results in the standard errors being underestimated and failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients (e.g., confidence intervals that are too narrow and p-values that are too small). To allow for overdispersion, PROC GENMOD provides options for estimating ϕ and making suitable adjustments to standard errors and test statistics. Strictly speaking, in these cases where ϕ is estimated, rather than assumed to be fixed, we no longer have a legitimate distribution for the response variable and the function that is maximized is referred to as a quasi-likelihood function rather than a likelihood function. Alternatively, an adjustment to the nominal standard errors to account for overdispersion can be made by basing standard errors on the so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; the "sandwich" estimator will be described in Chapter 11.

10.5 OVERVIEW OF GENERALIZED LINEAR MODELS*

In this section[†] we present a somewhat more technical and detailed overview of generalized linear models that supplements the material presented in Section 10.2. Generalized linear models are a broad class of regression models suitable for analyzing diverse types of univariate responses (e.g., continuous, binary, counts). As was mentioned in Section 10.2, a generalized linear model for Y_i has a three-part specification:

1. a distributional assumption,
2. a systematic component, and
3. a link function,

and we consider each of these three components in turn.

Distributional Assumption

Generalized linear models are an extended family of probability models for a univariate response variable, Y_i. The family of probability distributions, known as the exponential family, includes the normal distribution for a continuous response, the Bernoulli (or binomial) distribution for a binary response, and the Poisson distribution for counts. The exponential family also includes many other distributions, for example, the gamma, beta, and negative binomial distributions.

Any distribution that belongs to the exponential family can be expressed in the same general form. Before we describe that general form we want to emphasize that our motivation for doing so is three-fold. First, we want to demonstrate that probability distributions for seemingly quite different data types (e.g., continuous, binary, and count data) have much in common as members of the exponential family of distributions. Second, we want to emphasize the importance of the so-called canonical "location" parameter in exponential family distributions; the canonical location parameter is closely related to, but generally not equal to, the mean of the distribution. Third, we want to emphasize that the variance of many exponential family distributions depends on the mean, via a so-called "variance function". We caution the reader that the material in the remainder of this section is somewhat technical in nature, but we strongly encourage the reader to stay the course.

All distributions that belong to the exponential family can be expressed as follows,

$$f(y_i; \theta_i, \phi) = \exp\left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)\right], \qquad (10.1)$$

for some specific functions $a(\cdot)$ and $b(\cdot)$. The specific functions $a(\cdot)$ and $b(\cdot)$ associated with an exponential family distribution distinguish one member of the family

[†]This section provides a more technical presentation of generalized linear models and can be omitted without loss of continuity.

from another. For example, the normal, Bernoulli, and Poisson distributions can all be expressed in the same form, albeit with different functions $a(\cdot)$ and $b(\cdot)$. This expression for the exponential family has two parameters, θ_i and ϕ. The first parameter, θ_i, is a location parameter (and is sometimes referred to as the "canonical" location parameter); the second parameter, ϕ, is a scale or dispersion parameter. As these terms imply, θ_i is related to the mean of the distribution (but θ_i is not necessarily the mean), while ϕ is related to the variance. For many distributions for discrete data, is not a parameter that requires estimation but is a known constant; for other distributions ϕ is an unknown parameter. When ϕ is known, Y_i is said to have a one-parameter exponential family distribution, while when ϕ is unknown, it has a two-parameter exponential family distribution.

While many elegant statistical properties can be derived for distributions that belong to the exponential family, the main concept we want to emphasize in this section is that the exponential family provides some unification for distributions that are commonly assumed for seemingly diverse types of responses variables (e.g., probability distributions for continuous and binary responses).

To fix ideas, we will demonstrate how three of the most commonly encountered distributions in biomedical applications, the normal, Bernoulli, and Poisson distributions can be expressed in the exponential family form given in (10.1). Recall that the probability density function for the normal distribution (see Section 3.2) is usually written as,

$$f\left(y_i;\mu_i,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\left(y_i-\mu_i\right)^2/2\sigma^2\right\}.$$

However, it is possible to re-arrange the terms in this expression for the normal density to obtain

$$\begin{aligned} f\left(y_i;\mu_i,\sigma^2\right) &= \exp\left\{-1/2\log\left(2\pi\sigma^2\right)\right\}\exp\left\{-\left(y_i-\mu_i\right)^2/2\sigma^2\right\} \\ &= \exp\left\{-\left(y_i^2-2y_i\mu_i+\mu_i^2\right)/2\sigma^2 - 1/2\log\left(2\pi\sigma^2\right)\right\} \\ &= \exp\left[\left\{y_i\mu_i-\mu_i^2/2\right\}/\sigma^2 - 1/2\left\{y_i^2/\sigma^2+\log\left(2\pi\sigma^2\right)\right\}\right]. \end{aligned}$$

When expressed in this form, the normal distribution is seen to be an exponential family distribution with canonical location parameter, $\theta_i = \mu_i$, and scale parameter, $\phi = \sigma^2$ (with $v(\mu_i) = 1$). Also,

$$a\left(\theta_i\right) = \mu_i^2/2 = \theta_i^2/2,$$

and

$$\begin{aligned} b\left(y_i,\phi\right) &= -1/2\left\{y_i^2/\sigma^2+\log\left(2\pi\sigma^2\right)\right\} \\ &= -1/2\left\{y_i^2/\phi+\log\left(2\pi\phi\right)\right\}. \end{aligned}$$

Thus, for the normal distribution the location parameter, θ_i, happens to be the mean of the response and the scale parameter happens to be the variance.

Two important exponential family distributions for discrete response data are the Bernoulli and the Poisson distributions. The Bernoulli distribution is ordinarily expressed as

$$f\left(y_i;\mu_i\right) = \mu_i^{y_i}\left(1-\mu_i\right)^{1-y_i},$$

where $\mu_i = E(Y_i) = \Pr\left(Y_i = 1\right)$. At first glance, it is not obvious that the Bernoulli distribution also belongs to the exponential family. However, the Bernoulli distribution can also be re-expressed as

$$\begin{aligned} f\left(y_i;\mu_i\right) &= \mu_i^{y_i}\left(1-\mu_i\right)^{(1-y_i)} \\ &= \exp\left\{y_i\log\left(\mu_i\right)+\left(1-y_i\right)\log\left(1-\mu_i\right)\right\} \\ &= \exp\left[y_i\log\left\{\mu_i/\left(1-\mu_i\right)\right\}+\log\left(1-\mu_i\right)\right]. \end{aligned}$$

When expressed in this form, the Bernoulli distribution is seen to be a one-parameter exponential family distribution with location parameter,

$$\theta_i = \log\left\{\mu_i/\left(1-\mu_i\right)\right\} = \operatorname{logit}\left(\mu_i\right),$$

and $\phi = 1$ is simply a fixed and known constant. Finally, the Poisson distribution is ordinarily expressed as

$$f\left(y_i;\mu_i\right) = e^{-\mu_i}\mu_i^{y_i}/y_i!$$

but it too can be re-expressed as

$$\begin{aligned} f\left(y_i;\mu_i\right) &= e^{-\mu_i}\mu_i^{y_i}/y_i! \\ &= \exp\left\{y_i\log\mu_i-\mu_i-\log\left(y_i!\right)\right\}. \end{aligned}$$

When written in this form, the Poisson distribution is also a one-parameter exponential family distribution with location parameter,

$$\theta_i = \log\left(\mu_i\right),$$

and $\phi = 1$, a fixed and known constant.

The exponential family unifies many probability distributions for diverse types of response variables. Moreover, it is also possible to derive some elegant statistical properties for distributions belonging to this family. The two properties that we focus on here are the mean and variance of exponential family distributions. It can be shown (although it requires the use of calculus) that the mean of Y_i can be expressed as

$$E\left(Y_i\right) = \mu_i = \frac{\partial a\left(\theta_i\right)}{\partial\theta},$$

where $\frac{\partial a(\theta_i)}{\partial\theta}$ denotes differentiation of the function $a(\theta_i)$ with respect to θ. For readers unfamiliar with calculus, $\frac{\partial a(\theta_i)}{\partial\theta}$ can simply be thought of as another known function of θ_i. Thus, μ_i, the mean of Y_i, is simply a known function of θ_i, and vice

versa. The second property that we are interested in is the variance of exponential family distributions. The variance of Y_i can be expressed as

$$\text{Var}(Y_i) = \phi \frac{\partial^2 a(\theta_i)}{\partial \theta^2},$$

where $\frac{\partial^2 a(\theta_i)}{\partial \theta^2}$ (known in calculus as the second derivative of $a(\theta_i)$ with respect to θ) is simply another known function of θ_i. Thus the variance of Y_i for distributions belonging to the exponential family can be expressed as the product of ϕ, the dispersion parameter, and some known function of θ_i. The latter function is referred to as the "variance function". However, recall that θ_i can be expressed as some known function of the mean, μ_i (since earlier we showed that μ_i is a known function of θ_i). Because θ_i and μ_i are functionally related to one another, the variance of Y_i can be expressed as the product of ϕ and some known function of μ_i. When expressed in terms of the mean, the variance function is denoted by $v(\mu_i)$ and

$$\text{Var}(Y_i) = \phi v(\mu_i). \tag{10.2}$$

Thus, for distributions belonging to the exponential family, the variance of Y_i can be expressed in terms of a scale or dispersion parameter ϕ and some known function of the mean, $v(\mu_i)$. For the normal distribution, the variance of Y_i is

$$\text{Var}(Y_i) = \sigma^2 = \phi,$$

and $v(\mu_i) = 1$. For the Bernoulli distribution

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i(1 - \mu_i),$$

and $\phi = 1$; while for the Poisson distribution

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i,$$

and $\phi = 1$. For one-parameter exponential family distributions (e.g., Bernoulli and Poisson), the variance of Y_i is simply a known function of the mean, μ_i, that is, the variance is completely determined by the mean response.

In summary, generalized linear models assume that the response, Y_i, has a probability distribution that belongs to the "exponential family". This extended family of distributions includes, among others, the normal, Bernoulli and Poisson distributions. Some exponential family distributions (e.g., Bernoulli and Poisson) have only a single "location" (or canonical) parameter and this parameter is related to (but it is not necessarily) the mean of the distribution. For one-parameter exponential family distributions, the variance of Y_i is a known function of the mean, referred to as the variance function. For two-parameter exponential family distributions (e.g., the normal distribution), there is an additional "scale" parameter, often referred to as a dispersion parameter. In two-parameter exponential family distributions the variance can be expressed as the product of the scale parameter and a variance function, where the latter is a known function of the mean.

Systematic Component

The systematic component of the generalized linear model specifies that the effects of the covariates, X_i, on the mean of the distribution of Y_i can be expressed via the following "linear predictor":

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

The linear predictor is simply a linear combination of the unknown vector of regression coefficients, $\beta = (\beta_1, \ldots, \beta_p)'$ and the vector of covariates, X_i,

$$\eta_i = \sum_{k=1}^{p} \beta_k X_{ik}. \tag{10.3}$$

The term "linear", as used in this context, means that η_i must be linear in the regression parameters.

We remind the reader that the restriction that η_i be linear in the regression parameters does not preclude relationships between the mean response and covariates that are curvilinear or non-linear. This latter type of non-linearity can be accommodated by taking appropriate transformations of the covariates (e.g., log(X)) and/or by including a polynomial in X). The inclusion of transformed covariates does not violate in any way the requirement that η_i be linear in the regression parameters.

Link Function

Finally, the formulation of a generalized linear model is completed by specifying the connection between the random and systematic components of the model through a "link function". The link function describes the relation between μ_i, the mean of Y_i, and the linear predictor, η_i, given by (10.3). Specifically, the link function is some known function $g(\cdot)$ such that

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{10.4}$$

In the case of the standard linear regression model, the random and systematic components are directly related, with

$$E(Y_i) = \mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

When viewed as a generalized linear model, the standard linear regression model adopts an identity link function, $g(\mu_i) = \mu_i$.

The primary motivation for considering link functions other than the identity is to ensure that the linear predictor produces predictions of the mean response that are within the allowable range. For example, when analyzing a binary response, μ_i has interpretation in terms of the probability of "success". As a result, we must have $0 < \mu_i < 1$ and the identity link is not appealing since, for sufficiently large or small values of the covariates, it can yield predicted probabilities outside of the range from

0 to 1. It is preferable to use a link function that takes a non-linear transformation of μ_i, mapping the range of μ_i from [0,1] onto the unrestricted range $(-\infty, \infty)$.

In principle, any function $g(\cdot)$ can be chosen to link the mean of Y_i to the linear predictor. However, every distribution that belongs to the exponential family has a special link function called the *canonical* link function. The canonical link function is defined as that function $g(\cdot)$ such that

$$g(\mu_i) = \theta_i,$$

where θ_i is the canonical location parameter (recall that μ_i is a known function of θ_i, and vice versa). Although there is no *a priori* reason why the covariate effects should necessarily be additive (or linear) on the particular scale defined by the canonical link function, generalized linear models with canonical link functions produce the most widely used regression models in biomedical applications.

For example, the canonical link function for the normal distribution is the identity link function,

$$g(\mu_i) = \mu_i,$$

and this gives the standard linear regression model,

$$\mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

For counts from a Poisson distribution, where we must have $\mu_i > 0$, the canonical link function is the log link function,

$$g(\mu_i) = \log(\mu_i),$$

and this gives the log-linear regression model,

$$\log(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

For the Bernoulli distribution, where $0 < \mu_i < 1$, the canonical link function is the logistic or logit link function,

$$g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$$

and this gives the logistic regression model,

$$\log\{\mu_i/(1-\mu_i)\} = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

We can, however, choose other link functions when they seem appropriate to the application at hand. For example, when Y_i is Bernoulli, we would generally prefer a link function that transforms the interval [0, 1] on to the entire real line, $(-\infty, \infty)$. The complementary log-log link function,

$$g(\mu_i) = \log\{-\log(1-\mu_i)\},$$

and the probit link function,

$$g(\mu_i) = \Phi^{-1}(\mu_i),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, both have this property. In some applications, it may be of interest to consider a link function that does not transform the range of μ_i on to the entire real line $(-\infty, \infty)$. For example, in modelling the prevalence of a disease, and the impact of risk factors, it may be preferable on scientific grounds to use a log link function since the resulting regression coefficients, β, have interpretation in terms of the log relative risk of disease. The relative risk of disease is simply the ratio of the probability of disease when a risk factor is present to the probability of disease when a risk factor is absent. The relative risk is an index of association that is favored by many empirical researchers. Although, in principle, link functions that do not transform the range of μ_i on to $(-\infty, \infty)$ can be adopted, in practice, this can result in problems with predictions that are out of range (e.g., predicted probabilities less than zero or greater than 1) and problems with convergence of the model fitting algorithm.

In summary, the link function connects the random and systematic components of the generalized linear model. It relates the mean of Y_i to the linear predictor and determines the scale on which the additive effects of covariates have an impact on the mean response. Each distribution has a special link function called the canonical link function and adoption of the canonical link function gives rise to many of the widely used regression models in biomedical applications (e.g., linear regression for a normally distributed continuous response, logistic regression for a Bernoulli response, and log-linear regression for Poisson counts). In principle, other link functions can be selected and these link functions bear no relationship to the assumed distribution for the response. Instead, a non-canonical link function may be chosen because additivity of the covariate effects is more appropriate on that scale or because it yields regression coefficients, β, that have somewhat more useful interpretations.

Estimation

Next, we very briefly discuss estimation of the regression coefficients in a generalized linear model. This section is somewhat more technical and can be omitted on a first reading of this chapter. Recall that in the standard linear regression setting we estimate the linear regression coefficients using the method of least squares. The least squares criterion chooses values for the regression coefficients that minimize the sum of squared deviations of the observed Y_i from their predicted values, denoted by $\widehat{Y}_i = \widehat{\mu}_i$, under the assumed regression model,

$$\mu_i = \eta_i = \sum_{k=1}^{p} \beta_k X_{ik}.$$

The least squares method yields estimates of the regression coefficients that are also the *maximum likelihood* (ML) estimates when Y_i is assumed to have a normal distribution with constant variance. We use the more general method of maximum likelihood estimation for estimating the parameters of a generalized linear model.

Recall from Chapter 4 (Section 4.2) that the method of maximum likelihood estimation chooses values of the regression coefficients that are most likely (or most

probable) to have generated the observed data. This is achieved by maximizing the *likelihood function* for the data. Construction of the likelihood function requires an assumption about the probability distribution of Y_i. In generalized linear models the response is assumed to have a distribution belonging to the exponential family of distributions. Assuming independent observations of the response, Y_i, and the covariates $X_{i1}, X_{i2}, ..., X_{ip}$ available on N individuals, the joint probability of $(Y_1, ..., Y_N)$ is the product of the N probability (density) functions. Thus the likelihood function can be expressed as the product

$$L = \prod_{i=1}^{N} \exp\left[\left\{y_i\theta_i - a\left(\theta_i\right)\right\}/\phi + b\left(y_i, \phi\right)\right].$$

It is this function, or equivalently, the logarithm of this likelihood function, that must be maximized. Note that the likelihood is a function of the unknown regression coefficients, β, since θ_i is a known function of the mean, μ_i and

$$\mu_i = g^{-1}\left(\sum_{k=1}^{p} \beta_k X_{ik}\right),$$

where $g^{-1}(\cdot)$ denotes the inverse link function (e.g., if $g(\cdot) = \log(\cdot)$, then $g^{-1}(\cdot) = \exp(\cdot)$, etc.). The maximum likelihood estimates of β are obtained by substituting the above expression for μ_i into the likelihood function and finding those values of the regression coefficients that produce the largest value for the likelihood function. Ordinarily, the likelihood function has only a single maximum.

Instead of maximizing the likelihood, it is usually more convenient to maximize the log-likelihood. We maximize the log-likelihood with respect to β by taking the derivative of the log-likelihood with respect to β, and then finding the values of β that make those derivatives equal to 0. Given

$$l = \log L = \sum_{i=1}^{N} \left[\left\{Y_i\theta_i - a\left(\theta_i\right)\right\}/\phi + b\left(Y_i, \phi\right)\right],$$

the derivative of the log-likelihood with respect to β can be shown (with the aid of calculus) to be the vector,

$$\partial l/\partial\beta = \sum_{i=1}^{N} \left(\partial\theta_i/\partial\beta\right)\left(y_i - \mu_i\right)/\phi.$$

When a canonical link function, $g(\mu_i) = \theta_i = \eta_i$, has been assumed,

$$\partial l/\partial\beta = \sum_{i=1}^{N} X_i\left(y_i - \mu_i\right)/\phi.$$

Solving this set of equations,

$$\sum_{i=1}^{N} X_i\left(y_i - \mu_i\right) = 0,$$

yields the maximum likelihood estimates of β. In general, this requires an iterative procedure that has been implemented in many of the commercially available statistical software packages (e.g., PROC GENMOD in SAS). What is quite remarkable about ML estimation for generalized linear models (with canonical link functions) is that it requires the solution to the exact same set of equations, regardless of the type of response variable.

Finally, estimates of the standard errors of the estimated regression coefficients can readily be obtained using the method of maximum likelihood estimation; in addition, likelihood ratio tests can be constructed by comparing nested models. Interestingly, the solution to this set of equations, $\widehat{\beta}$, is consistent for β (i.e., with very high probability, $\widehat{\beta}$ is close to the population regression parameters β for sufficiently large N) even if the variance of Y_i is misspecified; the only requirement is that the model for the mean response (the link function and linear predictor) has been correctly specified. However, when the variance of Y_i is misspecified, standard errors for components of $\widehat{\beta}$ should be based on the so-called "sandwich" estimator of $\mathrm{Cov}(\widehat{\beta})$; the "sandwich" estimator is discussed in Chapter 11.

10.6 FURTHER READING

A general overview of logistic regression, Poisson regression, and generalized linear models can be found in Chapter 14 of Neter *et al.* (1996). The textbooks by Dobson (1990) and Gill (2000) provide excellent introductions to generalized linear models. Hosmer and Lemeshow (2000) provide an accessible and comprehensive description of logistic regression models for binary data.

Bibliographic Notes

Generalized linear models were introduced in a seminal paper by Nelder and Wedderburn (1972). McCullagh and Nelder (1989) is the definitive textbook on this topic, providing a comprehensive description of the theory and application of generalized linear models. Firth (1991) presents a concise but remarkably lucid review of generalized linear models; also see Chapter 2 of Fahrmeir and Tutz (2001) and Chapter 5 of McCulloch and Searle (2001).

Problems

10.1 In an experimental study of patients with bladder cancer conducted by the Veterans Administration Cooperative Urological Research Group (Byar and Blackard, 1978; Wei *et al.*, 1989), patients underwent surgery to remove tumors. Following surgery, patients were randomized to either placebo or treatment with thiotepa. Sub-

sequently patients were examined at 18, 24, 30 and 36 months. For this problem set, we focus only on the data for month 18. The response variable is binary, indicating whether or not there is a new tumor ($Y = 1$, if new tumor; $Y = 0$, if no new tumor) at the 18 month visit. The objective of the analysis is to determine the effect of treatment on tumor recurrence by month 18.

The raw data are stored in an external file: `tumor.dat`

Each row of the data set contains the following three variables:

ID Treatment Y

Note: The response variable Y is coded 1 = new tumor, 0 = no new tumor. The categorical variable Treatment is coded 1 = thiotepa, 0 = placebo.

10.1.1 Assuming a Bernoulli distribution for the recurrence of tumor at month 18, fit the following logistic regression model relating the mean or probability of recurrence (μ_i) to Treatment:

$$\text{logit}(\mu_i) = \beta_1 + \beta_2 \, \texttt{Treatment}_i.$$

10.1.2 What are the interpretations of β_1 and β_2?

10.1.3 From the results obtained in Problem 10.1.1, what can you conclude about the effect of treatment on tumor recurrence at month 18?

10.1.4 What is the *estimated* probability of recurrence of a new tumor among those who received placebo?

10.1.5 What is the *estimated* probability of recurrence of a new tumor among those who received thiotepa?

10.1.6 Construct a 95% confidence interval for the log odds ratio, comparing thiotepa to placebo.

10.1.7 Construct a 95% confidence interval for the odds ratio, comparing thiotepa to placebo.

10.2 In a clinical trial of patients suffering from epileptic seizures (Thall and Vail, 1990), patients were randomized to receive either a placebo or the drug progabide, in addition to standard therapy. A baseline count of the number of epileptic seizures in an 8-week period prior to randomization was obtained. In addition, counts of the number of epileptic seizures in each of four successive 2-week (post-baseline) treatment periods were obtained. For this problem set, we focus only on the data from the last 2-week treatment period. The goal of the analysis is to make a comparison between the two treatment groups in terms of the counts of the number of seizures

in the final 2-week period of the study. The question we want to address is whether treatment with progabide is effective in reducing epileptic seizures.

The raw data are stored in an external file: `seizure4.dat`

Each row of the data set contains the following four variables:

ID Treatment Age Y

Note: The response variable Y is a count of the number of epileptic seizures in a 2-week interval. The categorical variable Treatment is coded 1 = progabide, 0 = placebo. The variable Age is the age of each patient (in years) at baseline.

10.2.1 Assuming a Poisson distribution for the counts, fit the following model relating the mean number of seizures (μ_i) to Treatment:

$$\ln(\mu_i) = \beta_1 + \beta_2 \, \texttt{Treatment}_i.$$

10.2.2 What are the interpretations of β_1 and β_2?

10.2.3 From the results obtained in Problem 10.2.1, what can you conclude about the effect of progabide in reducing the number of epileptic seizures.

10.2.4 Construct a 95% confidence interval for the rate ratio, comparing progabide to placebo.

10.2.5 Redo the analysis in Problem 10.2.1, adjusting for the effect of baseline age of the patient:

$$\ln(\mu_i) = \beta_1 + \beta_2 \, \texttt{Treatment}_i + \beta_3 \, \texttt{Age}_i.$$

10.2.6 Based on the results of the analysis for Problem 10.2.5, construct a 95% confidence interval for the age-adjusted rate ratio, comparing progabide to placebo.

10.2.7 Redo the analysis in Problem 10.2.5, allowing for potential overdispersion (i.e., variability greater than that predicted by the Poisson distribution).

10.2.8 Construct a 95% confidence interval for the age-adjusted rate ratio, comparing progabide to placebo, after taking account of any potential overdispersion.

11

Marginal Models: Generalized Estimating Equations (GEE)

11.1 INTRODUCTION

In the previous chapter we reviewed generalized linear models for a single response variable. A straightforward application of these models to longitudinal data is not appropriate, owing to the lack of independence among repeated measures obtained on the same individual. There are, however, a number of ways to extend generalized linear models to handle longitudinal data. All of these procedures account for the within-subject correlation among the repeated measures, though they differ in approach. We shall see in Chapters 11–13 that the method of accounting for the within-subject association has important ramifications for the interpretation of the regression coefficients in non-linear models for discrete longitudinal data. For the linear regression models for continuous responses considered in Part II, the interpretation of the regression coefficients is independent of assumptions made about the correlation among the repeated measures. With discrete longitudinal data this is no longer necessarily the case. Instead, different assumptions about the source of the within-subject association can lead to regression coefficients with quite distinct interpretations. The need to distinguish models according to the interpretation of their regression coefficients has led to the use of the terms "marginal models" and "mixed effects models"; the former are often referred to as "population-average models", the latter as "subject-specific models"). For the former the target of inference is the population, for the latter the target of inference is the individual. In this chapter we focus on marginal models; the meaning of the term "marginal", as used in this context, will soon be apparent. Mixed effects models, specifically, generalized linear models with random effects, are the focus of Chapter 12.

Because the method of accounting for the within-subject association has consequences for the interpretation of the regression model parameters, the choice of method for analyzing discrete longitudinal data cannot be made through any automatic procedure. Rather, the choice must be made on subject-matter grounds. Different models for discrete longitudinal data have somewhat different targets of inference and thereby address subtly different scientific questions. We return to this important issue in Chapter 13.

In this chapter we consider an approach for extending generalized linear models to longitudinal data that leads to a class of regression models that are known as *marginal models*. The term *marginal* in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses. That is, the term marginal is used to emphasize that the model for the mean response at each occasion does not incorporate dependence on any random effects or previous responses. This is in contrast to *mixed effects models*, where the mean response depends not only on covariates but also on a vector of random effects. Marginal models provide a very natural way of extending generalized linear models to longitudinal data and they have frequently been applied in the biomedical and health sciences. Marginal models do not require distributional assumptions for the observations, only a regression model for the mean response. That is, marginal models provide a unified method for analyzing diverse types of longitudinal responses, which avoids making assumptions about the distribution of the vector of responses; the method relies solely on assumptions about the mean response. The avoidance of distributional assumptions leads to a method of estimation known as *generalized estimating equations* (GEE).

In our discussion of marginal models the main focus is on discrete response data, for example, binary responses and counts. However, we also point out connections between marginal models for a continuous response and the methods for longitudinal data analysis presented in Part II. In doing so, we can provide some rationale for why the multivariate normal distributional assumption made in Part II often can be relaxed.

11.2 MARGINAL MODELS FOR LONGITUDINAL DATA

We begin our discussion of marginal models by introducing some notation similar to that used in Part II. We assume that N subjects are measured repeatedly over time. We let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. The response variable could be continuous, binary, or a count. The nature of the response variable does have important implications for model specification; however, the notation does not distinguish between the different types of responses.

We do not require that subjects have the same number of repeated measures or that they are measured at a common set of occasions. To accommodate unbalanced data (i.e., repeated measurements that are not obtained at a common set of occasions), we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. Both the longitudinal data structure and

the notation are the same as that used in Chapter 8; the only difference is that the response variable is no longer assumed to be continuous. The response variables for the i^{th} subject can be grouped into an $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, ..., N;$$

where the vectors of responses, Y_i, are assumed to be independent of one another (but the repeated measures on the same subject are emphatically not assumed to be independent). Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \;\; j = 1, ..., n_i.$$

Each individual has a vector of covariates, X_{ij}, associated with the response at each occasion, Y_{ij}. Note that X_{ij} may include covariates whose values do not change throughout the duration of the study and covariates whose values change over time. The former are referred to as time-stationary or between-subject covariates (e.g., gender and fixed experimental treatments), whereas the latter are referred to as time-varying or within-subject covariates (e.g., time since baseline, current smoking status, and environmental exposures). In the former case, the same values of the covariates are replicated in the corresponding rows of X_{ij}, for $j = 1, ..., n_i$. In the latter case, the values taken by the covariates can vary over time (for at least some individuals) and the values in the corresponding rows of X_{ij} can be different at each occasion.

We can group the vectors of covariates into an $n_i \times p$ matrix of covariates

$$X_i = \begin{pmatrix} X_{i1}' \\ X_{i2}' \\ \vdots \\ X_{in_i}' \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}, \quad i = 1, ..., N,$$

where the rows of X_i correspond to the covariates associated with the responses at the n_i different measurement occasions, and the columns of X_i correspond to the p distinct covariates. So far, we have assumed that each subject has a vector of repeated responses, denoted by Y_i, and associated with each repeated measure there is a vector of p covariates which can be grouped into a matrix, X_i.

Marginal models are primarily used to make inferences about population means. As a result, marginal models for longitudinal data separately model the mean response and the within-subject association among the repeated responses. In a marginal

model, the goal is to make inferences about the former, whereas the latter is regarded as a nuisance characteristic of the data that must be accounted for to make correct inferences about changes in the population mean response.

A marginal model for longitudinal data has the following three-part specification:

1. The conditional expectation or mean of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, is assumed to depend on the covariates through a known link function

$$g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta.$$

2. The conditional variance of each Y_{ij}, given the covariates, is assumed to depend on the mean according to

$$\text{Var}(Y_{ij}) = \phi v(\mu_{ij}),$$

where $v(\mu_{ij})$ is a known "variance function" (i.e., a known function of the mean, μ_{ij}) and ϕ is a scale parameter that may be known or may need to be estimated. For balanced longitudinal designs, a separate scale parameter, ϕ_j, could be estimated at each occasion; alternatively, the scale parameter could depend on the tim s of measurement, with $\phi(t_{ij})$ being some parametric function of t_{ij}.

3. The conditional within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, α (and also depends upon the means, μ_{ij}). For example, the components of α might represent the pairwise correlations or log odds ratios among the repeated responses. The within-subject association among the responses is described in more detail below.

This three-part specification of a marginal model makes the extension of generalized linear models to longitudinal data more transparent. The first two parts of the marginal model correspond to the standard generalized linear model, albeit with no distributional assumptions about the responses. It is the third component, the incorporation of the within-subject association among the repeated responses from the same individual, that represents the main extension of generalized linear models to longitudinal data. In principle, this three-part specification of a marginal model can be extended by making full distributional assumptions about the vector of responses, Y_i. However, in Section 11.6, we will show that assumptions about the joint distribution of Y_i are not necessary for estimation of the parameters of the marginal model.

As noted above, the first two components of a marginal model specify the mean and variance of Y_{ij} following the standard generalized linear model formulation described in Chapter 10, the only difference being that we have a common link function relating the vector of mean responses to the covariates. The third component recognizes the characteristic lack of independence among longitudinal data by modelling the within-subject association among the repeated responses from the same individual. In describing the third component we have been careful to avoid the use of the term *correlation* for two reasons. First, with a continuous response variable, the correlation

is a very natural measure of the linear dependence among the repeated responses. Also, the correlations are independent of the mean response, in the sense that the correlations are free to vary from -1 to 1, regardless of the values of the vector of mean responses. However, this is not the case with discrete responses. With discrete responses, the correlations are constrained by the mean responses, and vice versa. The most extreme example arises when the response variable is binary. For binary responses, the correlations are restricted to ranges that are determined by the means of the responses (or the probabilities of success). For example, in the bivariate case, if $\mu_1 = E(Y_1) = \Pr(Y_1 = 1) = 0.2$ and $\mu_2 = E(Y_2) = \Pr(Y_2 = 1) = 0.8$, then $\rho_{12} = \text{Corr}(Y_1, Y_2) \leq 0.25$. That is, the correlation can be no larger than 0.25 when the probabilities of success are 0.2 and 0.8. As a result, with discrete responses the correlation is not the most natural measure of within-subject association. Instead, the odds ratio (or the log odds ratio) is a preferable metric for association among pairs of binary responses. Second, for a continuous response that has a multivariate normal distribution, the correlations, along with the variances and the means, completely specify the joint distribution of the vector of longitudinal responses. This is not the case with discrete data. That is, the vector of means and the covariance matrix (the variances and correlations) do not, in general, completely specify the joint distribution of discrete longitudinal responses. Instead, the joint distribution requires specification of pairwise (e.g., pairwise odds ratios) and higher-order associations among the responses.

In a certain sense marginal models are a very natural way to extend generalized linear models, developed for the analysis of independent observations, to the setting of correlated longitudinal responses. Marginal models specify a generalized linear model for the longitudinal responses but also include a model for the within-subject association among the responses. A crucial aspect of marginal models is that the mean response and within-subject association are modelled separately. This separation of the modelling of the mean response and the association among responses has important implications for interpretation of the regression parameters in the model for the mean response. In particular, the regression parameters, β, in the marginal model have *population-averaged* interpretations. That is, they describe features of the mean response in the population and how those features relate to covariates. For example, regression parameters in a marginal model might have interpretation in terms of contrasts of the changes in the mean responses in sub-populations (e.g., different treatment or exposure groups). Of note, the interpretation of β is not altered in any way by the assumptions made about the nature or magnitude of the within-subject association. We will return to this point later in the chapter.

Of note, marginal models do not require distributional assumptions for the observations, only a regression model for the mean response. The avoidance of distributional assumptions can be advantageous, since there is no convenient specification of the joint multivariate distribution of Y_i for marginal models when the responses are discrete. The avoidance of distributional assumptions for Y_i leads to a method of estimation known as *generalized estimating equations* (GEE). The GEE approach provides a convenient alternative to maximum likelihood estimation; the GEE approach for estimating the parameters of marginal models is described in Section 11.3.

In Section 11.6 we present a more detailed discussion of how assumptions about the joint distribution of Y_i are not required for estimation of the marginal model parameters and why it can be advantageous to avoid making distributional assumptions. The material is Section 11.6 is somewhat technical and can be omitted without loss of continuity. Next, we consider some examples of marginal models using the three-part specification given earlier.

Example 1: Marginal Model for a Continuous Response

The linear regression model for longitudinal data described in Part II is a special case of the marginal model. It is useful to consider its formulation within the framework and terminology of marginal models. By doing so, the extensions to other types of response variables will become more apparent.

Suppose that Y_{ij} is a continuous response and it is of interest to relate changes in the mean response over time to the covariates. An example of a marginal model for Y_{ij} is given by the following three-part specification:

1. The mean of Y_{ij} is related to the covariates by an identity link function,

$$\mu_{ij} = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the effects of the covariates, is ϕ and does not depend on the mean response. That is,

$$\text{Var}\,(Y_{ij}) = \phi\, v\,(\mu_{ij}) = \phi,$$

where $v\,(\mu_{ij}) = 1$ and ϕ is a scale parameter that needs to be estimated. Note that this model makes the strong, and often unrealistic, assumption that the variance is homogeneous over time. Alternatively, a separate scale parameter, ϕ_j, could be estimated at the j^{th} occasion if the longitudinal design is balanced on time.

3. The within-subject association among the vector of repeated responses is modelled by assuming a first-order autoregressive correlation pattern

$$\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha^{|k-j|},$$

where $0 \leq \alpha \leq 1$. In this example it is assumed that the within-subject associations do not depend upon the means, but only on a single correlation parameter, α. That is, α is used to model the pairwise correlations among the responses (which are assumed to be approximately equally separated in time).

This illustration of a marginal model for a continuous response is a special case of the linear regression models for longitudinal data considered in Part II. However, marginal models provide a much broader class of models for continuous responses. For example, the means can be related to the covariates by a link function other than the identity or the variances can be allowed to depend on some known function of

the means. Also, in this illustration the correlations among the components of Y_i have been specified as a function of the parameter α via a first-order autoregressive correlation pattern. Other models for the correlation (e.g., unstructured, exchangeable or equicorrelated correlation patterns) are also possible.

Example 2: Marginal Model for Counts

Next, suppose that Y_{ij} is a count and we wish to relate changes in the expected count (or expected rate) to the covariates. Counts are often modelled as Poisson random variables, using a log link function and a Poisson variance function. This motivates the following illustration of a marginal model for Y_{ij}:

1. The mean of Y_{ij} is related to the covariates through a log link function,

$$\log\,(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the effects of the covariates, depends on the mean response,

$$\text{Var}\,(Y_{ij}) = \phi\,\mu_{ij},$$

where ϕ is a time-invariant scale parameter that needs to be estimated.

3. The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise correlation pattern,

$$\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha_{jk}.$$

Here, a balanced longitudinal design is assumed and the vector of parameters α represents the pairwise correlations among the responses.

The marginal model specified above is a log-linear regression model, with an extra-Poisson variance assumption. The within-subject association is specified in terms of an unstructured pairwise correlation pattern. Of course, other choices for the link and variance functions are possible; similarly, other models for the correlation (e.g., first-order autoregressive correlation pattern) are also possible. In this example the extra-Poisson variance assumption allows the variance to be inflated by a factor ϕ (when $\phi > 1$). In many biomedical applications, count data have variability that far exceeds that predicted by the Poisson distribution; this phenomenon is referred to as *overdispersion*. Indeed, many statisticians believe that overdispersion is the rule, not the exception, when dealing with count data. The excess variability can be accounted for by including the scale factor ϕ in the specification of the variance.

Example 3: Marginal Model for a Binary Response

Finally, suppose that Y_{ij} is a binary response, taking values of 0 (denoting "failure") or 1 (denoting "success"), and it is of interest to relate changes in $E(Y_{ij}) = \Pr(Y_{ij} = 1)$ to the covariates. With a binary response, the distribution of each Y_{ij} is Bernoulli

and the probability of success is often modelled using a logit or probit link function. Recall that for a Bernoulli random variable, the variance is a known function of the mean. This motivates the following illustration of a marginal model for Y_{ij}:

1. The mean of Y_{ij}, or probability of success, is related to the covariates by a logit link function,
$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = X_{ij}'\beta.$$
2. The variance of each Y_{ij}, given the effects of the covariates, depends on the mean response,
$$\text{Var}\,(Y_{ij}) = \mu_{ij}(1-\mu_{ij}),$$
and $\phi = 1$.
3. The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise log odds ratio pattern,
$$\log \text{OR}\,(Y_{ij}, Y_{ik}) = \alpha_{jk},$$
where
$$\text{OR}(Y_j, Y_k) = \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}.$$

The marginal model specified above is a logistic regression model, with a Bernoulli variance assumption, $\text{Var}\,(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$, and an unstructured within-subject association specified in terms of pairwise log odds ratios rather than pairwise correlations.

The three examples of marginal models considered so far are purely illustrative. They demonstrate how the choices of the three components of a marginal model might differ according to the type of response variable. However, these three examples should not be considered prescriptions for constructing marginal models; in principle, any suitable link function can be chosen and alternative assumptions about the variances and within-subject associations can be made. The choices for the three components of a marginal model should reflect statistical and subject-matter considerations.

Finally, we note that there is an implicit assumption in the first component of a marginal model that is often overlooked. Marginal models assume that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends only on X_{ij}

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij}). \qquad (11.1)$$

With time-stationary covariates, this assumption poses no difficulties; it necessarily holds since $X_{ij} = X_{ik}$ for all occasions $k \neq j$. Also, with time-varying covariates that are fixed by design of the study (e.g., time since baseline, treatment group indicator in a crossover trial), the assumption also holds since values of the covariates at any occasion are determined *a priori* by study design and in a manner completely unrelated

to the longitudinal response. However, when a time-varying covariate varies randomly over time the assumption made in (11.1) may not hold. For example, the assumption will be violated when the current value of the response, say Y_{ij}, given the current covariates X_{ij}, predicts the subsequent value of X_{ij+1}. This might arise, for example, in a longitudinal observational study designed to assess the effects of physical exercise on reducing blood glucose levels. If study participants with elevated blood glucose levels, Y_{ij}, at the j^{th} occasion subsequently increase their amount of physical activity, X_{ij+1} (while those with normal blood glucose levels continue to maintain their usual level of physical activity), then the assumption made in (11.1) does not hold. As a result, somewhat greater care is required when fitting marginal models with time-varying covariates that are not fixed by design of the study. A more detailed discussion of this issue is postponed until Chapter 15 (see Section 15.3).

11.3 ESTIMATION FOR MARGINAL MODELS: GENERALIZED ESTIMATING EQUATIONS

Since there is no convenient specification of the joint multivariate distribution of Y_i for marginal models when the responses are discrete, we require an alternative to maximum likelihood estimation. The generalized estimating equations (GEE) approach provides that alternative. The GEE approach is based on the concept of "estimating equations" and provides a very general and unified approach for analyzing correlated responses that can be discrete or continuous. The essential idea behind the GEE approach is to generalize and extend the usual likelihood equations for a generalized linear model for a univariate response by incorporating the covariance matrix of the vector of responses, Y_i. For the case of linear models (i.e., marginal models with an identity link function), the generalized least squares (GLS) estimator of β discussed in Chapter 4 can be considered a special case of the GEE approach. For marginal models with non-linear link functions, this approach is known as "generalized estimating equations" (or GEE).

Suppose, as in Section 11.2, that the following marginal model has been assumed:

1. The marginal expectation of the response, $E\,(Y_{ij}) = \mu_{ij}$, depends on the covariates, X_{ij}, through a known link function
$$g\,(\mu_{ij}) = \eta_{ij} = X_{ij}'\beta.$$
2. The variance of each Y_{ij}, given the covariates, depends on the mean according to
$$\text{Var}\,(Y_{ij}) = \phi\, v\,(\mu_{ij}),$$
where $v\,(\mu_{ij})$ is a known "variance function" (i.e., a known function of the mean, μ_{ij}) and ϕ is a scale parameter that may be known or may need to be estimated. In principle, a separate scale parameter, ϕ_j, could be estimated at each occasion for balanced designs; alternatively, the scale parameter could

depend on the times of measurement, with $\phi(t_{ij})$ being some parametric function of t_{ij}. In practice, a limitation of many of the implementations of the GEE approach in widely available software is that they assume the scale parameter ϕ is time-invariant. This restriction on the scale parameter makes the GEE approach unappealing for analyzing longitudinal data when the response variable is continuous and the variance of the repeated measurements is not constant over the duration of the study.

3. The *pairwise* (or two-way) within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of the means, μ_{ij}, and an additional set of within-subject association parameters, α. For example, when the vector of parameters α represents the pairwise correlations among the responses, the covariances among the responses depend on $\mu_{ij}(X_{ij}'\beta)$, ϕ, and α. That is, given a model for the pairwise correlations, the corresponding covariance matrix can be constructed as the product of standard deviations and correlations

$$V_i = A_i^{\frac{1}{2}} \text{Corr}(Y_i) A_i^{\frac{1}{2}},$$

where A_i is a diagonal matrix with $\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$ along the diagonal (and $A_i^{\frac{1}{2}}$ is a diagonal matrix with the standard deviations, $\sqrt{\phi v(\mu_{ij})}$, along the diagonal), and $\text{Corr}(Y_i)$ is the correlation matrix (here a function of α). In the parlance of the GEE approach, V_i is known as a "working" covariance matrix to distinguish it from the true underlying covariance among the Y_i. That is, the term "working" acknowledges our uncertainty about the assumed model for the variances and within-subject associations; unless they have been correctly modelled, our model for the covariance matrix may not be correct.

Next, we provide some motivation for the GEE approach. Recall from Chapter 4 that the generalized least squares (GLS) estimator of β for the linear model minimizes the objective function

$$\sum_{i=1}^{N} (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta).$$

Using calculus, it can be shown that if a minimum of this function exists it must solve the following equations:

$$\sum_{i=1}^{N} X_i' \Sigma_i^{-1} (y_i - \mu_i) = 0,$$

where $\mu_i = \mu_i(\beta) = X_i\beta$. (Here, $\mu_i(\beta)$ simply denotes that the mean vector, μ_i, depends on β.) For the linear model, these equations have the following closed-form solution

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} y_i),$$

and $\widehat{\beta}$ is known as the GLS estimator of β. The GEE estimator of β for marginal models (or generalized linear models for longitudinal data) can be thought of as arising from minimizing the following objective function:

$$\sum_{i=1}^{N} \{y_i - \mu_i(\beta)\}' V_i^{-1} \{y_i - \mu_i(\beta)\}, \tag{11.2}$$

with respect to β, where V_i is treated as known (by ignoring its dependence on β through μ_i) and μ_i is the vector of mean responses, with elements

$$\mu_{ij} = \mu_{ij}(\beta) = g^{-1}(X_{ij}'\beta).$$

Using calculus, it can be shown that if a minimum of the function given by (11.2) exists it must solve the following *generalized estimating equations*:

$$\sum_{i=1}^{N} D' V_i^{-1} (y_i - \mu_i) = 0, \tag{11.3}$$

where V_i is the so-called "working" covariance matrix and $D_i = \partial\mu_i/\partial\beta$ is the "derivative" matrix (i.e., the matrix containing the derivative of μ_i with respect to the components of β). By "working" covariance matrix we mean that V_i approximates the true underlying covariance matrix for Y_i, that is, $V_i \approx \text{Cov}(Y_i)$, recognizing that $V_i \neq \text{Cov}(Y_i)$ unless the models for the variances and the within-subject associations are correct. As before, we let the true covariance matrix for Y_i be denoted by Σ_i. The matrix D_i can easily be derived using calculus and can be thought of as a matrix that transforms from the original units of Y_i (and μ_i) to the units of $g(\mu_{ij})$. Recall that $g(\mu_{ij})$ is the scale on which β has interpretation (e.g., the log odds scale rather than the probability scale when the Y_{ij} are binary and a logit link function has been assumed). The $n_i \times p$ "derivative" matrix D_i is given by

$$D_i = \begin{pmatrix} \partial\mu_{i1}/\partial\beta_1 & \partial\mu_{i1}/\partial\beta_2 & \cdots & \partial\mu_{i1}/\partial\beta_p \\ \vdots & \vdots & \ddots & \vdots \\ \partial\mu_{in_i}/\partial\beta_1 & \partial\mu_{in_i}/\partial\beta_2 & \cdots & \partial\mu_{in_i}/\partial\beta_p \end{pmatrix},$$

and is only a function of β (since the μ_{ij} only depend on β). On the other hand, V_i is a function of β, ϕ, and α, since the diagonal elements of V_i are the variances and the off-diagonal terms are the "working" covariances. That is, the variances depend upon the means, and hence β, via the variance function, $v(\mu_{ij})$ (they also depend upon ϕ); the covariances among the components of Y_i depend upon both β and α. As a result, the generalized estimating equations are functions of both β and α. For generalized linear models with non-identity link functions, the GEE have no closed-form solution; instead, the solution requires an iterative algorithm.

Because the GEE depend on both β and α, the following iterative two-stage estimation procedure is required:

1. Given current estimates of α and ϕ, V_i is estimated and an estimate of β is obtained as the solution to the generalized estimating equations given by (11.3).

2. Given the current estimate of β, estimates of α and ϕ are obtained based on the standardized residuals

$$e_{ij} = (Y_{ij} - \widehat{\mu}_{ij}) / \sqrt{v(\widehat{\mu}_{ij})}.$$

For example, ϕ can be estimated by

$$\widehat{\phi} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{n_i} e_{ij}^2}{\sum_{i=1}^{N} n_i}.$$

The pairwise association parameters, α, can be estimated in a similar way, depending on the model for the within-subject association in the third component of the marginal model. For example, in a balanced design, when the association is expressed in terms of unstructured correlations, $\alpha_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$ can be estimated by

$$\widehat{\alpha}_{jk} = \left(\frac{1}{\widehat{\phi}\, N}\right) \sum_{i=1}^{N} e_{ij} e_{ik}.$$

Finally, in this two-stage estimation procedure, we usually iterate between steps 1 and 2 until convergence has been achieved; starting or initial estimates of β are usually obtained from fitting a generalized linear model assuming independent observations. This algorithm is computationally quite simple and the GEE approach has been implemented in many general-purpose statistical software packages.

At convergence, $\widehat{\beta}$, the solution to the generalized estimating equations, has the following properties:

1. $\widehat{\beta}$ is a consistent estimator of β. That is, with very high probability, $\widehat{\beta}$ is close to the population regression parameters β in large samples (i.e., for sufficiently large N). Of note, $\widehat{\beta}$ is a consistent estimator of β whether or not the within-subject associations have been correctly modelled. That is, for $\widehat{\beta}$ to provide a valid estimate of β we only require that the model for the mean response has been correctly specified. This is an important robustness property of $\widehat{\beta}$ that makes the GEE approach very appealing in many applications.

2. In large samples, the sampling distribution of $\widehat{\beta}$ is multivariate normal with mean β and

$$\text{Cov}(\widehat{\beta}) = B^{-1} M B^{-1},$$

where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

$$M = \sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}(Y_i)\, V_i^{-1} D_i.$$

Note that B and M can be estimated by replacing α, ϕ, and β by their estimates, and by replacing $\text{Cov}(Y_i) = \Sigma_i$ in M by $(Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'$. That is, the expression for $\widehat{\text{Cov}}(\widehat{\beta})$ is given by

$$\left(\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} \widehat{D}_i\right)^{-1} \left\{\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)' \widehat{V}_i^{-1} \widehat{D}_i\right\} \left(\sum_{i=1}^{N} \widehat{D}_i' \widehat{V}_i^{-1} \widehat{D}_i\right)^{-1}. \quad (11.4)$$

This is known as the empirical or so-called "sandwich" estimator; the components B and M can be thought of as the "bread" and "meat" of this sandwich estimator of $\text{Cov}(\widehat{\beta})$. Finally, if we model V_i correctly, $V_i = \Sigma_i$, and $\text{Cov}(\widehat{\beta}) = B^{-1}$.

In summary, the GEE approach has a number of appealing properties for estimation of the regression parameters in marginal models. First, in many longitudinal designs the GEE estimator $\widehat{\beta}$ is almost as precise or efficient as the MLE. For example, it can be shown that the GEE has a similar expression to the likelihood equations for β in a linear model for continuous responses that are assumed to have a multivariate normal distribution. That is, the GLS estimator of β can be considered a special case of the GEE approach. The GEE also has an expression similar to the likelihood equations for β in certain models for discrete longitudinal data. As a result, for many longitudinal designs, there is little loss of precision when the GEE approach is adopted as an alternative to maximum likelihood. Second, the GEE estimator $\widehat{\beta}$ has a very appealing robustness property, yielding a consistent estimate of β even if the within-subject associations among the repeated measures have been misspecified. Although the GEE approach yields a consistent estimate of β under misspecification of the within-subject associations, the usual standard errors obtained under the misspecified model for the within-subject association are not valid. Fortunately, in many cases, valid standard errors for $\widehat{\beta}$ can be obtained using the empirical or so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$.

Finally, although the main emphasis of this chapter has been on longitudinal analysis of a discrete response, the GEE approach can be applied equally to continuous responses. That is, for the linear regression models described in Part II, the multivariate normal assumption is not crucial. Specifying a linear model for the longitudinal responses and a model for the covariance among the responses is sufficient for the purposes of estimating β using the GEE approach. As was mentioned above, the GLS estimator of β can be considered a special case of the GEE approach and the multivariate normal distribution assumption for the responses is not required. The validity of the GLE/GEE estimates of β rests only on having a correct model for the mean response. When the model for the covariance is misspecified, valid standard errors for $\widehat{\beta}$ can be obtained using the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$. Thus, although the GEE approach and the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$ are more widely used in marginal models for discrete data, they can also be applied in the linear models for continuous data described in Part II. However, because many of the implementations of the GEE approach in widely available software assume the scale parameter is time-invariant, we do not recommend their use for analyzing longitudinal data when the response variable is continuous. Instead, the GEE approach can be implemented

using existing software for the general linear model (e.g., PROC MIXED in SAS) that allows a much wider range of covariance pattern models and/or random effects covariance structures (coupled with the option of calculating standard errors for $\widehat{\beta}$ based on the "sandwich" estimator).

A Note on the "Sandwich" Estimator of Cov($\widehat{\beta}$)

An appealing property of the GEE estimator $\widehat{\beta}$ is that it yields a consistent estimate of β even if the assumed model for the covariances among the repeated measures is not correct. It only requires that the model for the mean response be correct. This robustness property of GEE is important because the usual focus of a longitudinal study is on changes in the mean response. Based either on theoretical grounds (e.g., randomization in an experiment) or subject-matter knowledge of similar data, the data analyst can often specify how changes in the mean response depend on the covariates. On the other hand, much less is usually known about the patterns of two- and higher-way associations among the responses; moreover, these so-called "higher-order moments" are increasingly difficult to estimate from the data.

For inferences about β, valid standard errors can be obtained from the so-called "sandwich" estimator of Cov($\widehat{\beta}$) given by (11.4). The remarkable property of the "sandwich" estimator is that it is also robust in the sense that it provides valid standard errors when the assumed model for the covariances among the repeated measures is not correct. That is, with large sample sizes, the "sandwich" estimator of Cov($\widehat{\beta}$) yields correct standard errors. Maintaining the culinary theme of this section, it would seem that we can have our cake and eat it: we can obtain valid estimates of β and its sampling variability, even if we have not modelled the within-subject association correctly. Indeed, some readers may see it as a delicious irony that we can disregard the model for the covariances among the repeated measures for the purposes of inference about β.

This raises an important issue. Why bother expending effort to model the within-subject association? For example, naively assuming the responses are independent (i.e., specifying the "working" covariance matrix as diagonal) yields valid estimates of β; valid standard errors can then be obtained using the "sandwich" estimator of Cov($\widehat{\beta}$). There are two main reasons for modelling the covariance. First, in general, the closer the "working" covariance matrix (V_i) approximates the true underlying covariance matrix (Σ_i), the greater the efficiency or precision with which β can be estimated. That is, a "working" covariance matrix that approximates the true underlying covariance matrix makes optimal use of the available data for estimation of β. Second, the robustness property of the "sandwich" estimator of Cov($\widehat{\beta}$) is a large sample (or asymptotic) property. In general, use of the "sandwich" estimator of Cov($\widehat{\beta}$) is best suited to balanced longitudinal designs where the number of subjects (N) is relatively large and the number of repeated measures (n) is relatively small. Moreover, the "sandwich" estimator is less appealing when the design is severely unbalanced and/or when there are few replications to estimate the true underlying covariance matrix. In applications, use of the "sandwich" estimator implicitly relies on

there being many replications of the vector of responses associated with each distinct set of covariate values. For example, in a longitudinal clinical trial with two treatment groups there will be many replications of Y_i associated with the two distinct set of covariate values (X_i) for the treatment and control groups. In that case, use of the "sandwich" estimator of Cov($\widehat{\beta}$) is justified because there is a sufficient number of replications (or number of subjects) to estimate the true underlying covariance matrix within each treatment group. In many observational studies, however, there may be few, if any, replications of Y_i associated with each distinct set of covariate values, especially when X_i includes many covariates and/or quantitative covariates. Similarly, if the longitudinal design is severely unbalanced, with each individual having a unique sequence of measurement occasions, $t_{i1}, \ldots, t_{in_i}$, there are no replications at each of the measurement occasions. In these cases, the use of the "sandwich" estimator is problematic. In particular, "sandwich"-based standard errors tend to be biased downward, that is, the nominal standard errors are too small and underestimate the covariance of $\widehat{\beta}$. In addition, the sampling variability of the "sandwich" estimator of Cov($\widehat{\beta}$) can be very large, resulting in an unstable estimate of variability.

In summary, reliance on the "sandwich" estimator of Cov($\widehat{\beta}$) is unappealing when the number of independent subjects is modest (relative to the number of repeated measures), the design is inherently unbalanced, or when subjects cannot be grouped on the basis of having identical covariate design matrices. For any of these cases it is advantageous to model the covariances among the responses and use a "model-based" estimator of Cov($\widehat{\beta}$). The model-based estimator is given by

$$\text{Cov}(\widehat{\beta}) = B^{-1},$$

where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

and can be estimated by replacing α, ϕ, and β by their estimates. This estimator of Cov($\widehat{\beta}$) is called a "model-based" estimator to remind us that it yields valid standard errors provided that the "working" covariance matrix, V_i, is a close approximation to the true underlying covariance matrix, Σ_i. That is, the "model-based" estimator does require that the model for the covariance, the "working" covariance, be correctly specified.

11.4 CASE STUDIES

Next, we illustrate the main ideas presented in this chapter by considering marginal models for analyzing longitudinal data from two different studies. The first illustration employs marginal models to analyze obesity data in a sample of school-age children from the Muscatine Coronary Risk Factor (MCRF) study. The second illustration considers marginal models for analyzing count data from a study comparing two antibiotics to a placebo for the treatment of leprosy.

Table 11.1 Percentage of children from the Muscatine Coronary Risk Factor study classified as obese in 1977, 1979, and 1981.

Gender	Age Cohort	Percentage Obese 1977	1979	1981
Males				
	5–7	7.9	15.4	21.2
	7–9	18.8	20.5	23.7
	9–11	21.2	22.7	22.5
	11–13	24.3	21.8	19.4
	13–15	19.2	21.1	18.2
Females				
	5–7	14.0	17.2	25.1
	7–9	16.5	24.0	24.9
	9–11	25.4	26.2	22.2
	11–13	23.8	22.1	19.9
	13–15	22.9	25.8	20.9

Muscatine Coronary Risk Factor Study

The Muscatine Coronary Risk Factor (MCRF) study was a longitudinal survey of school-age children in Muscatine, Iowa (Woolson and Clarke, 1984; Lauer *et al.*, 1997). The goal of the study was to examine the development and persistence of risk factors for coronary disease in children. In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years, were obtained biennially from 1977 to 1981. In total, data were collected on 4856 boys and girls. Although each child was eligible to participate in all three surveys, the data are incomplete for many children.

In this section we present longitudinal analyses of a binary response, indicating whether the child is obese. At each occasion, on the basis of a comparison of their weight to age-gender specific norms, children were classified as obese or not obese. The goal of the analyses is to determine whether the risk of obesity increases with age and whether patterns of change in obesity are the same for boys and girls. The percentages of the children classified as obese at each of the three measurement occasions are displayed in Table 11.1. These percentages were calculated based on the available data in each age-gender cohort at each occasion. These descriptive

statistics suggest that the rates of obesity increase from ages 6 to 12, but decline thereafter. They also suggest that the rates of obesity are higher for girls at all ages.

Initially, our analysis of these data assumes that there are no cohort effects. The marginal probability of obesity is modelled as a logistic function of the covariates: linear and quadratic age, gender, and the gender-age interactions. Here, age is the midpoint of the age cohort that a child belongs to (e.g., 6, 8, and 10 years at the first, second, and third occasions for the cohort of children initially aged 5–7 years). Letting $Y_{ij} = 1$ if the i^{th} child is classified as obese at the j^{th} occasion, and $Y_{ij} = 0$ otherwise, we assume that the marginal probability of obesity at each occasion follows the logistic model

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\mathtt{gender_i} + \beta_3\,\mathtt{age_{ij}} + \beta_4\,\mathtt{age_{ij}^2} + \beta_5\,\mathtt{gender_i} \times \mathtt{age_{ij}} + \beta_6\,\mathtt{gender_i} \times \mathtt{age_{ij}^2},$$

where $\mathtt{age_{ij}}$ = midpoint of age cohort at the j^{th} occasion $-$ 12 years; $\mathtt{gender_i} = 1$ if the i^{th} child is female, and $\mathtt{gender_i} = 0$ otherwise. This specifies the first component of a marginal model, the model for the mean response. It is assumed that the log odds of obesity changes curvilinearly with age (i.e., quadratic age trend), but the trend over time is allowed to be different for girls and boys. Next, we assume that

$$\mathrm{Var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij}).$$

This specifies the second component, the variance function and known scale parameter ($\phi_j = 1, j = 1, \ldots, 3$). Finally, we need to make assumptions about the pairwise within-subject associations among the binary responses. Because the response is binary, correlation is not the most appealing metric for association. As was mentioned in Section 11.2, with binary responses the correlations are constrained and must satisfy certain linear inequalities determined by the marginal probabilities. Instead, we specify the association in terms of pairwise log odds ratios, a more natural measure of association between pairs of binary responses. Specifically, the within-subject association among the three repeated binary responses is assumed to have the following unstructured pairwise log odds ratio pattern,

$$\log \mathrm{OR}\,(Y_{ij}, Y_{ik}) = \alpha_{jk},$$

where

$$\mathrm{OR}(Y_j, Y_k) = \frac{\Pr(Y_j=1, Y_k=1)\Pr(Y_j=0, Y_k=0)}{\Pr(Y_j=1, Y_k=0)\Pr(Y_j=0, Y_k=1)}.$$

The estimated regression coefficients and pairwise log odds ratios for the within-subject association obtained using the GEE approach are presented in Table 11.2. A test of the hypothesis that changes in the log odds of obesity are the same for boys and girls, $\mathrm{H_0}$: $\beta_5 = \beta_6 = 0$, can be constructed using a multivariate Wald statistic. This test produces a Wald statistic, $W^2 = 0.91$, with 2 df ($p > 0.60$) and the null hypothesis cannot be rejected at the 0.05 significance level. Thus, a marginal logistic

Table 11.2 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study.

Variable	Estimate	SE†	Z
Intercept	−1.2135	0.0506	−24.00
$\texttt{gender}_i$	0.1159	0.0711	1.63
$\texttt{age}_{ij}$	0.0378	0.0133	2.85
$\texttt{age}_{ij}^2$	−0.0175	0.0034	−5.19
$\texttt{gender}_i \times \texttt{age}_{ij}$	0.0075	0.0182	0.41
$\texttt{gender}_i \times \texttt{age}_{ij}^2$	0.0039	0.0046	0.85
α_{12}	3.1528	0.1280	24.63
α_{13}	2.5975	0.1353	19.20
α_{23}	2.9868	0.1236	24.17

†SE based on "sandwich" variance estimator.

regression model without the gender × age interactions is defensible. Note that the $\widehat{\alpha}_{jk}$ have interpretation in terms of the pairwise log odds ratio for the responses at the j^{th} and k^{th} occasions. The pairwise log odds ratios between adjacent occasions are very similar and approximately equal to 3, indicating that the odds ratio for within-subject association is approximately 20 (or e^3). As expected, there is strong positive association among the indicators of obesity status at the three measurement occasions.

Recall that these data on obesity are from five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years. Our analysis of the trends in the risk of obesity with age implicitly assumes that there are no cohort effects. That is, the logistic model for the probability of obesity assumes that the cross-sectional and longitudinal effects of aging are identical. Following the approach used in the analysis of the FEV_1 data in Section 8.8 (and discussed in greater detail in Chapter 15; see Section 15.4), we can conduct a formal test of equality of the cross-sectional and longitudinal effects of aging by including linear and quadratic effects of both baseline age and current age minus baseline age (and also their interactions with gender) in the logistic model,

$$\begin{aligned}\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} &= \beta_1 + \beta_2 \texttt{gender}_i + \beta_3 \texttt{age}_{i1} + \beta_4 \texttt{age}_{i1}^2 + \beta_5 \texttt{gender}_i \times \texttt{age}_{i1} \\ &\quad + \beta_6\, \texttt{gender}_i \times \texttt{age}_{i1}^2 + \beta_7\,(\texttt{age}_{ij} - \texttt{age}_{i1}) + \beta_8\,(\texttt{age}_{ij} - \texttt{age}_{i1})^2 \\ &\quad + \beta_9\, \texttt{gender}_i \times (\texttt{age}_{ij} - \texttt{age}_{i1}) + \beta_{10}\, \texttt{gender}_i \times (\texttt{age}_{ij} - \texttt{age}_{i1})^2.\end{aligned}$$

Table 11.3 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study, omitting gender × age and gender × age² interactions.

Variable	Estimate	SE†	Z
Intercept	−1.2283	0.0477	−25.75
$\texttt{gender}_i$	0.1449	0.0627	2.31
$\texttt{age}_{ij}$	0.0418	0.0091	4.60
$\texttt{age}_{ij}^2$	−0.0155	0.0023	−6.73
α_{12}	3.1496	0.1280	24.61
α_{13}	2.5931	0.1352	19.17
α_{23}	2.9878	0.1236	24.18

†SE based on "sandwich" variance estimator.

This model distinguishes between the cross-sectional effects of aging ($\beta_3, \beta_4, \beta_5$, and β_6) and the longitudinal effects of aging ($\beta_7, \beta_8, \beta_9$, and β_{10}). Note that, when $\beta_3 = \beta_7$, $\beta_4 = \beta_8$, $\beta_5 = \beta_9$, and $\beta_6 = \beta_{10}$, we obtain the logistic model considered previously. A test of equality of the cross-sectional and longitudinal effects of aging,

$$\text{H}_0\text{: } (\beta_3 - \beta_7) = (\beta_4 - \beta_8) = (\beta_5 - \beta_9) = (\beta_6 - \beta_{10}) = 0,$$

produces a (multivariate) Wald statistic, $W^2 = 7.62$, with 4 df, ($p > 0.10$). This suggests that the results for aging presented in Table 11.2 are not confounded by cohort effects.

Next, we consider a marginal logistic regression model without the gender × age interactions. Specifically, we consider the following model for the log odds of obesity:

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2 \texttt{gender}_i + \beta_3 \texttt{age}_{ij} + \beta_4 \texttt{age}_{ij}^2,$$

while retaining the same assumptions about the variances and pairwise log odds ratios. The estimated regression coefficients (and pairwise log odds ratios) for this model are presented in Table 11.3. The estimated effect of age² is significant at the 0.05 level and these results provide evidence that the log odds of obesity increases from 6 to 12 years, levels off between age 12 to age 14, and declines between 14 to 18 years. Although the rates of obesity are significantly higher for girls at all ages, the patterns of change in the rates of obesity over time do not depend on gender. To translate these

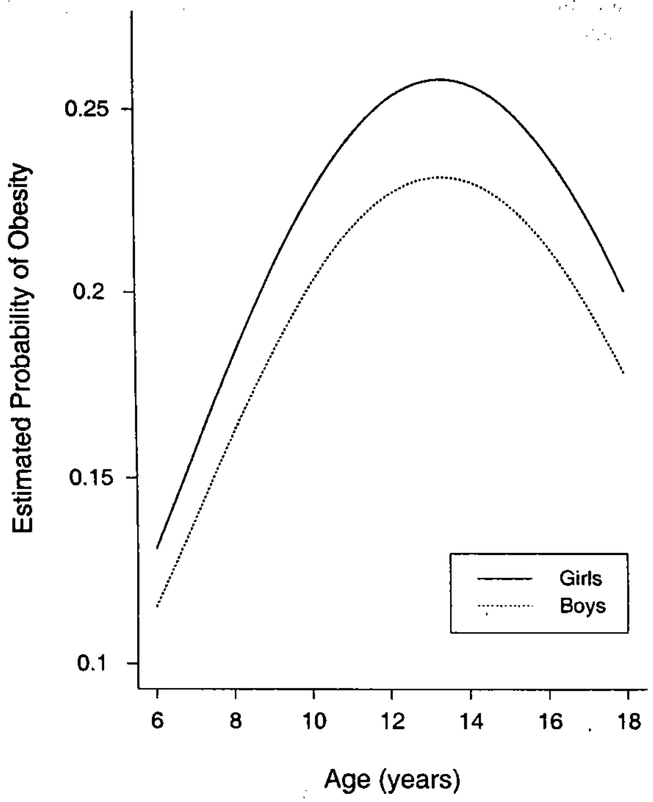

Fig. 11.1 Estimated probability of obesity versus age for boys and girls in the Muscatine Coronary Risk Factor study.

results on to a more interpretable scale, we can estimate the probability of obesity at each age for boys and girls,

$$\frac{e^{\widehat{\beta}_1+\widehat{\beta}_2\,\mathtt{gender}_i+\widehat{\beta}_3\,\mathtt{age}_{ij}+\widehat{\beta}_4\,\mathtt{age}_{ij}^2}}{1+e^{\widehat{\beta}_1+\widehat{\beta}_2\,\mathtt{gender}_i+\widehat{\beta}_3\,\mathtt{age}_{ij}+\widehat{\beta}_4\,\mathtt{age}_{ij}^2}}.$$

For example, the estimated probability of obesity for boys at ages 6, 10, 14, and 18 is 0.12, 0.20, 0.23, and 0.18, respectively; for girls, the estimated probability of obesity at ages 6, 10, 14, and 18 is 0.13, 0.22, 0.26, and 0.20, respectively (see Figure 11.1). Note that with the logistic model, an additive effect of gender does not translate into a constant difference over time in the probability of obesity. Potential confounding

Table 11.4 Parameter estimates and standard errors from marginal logistic regression model for the obesity data from the Muscatine Coronary Risk Factor study, omitting gender × age and gender × age^2 interactions.

Variable	Estimate	SE†	Z
Intercept	−1.2270	0.0477	−25.72
$\mathtt{gender}_i$	0.1445	0.0627	2.31
$\mathtt{age}_{ij}$	0.0416	0.0091	4.58
$\mathtt{age}_{ij}^2$	−0.0156	0.0023	−6.77
α_1	3.0684	0.0957	32.07
α_2	2.5929	0.1353	19.17

†SE based on "sandwich" variance estimator.

of these trends by cohort effects can be examined by including linear and quadratic effects of both baseline age and current age minus baseline age in the logistic model. A test of equality of the cross-sectional and longitudinal effects of aging produces a (multivariate) Wald statistic, $W^2 = 5.35$, with 2 df, ($p > 0.05$), suggesting that the results for aging presented in Table 11.3 are not confounded by cohort effects.

Finally, purely for illustrative purposes, we consider the same logistic model for the log odds of obesity,

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2\,\mathtt{gender}_i + \beta_3\,\mathtt{age}_{ij} + \beta_4\,\mathtt{age}_{ij}^2,$$

and retain the same assumptions about the variances, but assume a Toeplitz pattern for the log odds ratios:

$$\log \mathrm{OR}(Y_{ij}, Y_{ij-1}) = \alpha_1;$$
$$\log \mathrm{OR}(Y_{ij}, Y_{ij-2}) = \alpha_2.$$

The estimated parameters for this model are displayed in Table 11.4 and the estimates of the regression parameters (and their standard errors) are very similar to those reported in Table 11.3. The estimates of the within-subject log odds ratios display a characteristic decreasing time-dependence as the time separation increases. Finally, since the Toeplitz pattern for the within-subject log odds ratios is nested within the unstructured pairwise log odds ratio model it is possible to assess the goodness of fit of the Toeplitz model. That is, by appropriately reparameterizing the unstructured pairwise log odds ratio model,

Table 11.5 Mean count of leprosy bacilli at six sites of the body (and variance) pre- and post-treatment.

Treatment Group	Baseline	Post-Treatment
Drug A (Antibiotic)	9.3 (22.7)	5.3 (21.6)
Drug B (Antibiotic)	10.0 (27.6)	6.1 (37.9)
Drug C (Placebo)	12.9 (15.7)	12.3 (51.1)

$$\log \mathrm{OR}(Y_{i1}, Y_{i2}) = \alpha_1,$$
$$\log \mathrm{OR}(Y_{i1}, Y_{i3}) = \alpha_2,$$
$$\log \mathrm{OR}(Y_{i2}, Y_{i3}) = \alpha_1 + \alpha_3,$$

a 1-degree-of-freedom goodness-of-fit test (based on a Wald test) for the Toeplitz model can be constructed. The test of the null hypothesis, H_0: $\alpha_3 = 0$, produces a Wald $Z = -0.99$ ($p > 0.30$), indicating that the Toeplitz pattern is defensible for these data.

Clinical Trial of Antibiotics for Leprosy

Next, we consider count data from a placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines (Snedecor and Cochran, 1967). Participants in the study were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C). Prior to receiving treatment, baseline data on the number of leprosy bacilli at six sites of the body where the bacilli tend to congregate were recorded for each patient. After several months of treatment, the number of leprosy bacilli at six sites of the body were recorded a second time. The outcome variable is the total count of the number of leprosy bacilli at the six sites.

Before proceeding with the analysis, a feature of these data should be noted. These data display substantially greater variability than that predicted by the mean under a Poisson distribution assumption. The mean number of bacilli and the variance are displayed in Table 11.5. Although the sample sizes are relatively small, these

descriptive statistics reveal that the variances are substantially greater than the means. As a result, a Poisson assumption for the variance, with $\mathrm{Var}(Y_{ij}) = \mu_{ij}$, is not appropriate for these data. Instead, we consider

$$\mathrm{Var}(Y_{ij}) = \phi\, \mu_{ij},$$

where it is assumed that $\phi > 1$.

In this study, the question of main scientific interest is whether treatment with antibiotics (drugs A and B) reduces the abundance of leprosy bacilli at the six sites of the body when compared to placebo (drug C). To address this question we can compare the changes, from baseline to follow-up, in the average count of leprosy bacilli in the three treatment groups. This can be expressed in the following marginal model for the expected counts of leprosy bacilli

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2\, \mathtt{time_{ij}} + \beta_3\, \mathtt{time_{ij}} \times \mathtt{trt_{1i}} + \beta_4\, \mathtt{time_{ij}} \times \mathtt{trt_{2i}},$$

where Y_{ij} is the count of the number of leprosy bacilli for the i^{th} patient in the j^{th} period of observation ($j = 1, 2$). The variables $\mathtt{trt}_1$ and $\mathtt{trt}_2$ are indicator variables for drugs A and B respectively, with $\mathtt{trt}_1 = 1$ if a patient was randomized to drug A and $\mathtt{trt}_1 = 0$ otherwise, and $\mathtt{trt}_2 = 1$ if a patient was randomized to drug B and $\mathtt{trt}_2 = 0$ otherwise. The binary variable, `time`, denotes the baseline and post-treatment follow-up periods, with `time` $= 0$ for the baseline period (period 1) and `time` $= 1$ for the post-treatment follow-up period (period 2). Because patients were randomized to one of the three treatments, the model does not include main effects of treatment (since the mean count of the number of leprosy bacilli at baseline can be assumed to be equal in the three treatment groups). To complete the specification of the model, we must make assumptions about the variances of the counts and the within-subject association among the repeated counts. Because of the discernible *overdispersion* in these data (relative to Poisson variability), we assume that the variance of Y_{ij} is given by

$$\mathrm{Var}(Y_{ij}) = \phi\, \mu_{ij},$$

where ϕ can be thought of as an overdispersion factor. Finally, the within-subject association is accounted for by assuming a common correlation,

$$\mathrm{Corr}(Y_{i1}, Y_{i2}) = \alpha.$$

In this marginal model for the expected number of leprosy bacilli, all of the covariates are dichotomous and the log-linear regression parameters can be given interpretations in terms of (log) rate ratios. In Table 11.6 we summarize the interpretation of β in terms of the log expected counts in the three groups at baseline and during post-treatment follow-up. So, for example, the expected count of leprosy bacilli at the six sites of the body at baseline in the placebo group (drug C) is e^{β_1}; while the expected count during the follow-up period is $e^{\beta_1+\beta_2}$. Thus, e^{β_2} is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the placebo group (drug C). Similarly, $e^{\beta_2+\beta_3}$ is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the group randomized to drug A. Finally, $e^{\beta_2+\beta_4}$ is the rate ratio of

Table 11.6 Parameters of the marginal log-linear regression model for the leprosy bacilli data.

Treatment Group	Period	$\log(\mu_{ij})$
Drug A (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_3$
Drug B (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_4$
Drug C (Placebo)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2$

leprosy bacilli, comparing the follow-up period to baseline, in the group randomized to drug B.

As a result, a direct comparison of the three treatment groups in terms of changes in the expected rates of leprosy bacilli is expressible in terms of β_3 and β_4. That is, β_3 and β_4 represents the difference between the changes in the log expected rates, comparing drug A and B to the placebo (drug C). For example, a value of $\beta_3 < 0$ indicates a greater reduction in the rate of bacilli from baseline in the group randomized to drug A (when compared to the placebo group).

The estimated regression coefficients are displayed in Table 11.7 (with standard errors based on the "sandwich" estimator). A test of the null hypothesis, H_0: $\beta_3 = \beta_4 = 0$, produces a (multivariate) Wald statistic, $W^2 = 6.99$, with 2 degrees of freedom ($p < 0.05$). This indicates that treatment with antibiotics significantly reduces the abundance of leprosy bacilli at the six sites of the body. A test of the null hypothesis that both antibiotics are equally effective, H_0: $\beta_3 = \beta_4$, produces a Wald statistic, $W^2 = 0.08$, with 1 degree of freedom ($p > 0.7$). Thus, we cannot reject the null hypothesis that the two antibiotics are equally effective in reducing the number of leprosy bacilli. To obtain a common estimate of the log rate ratio, comparing both antibiotics (drugs A and B) to placebo, we can fit the reduced model

$$\log E(Y_{ij}) = \log \mu_{ij} = \beta_1 + \beta_2 \texttt{time}_{ij} + \beta_3 \texttt{time}_{ij} \times \texttt{trt}_i,$$

where the variable `trt` is an indicator variable for antibiotics, with `trt` $= 1$ if a patient was randomized to either drug A or B and `trt` $= 0$ otherwise. We retain the same assumptions about the variance and correlation as before.

The estimated regression coefficients are displayed in Table 11.8. The common estimate of the log rate ratio, comparing post-treatment rates of bacilli in the antibiotics

Table 11.7 Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\texttt{time}_{ij}$	−0.0138	0.1573	−0.09
$\texttt{time}_{ij} \times \texttt{trt}_{1i}$	−0.5406	0.2186	−2.47
$\texttt{time}_{ij} \times \texttt{trt}_{2i}$	−0.4791	0.2279	−2.10

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.45$.
Estimated working correlation: $\widehat{\alpha} = 0.797$.

Table 11.8 Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data, with common estimate of the effect of antibiotics.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\texttt{time}_{ij}$	−0.0108	0.1572	−0.07
$\texttt{time}_{ij} \times \texttt{trt}_i$	−0.5141	0.1966	−2.62

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.41$.
Estimated working correlation: $\widehat{\alpha} = 0.780$.

group (drugs A and B) to placebo, is −0.5141. Thus, the rate ratio is 0.60 (or $e^{-0.5141}$), with 95% confidence interval, 0.41 to 0.88, indicating that treatment with antibiotics significantly reduces the average number of bacilli when compared to placebo. In the placebo group, there is a non-significant reduction in the average number of bacilli of approximately 1% (or $[1 - e^{-0.0108}] \times 100\%$), while in the antibiotics group there is a significant reduction of approximately 40% (or $[1 - e^{-0.0108-0.5141}] \times 100\%$).

Finally, the estimated pairwise correlation is relatively large (approximately 0.8), suggesting that there may be substantial heterogeneity among patients in their disease severity (as indicated by the counts of the number of bacilli). Of note, the estimated scale parameter is approximately 3.4, revealing overdispersion, relative to that predicted by Poisson variability, in these data.

Table 11.9 Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of log odds ratios, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
    CLASS id group;
    MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;
    REPEATED SUBJECT=id / WITHINSUBJECT=time LOGOR=FULLCLUST;
```

Table 11.10 Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD;
    CLASS id group;
    MODEL y=group time group*time / DIST=POISSON LINK=LOG;
    REPEATED SUBJECT=id / WITHINSUBJECT=time TYPE=UN;
```

11.5 COMPUTING: GENERALIZED ESTIMATING EQUATIONS USING PROC GENMOD IN SAS

To fit marginal models using the generalized estimating equations approach, we can use an enhanced option for repeated measures data in the PROC GENMOD procedure in SAS. Although PROC GENMOD is primarily a procedure for fitting generalized linear models to a single response, the use of a REPEATED statement in PROC GENMOD allows for the fitting of marginal models to correlated responses using the GEE approach.

For example, to fit a marginal logistic regression model to longitudinal data from two groups, with the within-subject associations specified in terms of log odds ratios, we can use the illustrative SAS commands given in Table 11.9. Similarly, to fit a marginal log-linear regression model to longitudinal data from two groups, with the within-subject associations specified in terms of correlations, we can use the illustrative SAS commands given in Table 11.10. Next, we describe the most salient parts of the command syntax required for fitting marginal models to longitudinal data using the GEE approach within PROC GENMOD in SAS.

PROC GENMOD <options>;

This statement calls the procedure GENMOD in SAS. It can also include an option for specifying the level of the response variable that is modelled. By default, the lower response level is modelled. For a binary response, coded (0,1), it is the probability that $Y = 0$ that is modelled. Use of the DESCENDING option reverses the default ordering of the response levels, resulting in the highest response level being modelled (i.e., the probability that $Y = 1$ for binary data that are coded as 0 and 1).

CLASS variables;

The CLASS statement is used to define all variables that are to be regarded as categorical or factors. By default, this statement will create indicator variables for each factor using a reference group coding, with the last level (where "last" here refers to the level with the largest alpha-numeric value) regarded as the reference group. Different sort orders for the CLASS variables can be requested by the ORDER=<option> on the PROC GENMOD statement.

MODEL response = <effects> / <options>;
MODEL events/trials = <effects> / <options>;

The MODEL statement specifies the response variable and the covariate effects. The second form of the MODEL statement, with the `events/trials` syntax, allows the response to be in the form of a ratio of two variables (e.g., counts of the number of successes and the number of trials) and is used for binomial response data. The linear predictor can include both discrete (defined in the CLASS statement) and quantitative (excluded from the CLASS statement) covariates. By default, PROC GENMOD includes a column of 1's for the intercept in the model.

The option that ordinarily is used to specify the distribution of a single univariate response has a somewhat different role when fitting a marginal model using the GEE approach. The option DIST=*keyword* does not specify a distribution for the vector of correlated responses; instead it specifies the default canonical link function and variance function that happen to be associated with particular exponential family distributions. For example, the option DIST=POISSON does not specify that the response vector (or even its separate components) has a Poisson distribution; instead it specifies that the mean of the response vector is related to the covariates via a log link function (the canonical link for the Poisson distribution) and the mean and variance of the responses are related by $\text{Var}(Y) = E(Y) = \mu$ (i.e., the variance function is $v(\mu) = \mu$).

Note that PROC GENMOD also provides a wide choice of options for the inclusion of a dispersion parameter, ϕ. However, the scale parameter ϕ is assumed to be time-invariant. This restriction on the scale parameter is a limitation of the implementation of the GEE approach that makes it unappealing for analyzing longitudinal data when the response variable is continuous and the variance of the repeated measurements is not constant over the duration of the study (e.g., the variance of baseline measurements is often discernibly

smaller than post-baseline measurements).

The LINK=*keyword* specifies the choice of built-in link function relating the mean response to the linear predictor. If the LINK=*keyword* is omitted, the default link function is the canonical link function associated with the particular exponential family distribution specified on DIST=*keyword*.

A final option often required when modelling count data is an offset. The OFFSET=*variable* specifies a variable to be used as an offset. For example, in modelling count data the rate is often of more direct interest and the denominator for the counts or "population at risk" can be included as an offset. Note that this variable cannot be a CLASS variable and it should not be included as one of the covariates listed on the MODEL statement.

REPEATED SUBJECT=subject-effect / <options>;

The REPEATED statement distinguishes the fitting of a generalized linear model for a single univariate responses via maximum likelihood from the fitting of a marginal model to a vector of correlated responses using the GEE approach. The REPEATED statement is used to specify the assumed structure of the within-subject association among the repeated measurements.

In particular, the REPEATED statement defines a variable that determines the clustering of observations within an individual. The latter is achieved by including a subject identifier, that distinguishes clusters of correlated responses, on the SUBJECT=*subject-effect*; this is not optional, a *subject-effect* must be included with the REPEATED statement and this variable must be listed in the CLASS statement. By including a subject identifier, pairs of observations with the same value of that variable are regarded as correlated (by virtue of arising from the same subject) while pairs of observations with distinct values are regarded as independent.

A useful option on the REPEATED statement is the WITHINSUBJECT=*within-subject effect*. With this option a variable denoting the "repeated effect" can be included and this identifies the order of the repeated measurements within subjects. In the context of longitudinal data, the "repeated effect" identifies the measurement occasions. While it is not always necessary to include this variable, failure to do so may have unforseen consequences when there are vectors of repeated measures of different length and/or when the vector of responses are not in the same order for all subjects. To avoid any potential problems, this variable should be included on the REPEATED statement, whenever possible, to ensure that the within-subject association is estimated appropriately.

While the REPEATED statement in PROC GENMOD has a similar function to the REPEATED statement in PROC MIXED, the order in which the *subject-effect* and the *within subject-effect* appear in the REPEATED statement are reversed (for reasons perhaps best known only to the developers at SAS Institute).

By default, PROC GENMOD produces a table of regression parameter estimates, standard errors, and Z statistics. The standard errors and Z statistics are based on the empirical or "sandwich" estimator of $\mathrm{Cov}(\widehat{\beta})$ described in Section 11.3. Use of the REPEATED statement with the MODELSE option produces the corresponding table based on the "model-based" estimator of $\mathrm{Cov}(\widehat{\beta})$.

Finally, two additional options are used for specifying assumptions about the structure of the working correlation matrix or the log odds ratios (for binary responses only) among the repeated measurements. The TYPE=*correlation-structure* specifies the working correlation structure. PROC GENMOD provides a number of build-in correlation structures, including unstructured (UN), m-dependent (MDEP(m), where m is the order of dependence), first-order autoregressive (AR), and exchangeable (analogous to "compound symmetry") or equicorrelated (EXCH/CS). For binary responses only, the structure of the within-subject association among the responses can be specified in terms of log odds ratios using the LOGOR=*log odds ratio structure* option. PROC GENMOD allows a very flexible regression structure for the log odds ratios. Note that either the TYPE=*correlation-structure* or the LOGOR=*log odds ratio structure* option should be specified, but not both. By default, a working independence structure is assumed.

Of note, the initial output produced by PROC GENMOD is the standard output from a generalized linear model assuming that all observations are independent. The resulting estimates of the regression coefficients are used as initial values for the generalized estimating equations algorithm. However, the reader is cautioned that this initial output should be ignored. In particular, the reported value of the log-likelihood and various likelihood-based goodness of fit statistics should not be conside. d part of the GEE output.

11.6 DISTRIBUTIONAL ASSUMPTIONS FOR MARGINAL MODELS*

In Section 11.2, a marginal model was defined in terms of a three-part formulation. This formulation highlights how generalized linear models have been extended to handle longitudinal data. In this section† we consider making additional distributional assumptions about the vector of responses, Y_i. Previously we mentioned that specification of the mean vector and the covariance (or the variance and pairwise associations) does not, in most cases, determine the joint distribution of discrete longitudinal data. That is, the three-part marginal model specification does not determine the joint distribution of Y_i. As a result, the method of maximum likelihood (ML or REML) cannot be used for estimation of the parameters in the marginal model without further distributional assumptions. This presents two alternative ways to proceed.

The first is to attempt to enrich the formulation of the marginal model so that full distributional assumptions about Y_i have been made. Then, the likelihood can be specified and the method of maximum likelihood can be used for estimation and inference. However, this poses a number of difficulties. First, unlike the multivariate

†This section provides a rationale for the use of the generalized estimating equations (GEE) approach for marginal models presented in Section 11.3. The content of this section is somewhat technical and can be omitted without loss of continuity.

normal distribution for a continuous response, the joint distribution of Y_i is not usually specified by the mean vector and covariance matrix. That is, with discrete longitudinal data there is no simple analogue of the multivariate normal distribution. Instead, the joint distribution of Y_i requires specification of the mean vector and pairwise (or two-way) associations, as well as the three-, four- and higher-way associations among the responses. As the number of responses increases, the number of association parameters proliferates rapidly. This is best exemplified in the case where Y_i is a vector of binary responses. When the number of repeated measures $n_i = 10$ the joint distribution of Y_i has 1013 (or $2^{10} - 10 - 1$) two-way, three-way, four-way and higher-way association parameters. This excessive number of within-subject association parameters will often far exceed the number of subjects enrolled in a longitudinal study. As a result, specification of the joint distribution for discrete longitudinal data is inherently difficult. In addition, even in cases where it might be possible to specify the joint distribution of Y_i, the likelihood is often intractable and maximum likelihood estimation is computationally infeasible. Furthermore, procedures for ML estimation of marginal models are not currently incorporated in commercially available general-purpose statistical software packages.

The second alternative is to avoid distributional assumptions about Y_i altogether and specify the marginal model solely in terms of assumptions about the mean response, the variances and the pairwise (or two-way) within-subject association. This corresponds to the three components in the formulation given in Section 11.2. This alternative approach has the following three advantages. First, it leads to a method for estimation and inference that does not require any distributional assumptions on Y_i. As a result, the empirical researcher does not have to be concerned that the distribution of Y_i closely approximates some multivariate distribution. Put another way, there may be a gain in robustness because distributional assumptions on Y_i are not required. Second, it circumvents the need to specify models for the three-way, four-way and higher-way associations among the responses. Modelling three-way, four-way and higher-way associations among the responses is conceptually very difficult, and ordinarily requires a relatively large sample size. Third, it leads to a method of estimation, known as generalized estimating equations (GEE). The GEE approach has become an extremely popular method for analyzing longitudinal data, and for good reasons too. It provides a flexible approach for modelling the mean and the pairwise within-subject association structure. It can handle inherently unbalanced designs and missing data with ease. Finally, the GEE approach is computationally straightforward and has been implemented in existing, widely available statistical software. The one potential drawback that must be acknowledged is that avoidance of distributional assumptions will usually result in some loss of efficiency for estimation of β relative to the optimal, but intractable, likelihood-based estimates. In addition there are some implications for the assumptions made about missing responses; the latter issue will be addressed in Chapter 14. However, given that the distinct advantages of this alternative approach far outweigh its drawbacks, this is the approach that we emphasize throughout this chapter.

11.7 FURTHER READING

Burton *et al.* (1998) provide an accessible introduction to generalized estimating equations. A more comprehensive description of generalized estimating equations can be found in Chapter 6 of the textbook by Myers *et al.* (2001).

Bibliographic Notes

The early foundations for statistical methods for the analysis of repeated categorical responses can be traced to a general approach developed by Grizzle, Starmer and Koch (1969); this approach became known as the GSK method. Koch *et al.* (1977) applied the GSK method to the analysis of repeated measurements. However, the application of the GSK method was limited to categorical covariates. The GEE approach overcame many of the limitations of the GSK method.

The theoretical foundation for the generalized estimating equations approach can be found in Godambe (1960) and Durbin (1960); also see Huber (1967, 1981) and White (1982). Liang and Zeger (1986) and Zeger and Liang (1986), in companion papers, proposed a class of generalized estimating equations for repeated measures and longitudinal data; see Liang and Zeger (1995) for a historical perspective on generalized estimating equations. Connections between the GEE approach and likelihood-based methods were made by Zhao *et al.* (1992), Fitzmaurice and Laird (1993), and Fitzmaurice *et al.* (1993).

The "sandwich" variance estimator was derived in Huber (1967), White (1982), Gourieroux *et al.* (1984), and Royall (1986); see Hinkley and Wang (1991) and Kauermann and Carroll (2001) for a discussion of properties of the "sandwich" variance estimator. For finite samples, simulation studies have shown that Wald tests using the sandwich estimator tend to be liberal, that is, have nominal p-values that are too small (see Lin and Wei, 1989; Emrich and Piedmonte, 1992; Gunsolley *et al.*, 1995; Fay *et al.*, 1998; Mancl and DeRouen, 2001; Fay and Graubard, 2001).

Problems

11.1 In a clinical trial of patients with respiratory illness, 111 patients from two different clinics were randomized to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined. These data are from Koch *et al.* (1990), and are reported in Davis (1991) and Stokes *et al.* (1995). The main objective of the analyses is to understand the joint effects of treatment and time on the probability that respiratory status is classified as good. It is also of interest to determine whether the effect of treatment is the same for patients from the two clinics.

The raw data are stored in an external file: `respir.dat`

Each row of the data set contains the following eight variables:

ID Clinic Treatment Y_0 Y_1 Y_2 Y_3 Y_4

Note: The respiratory status response variable Y_j is coded 1 = good, and 0 = poor, at the j^{th} occasion. The categorical (character) variable Treatment is coded A = Active drug, P = Placebo. The categorical variable Clinic is coded 1 = clinic 1, 2 = clinic 2.

11.1.1 Ignoring the clinic variable, consider a model for the log odds that respiratory status is classified as good, including the main effects of treatment and time (where time is regarded as a categorical variable with 5 levels), and their interaction.
Use generalized estimating equations (GEE), assuming separate pairwise log odds ratios (or separate pairwise correlations, if available software does not permit the within-subject association to be parameterized in terms of log odds ratios) among the five binary responses. Construct a test of the null hypothesis of no effect of treatment on *changes* in the log odds that respiratory status is classified as good based on the empirical standard errors.

11.1.2 What conclusions do you draw about the effect of treatment on changes in the log odds? Provide results that support your conclusions.

11.1.3 Patients in this trial were drawn from two separate clinics. Repeat the analysis for Problem 11.1.1, allowing the effects of treatment (and, possibly, time) to depend upon clinic.

(a) Is the effect of treatment the same in the two clinics? Present results to support your conclusion.

(b) Find a parsimonious model that describes the effects of clinic, treatment, and time, on the log odds that respiratory status is classified as good. For the model selected, give a clear interpretation of the estimated regression parameters for the final model selected.

11.1.4 For the final model selected in Problem 11.1.3, construct a table of the estimated probabilities that respiratory status is classified as good as a function of both time and treatment group (and, possibly, clinic). What do you conclude from this table?

11.2 In a clinical trial of patients suffering from epileptic seizures (Thall and Vail, 1990), patients were randomized to receive either a placebo or the drug progabide, in addition to standard therapy. A baseline count of the number of epileptic seizures in an 8-week period prior to randomization was obtained. In addition, counts of the number of epileptic seizures in each of four successive 2-week (post-baseline) treatment periods were obtained. The goal of the analysis is to make a comparison between the two treatment groups in terms of changes in the rates of epileptic seizures throughout the duration of the study.

The raw data are stored in an external file: `epilepsy.dat`

Each row of the data set contains the following eight variables:

ID Y_1 Y_2 Y_3 Y_4 Treatment Y_0 Age

Note: The response variable Y_0 is a baseline count of the number of epileptic seizures in an 8–week interval. The response variables Y_j are counts of the number of epileptic seizures in the four successive 2–week (post-baseline) treatment intervals, for $j = 1, ..., 4$. The categorical variable Treatment is coded 1 = Progabide, 0 = Placebo. The variable Age is the age of each patient (in years) at baseline.

11.2.1 Consider a model for the log seizure rate that includes the main effects of treatment and time (where time is regarded as a categorical variable with 5 levels), and their interaction.
Use generalized estimating equations (GEE), assuming separate pairwise correlations among the five responses. Construct a test of the null hypothesis of no effect of treatment on changes in the log seizure rate based on the empirical standard errors.

11.2.2 What conclusions do you draw about the effect of treatment on *changes* in the log seizure rate?

11.2.3 Construct a new variable, Ptime, where:

Ptime = 0 if baseline, and Ptime = 1 if post-baseline (any of the four successive 2-week intervals).

Repeat the analysis for Problem 11.2.1 using Ptime (instead of time as a categorical variable with 5 levels). Construct a test of the null hypothesis of no effect of treatment on changes in the log seizure rate based on the empirical standard errors.

11.2.4 From the results of the analysis for Problem 11.2.3, what conclusions do you draw about the effect of treatment on changes in the log seizure rate?

11.2.5 Patient 49 (ID = 49) is a potential outlier. This patient reported 151 seizures during the 8-week baseline interval and 302 (102+65+72+63) seizures during the four successive 2-week intervals. Repeat all of the analyses in Problems 11.2.1 to 11.2.4, excluding all of the repeated count data from patient 49. When the data from patient 49 are excluded, what conclusions do you draw about the effect of treatment on changes in the log seizure rate?

12

Generalized Linear Mixed Effects Models

12.1 INTRODUCTION

In the previous chapter we described marginal models for longitudinal data. Marginal models can be considered an extension of generalized linear models that *directly* incorporate the within-subject association among the repeated measurements. To estimate the regression parameters in a marginal model, we made some assumptions about the marginal distribution of the response at each occasion (e.g., assumptions about the mean, and its dependence on the covariates, and the variance of each Y_{ij}). We also made assumptions about the pairwise within-subject associations among the responses, thereby linking repeated observations of the same subject. A notable feature of marginal models is that the mean response and the covariance are modelled separately. This separation ensures that the interpretation of the regression coefficients in a marginal model does not rely on the assumed model for the covariance among the responses. In specifying the marginal means, variances, and pairwise associations, we did not fully specify the joint distribution of the vector of responses. However, these assumptions were sufficient for estimating and constructing confidence intervals for the regression parameters using the GEE approach.

An alternative approach for accounting for the within-subject association is via the introduction of random effects. In Chapter 8 we saw how the incorporation of random effects at the individual level induces correlation among the repeated measures at the population level. In this chapter we describe how generalized linear models can be extended to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual to another. These models are known as *generalized linear mixed effects models* and they extend in a natural way the concep-

tual approach represented by the linear mixed effects models discussed in Chapter 8. However, we must caution the reader at the outset that the introduction of random effects in generalized linear models produces a greater degree of conceptual and analytic complexity relative to marginal models or to random effects in linear models. Although both classes of models account for the within-subject association among repeated measurements, the manner in which they do so has important implications for the interpretation of the regression parameters. In Chapter 13 we highlight the major distinctions between the regression coefficients in marginal and generalized linear mixed models and consider various aspects of interpretation of the regression effects in these two classes of models for longitudinal data.

12.2 INCORPORATING RANDOM EFFECTS IN GENERALIZED LINEAR MODELS

The basic premise underlying the generalized linear mixed effects model for longitudinal data is the assumption of heterogeneity across individuals in the study population in a subset of the regression coefficients from a generalized linear model. That is, a subset of the regression coefficients (e.g., intercepts in a logistic regression model) are assumed to vary across individuals according to some distribution. The random effects can be thought of as reflecting natural heterogeneity due to many unmeasured factors. For mathematical and computational convenience, we ordinarily assume that the random effects have a multivariate normal distribution. Then, conditional on the random effects, we assume that the responses for any particular individual are independent observations from a distribution belonging to the exponential family (e.g., the Bernoulli distribution if Y_{ij} is binary or the Poisson distribution if Y_{ij} is a count). The latter assumption is completely analogous to the "conditional independence" assumption ($R_i = \sigma^2 I_{n_i}$) made in the linear mixed effects model described in Chapter 8. In fact, the linear mixed effects model is simply a special case of the generalized linear mixed effects model where the conditional mean, given the random effects, is related to the covariates via an identity link function and the conditional distribution of the responses is assumed to be normal. Because the linear mixed effects model is a special case, it is useful for pedagogical purposes to consider its formulation within the framework and terminology of generalized linear models. By doing so, the extensions to other types of response variables will become more apparent.

Linear Mixed Effects Models

In this section we consider the linear mixed effects model as a generalized linear model, albeit one with both fixed and random effects. Recall that the standard generalized linear model formulation requires a three-part specification: i) a distributional assumption; ii) a systematic component; and iii) a link function. In the linear mixed effects model it is assumed that the *conditional* distribution of each Y_{ij}, given a vector of random effects b_i, has a normal distribution, with $\text{Var}(Y_{ij}|b_i) = \sigma^2$ (i.e., $\phi = \sigma^2$ and $v(\mu_{ij}) = 1$). In addition, given the random effects b_i, it is assumed that the Y_{ij} are independent of one another (i.e., given b_i, Y_{ij} and Y_{ik} are assumed to be independent of each other). This completes the distributional assumptions on the Y_{ij}. Next, the conditional mean of Y_{ij} is assumed to depend upon both fixed and random effects via the following extended definition of the linear predictor:

$$\eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i,$$

where the definition of the linear predictor has been extended to incorporate both population (or fixed) and subject-specific (or random) effects. In addition, the random effects, b_i, are assumed to have a multivariate normal distribution. This specifies the systematic component. Finally, an identity link function relates the conditional mean of Y_{ij} to the linear predictor,

$$E(Y_{ij}|b_i) = \eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i.$$

That is, for the identity link, $\eta_{ij} = g\{E(Y_{ij}|b_i)\} = E(Y_{ij}|b_i)$ and hence,

$$E(Y_{ij}|b_i) = X'_{ij}\beta + Z'_{ij}b_i.$$

In the linear mixed effects model, the response for the i^{th} subject at the j^{th} occasion is assumed to differ from the population mean, $X'_{ij}\beta$, by a subject-specific effect, $Z'_{ij}b_i$, and a within-subject measurement error, e_{ij}. The within-subject measurement errors are independently normally distributed, with zero mean and variance σ^2.

When collected in a vector, $e_i \sim N(0, R_i)$, where $R_i = \sigma^2 I_{n_i}$ (the "conditional independence" assumption). Recall that $R_i = \text{Cov}(e_i)$ describes the covariance among observations when we focus on the mean response profile of any *individual*, that is, it is the covariance of the i^{th} subject's deviations from his/her mean response profile, $X_i\beta + Z_ib_i$. Also, the b_i are assumed to vary independently from one individual to another, with $b_i \sim N(0, G)$.

When expressed in vector and matrix notation, the linear mixed effects model is

$$Y_i = X_i\beta + Z_ib_i + e_i,$$

where the vector of regression parameters β (the fixed effects) is assumed to be the same for all individuals and the vector of subject-specific regression coefficients b_i (the random effects) describes how the i^{th} individual's mean response profile deviates from the overall population trend. A distinctive feature of the linear mixed effects model is that it yields simple expressions for both the conditional mean response (for any individual),

$$E(Y_i|b_i) = X_i\beta + Z_ib_i,$$

and the marginal mean response (for the population), averaged over all individuals,

$$E(Y_i) = X_i\beta.$$

Thus the regression coefficients β have population-averaged interpretations in terms of how the mean response changes over time and how these changes relate to covariates.

Finally, the conditional covariance of the responses, given the random effects b_i, is assumed to be a diagonal matrix with

$$\text{Cov}\,(Y_i|b_i) = \text{Cov}\,(e_i) = R_i = \sigma^2 I_{n_i}.$$

On the other hand, the marginal covariance of the responses (the covariance among deviations of the i^{th} individual's responses from the population mean, $X_i\beta$),

$$\begin{aligned}\text{Cov}\,(Y_i) &= \text{Cov}\,(Z_i b_i) + \text{Cov}\,(e_i) \\ &= Z_i G Z_i' + R_i \\ &= Z_i G Z_i' + \sigma^2 I_{n_i},\end{aligned}$$

is certainly not diagonal. Thus, the introduction of random effects, b_i, in the linear mixed effects model induces correlation (marginally) among the Y_i. This consequence of introducing random effects extends more generally and, in a very natural way, to any generalized linear model with random effects. That is, the correlations among the repeated observations on an individual can be thought of as arising from sharing a set of underlying random effects.

Generalized Linear Mixed Effects Models

Next we consider how the ideas underlying the linear mixed effects model can be extended to generalized linear models. Once again, we can formulate the generalized linear mixed model using a three-part specification:

1. We assume that the conditional distribution of each Y_{ij}, given a $q \times 1$ vector of random effects b_i, belongs to the exponential family of distributions and that $\text{Var}(Y_{ij}|b_i) = v\{E(Y_{ij}|b_i)\}\,\phi$, where $v(\cdot)$ is a known variance function, a function of the conditional mean, $E(Y_{ij}|b_i)$. In addition, given the random effects b_i, it is assumed that the Y_{ij} are independent of one another; this is the so-called "conditional independence" assumption.

2. The conditional mean of Y_{ij} is assumed to depend upon fixed and random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}'\beta + Z_{ij}'b_i,$$

with

$$g\{E(Y_{ij}|b_i)\} = \eta_{ij} = X_{ij}'\beta + Z_{ij}'b_i$$

for some known link function, $g(\cdot)$.

3. Finally, the random effects are assumed to have some probability distribution. In principle, any multivariate distribution can be assumed for the b_i; in practice, it is common to assume that the b_i have a multivariate normal distribution, with

zero mean and $q \times q$ covariance matrix, G. In addition, the random effects, b_i, are assumed to be independent of the covariates, X_i.

These three components completely specify a broad class of generalized linear mixed models. Note that in Chapter 11 we extended generalized linear models by making assumptions about the mean and covariance of Y_i; in particular, we did not make full distributional assumptions about Y_i. In contrast, the three components of a generalized linear mixed model given above completely specify the joint distribution of Y_i. To fix ideas, consider the following three illustrative examples of generalized linear mixed effects models using this three component specification.

Example 1: Generalized Linear Mixed Model for a Continuous Response

Suppose that Y_{ij} is a continuous response and it is of interest to relate changes in the mean response over time to the covariates. An example of a linear mixed effects model for Y_{ij} is given by the following three-part specification:

1. Conditional on a vector of random effects b_i, the Y_{ij} are independent and assumed to have a normal distribution, with $\text{Var}(Y_{ij}|b_i) = \sigma^2$ (i.e., $\phi = \sigma^2$ and the variance does not depend on the conditional mean).

2. The conditional mean of Y_{ij} depends upon fixed and random effects via the following linear predictor,

$$\eta_{ij} = X_{ij}'\beta + Z_{ij}'b_i,$$

where $X_{ij}' = Z_{ij}' = (1, t_{ij})$, with

$$\begin{aligned}E(Y_{ij}|b_i) = \eta_{ij} &= X_{ij}'\beta + Z_{ij}'b_i \\ &= \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i}t_{ij} \\ &= (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij}.\end{aligned}$$

That is, the conditional mean of Y_{ij} is related to the linear predictor by an identity link function, $\eta_{ij} = g\{E(Y_{ij}|b_i)\} = E(Y_{ij}|b_i)$.

3. The random effects are assumed to have a bivariate normal distribution, with zero mean and 2×2 covariance matrix G.

This illustration of a generalized linear mixed effects model is simply a random intercepts and slopes model and is a special case of the linear mixed effects models considered in Chapter 8. However, when it is viewed as a generalized linear mixed effects model, a much broader class of models for continuous responses can, in principle, be entertained. For example, the mean can be related to the linear predictor by a link function other than the identity. Thus, if the effects of covariates are thought to act multiplicatively on the mean response, a log link function might be more appropriate. Alternatively, the variance can be allowed to depend upon any known function of the mean response.

Example 2: Generalized Linear Mixed Model for Counts

Next, suppose that Y_{ij} is a count. An example of a generalized linear mixed model for Y_{ij} is given by the following three-part specification:

1. Conditional on a vector of random effects b_i, the Y_{ij} are independent and have a Poisson distribution, with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends upon fixed and random effects via the following linear predictor:

$$\eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i,$$

   where $X'_{ij} = Z'_{ij} = (1, t_{ij})$, with

$$\log\{E(Y_{ij} = 1|b_i)\} = \eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i.$$

   That is, the conditional mean of Y_{ij} is related to the linear predictor by a log link function; this is an example of a log-linear mixed effects model.

3. The random effects are assumed to have a bivariate normal distribution, with zero mean and 2×2 covariance matrix G.

In this example, the model is a log-linear regression model with randomly varying intercepts and slopes. This model posits that there is natural heterogeneity among individuals in both their baseline level and changes in the expected counts over time.

Example 3: Generalized Linear Mixed Model for a Binary Response

Finally, suppose that Y_{ij} is a binary response, taking values of 0 or 1. A logistic mixed effects model for Y_{ij} is given by the following three-part specification:

1. Conditional on a single random effect b_i, the Y_{ij} are independent and have a Bernoulli distribution, with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)\{1 - E(Y_{ij}|b_i)\}$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends upon fixed and random effects via the following linear predictor:

$$\eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i = X'_{ij}\beta + b_i,$$

   where $Z_{ij} = 1$ for all $i = 1, ..., N$, and $j = 1, ..., n_i$, with

$$\log\left\{\frac{\Pr(Y_{ij} = 1|b_i)}{\Pr(Y_{ij} = 0|b_i)}\right\} = \eta_{ij} = X'_{ij}\beta + b_i.$$

   That is, the conditional mean of Y_{ij} is related to the linear predictor by a logit link function.

3. The single random effect b_i is assumed to have a univariate normal distribution, with zero mean and variance g_{11}.

In this example, the model is a simple logistic regression model with randomly varying intercepts. This model can be considered a discrete data analogue of the "compound symmetry" model discussed in Chapters 7 and 8. The model posits that there is natural heterogeneity in individuals' propensity to respond positively that persists throughout all binary responses obtained on any individual.

Finally, although in all three examples we have chosen canonical link functions to relate the conditional mean Y_{ij} to η_{ij}, in principle, any suitable link function can be chosen. The three examples of generalized linear mixed effects models considered so far are purely illustrative. They demonstrate how the choices of the three components might differ according to the type of response variable. However, these three examples should not be considered prescriptions for constructing generalized linear mixed effects models.

12.3 INTERPRETATION OF REGRESSION PARAMETERS

Although the introduction of random effects can simply be thought of as a means of accounting for the correlation among longitudinal responses, it has important implications for the interpretation of the regression coefficients in generalized linear mixed models. The regression parameters, β, have somewhat different interpretations than the regression parameters in the marginal models considered in Chapter 11. In generalized linear mixed models the regression coefficients have subject-specific interpretations. That is, they represent the influence of covariates on a *specific* subject's mean response. In particular, the regression coefficients are interpreted in terms of the effects of covariates on changes in an individual's transformed mean response, while holding the remaining covariates constant. This interpretation for β can be better appreciated by considering the following simple example of a logistic regression model with randomly varying intercepts:

$$\log\left\{\frac{\Pr(Y_{ij} = 1|b_i)}{\Pr(Y_{ij} = 0|b_i)}\right\} = X'_{ij}\beta + b_i,$$

where b_i is assumed to have a univariate normal distribution, with zero mean and variance g_{11}. The interpretation of a component of β, say β_k, is in terms of changes in any given *individual's* log odds of response for a unit change in the corresponding covariate, say X_{ijk}. That is, when X_{ijk} takes on some value x, the log odds of a positive response is

$$\log\left\{\frac{\Pr(Y_{ij} = 1|b_i, X_{ij1}, ..., X_{ijk} = x, ..., X_{ijp})}{\Pr(Y_{ij} = 0|b_i, X_{ij1}, ..., X_{ijk} = x, ..., X_{ijp})}\right\}$$

$$= b_i + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

Similarly, when X_{ijk} now takes on some value $x+1$, but all other covariate values are held fixed, the log odds of a positive response is,

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_i,X_{ij1},...,X_{ijk}=x+1,...,X_{ijp})}{\Pr(Y_{ij}=0|b_i,X_{ij1},...,X_{ijk}=x+1,...,X_{ijp})}\right\}$$

$$= b_i+\beta_1X_{ij1}+\cdots+\beta_k(x+1)+\cdots+\beta_pX_{ijp}.$$

Thus, for any individual, the log odds of a positive response for a unit increase in X_{ijk} is simply β_k (obtained by subtracting the former log odds from the latter). That is, β_k measures the change in the log odds of response per unit increase in X_{ijk}, for any given individual having some unobservable underlying propensity to respond positively, b_i. Also, note that this subject-specific interpretation of β_k is far more natural for a covariate that varies within an individual (i.e., a within-subject or time-varying covariate). In that case, β_k has interpretation as the change in an individual's log odds of response for a unit increase in X_{ijk}, while holding that individual's covariates fixed. Because the components of the fixed effects, β, have interpretations that depend upon holding b_i, the i^{th} individual's random effects, fixed, they are often referred to as *subject-specific* regression coefficients. As a result, generalized linear mixed models are most useful when the main scientific objective is to make inferences about individuals rather than the population averages; the population averages are the targets of inference in marginal models.

When there are between-subject (or time-invariant) covariates in the model, the interpretation of the corresponding components of β is somewhat less transparent and potentially misleading. If X_{ijk} is a between-subject covariate (e.g., gender, treatment or exposure group) it is misleading to give it a subject-specific interpretation in terms of the change in the log odds of response for a unit increase in X_{ijk} since there are simply no data that provide any information about such an effect. In a sense, this interpretation of β_k would be a complete extrapolation beyond the observed data. Problems of interpretation with a between-subject covariate arise because a change in the value of the covariate requires also a change in the index i of X_{ijk} to, say, $X_{i'jk}$ (for $i \neq i'$). However, β_k then becomes confounded with $b_i - b_{i'}$, the difference between the unobserved random effect for the two individuals indexed by i and i', respectively. To circumvent this problem we must assume that $b_i = b_{i'}$. That is, β_k must be given an interpretation in terms of a contrast of the log odds of response for two different individuals who happen to have the same value for the unobserved random effects (i.e., $b_i = b_{i'}$), but who differ by one unit in the covariate X_{ijk} (e.g., one individual is exposed, the other is unexposed, or one individual is randomized to treatment, the other to placebo). Because the random effects are latent, unobserved variables, this effect of the covariate is not directly observable from the data at hand. As a result, it is somewhat unclear where the information about β_k is obtained when X_{ijk} is a between-subject covariate. Because the estimate of β_k is a model-based extrapolation, it may be more sensitive to assumptions concerning the random effects distribution that are difficult to check from the data.

The distinction between the regression coefficients in generalized linear mixed models and marginal models is best understood in terms of the targets of inference.

In generalized linear mixed models the target of inference is the individual, since the regression coefficients have interpretation in terms of contrasts of the transformed conditional means,

$$E(Y_{ij}|X_{ij},b_i).$$

By conditioning on the unobserved random effects, b_i, the target of inference has shifted from the population to the individual. In contrast, in marginal models the target of inference is the population, since the regression coefficients in marginal models have interpretation in terms of contrasts of the transformed population means,

$$E(Y_{ij}|X_{ij})$$

and describe how the average response varies across different subsets of the study population defined by the covariates (e.g., gender, exposure groups, treatment groups). Note that the population means,

$$E(Y_{ij}|X_{ij})$$

in marginal models are averaged over the natural individual-to-individual heterogeneity in the study population (as well as over the measurement or sampling variability in the response).

For the special case where an identity link function has been adopted (i.e., for the special case of linear mixed effects models), the regression coefficients in the model for the conditional means,

$$E(Y_{ij}|X_{ij},b_i)=X_{ij}'\beta+Z_{ij}'b_i,$$

also happen to have interpretation in terms of the population means, since

$$\begin{aligned}E(Y_{ij}|X_{ij}) &= E\{E(Y_{ij}|X_{ij},b_i)\}\\ &= E\left(X_{ij}'\beta+Z_{ij}'b_i\right)\\ &= X_{ij}'\beta+Z_{ij}'E(b_i)\\ &= X_{ij}'\beta,\end{aligned}$$

when averaged over all individuals in the study population. That is, averaged over the distribution of the random effects, the population means also follow a linear model with regression coefficients β. However, in general, for the non-linear link functions usually adopted for discrete data, this relationship no longer holds. That is, if

$$g\{E(Y_{ij}|X_{ij},b_i)\}=X_{ij}'\beta+Z_{ij}'b_i,$$

where $g(\cdot)$ is a non-linear link function (e.g., logit$(\cdot)$ or log$(\cdot)$), then

$$g\{E(Y_{ij}|X_{ij})\}\neq X_{ij}'\beta,$$

for all β, when averaged over the distribution of the random effects. Thus, for non-linear (or non-identity) link functions, the regression coefficients in generalized linear

Table 12.1 Hypothetical data on the true propensity for disease, at baseline and post-baseline, for three individuals with heterogeneous propensities for disease.

Individual	Baseline	Post-Baseline	Difference	Log(Odds Ratio)
A	0.80	0.67	−0.13	−0.68
B	0.50	0.33	−0.17	−0.71
C	0.20	0.11	−0.09	−0.70
Population Average	0.50	0.37	−0.13	

mixed effects and marginal models have quite distinct interpretations and these two classes of regression models have different targets of inference. In short, these two classes of models address different scientific questions. Marginal models address scientific questions that are concerned with changes in the (transformed) mean response over time in the study population, and the impact of covariates on these changes. In contrast, generalized linear mixed effects models address scientific questions that are concerned with changes in the mean response for any individual, and the impact of covariates on these changes.

The following simple illustration helps to highlight the main distinction between the regression coefficients in marginal and generalized linear mixed models. Consider the hypothetical data presented in Table 12.1. It displays the (usually unobserved) true propensity for disease, $\Pr(Y_{ij} = 1|b_i)$, for three individuals measured at baseline and following treatment with a new drug intended to reduce the risk of disease. The three individuals are discernibly different in terms of their underlying propensity for disease at baseline. This heterogeneity can be expressed in terms of random effects, b_i. In a sense, individuals A, B, and C have "high", "medium" and "low" underlying risk for disease. Also, let us assume that the entire population is comprised of an equal number of individuals that fall into these three distinct risk groups. Based on this assumption, the final row of Table 12.1 contains the population averages (obtained as equally weighted means).

If we considered a linear model for the probability of disease, the risk difference, or difference between the probabilities of disease at baseline and post-baseline, provides a measure of the effectiveness of the new drug. These differences (post-baseline − baseline) are displayed in the fourth column of Table 12.1 and vary from −0.09 to −0.17. These can be thought of as subject-specific effects of the drug. We can then consider two possible ways to produce a single number summary of the effectiveness of the drug. The first summary can be obtained by taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.13 - 0.17 - 0.09}{3} = -0.13.$$

Alternatively, the average propensity for disease at baseline (0.5) can be compared to the average propensity for disease post-baseline (0.37). The latter can be thought of as a contrast of population averages and this comparison also yields

$$(0.37 - 0.50) = -0.13.$$

That is, the difference (post-baseline − baseline) between the population averages is identical to the population average of the individual-specific differences. As such, the "difference of the averages" is equal to the "average of the differences". This simple numerical illustration confirms the remark that was made earlier about how the fixed effects regression coefficients in the linear mixed effects model (with identity link function) also happen to have interpretation in terms of population averages.

Next, let us consider a non-linear function of the propensity for disease (this corresponds to adopting a non-linear link function for the probability of disease). The log odds ratio provides a natural measure of the effectiveness of the drug in reducing the risk of disease from baseline. The log odds ratios (comparing the odds of disease post-baseline to the odds of disease at baseline) for individuals A, B, and C are displayed in the fifth column of Table 12.1. For example, the log odds ratio for individual A is

$$\log\left\{\frac{0.67/(1-0.67)}{0.8/(1-0.8)}\right\} = -0.68.$$

The log odds ratios are all very similar in magnitude, ranging from −0.68 to −0.71. Once again, these can be thought of as subject-specific effects of the drug. We can then consider two possible ways to produce a single number summary of the effectiveness of the drug. The first can be obtained by taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.68 - 0.71 - 0.70}{3} = -0.697.$$

This indicates that the effect of the drug on any individual is to approximately halve the odds of disease (since $e^{-0.697} \approx 0.5$). Alternatively, the effectiveness of the drug can be assessed by comparing the log odds of disease in the population at baseline, $\log(0.5/0.5) = 0$, with the log odds of disease in the population post-baseline, $\log(0.37/0.63) = -0.532$. The latter can be thought of as a contrast of population log odds and this comparison yields a measure of effect, −0.532, which is approximately 25% smaller than the summary of the subject-specific effect, −0.697. That is, the comparison of population logs odds results in a discernibly different measure of the effectiveness of the drug than was found in the comparison of subject-specific effects. With a non-linear function of the propensity for disease, a "non-linear contrast of the averages" is not equal to the "average of the non-linear contrasts". This highlights the main differences between these two approaches when a non-linear link function is adopted.

For this simple numerical illustration the reader may be curious about which of the two summary statistics, -0.697 or -0.532, provides the most *realistic* estimate of the effectiveness of the drug. The answer is that they both do, although they address somewhat different scientific questions. The estimate of -0.697 provides a measure of the expected change in the odds of disease for any individuals treated with the drug. That is, there is an approximately 50% reduction in the odds of disease (since $1 - e^{-0.697} \approx 0.5$) for any individuals treated with the drug. This is the estimate that will be of most interest to an individual and his/her physician in the physician–patient context. On the other hand, the estimate of -0.532 provides a measure of the expected change in prevalence of disease in the study population if everyone were to be treated with the drug. That is, there would be an expected reduction in the odds of disease in the population of approximately 40% (since $1 - e^{-0.532} \approx 0.4$). This is the estimate that will be of most interest to public health researchers who are considering th potential benefits of the drug for the study populati n as a whole.

To provide further intuition for why the regression coefficients in generalized linear mixed models and marginal models differ, consider the following example of a logistic regression model, with normally distributed random intercepts:

$$\text{logit}\{E(Y_{ij}|b_i)\} = \beta_1^* + \beta_2^* t_{ij} + b_i,$$

where $t_{ij} = 0$ at baseline and $t_{ij} = 1$ post-baseline. Similar to the illustration in Table 12.1, we assume that individuals are measured at baseline and following treatment with a new drug intended to reduce the risk of disease. Individuals in the population differ in terms of their underlying propensity for disease at baseline; this heterogeneity is expressed in terms of the random effect, b_i. For a "typical" individual from the population (where a "typical" individual is one with unobserved random effect $b_i = 0$, the mean and median of the distribution of b_i), the log odds of disease at baseline is β_1^*; the log odds of disease following treatment with the new drug is $\beta_1^* + \beta_2^*$.

The log odds of disease at baseline and post-baseline are displayed in Figure 12.1, for the case where $\beta_1^* = 1.5$, $\beta_2^* = -3.0$, and $\text{Var}(b_i) = 1.0$. At baseline, the log odds has a normal distribution with mean and median of 1.5. (See the shaded density for the log odds in Figure 12.1.) From Figure 12.1 it is clear that there is heterogeneity in risk of disease, with approximately 95% of individuals having a baseline log odds of disease that varies from -0.46 to 3.46 (or $1.5 \pm 1.96\sqrt{1.0}$). When the risk of disease is translated from the log odds scale to the probability scale, the baseline probability of disease for a typical individual from the population is approximately 0.82. Furthermore, approximately 95% of individuals have a baseline probability of disease that varies from 0.39 to 0.97.

From Figure 12.1 it is transparent that the symmetric, normal distribution for the baseline log odds does not translate into a corresponding symmetric, normal distribution for the probability of disease. Instead, the subject-specific probabilities of disease have a negatively skewed distribution with median, but not mean, of 0.82. (See solid line in Figure 12.1.) Because of the skewness, the mean of the distribution of subject-specific baseline probabilities is pulled toward the tail and is equal to 0.7785. (See dashed line in Figure 12.1.) Thus, the probability of disease for a "typical"

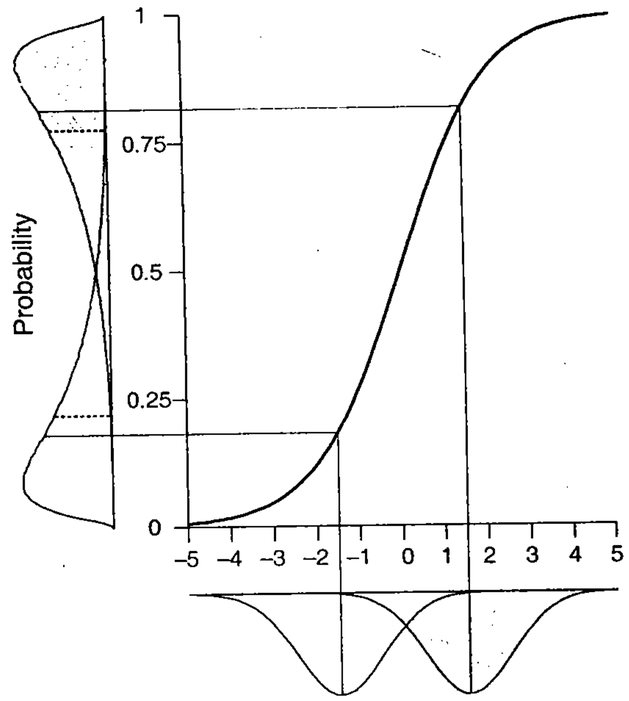

Fig. 12.1 Subject-specific probability of disease as a function of subject-specific log odds of disease at baseline (shaded densities) and post-baseline (unshaded densities). Solid lines represent medians of the distributions; dashed lines represent means of the distributions.

individual from the population (0.82) is not the same as the prevalence of disease in the same population (0.78), due to the non-linearity of the relationship between subject-specific probabilities and log odds.

Similarly, the log odds of disease post-baseline has a normal distribution with mean and median of -1.5. (See the unshaded density for the log odds in Figure 12.1); approximately 95% of individuals have a post-baseline log odds of disease that varies from -3.46 to 0.46 (or $-1.5 \pm 1.96\sqrt{1.0}$.) This shift in the log odds corresponds to a 20-fold decrease (since $1 - e^{-3.0} \approx 0.95$) in the subject-specific odds of disease. When the risk of disease is translated from the log odds scale to the probability scale, the post-baseline probability of disease for a typical individual from the population is approximately 0.18. Furthermore, approximately 95% of individuals have a post-baseline probability of disease that varies from 0.03 to 0.61. From Figure 12.1 it is apparent that the subject-specific post-baseline probabilities of disease have a positively skewed distribution with median, but not mean, of 0.18. (See solid line

in Figure 12.1.) Because of the skewness, the mean is pulled toward the tail and is equal to 0.2215. (See dashed line in Figure 12.1.)

Figure 12.1 highlights how the effect of treatment on the log odds of disease for a typical individual from the population, $\beta_2^* = -3.0$, is not the same as the contrast of population log odds. The latter is what is estimated in a marginal model, say

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2 t_{ij},$$

and can be obtained by comparing the log odds of disease in the population at baseline, $\log(0.7785/0.2215) = 1.257$, with the log odds of disease in the population post-baseline, $\log(0.2215/0.7785) = -1.257$. This yields a population-averaged measure of effect, $\beta_2 = -2.514$, which is approximately 15% smaller than β_2^*, the subject-specific effect of treatment.

12.4 ESTIMATION AND INFERENCE

Unlike marginal models, where specification of the marginal means, variances and pairwise associations does not fully specify the joint distribution of the vector of responses, with generalized linear mixed effects models the joint distributions of both the vector of responses and the vector of random effects are fully specified. As a result, we can base estimation and inference on the likelihood function. In this section we briefly describe maximum likelihood estimation of the fixed effects, β, and the random effects covariance parameters, G. We also discuss prediction of the random effects. Although ML estimation is far less straightforward for generalized linear mixed effects models than it is for the linear mixed effects models considered in Chapter 8, a variety of numerical methods for maximizing the likelihood have recently been implemented in commercial software packages. In this section we discuss the use of quadrature methods; quadrature methods are simply numerical methods that can be made highly accurate, albeit with substantial computational overhead.

Given the three-part specification of a generalized linear mixed effects model, the joint probability for Y_i and b_i can be expressed as:

$$f(Y_i, b_i) = f(Y_i|b_i)f(b_i),$$

where

$$f(Y_i|b_i) = f(Y_{i1}|b_i)f(Y_{i2}|b_i)\cdots f(Y_{in_i}|b_i),$$

under the "conditional independence" assumption. Furthermore, $f(Y_{ij}|b_i)$ is assumed to have an exponential family distribution, whereas $f(b_i)$ is assumed to have a multivariate normal distribution, with zero mean and covariance matrix G. Since the random effects b_i are unobserved, inference about β and G is based on the so-called marginal or integrated likelihood function:

$$L(\beta, \phi, G) = \prod_{i=1}^{N} \int f(Y_i|b_i)f(b_i)db_i,$$

obtained by integrating out or averaging over the distribution of the unobserved random effects, b_i. An integral appears in the marginal likelihood and this integral denotes the averaging over the distribution of b_i. Since the marginal likelihood has averaged over the b_i, the resulting marginal likelihood function depend only on β, ϕ, and G. That is, the marginal likelihood depends on the covariance of b_i, but not on the unobserved b_i.

The ML estimates of β, ϕ, and G are simply those values of β, ϕ, and G that maximize this likelihood function. However, unlike the case of the linear mixed effects model, there are no simple, closed-form solutions. Instead, numerical integration techniques are required for maximizing the likelihood function. Numerical integration techniques, known as Gaussian quadrature, simply approximate the integral appearing in the marginal likelihood function as a weighted sum,

$$L(\beta, \phi, G) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} f(Y_i|b_i = v_k)w_k,$$

where the known quadrature points (the weights, w_k, and the evaluation points, v_k) are chosen to provide an accurate numerical approximation. The number of quadrature points determines the degree of accuracy of the approximation involved in replacing the integral by a weighted sum. The number of quadrature points, K, can be increased or decreased as desired. The more quadrature points used, the more accurate the approximation will be. However, the computational burden also increases with the number of quadrature points, and grows exponentially with the number of random effects. As a result, there is a trade off that must be carefully balanced between the computational burden of quadrature methods and the desired accuracy of the results. In general, computational time is negligible when compared to the time expended in collecting longitudinal data. As a result, we recommend increasing the number of quadrature points until there is evidence that all parameter estimates and standard errors are stable.

Given ML estimates of β, ϕ, and G, the random effects b_i for any particular subject can be predicted as follows:

$$\widehat{b}_i = E(b_i|Y_i; \widehat{\beta}, \widehat{\phi}, \widehat{G}).$$

That is, the predicted random effects for the i^{th} subject are simply "estimated" as the conditional mean of b_i given Y_i (and $\widehat{\beta}, \widehat{\phi}, \widehat{G}$); this coincides with the empirical Bayes or BLUP used for prediction of b_i in Chapter 8. Note that $E(b_i|Y_i; \widehat{\beta}, \widehat{\phi}, \widehat{G})$, being a conditional mean, also requires integrating (or averaging) over the distribution of the unobserved random effects, b_i. As a result, simple analytic solutions for $\widehat{b}_i$ are rarely available and numerical integration techniques must also be used here.

Finally, in our discussion of generalized linear mixed models we have assumed the distribution of the random effects is multivariate normal. Distributional assumptions about the random effects are difficult to assess from the data at hand. In particular, when the response variable is discrete, the data often contain little information to distinguish between competing distributions for the random effects. As mentioned at

the end of Section 9.4, predictions of the random effects (i.e., the empirical BLUPs) are known to be heavily influenced by the normal distribution assu nption for the random effects. As a result, histograms and normal quantile plots of the empirical BLUPs cannot be relied upon for assessing the adequacy of the normal distribution assumption for the random effects. However, in general, the estimates of the fixed effects are much less sensitive to misspecification of the random effects distribution. That is, assuming the random effects have a normal distrib tion when the true distribution of the random effects is non-normal (e.g., a skewed distribution) does not produce discernibly biased estimates of the fixed effects. The fixed effects estimates are, however, sensitive to a different kind of misspecification of the random effects distribution. When the assumption that the random effects are independent of the covariates, X_i, does not hold, the estimates of the fixed effects can be severely biased. This type of misspecification might arise, for example, in a study where one exposure group is more heterogeneous than another (i.e., the variance of the random effects depends upon exposure group).

12.5 CASE STUDIES

Next, we illustrate the main ideas presented in this chapter by considering generalized linear mixed effects models for analyzing longitudinal data from two studies. The first illustration considers a logistic regression model, with random effects, for analyzing data on amenorrhea from a randomized clinical trial of contracepting women. The second illustration considers a Poisson regression model, with random effects, for analyzing count data on epileptic seizures from a clinical trial of the anti-epileptic drug, progabide.

Clinical Trial of Contracepting Women

The first example is from a longitudinal clinical trial of contracepting women reported by Machin *et al.* (1988). In this trial women received an injection of either 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA) on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up visit 90 days after the fourth injection, that is, one year after the first injection. Throughout the study each woman completed a menstrual diary that recorded any vaginal bleeding pattern disturbances. The diary data were used to determine whether a women experienced amenorrhea, the absence of menstrual bleeding for a specified number of days.

A total of 1151 women completed the menstrual diaries and the diary data were used to generate a binary sequence for each woman, according to whether or not she had experienced amenorrhea in the four successive three-month intervals. A feature of this clinical trial is that there was substantial dropout. More than one-third of the women dropped out before the completion of the trial; 17% dropped out after receiving only one injection of DMPA, 13% dropped out after receiving

only two injections, and 7% dropped out after receiving three injections. For women who dropped out before the end of the 90-day injection interval, a determination of whether or not they experienced amenorrhea was made, on a proportionate basis, using their existing menstrual diary data for that interval. Statistical issues concerning the potential impact of missing data on the analysis are discussed in Chapter 14.

In clinical trials of modern hormonal contraceptives, pregnancy is exceedingly rare (and would be regarded as a failure of the contraceptive method), and is not the main outcome of interest in this study (Machin *et al.*, 1988). Instead, the outcome of interest is a binary response indicating whether a woman experienced amenorrhea in the four successive three-month intervals. The goal of the analyses presented here is to determine subject-specific changes in the risk of amenorrhea over the course of the study (12 months), and the influence of dosage of DMPA on changes in a woman's risk of amenorrhea. Of note, the treatment covariate (high versus low dosage of DMPA) is time-invariant.

Let $Y_{ij} = 1$ if the i^{th} woman experienced amenorrhea in the j^{th} injection interval ($j = 1, \ldots, 4$), and $Y_{ij} = 0$ otherwise. The following mixed effects logistic regression model for Y_{ij} was fit to the data:

$$\begin{aligned}\text{logit}\{E(Y_{ij}|b_i)\} &= \beta_1 + \beta_2\,\texttt{time}_{ij} + \beta_3\,\texttt{time}_{ij}^2 + \beta_4\,\texttt{dose}_i \times \texttt{time}_{ij} \\ &\quad + \beta_5\,\texttt{dose}_i \times \texttt{time}_{ij}^2 + b_{1i},\end{aligned}$$

where $\texttt{time} = 1, 2, 3, 4$ for the four consecutive 90-day injection intervals, and $\texttt{dose} = 1$ if randomized to 150mg of DMPA, and $\texttt{dose} = 0$ otherwise. Note that there is no baseline measure of amenorrhea prior to receiving the first contraceptive injection. However, due to randomization, we assume that the baseline risk (at $\texttt{time} = 0$) is the same in both dosage groups and omit a main effect of dose from the model.

Given b_{1i}, it is assumed that the Y_{ij} are independent and have a Bernoulli distribution, with $\text{Var}(Y_{ij}|b_{1i}) = E(Y_{ij}|b_{1i})\,(1 - E(Y_{ij}|b_{1i}))$, and $\phi = 1$. Finally, we assume that the single random effect b_{1i} has a univariate normal distribution, with zero mean and variance g_{11}, $b_{1i} \sim N(0, g_{11})$. This mixed effects model posits that there is natural heterogeneity in women's propensity or underlying risk of amenorrhea that persists throughout all binary responses obtained over the duration of the study.

The ML estimates of the regression parameters for this model are presented in Table 12.2. These results provide evidence that the subject-specific log odds of amenorrhea increase over the 12 months of the trial, and that subject-specific changes in the risk of amenorrhea depend on the dose of DMPA. For example, for a woman assigned to the low dose of DMPA, the log odds of amenorrhea increases approximately linearly, with an increase in the log odds of 1.09 (or $1.1332 - 0.0419$) at 3 months, 2.10 (or $2 \times 1.1332 - 4 \times 0.0419$) at 6 months, 3.02 (or $3 \times 1.1332 - 9 \times 0.0419$) at 9 months, and 3.86 (or $4 \times 1.1332 - 16 \times 0.0419$) at 12 months. These increases in risk correspond to odds ratios of 3.0 (or $e^{1.09}$), 8.2 (or $e^{2.10}$), 20.5 (or $e^{3.02}$), and 47.5 (or $e^{3.86}$) at 3, 6, 9, and 12 months, respectively. On the other hand, for a women assigned to the high dose of DMPA, the log odds of amenorrhea increases quadratically, with an increase of 1.55 at 3 months, 2.79 at 6 months, 3.73 at 9 months, and 4.37 at

Table 12.2 Parameter estimates and standard errors from a mixed effects logistic regression model, with randomly varying intercepts, for the amenorrhea data.

Variable	Estimate	SE	Z
Intercept	−3.8057	0.3050	−12.48
$\texttt{time}_{ij}$	1.1332	0.2682	4.22
$\texttt{time}^2_{ij}$	−0.0419	0.0548	−0.76
$\texttt{dose}_i \times \texttt{time}_{ij}$	0.5644	0.1922	2.94
$\texttt{dose}_i \times \texttt{time}^2_{ij}$	−0.1095	0.0496	−2.21
g_{11}	5.0646	0.5840	8.67

ML estimation based on 50-point adaptive Gaussian quadrature.

12 months. That is, the early linear trend shows a decline toward the end. These increases in risk correspond to odds ratios of 4.7 (or $e^{1.55}$), 16.3 (or $e^{2.79}$), 41.7 (or $e^{3.73}$), and 79.0 (or $e^{4.37}$) at 3, 6, 9, and 12 months, respectively. For both groups, all of these increases in the odds ratios are significant at the 0.05 level.

Because treatment (low versus high dose of DMPA) is a between-subject variable, this makes the interpretation of the fixed effects for the `dose` × `time` interactions more difficult. The interaction effects must be given an interpretation in terms of a contrast of the increases in log odds of amenorrhea (or the odds ratio) for two different women, who happen to have the same underlying risk of experiencing amenorrhea prior to randomization, but who differ in terms of dose (i.e., one is assigned to low dose and the other to high dose). From the estimates of the fixed effects in Table 12.2, the ratio of the increased odds of amenorrhea at 12 months for a woman assigned to the high dose, versus another woman with the same risk of amenorrhea prior to randomization who was assigned to the low dose, is 1.66 (or $e^{4.37-3.86}$), with 95% confidence interval: 1.03 to 2.66.

The estimated variance of the random intercepts is relatively large, $\widehat{g}_{11} = 5.065$. This implies that there is substantial variability in the propensity to experience amenorrhea, since approximately 95% of the women have a baseline risk of amenorrhea that varies from

$$\frac{\exp(-3.8057 - 1.96\sqrt{5.0646})}{1 + \exp(-3.8057 - 1.96\sqrt{5.0646})}$$

to

$$\frac{\exp(-3.8057 + 1.96\sqrt{5.0646})}{1 + \exp(-3.8057 + 1.96\sqrt{5.0646})}$$

or 0.03% to 64.68%. Alternatively, we can interpret $\widehat{g}_{11}$ by appealing to the notion of a latent variable distribution (see Section 10.3). That is, we can assume a linear mixed effects model for the latent variable U_{ij},

$$\begin{aligned} U_{ij} &= \beta_1 + \beta_2\,\texttt{time}_{ij} + \beta_3\,\texttt{time}^2_{ij} + \beta_4\,\texttt{dose}_i \times \texttt{time}_{ij} \\ &\quad + \beta_5\,\texttt{dose}_i \times \texttt{time}^2_{ij} + b_{1i} + e_{ij}, \end{aligned}$$

where b_{1i} has a normal distribution, with zero mean and variance g_{11}, and the e_{ij} have a standard logistic distribution, with mean zero and variance $\pi^2/3$. Without loss of generality, we can assume the threshold for categorizing U_{ij} is zero, with

$$Y_{ij} = 1 \quad \text{if} \quad U_{ij} > 0,$$

$$Y_{ij} = 0 \quad \text{if} \quad U_{ij} \leq 0.$$

This model for the latent variable implies the mixed effects logistic regression model for Y_{ij},

$$\begin{aligned} \text{logit}\{E(Y_{ij}|b_i)\} &= \beta_1 + \beta_2\,\texttt{time}_{ij} + \beta_3\,\texttt{time}^2_{ij} + \beta_4\,\texttt{dose}_i \times \texttt{time}_{ij} \\ &\quad + \beta_5\,\texttt{dose}_i \times \texttt{time}_{ij} + b_{1i}. \end{aligned}$$

Thus, using the notion of an underlying latent variable distribution, we can compare the magnitudes of the between-subject and within-subject sources of variability of the U_{ij} in terms of the intra-subject correlation (often referred to as the *intra-class correlation*)

$$\rho = \text{Corr}(U_{ij}, U_{ik}) = \frac{g_{11}}{g_{11} + \pi^2/3}.$$

The estimated intra-subject correlation for the repeated latent responses is

$$\widehat{\rho} = \frac{\widehat{g}_{11}}{\widehat{g}_{11} + \pi^2/3} = \frac{5.065}{5.065 + 3.290} = 0.61,$$

indicating that there is substantial heterogeneity in the underlying propensity to experience amenorrhea. Note that ρ is the marginal correlation (averaged over the distribution of the random effects) among the unobserved U_{ij}; it is not the marginal correlation among the Y_{ij}.

The mixed effects model considered above includes only a single random effect, b_{1i}. With binary data, and measurements at only four occasions, greater care must be exercised in the specification of the random effects as the limited amount of data may not support estimation of more than a single variance component. Inclusion of both randomly varying intercepts and slopes for the linear time trend in the logistic regression model for the amenorrhea data resulted in convergence problems during estimation. It should not be too surprising that problems might arise when fitting this model to the data. Intuitively, attempts to fit a series of logistic regressions to data on at most 4 observations (the number of repeated measurements on each woman with complete data; recall that due to dropout 37% of the women had fewer than 4

Table 12.3 Sensitivity of estimate of variance component to number of quadrature points: mixed effects logistic regression model, with random intercepts, for the amenorrhea data.

Quadrature Points	Log-Likelihood	Estimate of g_{11}	CPU Time
1	−1957.303	4.3366	23.72
2	−1957.246	3.9812	26.21
3	−1944.099	4.3766	26.66
4	−1933.495	5.1992	28.79
5	−1936.213	4.8369	33.05
10	−1934.514	5.0540	49.14
20	−1934.465	5.0648	82.34
30	−1934.465	5.0646	113.72
40	−1934.465	5.0646	145.69
50	−1934.465	5.0646	176.60
100	−1934.465	5.0646	329.60

†CPU time using PROC MIXED in SAS on a Sun Enterprise 5500 computer.

measurements) are likely to result in numerical problems and/or produce unstable estimates.

This highlights an important feature of longitudinal binary data: there is usually not much information available about random effects, beyond a random subject effect (or random intercept), when the number of repeated measurement is relatively small. Thus convergence problems during estimation are often encountered when random effects beyond a random subject effect are included in logistic regression models for longitudinal data.

The estimates of the fixed effects and variance component reported in Table 12.2 were obtained by maximizing an approximate integrated likelihood, where the integration over the distribution of the random effects was achieved using numerical quadrature. Choice of the number of quadrature points determines the degree of accuracy of the approximation. In Table 12.3 we display the estimate of the variance component, g_{11}, and the value of the maximized log-likelihood for increasing number of quadrature points. The results in Table 12.3 indicate that 5–10 quadrature points do not provide sufficient numerical accuracy for the amenorrhea data; this provides an illustration of the dangers of using too few quadrature points. The value for the maximized log-likelihood and the estimate of g_{11} become stable once the number of quadrature points exceeds 30. Table 12.3 also provides the CPU time required for fitting the model. As expected, the computational burden increases with the number

of quadrature points. In this example, there is only a single random effect and the increase in computational burden is relatively minor; however, in general, the computations grow exponentially with the number of random effects. When compared to the time expended in collecting longitudinal data, we regard the time required to accurately fit generalized linear mixed models to be negligible. So, in general, we recommend repeating analyses, with increasing number of quadrature points, until all estimates and standard errors become stable.

Finally, note that the estimates of the fixed effects in the mixed effects logistic regression model are larger than those obtained in similar analysis using a marginal model. For illustrative purposes, we fit the following marginal model to the amenorrhea data:

$$\begin{aligned}\text{logit}\{E(Y_{ij})\} &= \beta_1 + \beta_2\,\texttt{time}_{ij} + \beta_3\,\texttt{time}_{ij}^2 + \beta_4\,\texttt{dose}_i \times \texttt{time}_{ij} \\ &\quad + \beta_5\,\texttt{dose}_i \times \texttt{time}_{ij}^2,\end{aligned}$$

and assumed an unstructured log odds ratio pattern for the within-subject association,

$$\log \text{OR}\,(Y_{ij}, Y_{ik}) = \alpha_{jk},$$

where

$$\text{OR}(Y_j, Y_k) = \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}.$$

The estimated regression coefficients and pairwise log odds ratios for the within-subject association, obtained using the GEE approach, are presented in Table 12.4. Note that the estimated logistic regression coefficients are smaller (in absolute value) than the estimated fixed effects in Table 12.2. Furthermore, from the estimates of the fixed effects in Table 12.4, the ratio of the population odds of amenorrhea at 12 months for women on the high dose versus low dose is 1.30, with 95% confidence interval: 0.98 to 1.71. However, these differences in the estimated coefficients and odds ratios are due to the different interpretations of β in the two classes of models, that is, these two classes of models estimate parameters that address substantively different questions. The estimates of the fixed effects of dose in the mixed effects logistic regression model describe the effect of dose on a specific woman's risk of amenorrhea. The corresponding effects in the marginal logistic regression model describe the effects of dose on the prevalence of amenorrhea in the population of women assigned to high versus low doses of DMPA. Finally, although the regression parameters for dose have distinct interpretations, their values coincide when there is no effect of dose. That is, at the null value the same hypotheses concerning the dependence of the risk of amenorrhea on dose is being tested. For example, a multivariate Wald test of $H_0 : \beta_4 = \beta_5 = 0$ based on the marginal model parameter estimates produces $W^2 = 12.3$, with 2 df ($p < 0.005$). The corresponding test from the mixed effects logistic regression parameter estimates produces $W^2 = 12.4$, with 2 df ($p < 0.005$).

Table 12.4 Parameter estimates and standard errors, obtained using GEE approach, from marginal logistic regression model for the amenorrhea data.

Variable	Estimate	SE	Z
Intercept	−2.2461	0.1765	−12.72
$\texttt{time}_{ij}$	0.7030	0.1581	4.45
$\texttt{time}_{ij}^2$	-0.0323	0.0318	−1.02
$\texttt{dose}_i \times \texttt{time}_{ij}$	0.3380	0.1097	3.08
$\texttt{dose}_i \times \texttt{time}_{ij}^2$	−0.0683	0.0284	−2.40
α_{12}	1.8475	0.1810	10.21
α_{13}	1.4851	0.1985	7.48
α_{14}	1.7605	0.2482	7.09
α_{23}	2.1610	0.1761	12.27
α_{24}	2.0665	0.2034	10.16
α_{34}	2.2783	0.1827	12.47

Clinical Trial of an Anti-Epileptic Drug

Next we consider data from the placebo-controlled clinical trial of 59 epileptic patients, conducted by Leppik *et al.* (1987). Patients with partial seizures were enrolled in a randomized clinical trial of the anti-epileptic drug, progabide. Progabide is an anti-epileptic drug whose primary mechanism of action is to enhance gamma-aminobutyric acid (GABA) content; GABA is the primary inhibitory neurotransmitter in the brain.

Participants in the study were randomized to either progabide or a placebo, as an adjuvant to the standard anti-epileptic chemotherapy. Prior to receiving treatment, baseline data on the number of epileptic seizures during the preceding 8-week interval were recorded. Counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visit were recorded. The average rates of seizures (per week), at baseline and in the four post-randomization visits are displayed in Figure 12.2.

These data contain an extreme observation or outlier: one of the patients (patient 49) reported 151 seizures in the 8-week baseline interval and 302 (102+65+72+63) seizures during the four successive 2-week intervals. This patient was assigned to the progabide group. Since this patient could potentially have an inordinate impact on the analysis, we present results that include and exclude data from this patient.

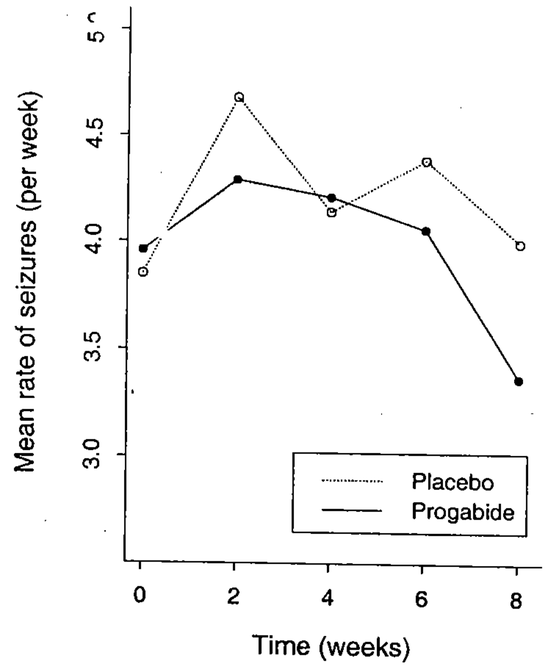

Fig. 12.2 Mean rate of seizures (per week) at baseline, week 2, week 4, week 6, and week 8 in the progabide and placebo groups.

We consider an analysis that addresses the question of whether or not treatment with progabide reduces the rate of epileptic seizures (when compared to placebo). To address this question we can compare the subject-specific changes, from baseline to follow-up, in the rate of seizures for patients in the two treatment groups. We consider the following mixed effects log-linear regression model for the subject-specific expected counts (or rates) of seizures,

$$\begin{aligned}\log E(Y_{ij}|b_i) &= \log(\texttt{T}_{ij}) + \beta_1 + \beta_2\,\texttt{time}_{ij} + \beta_3\,\texttt{trt}_i + \beta_4\,\texttt{trt}_i \times \texttt{time}_{ij} \\ &\quad + b_{i1} + b_{i2}\,\texttt{time}_{ij} \\ &= \log(\texttt{T}_{ij}) + (\beta_1 + b_{i1}) + (\beta_2 + b_{i2})\,\texttt{time}_{ij} + \beta_3\,\texttt{trt}_i \\ &\quad + \beta_4\,\texttt{trt}_i \times \texttt{time}_{ij},\end{aligned}$$

where Y_{ij} is the number of epileptic seizures for the i^{th} patient in the j^{th} period of observation ($j = 0, \ldots, 4$), and $\texttt{T}_{ij}$ is the length of period j (where $\texttt{T}_{ij} = 8$ if $j = 0$ and $\texttt{T}_{ij} = 2$ if $j = 1, 2, 3, 4$). The offset, $\log(\texttt{T}_{ij})$ is included because the "time at risk" is not the same in the baseline (8 weeks) and four successive follow-up periods (2 week intervals). The variable `trt` is an indicator variable for treatment

Table 12.5 Subject-specific log expected seizure rates in the two groups at baseline and during post-baseline follow-up.

Treatment Group	Period	$\log\left(\frac{E(Y_{ij}\|b_i)}{T_{1j}}\right)$
Placebo	Baseline	$\beta_1 + b_{1i}$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i})$
Progabide	Baseline	$(\beta_1 + b_{1i}) + \beta_3$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) + \beta_3 + \beta_4$

group, with $\texttt{trt} = 0$ if an individual was randomized to the placebo group and $\texttt{trt} = 1$ if randomized to the progabide group. The binary variable time denotes the baseline and follow-up periods, with $\texttt{time} = 0$ for the baseline period and $\texttt{time} = 1$ for the follow-up periods (periods 1–4). Given b_i, it is assumed that the Y_{ij} are independent and have a Poisson distribution, with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)$, (i.e., $\phi = 1$). Finally, we assume that the random intercepts and slopes, b_i, have a bivariate normal distribution, with zero mean and 2×2 covariance matrix G. This mixed effects log-linear regression model posits that there is not only natural heterogeneity among patients in terms of their baseline expected rate of seizures, but also heterogeneity in the changes in the expected rates of seizures over time. Unlike the previous example, where there was not much information available in the four repeated binary response about random effects beyond a random subject effect, with these repeated count data there is sufficient information to estimate both random intercepts and slopes.

In Table 12.5 we summarize the interpretation of β in terms of the subject-specific log expected seizure rates (per week) in the two groups at baseline and during post-baseline follow-up. Because all of the covariates in the model are dichotomous, the log-linear fixed effects regression parameters can be given interpretations in terms of subject-specific (log) rate ratios. So, for example, e^{β_2} is the rate ratio of seizures, comparing the follow-up periods to baseline, for a "typical" patient in the placebo group (a "typical" patient is one with unobserved random slope $b_{2i} = 0$, the mean and median of the distribution of b_{2i}). Similarly, $e^{\beta_2+\beta_4}$ is the rate ratio of seizures, comparing the follow-up periods to baseline, for a "typical" patient in the progabide group (with unobserved random slope $b_{2i} = 0$). A direct comparison of the two treatments in terms of changes in the expected rates of seizures is expressible in terms of β_4. That is, β_4 represents the difference between the changes in the log expected rates, comparing a patient from the progabide group to a patient from the placebo group, when the two patients are chosen so that they have the same value for the unobserved slope b_{2i}. That is, e^{β_4} is a ratio of rate ratios. If $\beta_4 < 0$, this indicates a greater reduction (or, alternatively, a smaller increase) in the seizure rate

Table 12.6 Parameter estimates and standard errors from mixed effects log-linear regression model for the seizure data.

Variable	Estimate	SE	Z
Intercept	1.0707	0.1406	7.62
$\texttt{time}_{ij}$	−0.0004	0.1097	−0.00
$\texttt{trt}_i$	0.0513	0.1931	0.27
$\texttt{trt}_i \times \texttt{time}_{ij}$	−0.3065	0.1513	−2.03
$g_{11} = \text{Var}(b_{i1})$	0.5010	0.1010	4.96
$g_{22} = \text{Var}(b_{i2})$	0.2334	0.0608	3.84
$g_{12} = \text{Cov}(b_{i1}, b_{i2})$	0.0541	0.0559	0.97

ML estimation based on 50-point adaptive Gaussian quadrature.

from baseline for the patient assigned to the progabide group (when compared to the patient assigned to the placebo group).

For the full sample (N=59) the estimated fixed effects and covariance parameters from the log-linear model are displayed in Table 12.6. A test of the null hypothesis, H_0: $\beta_4 = 0$, indicates that there is a significant time × treatment interaction at the 0.05 level. These results suggest that there are differences between the two treatments in terms of subject-specific changes in the expected rates of seizures. In particular, there is a greater reduction in the expected seizure rate from baseline for patients treated with progabide (when compared to patients treated with a placebo). For a patient receiving a placebo, there is no expected change in the rate of seizures (or $e^{-0.0004} \approx 1.0$), while for a patient treated with progabide the expected decrease in seizures is approximately 26% (or $e^{-0.0004-0.3065} = e^{-0.3069} \approx 0.74$).

The estimated covariance parameters for the random intercepts and slopes indicate that there is substantial variability in the baseline seizure rate in the study population and also substantial variability in the patient-to-patient changes in the seizure rates in response to treatment. For example, the estimated variance of the random intercepts, $\widehat{g}_{11} = 0.501$, implies that there is substantial patient-to-patient variability in terms of their baseline rate of seizures, since approximately 95% of the patients have a baseline seizure rate that varies from

$$\exp(1.071 - 1.96\sqrt{0.501}) \text{ to } \exp(1.071 + 1.96\sqrt{0.501}),$$

or 0.8 to 12.0 seizures per week. Similarly, there is discernible heterogeneity in the patient-to-patient changes in the seizure rates. For example, approximately 95% of

Table 12.7 Parameter estimates and standard errors from mixed effects log-linear regression model for the seizure data, excluding patient 49.

Variable	Estimate	SE	Z
Intercept	1.0692	0.1344	7.96
$\texttt{time}_{ij}$	0.0078	0.1070	0.07
$\texttt{trt}_i$	−0.0079	0.1860	−0.04
$\texttt{trt}_i \times \texttt{time}_{ij}$	−0.3461	0.1489	−2.33
$g_{11} = \text{Var}(b_{i1})$	0.4529	0.0935	4.84
$g_{22} = \text{Var}(b_{i2})$	0.2163	0.0587	3.68
$g_{12} = \text{Cov}(b_{i1}, b_{i2})$	0.0151	0.0529	0.29

ML estimation based on 50-point adaptive Gaussian quadrature.

patients treated with progabide have changes in the rates of seizures that vary from

$$\exp(-0.307 \pm 1.96\sqrt{0.233}),$$

or changes that vary from a decrease in seizures of 71% to an increase in seizures of 88%. Finally, the correlation among the random intercepts and slopes is weak, indicating that the expected change in the seizure rates in not directly related to the baseline rate of seizures.

As was noted earlier, patient 49 is an outlier with extreme counts at all occasions. While the observations on this patient are likely to inflate the variance of the random effects, especially the variance of b_{1i}, they might also have an inordinate influence on the estimates of the fixed effects parameters. To assess the impact this patient has on the results, we repeated the analysis excluding observations on this patient (N=58). The results of this analysis are displayed in Table 12.7. A test of the null hypothesis, H_0: $\beta_4 = 0$, indicates that there is still a significant time × treatment interaction at the 0.05 level. These results indicate that there is a greater reduction in the expected seizure rate from baseline for patients treated with progabide (when compared to patients treated with a placebo). For a patient receiving a placebo, there is no expected change in the rate of seizures (or $1 - e^{0.0078} \approx 0$), while for a patient treated with progabide the expected decrease in seizures is approximately 30% (or $1 - e^{0.0078-0.3461} = 1 - e^{-0.3383} \approx 0.29$). Qualitatively, the results in Table 12.7 are very similar to those obtained in Table 12.6. As might be expected, the exclusion of patient 49 results in a noticeably smaller estimate of $\text{Var}(b_{1i})$.

Table 12.8 Illustrative commands for a mixed effects logistic regression, with randomly varying intercepts, using PROC NLMIXED in SAS.

```
PROC NLMIXED QPOINTS=50;
   PARMS beta1=-3.0 beta2=-0.2 beta3=0.5 beta4=0.1 g11=0 to 5 by 0.5;
   eta = beta1 + beta2*time + beta3*group + beta4*group*time + b1;
   mu = exp(eta)/(1+exp(eta));
   MODEL y ~ BINARY(mu);
   RANDOM b1 ~ NORMAL(0, g11) SUBJECT=id;
   PREDICT mu OUT=predmean;
```

12.6 COMPUTING: FITTING GENERALIZED LINEAR MIXED MODELS USING PROC NLMIXED IN SAS

Until recently, a potential limitation of generalized linear mixed models was their computational burden. Because there is no simple closed-form solution for the marginal likelihood, numerical integration techniques are required. Maximum (marginal) likelihood estimation has only recently been implemented in standard statistical software, for example, PROC NLMIXED in SAS.

To fit generalized linear mixed models we use the PROC NLMIXED procedure in SAS. PROC NLMIXED directly maximizes an approximate integrated likelihood, where the integration over the random effects is achieved using numerical quadrature. For example, to fit a logistic regression model with randomly varying intercepts to longitudinal data from two groups (coded 0 and 1), we can use the illustrative SAS commands given in Table 12.8. Similarly, to fit a mixed effects log-linear regression, with randomly varying intercepts and slopes, we can use the illustrative SAS commands given in Table 12.9.

We note that PROC NLMIXED in SAS is a very general and versatile procedure for fitting non-linear mixed effects models to data from a wide variety of applications. Here we focus on the use of PROC NLMIXED to fit generalized linear mixed models to longitudinal data. No attempt is made here to give a comprehensive review of the main features of PROC NLMIXED. Instead, we present the source code for mixed effects logistic and log-linear regression in general terms (see Tables 12.8 and 12.9) and then describe the most salient parts of the command syntax for these two illustrations. Next, we present a brief description of each of the command statements used in Tables 12.8 and 12.9.

Table 12.9 Illustrative commands for a mixed effects log-linear regression, with randomly varying intercepts and slopes, using PROC NLMIXED in SAS.

```
PROC NLMIXED QPOINTS=50;
   PARMS beta1=1.0 beta2=0.0 beta3=0.0 beta4=-0.5 g11=0 to 2 by 0.5
         g22=0 to 2 by 0.5 g12=-1 to 1 by 0.25;
   eta = beta1 + beta2*time + beta3*group + beta4*group*time + b1 + b2*time;
   mu = exp(eta);
   MODEL y ~ POISSON(mu);
   RANDOM b1 b2 ~ NORMAL([0,0], [g11, g12, g22]) SUBJECT=id;
   PREDICT beta2+b2 OUT=slopes;
```

PROC NLMIXED <options>;

The PROC NLMIXED statement calls the procedure NLMIXED in SAS. It can also include an option for the number of quadrature points used during evaluation of integrals for the marginal likelihood. For example, QPOINTS=50 specifies that 50 quadrature points be used for each random effect, resulting in a total of 50^q quadrature points (where q is the number of random effects, the dimension of b_i). A note of caution, the likelihood approximation may not be accurate if too few quadrature points are used.

PARMS <name list>;

The PARMS statement lists the names of all the parameters in the model (the fixed effects and the covariance parameters for the random effects). The PARMS statement is also used to specify initial values (or a grid of initial values) for the parameters. Parameters not listed on the PARMS statement are assigned an initial value of 1. The latter will often be a very poor choice of starting value and may lead to convergence problems. Consequently we recommend that accurate initial values always be chosen. Accurate initial values for the fixed effects can be obtained from a prior analysis that assumes there are no random effects (a "working independence" assumption). Accurate initial values for the covariance parameters for the random effects can be obtained by specifying a grid of feasible values. When given a grid of values, PROC NLMIXED will first evaluate the marginal likelihood at each grid value and select the grid point that produces the largest value of the marginal likelihood as the initial values for the covariance parameters for the subsequent maximization of the marginal likelihood.

Program statements

The program statements are used to define the linear predictor (the fixed and random effects) and to relate the mean of the distribution of the response to the linear predictor. PROC NLMIXED allows multiple program statements.

MODEL response ~ distribution;

The MODEL statement specifies the response variable and the conditional distribution of the response given the random effects. PROC NLMIXED includes options for the following distributions from the exponential family:

NORMAL(m, v): specifies a normal distribution with mean m and variance v.

BINARY(p): specifies a Bernoulli distribution with probability of success p.

BINOMIAL(n, p): specifies a binomial distribution with n trials and probability of success p.

POISSON(m): specifies a Poisson distribution with mean m.

Note that the parameters of the distribution will ordinarily appear in the program statements and/or on the PARMS statement. For example, in Tables 12.8 and 12.9, mu is the mean parameter for the Bernoulli and Poisson distributions and is specified in the program statements. For Poisson and binomial response data, it is not possible to include a dispersion parameter. As a result, the introduction of random effects is the only mechanism for accounting for overdispersion in the data.

RANDOM effects ~ distribution SUBJECT=variable;

The random statement defines the random effects and a variable that determines the clustering of observations within an individual. The latter is achieved with the SUBJECT option which is used to denote a variable that distinguishes clusters of correlated responses. By including a variable on the SUBJECT option (e.g., a subject identifier), pairs of observations with the same value of that variable are regarded as correlated (by virtue of sharing the same random effects) while pairs of observations with distinct values are regarded as independent. Because PROC NLMIXED assumes a new realization of the random effects occurs whenever the SUBJECT=variable changes, the data must be sorted by this variable. This can be accomplished by sorting the data by the SUBJECT=variable prior to calling PROC NLMIXED. All random effects are assumed to have a normal distribution, normal(m, v), with mean (vector) m and variance (covariance matrix) v. For a single random effect the syntax is:

```
RANDOM b1 ~ NORMAL(0, g11) SUBJECT=id;
```

For two random effects the corresponding syntax requires the use of brackets for the mean vector and covariance matrix:

RANDOM b1 b2 ~ NORMAL([0,0], [g11, g12, g22]) SUBJECT=id;

Only the non-redundant, lower triangle of the covariance matrix is included in the parameters of the multivariate normal distribution for the random effects.

PREDICT expression OUT=SAS data-set;

The PREDICT statement is used to obtain predictions of any specified expression. Predicted values are based on the ML estimates of the fixed effects and empirical Bayes estimates of the random effects. The resulting predictions are placed in a SAS data-set specified with the OUT=SAS data-set option.

Of note, PROC NLMIXED does not have a CLASS statement that can be used to identify all variables that are to be regarded as categorical or factors. Instead, the user must create indicator variables for each factor, using appropriate coding.

Finally, a word of caution concerning the use of PROC NLMIXED. Our limited experience with this procedure indicates that it can be very sensitive to poor choices of starting values and/or the numerical accuracy of the quadrature used. Convergence of the algorithm implemented in PROC NLMIXED should never be taken for granted; neither should convergence to a global maximum be assumed. Instead, we recommend that users of this procedure provide different initial values and/or consider a grid search of values to ensure that a global maximum has been obtained. Problems with convergence and computation are likely to arise when the model parameters are on scales that vary by more than a few orders of magnitude. The latter problem can be circumvented by appropriately rescaling each parameter. For example, when the variances of random intercepts and slopes differ by more than a few orders of magnitude, the variance of the slopes can be rescaled by multiplying the variable for `time` by an appropriate constant.

12.7 FURTHER READING

A relatively non-technical discussion of random effects models for binary data can be found in Section 6.7 of Collett (1991). Chapter 12 of Agresti (2002) provides a detailed, although more mathematically challenging, description of generalized linear mixed effects models for categorical data.

Bibliographic Notes

The theoretical foundation for generalized linear mixed effects models can be found in Skellam's (1948) introduction of the beta-binomial distribution. Since then, the statistical literature on generalized linear mixed effects models has grown rapidly.

Some key references in the literature include: Cox (1970), Pierce and Sands (1975), Williams (1982), Stiratelli *et al.* (1984), Anderson and Aitkin (1985), Gilmour *et al.* (1985), Wong and Mason (1985), Schall (1991), Zeger and Karim (1991), Breslow and Clayton (1993), and Hedeker and Gibbons (1994).

The marginal maximum likelihood method described previously is based on a numerically integrated likelihood function and requires the computation of the integral over the random effects. A method known as adaptive Gaussian quadrature is commonly used for computing this integral and is described in detail in Pinheiro and Bates (1995); also see Anderson and Aitkin (1985) and Hedeker and Gibbons (1994, 1996). An alternative approximation, leading to an approach known as penalized quasi-likelihood (PQL), was proposed by Stiratelli *et al.* (1984). A more accurate approximation, based on higher-order Laplace approximations, is described in Breslow and Lin (1995) and Lin and Breslow (1996).

Finally, Lesaffre and Spiessens (2001) provide a striking example of the dangers of using too few quadrature points when fitting generalized linear mixed effects models; also, see Problem 12.1.10.

Problems

12.1 In a randomized, double-blind, parallel-group, multicenter study comparing two oral treatments (denoted A and B) for toe-nail infection (De Backer *et al.*, 1998; also see Lesaffre and Spiessons, 2001), patients were evaluated for the degree of onycholysis (the degree of separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48 thereafter. The onycholysis outcome variable is binary (none or mild versus moderate or severe). The binary outcome was evaluated on 294 patients comprising a total of 1908 measurements. The main objective of the analyses is to compare the effects of oral treatments A and B on changes in the probability of the binary onycholysis outcome over the duration of the study.

The raw data are stored in an external file: `toenail.dat`

Each row of the data set contains the following five variables:

ID Y Treatment Month Visit

Note: The binary onycholysis outcome variable Y is coded 0 = none or mild, 1 = moderate or severe. The categorical variable Treatment is coded 1 = oral treatment A, 0 = oral treatment B. The variable Month denotes the exact timing of measurements in months. The variable Visit denotes the visit number (visit numbers 1–7 correspond to scheduled visits at 0, 4, 8, 12, 24, 36, and 48 weeks).

12.1.1 First, consider a *marginal* model for the log odds of moderate or severe onycholysis. Using GEE, fit a model that assumes linear trends for the log odds over time, with common intercept for the two treatment groups, but different slopes:

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2\,\texttt{Month}_{ij} + \beta_3\,\texttt{Treatment}_i \times \texttt{Month}_{ij}.$$

Assume "exchangeable" log odds ratios (or "exchangeable" correlations, if available software does not permit the within-subject association to be parameterized in terms of log odds ratios) for the association among the repeated binary responses.

12.1.2 What is the interpretation of β_2 in this model?

12.1.3 What is the interpretation of β_3 in this model?

12.1.4 From the results of the analysis for Problem 12.1.1, what conclusions do you draw about the effect of treatment on changes in the log odds of moderate or severe onycholysis over time? Provide results that support your conclusions.

12.1.5 Next, consider a generalized linear mixed model, with randomly varying intercepts, for the patient-specific log odds of moderate or severe onycholysis. Using maximum likelihood (ML), fit a model with linear trends for the log odds over time and allow the slopes to depend on treatment group:

$$\text{logit}\{E(Y_{ij}|b_i)\} = (\beta_1 + b_i) + \beta_2\,\texttt{Month}_{ij} + \beta_3\,\texttt{Treatment}_i \times \texttt{Month}_{ij},$$

where, given b_i, Y_{ij} is assumed to have a Bernoulli distribution. Assume that $b_i \sim N(0, \sigma_b^2)$.

Suggestion: Use the GEE estimates of β from Problem 12.1.1 as starting values for the ML estimation routine.

12.1.6 What is the estimate of σ_b^2? Give an interpretation to the magnitude of the estimated variance.

12.1.7 What is the interpretation of the estimate of β_2?

12.1.8 What is the interpretation of the estimate of β_3?

12.1.9 Compare and contrast the estimates of β_3 from the marginal and mixed effects models. Why might they differ?

12.1.10 Repeat the analysis from Problem 12.1.5 sequentially increasing the number of quadrature points used. Compare the estimates and standard errors of the model parameters when the number of quadrate points is 2, 5, 10, 20, 30, and 50. Do the results depend on the number of quadrature points?

12.2 The Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled clinical trial of beta carotene to prevent non-melanoma skin cancer in high-risk subjects (Greenberg *et al.*, 1989, 1990; also see Stukel, 1993). A total of 1805 subjects were randomized to either placebo or 50 mg of beta carotene per day for 5 years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The outcome variable is a count of the number of new skin cancers per year. The outcome was evaluated on 1770 subjects comprising a total of 7081 measurements. The main objective of the analyses is to compare the effects of beta carotene on skin cancer rates.

The raw data are stored in an external file: `skin.dat`

Each row of the data set contains the following 9 variables:

ID Center Age Skin Gender Exposure Y Treatment Year

Note: The outcome variable Y is a count of the of the number of new skin cancers per year. The categorical variable Treatment is coded 1 = beta carotene, 0 = placebo. The variable Year denotes the year of follow-up. The categorical variable Gender is coded 1 = male, 0 = female. The categorical variable Skin denotes skin type and is coded 1 = burns, 0 = otherwise. The variable Exposure is a count of the number of previous skin cancers. The variable Age is the age (in years) of each subject at randomization.

12.2.1 Consider a generalized linear mixed model, with randomly varying intercepts, for the subject-specific log rate of skin cancers. Using maximum likelihood (ML), fit a model with linear trends for the log rate over time and allow the slopes to depend on treatment group:

$$\log\{E(Y_{ij}|b_i)\} = (\beta_1 + b_i) + \beta_2\,\texttt{Year}_{ij} + \beta_3\,\texttt{Treatment}_i \times \texttt{Year}_{ij},$$

where, given b_i, Y_{ij} is assumed to have a Poisson distribution. Assume that $b_i \sim N(0, \sigma_b^2)$.

Suggestion: Use GEE estimates of the fixed effects as starting values for β in the ML estimation routine.

12.2.2 What is the estimate of σ_b^2? Give an interpretation to the magnitude of the estimated variance.

12.2.3 What is the interpretation of the estimate of β_2?

12.2.4 What is the interpretation of the estimate of β_3?

12.2.5 From the results of the analysis for Problem 12.2.1, what conclusions do you draw about the effect of beta carotene on the log rate of skin cancers? Provide results that support your conclusions.

12.2.6 Obtain the predicted (empirical BLUP) random effect for each subject.

(a) Calculate the sample variance of the predictions. How does it compare to the estimate of σ_b^2 obtained in Problem 12.2.2? Why might they differ?

(b) Plot the predictions against age and the count of the number of previous skin cancers. What do you conclude?

12.2.7 Repeat the analysis from Problem 12.2.1 adjusting for skin type, age, and the count of the number of previous skin cancers. What conclusions do you draw about the effect of beta carotene on the adjusted log rate of skin cancers?

13

Contrasting Marginal and Mixed Effects Models

13.1 INTRODUCTION

In this chapter we compare and contrast marginal and mixed effects models for longitudinal data. There are a number of important distinctions between these two broad classes of models that go beyond simple differences in approaches to accounting for the within-subject association. We want to emphasize that these two classes of models have somewhat different targets of inference and therefore address subtly different questions regarding longitudinal change in the response. In this chapter we highlight the main distinctions and discuss the types of scientific questions addressed by each of the two classes of models.

13.2 LINEAR MODELS: A SPECIAL CASE

In Part II we focused on linear models for longitudinal data where the model for the mean response vector can be expressed as

$$E(Y_i) = X_i\beta. \tag{13.1}$$

To account for the positive correlation among the repeated measurements we described two broad approaches. The first approach is to adopt a covariance pattern model (e.g., autoregressive, Toeplitz) for $\Sigma_i = \text{Cov}(Y_i)$. The second approach is to introduce random effects in the model for the mean response,

$$E(Y_i|b_i) = X_i\beta + Z_ib_i, \tag{13.2}$$

where b_i is a vector of random effects that vary from individual to individual according to a probability distribution (commonly assumed to be multivariate normal). The introduction of random effects induces a random effects covariance structure for Σ_i,

$$\Sigma_i = Z_i G Z_i' + \sigma^2 I_{n_i},$$

where $G = \text{Cov}(b_i)$ and σ^2 is the variance of the measurement or sampling errors.

When discussing these two different approaches for accounting for the within-subject association, issues concerning the interpretation of β in (13.1) and (13.2) simply did not arise because β has the same interpretation in both models. That is, β describes how the mean response in the study population changes with time and how these changes are related to the covariates. This interpretation of β is transparent when (13.1) is considered. However, the linear mixed effects model given by (13.2) implies the exact same model for the marginal mean response when averaged over the distribution of the random effects. That is,

$$\begin{aligned} E(Y_i) &= E\{E(Y_i|b_i)\} \\ &= E(X_i\beta + Z_i b_i) \\ &= X_i\beta + Z_i E(b_i) \\ &= X_i\beta, \end{aligned}$$

since the vector of random effects has mean zero (i.e., $E(b_i) = 0$). In the process of taking the expectation or average over the distribution of the random effects we have implicitly used the property that expectation is a linear operation. This means that the expectation of any linear function of b_i can be easily evaluated. That is,

$$E(X_i\beta + Z_i b_i) = X_i\beta + Z_i E(b_i),$$

for any constants $X_i\beta$ and Z_i. Thus β has the same interpretation in (13.1) and (13.2) because (13.2) is a *linear* mixed effects model (i.e., the right-hand side of (13.2) is a linear function of b_i). Put more simply, in the linear mixed effects model β has a marginal interpretation because the average of the linear rates of change over time for individuals is the same as the linear rate of change over time in the population mean response. However, as we shall see in the next section, when evaluating expectations of any *non-linear* functions of b_i we can no longer proceed in this manner. That is, for any non-linear function of b_i, say $h(X_i\beta + Z_i b_i)$,

$$E\{h(X_i\beta + Z_i b_i)\} \neq h\{X_i\beta + Z_i E(b_i)\}.$$

13.3 GENERALIZED LINEAR MODELS

Next we consider the comparison of marginal and mixed effects generalized linear models for longitudinal data. Recall that one of the components in the specification of a generalized linear model is the link function, $g(\mu_i)$, which relates the mean of Y_i to the linear predictor. In the previous discussion of linear models, the link function was

the identity function, $g(\mu_i) = \mu_i$. For the special case of an identity link function, and hence linear models, the regression parameters β have the same interpretation in both marginal and mixed effects models. In this section, we focus on non-linear (or non-identity) link functions and compare the regression parameters in marginal and generalized linear mixed effects models.

Recall that a marginal model for the mean response vector is given by

$$g(\mu_i) = g\{E(Y_i)\} = X_i\beta, \tag{13.3}$$

where $g(\cdot)$ is an appropriate non-linear link function (e.g., logit or log). The regression parameters β in a marginal model have interpretation in terms of changes in the transformed mean response in the study population, and their relation to covariates. For example, when the components of Y_i are binary and a logit link function is adopted, with

$$\text{logit}(\mu_i) = X_i\beta,$$

the regression parameters have interpretation in terms of changes in the log odds of success in the study population. For any known link function $g(\cdot)$, the population means can be expressed in terms of the inverse link function, say $h(\cdot) = g^{-1}(\cdot)$,

$$h\{g(\mu_i)\} = \mu_i = E(Y_i) = h(X_i\beta). \tag{13.4}$$

For example, when the component of Y_i are binary and a logit link function has been adopted, the model for μ_i is

$$\mu_i = h(X_i\beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)},$$

where $h(\cdot)$ is the inverse-logit link function. Whether expressed as (13.3) or (13.4), the regression parameters β in a marginal model describe changes in the transformed population mean response vector, μ_i. Note also that μ_i depends on the index i only via fixed and known covariate values.

Next, consider generalized linear mixed models where the conditional mean of Y_i, given a vector of random effects b_i, is

$$g\{E(Y_i|b_i)\} = X_i\beta^* + Z_i b_i, \tag{13.5}$$

where the random effects b_i have a distribution with mean zero and covariance matrix G. Here we denote the fixed effects by β^* to clearly distinguish them from the corresponding regression parameters in the marginal model in (13.3). The regression coefficients β^* have subject-specific interpretations in terms of changes in the transformed mean response for any individual. That is, to interpret any component of β^* we must consider a unit change in the corresponding covariate while holding b_i fixed. However, the most natural way to hold b_i fixed at a particular value is to focus on the conditional mean response vector of any given individual. Alternatively, we can compare two individuals that have the same values for b_i, but who differ by a single unit in the corresponding covariate. The former interpretation of any component of

β^* is most natural when the covariate is time-varying; the latter interpretation is more natural when the covariate is time-invariant.

Thus, unlike β in marginal models, β^* has interpretation in terms of changes in the transformed mean response for any individual (or the notional comparison of individuals with the same values for b_i). The regression coefficients β^* do not describe changes in the transformed mean response in the study population. The implied model for the marginal means can only be obtained by averaging over the distribution of the random effects. This involves taking an expectation of a non-linear function of b_i,

$$\begin{aligned} \mu_i &= E(Y_i) \\ &= E\{E(Y_i|b_i)\} \\ &= E\{h(X_i\beta^* + Z_i b_i)\}. \end{aligned} \tag{13.6}$$

(Note that μ_i depends on the index i only via fixed and known covariate values.)

The expression given in (13.6) is the expectation of a non-linear function of b_i. It must be evaluated from the definition of expectation as a weighted average, weighted according to the distribution of the random effects,

$$\mu_i = E(Y_i) = E\{h(X_i\beta^* + Z_i b_i)\} = \int_{-\infty}^{\infty} h(X_i\beta^* + Z_i b_i) f(b_i) db_i, \tag{13.7}$$

where the integration denotes summation or averaging and $f(b_i)$ is the probability density function for b_i (or the "weights" used in the process of averaging). However, the expression for $E(Y_i)$ given by (13.7) does not, in general, have a closed-form expression and, moreover, as noted in the previous section,

$$E(Y_i) \neq h(X_i\beta),$$

for any β. For example, consider the logistic regression model with a randomly varying intercept,

$$\text{logit}\{E(Y_i|b_i)\} = X_i\beta^* + b_i,$$

where $b_i \sim N(0, \sigma_b^2)$. The implied model for the marginal mean or marginal probability of success is

$$\begin{aligned} \mu_i &= E(Y_i) \\ &= E\{E(Y_i|b_i)\} \\ &= E\left\{\frac{e^{(X_i\beta^*+b_i)}}{1+e^{(X_i\beta^*+b_i)}}\right\} \\ &= \int_{-\infty}^{\infty} \frac{e^{(X_i\beta^*+b_i)}}{1+e^{(X_i\beta^*+b_i)}} \frac{1}{\sqrt{2\pi\sigma_b^2}} e^{-\frac{1}{2}b_i^2/\sigma_b^2}\, db_i. \end{aligned}$$

This expression cannot be evaluated in closed-form, and moreover is not of the logistic regression form,

$$\frac{e^{(X_i\beta)}}{1+e^{(X_i\beta)}},$$

for any choice of β.

For the special case of the logistic regression model with only a single randomly varying intercept (or subject effect),

$$\text{logit}\{E(Y_i|b_i)\} = X_i\beta^* + b_i,$$

where $b_i \sim N(0, \sigma_b^2)$, the following approximate relationship holds:

$$\text{logit}\{E(Y_i)\} \approx (1 + k^2\sigma_b^2)^{-\frac{1}{2}} X_i\beta^*,$$

where $k = \frac{16\sqrt{3}}{15\pi}$. The derivation of this approximation is not important. What this approximate relationship highlights is how the logistic regression coefficients in the marginal model are attenuated relative to the corresponding fixed effects in the logistic regression model with a randomly varying intercept,

$$\beta \approx \frac{\beta^*}{\sqrt{1 + 0.346\sigma_b^2}},$$

where $k^2 = 0.346$. Thus, when $\text{Var}(b_i) = \sigma_b^2 > 0$ the marginal logistic regression model parameters, β, are smaller in absolute value than the fixed effects, β^*, in the mixed effects model. In addition, the discrepancy between β and β^* increases with increasing σ_b^2.

Thus, for non-linear link functions, the fixed effects β^* in generalized linear mixed models are not comparable to the regression parameters β in marginal models. The lack of comparability reflects the distinct targets of inference associated with generalized linear mixed models and marginal models. That is, the fixed effects β^* describe the effects of covariates on changes in an individual's response over time while the regression parameters β describe the effects of covariates on changes in the population mean response over time.

In addition to the special case of an identity function (i.e., linear mixed effects models), there happens to be one exceptional case where β^* and β are almost comparable. When a log link function is adopted and the model has only a single randomly varying intercept (or subject effect),

$$\log\{E(Y_i|b_i)\} = X_i\beta^* + b_i,$$

then

$$\begin{aligned} \mu_i &= E(Y_i) \\ &= E\{E(Y_i|b_i)\} \\ &= E\{e^{(X_i\beta^*+b_i)}\} \\ &= e^{X_i\beta^*} E(e^{b_i}) \\ &= e^{X_i\beta^* + \log E(e^{b_i})}. \end{aligned}$$

As a result,

$$\log\{E(Y_i)\} = \log(\mu_i) = X_i\beta^* + \log E(e^{b_i}),$$

where $\log E(e^{b_i})$ is simply a constant (e.g., if b_i is assumed to have a normal distribution, with zero mean and variance g_{11}, then $\log E(e^{b_i}) = g_{11}/2$) and only alters the intercept term. Thus, for the special case of a log link function and a single randomly varying intercept, the fixed effects β^* are directly comparable to the marginal model regression parameters β (with the exception of the intercept).

Finally, although it is possible, in principle, to obtain estimates of the marginal means from a generalized linear mixed effects model, the assumed form for the regression model for the conditional means given b_i (e.g., logistic or log-linear) no longer holds for the resulting marginal means when averaged over the distribution of the random effects. As a result, a set of regression parameters for μ_i, describing the dependence of the population mean response on the covariates, is not immediately available from a generalized linear mixed effects model, even after averaging over the distribution of the random effects. The practical consequence is that it is not possible to describe parsimoniously the effects of covariates on the population means in terms of regression coefficients. This may not be so problematic in the setting of a randomized longitudinal clinical trial where the parameter of interest is often a simple difference or contrast of treatment means (or changes in treatment means from baseline). In the latter setting there is only a single covariate of interest (e.g., treatment group) and it is discrete; furthermore, a suitable contrast of the *marginal* mean response profiles in the treatment groups can be estimated. However, when one or more of the covariates of interest is quantitative and/or when there are potential confounding variables that need to be controlled for in the analysis, no simple summaries of the effects of covariates on μ_i are readily available from generalized linear mixed effects models.

13.4 SIMPLE NUMERICAL ILLUSTRATION

To reinforce the distinctions made in the previous section, consider the following simple numerical illustration. Suppose that Y_i is a vector of binary responses and it is of interest to describe changes in the log odds of success over time. For simplicity, we assume that there are no covariates other than the times of measurement. A logistic regression model, with randomly varying intercepts, is given by

$$\text{logit}\{E(Y_{ij}|b_i)\} = \beta_1^* + \beta_2^* t_{ij} + b_i,$$

where b_i is assumed to have a normal distribution with zero mean and variance $g_{11} = \text{Var}(b_i)$. Figure 13.1 displays a plot of $E(Y_{ij}|b_i)$ versus t_{ij} for a random sample of b_i from a normal distribution with zero mean and variance $g_{11} = 4$ (with $\beta_1^* = -1.5$ and $\beta_2^* = 0.75$; $t_{ij} \in [-4, 8]$). Also displayed in Figure 13.1 is a plot of the marginal probability of success, averaged over the distribution of b_i. When the subject-specific logistic curves are compared to the population average curve, it is apparent that the slopes of the former (determined by β_2^*) are steeper than the slope of the latter. Focusing on the range of probabilities from 0.3 to 0.7, where the logistic curves are approximately linear, the slopes for the subject-specific curves rise faster than the slope for the marginal success probabilities. This reinforces the notion that

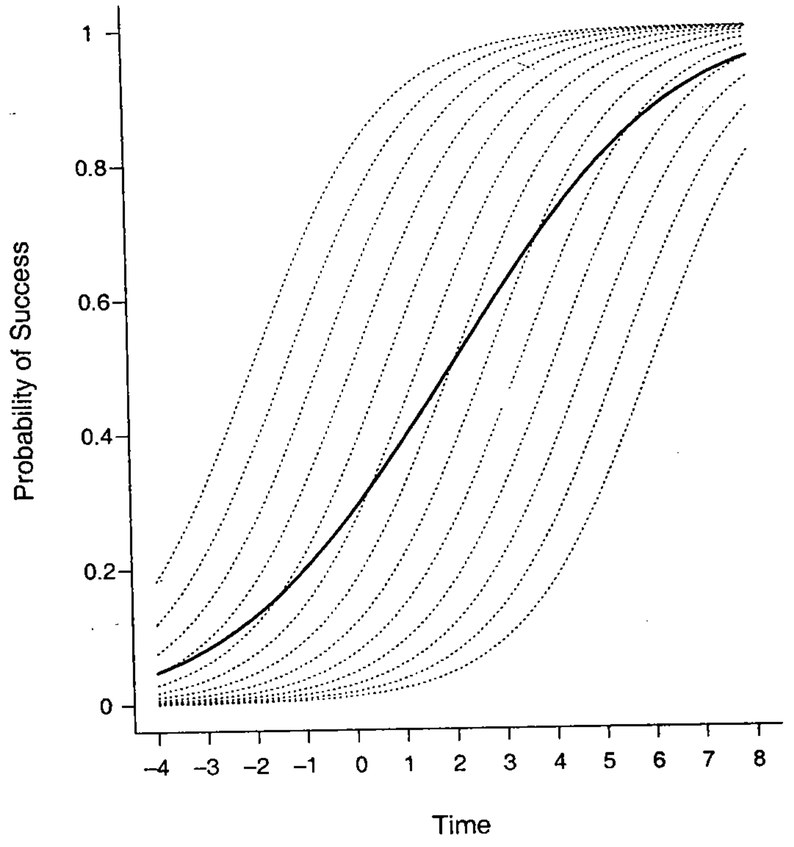

Fig. 13.1 Comparison of conditional probabilities of success (dotted lines) and marginal probability of success (solid line), averaged over the distribution of the random effects.

β^* does not characterize aspects of the population log odds of response, but instead describes changes in the log odds of success for an individual from the population.

13.5 CASE STUDY

This section highlights aspects of interpretation of the regression coefficients in marginal and generalized linear mixed effects models using safety data from a crossover trial on the disease cerebrovascular deficiency, that is, the variable we analyze is not a trial endpoint but rather a potential side effect. In this two-period crossover trial, comparing the effects of active drug to placebo, 67 patients were randomly allocated

Table 13.1 Data from a two-period crossover trial comparing the effects of active drug to placebo. The response indicates whether an electrocardiogram (ECG) was normal (0) or abnormal (1).

	Response (Period 1, Period 2)			
Sequence	(1,1)	(1,0)	(0,1)	(0,0)
Sequence 1 (P → A)	6	0	6	22
Sequence 2 (A → P)	9	4	2	18

P: Placebo; A: Active drug.

Source: Reprinted with permission from Table 3.1 of Jones and Kenward (1989).

to the two treatment sequences, with 34 patients receiving placebo → active, and 33 patients receiving active → placebo. The response variable is binary, indicating whether an electrocardiogram (ECG) was abnormal ($Y = 1$) or normal ($Y = 0$). Thus, each patient has a bivariate binary response vector, $Y_i = (Y_{i1}, Y_{i2})$, where Y_{ij} denotes the response for the i^{th} subject in the j^{th} period (for $i = 1, ..., 67; j = 1, 2$). In Table 13.1, the data are summarized in the form of a 2×4 contingency table.

First, we analyze these data using a marginal model. The marginal mean of the response (or probability of an abnormal ECG) is modelled as a logistic function of the covariates, $\texttt{Treatment}_{ij}$ (0 = Placebo, 1 = Active drug) and $\texttt{Period}_{ij}$ (0 = Period 1, 1 = Period 2),

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 \, \texttt{Treatment}_{ij} + \beta_3 \, \texttt{Period}_{ij}.$$

The within subject association between the two responses is modelled in terms of a common log odds ratio, α,

$$\log \text{OR}(Y_{i1}, Y_{i2}) = \log \left\{ \frac{\Pr(Y_{i1} = 1, Y_{i2} = 1) \Pr(Y_{i1} = 0, Y_{i2} = 0)}{\Pr(Y_{i1} = 1, Y_{i2} = 0) \Pr(Y_{i1} = 0, Y_{i2} = 1)} \right\} = \alpha.$$

The results, obtained using the GEE approach, are presented in Table 13.2. These results indicate that treatment with the active drug is harmful, increasing the rates of abnormal electrocardiograms. The odds of an abnormal electrocardiograms is 1.77 (or $e^{0.57}$) times higher when treated with active drug versus placebo. The estimate of the within-subject association is $\widehat{\alpha} = 3.56$, indicating very strong positive association. That is, the odds of an abnormal electrocardiogram at the second occasion are approximately 35 times higher if the electrocardiogram at the first occasion is abnormal rather than normal.

Table 13.2 Parameter estimates and standard errors from marginal logistic regression model for the ECG data.

Variable	Estimate	SE	Z
Intercept	−1.2433	0.2999	−4.15
Treatment	0.5689	0.2335	2.44
Period	0.2951	0.2319	1.27
log OR(α)	3.5617	0.8148	4.37

Next, we analyze these data using a generalized linear mixed model. In particular, we consider the following logistic regression model for the conditional mean of the response, given a random patient effect,

$$\text{logit}\{E(Y_{ij}|b_i)\} = \beta_1^* + \beta_2^* \, \texttt{Treatment}_{ij} + \beta_3^* \, \texttt{Period}_{ij} + b_i,$$

where the random effect b_i is assumed to have a normal distribution with zero mean and variance, $\text{Var}(b_i) = g_{11}$. In this model each patient is assumed to have some underlying propensity for an abnormal electrocardiogram given by b_i. Then a patient's odds of an abnormal electrocardiogram is multiplied by a common factor $e^{\beta_2^*}$ if they are treated with the active drug, regardless of their underlying propensity. Thus, $e^{\beta_2^*}$ has interpretation in terms of the ratio of a patient's odds of an abnormal electrocardiogram, when treated with the active drug versus placebo.

The ML estimates of the fixed effects and variance component are presented in Table 13.3. These results also indicate that treatment with the active drug is harmful, increasing the patient-specific rates of abnormal electrocardiograms. In particular, a patient's odds of an abnormal electrocardiogram is 6.4 (or $e^{1.86}$) times higher when treated with active drug than when treated with the placebo. This common treatment effect is the same, regardless of the patient's underlying propensity for an abnormal electrocardiogram. The estimate of the variance of b_i, $\widehat{g}_{11} = 24.4$, indicates that there is very substantial between-patient variability in the propensity for abnormal electrocardiograms and this is consistent with the very strong within-subject association found in the results from the marginal model reported in Table 13.2. Note, however, that the variance of b_i is poorly estimated (as evidenced by the very large standard error) since there are only two observations per patient. The estimates of the fixed effects presented in Table 13.3 are all substantially larger than the corresponding estimates in Table 13.2 (approximately three to four times larger), but so too are the standard errors (approximately four times larger). As a result, Wald test statistics for null hypotheses concerning the fixed effects are often quite similar in magnitude for

Table 13.3 Parameter estimates and standard errors from mixed effects logistic regression model for the ECG data.

Variable	Estimate	SE	Z
Intercept	−4.0817	1.6711	−2.44
Treatment	1.8631	0.9269	2.01
Period	1.0376	0.8189	1.27
$g_{11} = \text{Var}(b_i)$	24.4365	18.8500	1.30

ML estimation based on 100-point adaptive Gaussian quadrature.

the two classes of models. In general, it can be shown that the discrepancy between β^* from the logistic regression model with random subject effect and β from a marginal model (with corresponding fixed effects) increases with the variance of the random subject effect. That is, the greater the underlying heterogeneity among subjects, the greater the discrepancy between β^* and β, with $|\beta^*| > |\beta|$ for $\text{Var}(b_i) > 0$.

Comparison of the two estimates of treatment, $e^{\widehat{\beta}_2} = 1.8$ and $e^{\widehat{\beta}_2^*} = 6.4$, from the marginal and mixed effects logistic regression models highlights the distinction between these two analytic approaches. The estimated treatment effect from the marginal model describes how the average rates (expressed in terms of odds) of abnormal electrocardiograms would increase in the study population if patients were treated with the active drug. In contrast, the estimated treatment effect from the mixed effects model describes how the odds of an abnormal electrocardiograms increases for any patient treated with the active drug. As a result, the answer to the question "how harmful is the active drug" will depend on whether scientific interest is in its impact on the study population or on an individual drawn at random from that population.

Finally, note that the first row of Table 13.1 contains a sampling zero. To assess whether the sampling zero has an inordinate effect on the estimate of the treatment effect, a small constant was added to that cell of the table. Specifically, we added $\frac{1}{4}$ to the cell with the sampling zero and repeated the two analyses reported in Tables 13.2 and 13.3. The estimated treatment effect from the marginal model was $e^{\widehat{\beta}_2} = 1.7$, almost identical to the results found in Table 13.2. The estimated treatment effect from the mixed effects model was $e^{\widehat{\beta}_2^*} = 5.2$, somewhat smaller than the estimated treatment effect in Table 13.3. However, from a subject-matter point of view, the substantive conclusions do not change and the difference between the subject-specific and population-averaged effects of treatment is of the same order of magnitude as reported in Tables 13.2 and 13.3.

13.6 CONCLUSION

In Chapters 11 and 12 we considered two types of extensions of generalized linear models to longitudinal data: marginal models and generalized linear mixed models. These two quite different analytic approaches arise from different specifications of, or assumptions about, the joint distribution of Y_i and the source of the correlation among the repeated measures on the same individual. Unlike the linear models for continuous responses considered in Part II, with generalized linear models (and non-linear link functions) for discrete responses different assumptions about the source of the correlation can lead to regression coefficients with quite distinct interpretations.

The basic premise of marginal models is to make inferences about population means, albeit on a transformed scale (e.g., logit or log). The term "marginal" is used to emphasize that the mean response modelled is conditional only on the covariates and not on unobserved random effects or on previous responses. A distinctive feature of marginal models is that the regression models for the mean response and the models for the within-subject association are specified separately. This separation of the model for the mean response from the model for the within-subject association ensures that the marginal model regression coefficients have interpretation that does not depend on the assumptions made about the within-subject association. Specifically, the regression coefficients in marginal models describe the effects of covariates on the population mean response.

In contrast, the basic premise of generalized linear mixed effects models is that there is natural heterogeneity across individuals in the study population in a subset of the regression parameters. That is, a subset of the regression parameters (e.g. intercepts and slopes) are assumed to vary across individuals according to some underlying distribution. But, conditional on the random effects, it is assumed that the repeated measurements for any given individuals are independent observations. Generalized linear mixed models extend the conceptual approach of the linear mixed effects model in a very natural way. The correlation among repeated measurements arises from their sharing of common random effects. Unlike the linear mixed effects model, the regression parameters in generalized linear mixed models have subject-specific, rather than population-averaged, interpretations. That is, due to the non-linear link functions that are usually adopted for discrete responses, the fixed effects do not describe changes in the mean response in the study population. Instead, they describe changes in an individual's mean response and the relation of these changes to covariates. As a result, generalized linear mixed models are most useful when the scientific objective is to make inferences about individuals rather than the study population. For example, the regression parameters in a logistic mixed effects model describe how the log odds of response changes over time, and how these changes relate to covariates, for any individual. Unlike marginal models, they do not describe changes in the log odds of response, averaged over a population of individuals. In summary, with generalized linear mixed models the main focus is on inferences about each individual, while with marginal models the main focus is on inferences about the study population.

The choice between marginal and generalized linear mixed models for longitudinal data can only be made on subject-matter grounds. We have emphasized the different targets of inference for these two classes of models. For any given longitudinal study, different scientific questions will usually demand different analytic models. For example, a physician considering the potential benefits of a novel treatment for one of her patients might be more interested in the subject-specific effect of treatment. On the other hand, public health researchers or health insurance assessors considering the potential reduction in morbidity or mortality in the population if patients receive the novel treatment would be more interested in the population-averaged effect of treatment. When the answers to both of these questions are of interest, there is no contradiction in reporting estimates of both the subject-specific and population-averaged effects.

In summary, we do not prescribe (or proscribe for that matter) one class of models over another. While there has been much debate in the statistical literature concerning the appropriateness of these two classes of models for analyzing longitudinal data, much of the discussion has generated more heat than light. From a purely probabilistic point of view, generalized linear mixed models might appear to have a distinct advantage over marginal models since the marginal distribution of Y_i, the target of inference for marginal models, can, in principle, be derived from the generalized linear mixed effects model by averaging over the distribution of the random effects. However, this apparent advantage is somewhat illusory because the induced marginal model does not, in general, retain the same form. For example, consider a logistic regression model with random effects. The implied model for the marginal mean, averaged over the distribution of the random effects, cannot be a logistic regression model, that is, a logistic regression model with random effect is simply not compatible with a logistic regression model for the marginal means (when averaged over the distribution of the random effects). As a result, regression coefficients that parsimoniously summarize the covariate effects of interest are not readily available. If the goal is to make an inference about the population mean of Y_i, a marginal model should be adopted, thereby avoiding the aforementioned problem, the need to specify the conditional distribution of Y_i given b_i, the marginal distribution of b_i, and the computational demands of integrating over the distribution of the random effects. Thus we find ourselves in substantial agreement with Drum and McCullagh (1993, p. 300) when they comment that:

> "...the megalomaniacal strategy of fitting a grand unified model, supposedly capable of answering any conceivable question that might be posed, is, in our view, dangerous, unnecessary and counterproductive."

The answers to different scientific questions concerning longitudinal change will invariably demand that different statistical models have to be applied to the data at hand. In short, one size does not fit all.

13.7 FURTHER READING

A useful discussion of the distinct interpretations of the regression parameters in marginal and mixed effects models for binary data can be found in Section 12.2 of Agresti (2002).

Bibliographic Notes

Neuhaus *et al.* (1991) compare marginal and mixed effects models for analyzing correlated binary data; also, see Zeger *et al.* (1988), Graubard and Korn (1994), and Section 7.4 of Diggle *et al.* (2002).

Part IV

Advanced Topics for Longitudinal and Clustered Data

14

Missing Data and Dropout

14.1 INTRODUCTION

A major challenge for the analysis of longitudinal data is the problem of missing data. Although most longitudinal studies are designed to collect data on every individual in the sample at each time of follow-up, many studies have some missing observations. In longitudinal studies in the health sciences, missing data are the rule, not the exception. With longitudinal studies problems of missing data are far more acute than in cross-sectional studies, since non-response can occur at any occasion. An individual's response can be missing at one follow-up time and then be measured at a later follow-up time, resulting in a large number of distinct missingness patterns. Alternatively, longitudinal studies often suffer from the problem of attrition or "dropout", that is, some individuals "drop out" or withdraw from the study before its intended completion. In either case, the term "missing data" is used to indicate that an intended measurement could not be obtained.

Missing data have three important implications for longitudinal analysis. First, when longitudinal data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. As a result, missing data create complications for methods of analysis that require balanced data. The imbalance created by missingness is not of any concern for the regression methods described in Parts II and III. Second, when there are missing data there must necessarily be some loss of information. That is, missing data cause a reduction in efficiency or a drop in the precision with which changes in the mean response over time can be estimated. Not surprisingly, the reduction in precision is directly related to the amount of missing data, that is, the greater the

amount of missing data the greater the decrease in precision. The precision can also depend to some extent on the method of analysis; for example, analyses restricted to subjects with complete data will generally be less efficient than methods which use all available data. However, the location of the missing data (e.g., missingness spread sporadically over many subjects, or concentrated at a specific set of time points in a few subjects), and how highly correlated the missing data are with the observed data, will also affect loss of precision. Finally, under certain circumstances missing data can introduce bias and thereby lead to misleading inferences about changes in the mean response. It is this last factor, the potential for serious bias, that complicates the analysis of partially missing longitudinal data. As a result, the reasons for any missing data, often referred to as the *missing data mechanism*, must be carefully considered.

This is the basis for an important theme that will be emphasized throughout this chapter: when data are missing we must carefully consider the reasons for missing data. Some reasons for missing data are relatively benign and do not complicate the analysis, other reasons are not and can potentially introduce bias in the estimates of the regression parameters. The following two examples of partially missing longitudinal data will help to illustrate this point.

The first example is from the Six Cities Study of Air Pollution and Health, discussed in Section 8.8. This was a longitudinal study designed to characterize lung function growth as measured by changes in pulmonary function in children and adolescents. Most of the children were enrolled in the first or second grade (between the ages of six and seven) and measurements were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed. Due to late entry into the study and loss to follow-up or attrition, the number of measurements of pulmonary function of study children varied from a minimum of one to a maximum of twelve. The major reason for late entry or attrition was moving in or out of the school district. Let us focus on this main reason for missing data. If a child changed school district because of employment relocation by her parents, then the missing data mechanism can be thought of as unrelated to the child's pulmonary function. On the other hand, if a child moved out of the school district because she developed respiratory problems (e.g., relocating to an area with either better air quality or improved access to health care), then missingness is related to the child's pulmonary function.

The second example is from the Muscatine Coronary Risk Factor (MCRF) study, discussed in Section 12.5. This was a longitudinal survey of school-age children in Muscatine, Iowa, examining the development and persistence of risk factors for coronary disease. In the MCRF study, weight and height measurements of five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years, were obtained biennially from 1977 to 1981. On the basis of a comparison of their weight to age-gender specific norms, children were classified as obese or not obese. The study protocol required parental consent prior to each measurement. One objective of the MCRF study was to determine whether the risk of obesity increased with age and whether patterns of change in obesity were the same for boys and girls. Although each child was eligible to participate in all three surveys, there was a substantial amount of missing data on obesity, with less than 40% of the children providing complete data at all three measurement occasions. The two main reasons for missing data were: (1) no consent form signed by the parent was received, and (2) the child was not in school on the day of examination. Let us focus on these two reasons for missing data. Suppose that parents of children who were obese were either more or less likely to sign the consent form than parents of children who were not obese. For example, parents of children who were obese may have been more likely to sign the consent form due to concerns about the adverse health effects of obesity; conversely, they may have been less likely to sign the consent form due to concerns that participation in the study could be a source of embarrassment for their children. In either case, the reason for missing data on weight and height is related to the obesity status of the child. Similarly, missingness is related to the obesity status of the child if children who were obese were more likely to be absent on the day of examination (e.g., due to embarrassment about being overweight). On the other hand, suppose a child was absent on the day of examination because of employment relocation by her parents (completely unrelated to the health of the child). Then missingness does not depend on the child's obesity status.

These two examples show that there can be more than a single reason for missing data and that reasons for missing data may or may not be related to the health outcome of interest. When it is unrelated to the health outcome of interest, the impact of missing data is relatively benign and does not complicate the analysis. When it is related to the outcome, somewhat greater care is required because there is potential for bias when individuals with missing data differ in important ways from those with complete data.

In this chapter we review three general types of missing data mechanisms that can be distinguished. The three mechanisms differ in terms of assumptions concerning whether missingness is related to observed and unobserved responses. We also discuss the implicit assumptions about the missing data mechanism that underlie the methods for longitudinal analysis described in Parts II and III. We illustrate the main distinctions between the three types of missing data mechanisms for the common problem of dropout. Finally, we briefly review some alternative methods for handling dropout in longitudinal studies.

14.2 HIERARCHY OF MISSING DATA MECHANISMS

To obtain valid inferences from partially missing longitudinal data, we must consider the nature of the "missing data mechanism". Ordinarily, the missing data mechanism is not under the control of the investigators and often is not well understood. Instead, assumptions are made about the missing data mechanism and the validity of the analysis depends on whether these assumptions hold. When reporting the results of a longitudinal analysis, it is important to be explicit about the assumptions made regarding the reasons for missing data.

The missing data mechanism can be thought of as a probability model for the distribution of a set of response indicator variables. These response indicator variables take the value 1 when an intended response is obtained and the value 0 otherwise. For example, suppose we intend to take n repeated measures of the response variable on

Table 14.1 Schematic representation of R, the vector of response indicators, as a stratification variable.

Response Indicators						Response Vector†					
R_1	R_2	R_3	R_4	$\cdots$	R_n	Y_1	Y_2	Y_3	Y_4	$\cdots$	Y_n
1	1	1	1	$\cdots$	1	y_1	y_2	y_3	y_4	$\cdots$	y_n
1	0	1	1	$\cdots$	1	y_1	*	y_3	y_4	$\cdots$	y_n
1	1	0	1	$\cdots$	1	y_1	y_2	*	y_4	$\cdots$	y_n
1	1	1	0	$\cdots$	1	y_1	y_2	y_3	*	$\cdots$	y_n
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.	.	.	.	.	$\cdots$	.
1	0	0	0	$\cdots$	1	y_1	*	*	*	$\cdots$	y_n
1	0	0	0	$\cdots$	0	y_1	*	*	*	$\cdots$	*

†*Note*: * denotes missing value.

the same individual. A subject with a *complete* set of responses has an $n \times 1$ response vector denoted by

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})';$$

this is a slight abuse of the notation adopted in previous chapters where Y_i contained not the possible set of observations, but what was actually observed. Because of missing data, some of the components of Y_i are not observed for at least some individuals. We let R_i be an $n \times 1$ vector of response indicators

$$R_i = (R_{i1}, R_{i2}, ..., R_{in})',$$

with $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. In addition, associated with Y_i is ar $n \times p$ matrix of covariates, X_i. We do not consider missingness in the covariates, that is, we assume that any time-varying covariates are fixed by the study design. Given R_i, the complete set of responses, $Y_i = (Y_{i1}, \ldots, Y_{in})'$, can be partitioned into two components Y_i^O and Y_i^M, corresponding to those responses that are observed and missing, respectively. That is, Y_i^O denotes the vector of *observed* responses on the i^{th} subject, and Y_i^M denotes the complementary set of responses that are missing. The random vector R_i is recorded for all individuals. Also, given R_i, the target population of interest can be divided or stratified into a number of distinct sub-populations defined by the missing data patterns (including the sub-population of the so-called "completers"). Thus R_i can also be thought of as a stratification variable that divides the target population into a number of sub-populations. This is illustrated in Table 14.1, where the first response, Y_1, perhaps denoting a baseline response, is fully observed, but $Y_2, \ldots, Y_n$ are missing intermittently.

A hierarchy of three different types of missing data mechanisms can be distinguished by considering how R_i is related to Y_i:

(i) *Missing Completely at Random* (MCAR);

(ii) *Missing at Random* (MAR); and

(iii) *Not Missing at Random* (NMAR).

The hierarchy of missing data mechanisms is useful because the type of missing data mechanism determines the appropriateness of different methods of analyses, for example, maximum likelihood, GLS, and GEE. We discuss this topic later in the chapter. However, the nomenclature is not intuitive and leads to much confusion among statisticians and practitioners alike. A major objective of this chapter is to explain these mechanisms in a more intuitive manner so that the reader gains a better appreciation for their usage.

Much of the remainder of this section is devoted to a detailed explanation, with concrete examples, of this classification of missing data mechanisms in the context of longitudinal studies. We begin each description with the formal definition expressed as conditions on the probability distribution of the response indicators, R_i. We then provide some examples and explain the consequences of each type of missingness for the distribution of the observed data. Once the main distinctions are understood, it is then possible to describe the implicit assumptions about the missing data mechanism made by different methods for analyzing longitudinal data.

Missing Completely at Random (MCAR)

Data are said to be missing completely at random (MCAR) when the probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained or the set of observed responses. That is, longitudinal data are MCAR when R_i is independent of both Y_i^O and Y_i^M, the observed and unobserved components of Y_i, respectively. To better understand this missing data mechanism consider the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is assumed to be fully observed, and Y_{i2} is sometimes missing. In that case we require only a single response indicator, with $R_{i2} = 1$ if Y_{i2} is observed and $R_{i2} = 0$ if Y_{i2} is missing. If Y_{i2} is MCAR then

$$\Pr(R_{i2} = 1 | Y_{i1}, Y_{i2}, X_i) = \Pr(R_{i2} = 1 | X_i),$$

and the probability that Y_{i2} is missing does not depend on the observed value of Y_{i1} or the value of Y_{i2} that, in principle, should have been obtained. Missingness in Y_{i2} is simply the result of a chance mechanism that does not depend on observed or unobserved components of Y_i.

An example where partially missing longitudinal data are MCAR is the "rotating panel" study design. In this study design, which is commonly used in health surveys to reduce response burden, individuals rotate in and out of the study after providing a pre-determined number of repeated measures. However, the number and timing

of the measurements is determined by design. The decision about whether to obtain a measurement on an individual at any specific occasion is decided *a priori* by the investigators and is not related to the vector of responses, that is, R_i is unrelated to Y_i. In this example the missing data mechanism is under the control of the investigators and is well understood. Another example where missing data are MCAR is in the Six Cities Study of Air Pollution and Health, when children changed school district because of employment relocation by their parents (for reasons completely unrelated to the health of their children). Here, the reason for missing data is unrelated to the children's pulmonary function.

In the definition of MCAR given above, missingness can depend on the covariates, X_i. This raises a subtle, but important, point. Under MCAR, the response vector Y_i is conditionally independent of R_i, given the covariates X_i. However, this conditional independence of Y_i and R_i may not hold when conditioning on only a subset of the covariates. This has the following important implication. When an analysis is based on a subset of X_i that excludes a covariate that is predictive of R_i, the missing data can no longer be considered MCAR. When

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i|X_i), \tag{14.1}$$

the data are said to have *covariate-dependent* missingness and use of the term MCAR is sometimes restricted to the case where

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i). \tag{14.2}$$

In our discussion of missing data mechanisms we do not make this subtle distinction. Instead, we define MCAR using (14.1) and simply assume that X_i in (14.1) contains all relevant covariates for predicting both Y_i and R_i.

The essential feature of MCAR is that the observed data can by thought of as a random sample of the complete data. The result is that the moments (e.g., means, variances, and covariances) and, indeed, the distributions of the observed data do not differ from the corresponding moments or distributions of the complete data. Thus, "completers" can be regarded as a random sample from the target population, albeit with a smaller sample size than intended. This has important implications for the analysis of longitudinal data restricted to those subjects with complete response vectors. In particular, any method of analysis that yields valid inferences in the absence of missing data will also yield valid inferences when missing data are MCAR and the analysis is restricted to the "completers". The latter is often referred to as a "complete-case" analysis. With data MCAR it is legitimate (although possibly wasteful) to remove subjects with any missing data from the analysis since we can regard them as randomly chosen without regard to their data values. This feature of MCAR allows one to do a complete-case analysis without being concerned that the results might be biased by excluding those with missing data.

A similar result holds for subjects with some missing data. The responses actually obtained, Y_i^O, have the same distribution as the corresponding elements of the completers. As a result, all available data can be used to give valid estimates of moments such as means, variances and covariances. For example, if we modify our bivariate

example to allow data to be MCAR either at time 1 or time 2, then subjects with only one observation can be used along with the complete cases to estimate means and variances. Only the complete cases can here be used to estimate the covariance, but with more observations per subject, the observed cases with at least two observations can be used for covariance estimation. As a result, methods for longitudinal analysis that incorporate all of the available observations will yield valid inferences when missing data are MCAR. This includes all of the methods that were discussed in Parts II and III of this book.

These properties of MCAR follow directly from the definitions in (14.1) and (14.2). They can be used to show that, when the missing data mechanism is MCAR, the distribution of Y_i (given X_i) is the same in each of the distinct sub-populations defined by the missing data patterns (including the sub-population of "completers" or subjects with no missing responses). It also implies that these distributions coincide with the distribution of Y_i (given X_i) in the target population of interest. Moreover, when the missing data mechanism is MCAR, the distribution of the observed components Y_i^O for any distinct sub-population defined by the missing data patterns coincides with the distribution of the same components of Y_i in the target population.

Finally, we note that with MCAR, the distribution of Y_i^M for any distinct sub-population defined by the missing data patterns coincides with the distribution of the same components of Y_i for the "completers". For example, in the bivariate case, MCAR implies that the distribution of Y_{i2} for those missing Y_{i2} is the same as the distribution of Y_{i2} for those with no missing responses.

Missing at Random (MAR)

Data are said to be missing at random (MAR) when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained. In particular, longitudinal data are MAR when R_i is conditionally independent of Y_i^M, given Y_i^O, that is,

$$\Pr(R_i|Y_i^O, Y_i^M, X_i) = \Pr(R_i|Y_i^O, X_i). \tag{14.3}$$

Let us return to the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is fully observed, and Y_{i2} is sometimes missing. If Y_{i2} is MAR then

$$\Pr(R_{i2} = 1|Y_{i1}, Y_{i2}, X_i) = \Pr(R_{i2} = 1|Y_{i1}, X_i),$$

and the probability that Y_{i2} is missing depends on the observed value of Y_{i1}. However, given Y_{i1}, the probability that Y_{i2} is missing does not depend on the value of Y_{i2} that should have been obtained. Put another way, if subjects are stratified on the basis of similar values for Y_{i1}, missingness in Y_{i2} is simply the result of a chance mechanism that does not depend on values of Y_{i2}.

An example where longitudinal data are MAR might arise when a study protocol requires that a subject be removed from the study once the value of an outcome variable falls outside of a certain clinical range of values. In that case missingness in Y_i is under the control of the investigator and is related to observed components of Y_i only.

Another example where missing data are MAR is in the Six Cities Study of Air Pollution and Health, when children moved out of the school district because they developed respiratory problems. If the decision to relocate could be predicted based only on the recorded history of pulmonary function measurements (i.e., the observed components of Y_i only), then the missing data are MAR. However, the MAR assumption would not hold if the decision to relocate was based on some extraneous variable, unavailable to the investigators, that was predictive of the future but unobserved, pulmonary function measurements.

Because the missing data mechanism now depends upon Y_i^O, the distribution of Y_i in each of the distinct sub-populations defined by the missing data patterns is not the same as the distribution of Y_i in the target population. This has important consequences for analysis. One is that a "complete-case" analysis is not valid and can produce biased estimates of change in the mean response over time. Furthermore, the distribution of Y_i^O, the observed components of Y_i, in these sub-populations does not coincide with the distribution of the same components of Y_i in the target population. Therefore, the sample means, variances and covariances based on the available data are biased estimates of the corresponding parameters in the target population. This feature of MAR will be illustrated in the context of dropout in Section 14.4.

With MAR, the observed data cannot be viewed as a random sample of the complete data, but there is an important implication for the distribution of the missing data. The distribution of each individual's missing values, Y_i^M, conditioned on the observed values, Y_i^O, is the same as the conditional distribution of the corresponding observations among the complete cases, conditional on those complete cases having the same values as Y_i^O. In other words, if we stratify on values of Y_i^O, the distribution of Y_i^M is the same as the distribution of the corresponding observations in the complete-case and target populations. As a result, missing values can be validly predicted using the observed data (and a model for the joint distribution). However, the validity of the predictions of the missing values rests upon having correctly specified both the model for the mean and the model for the covariance (when the responses have a multivariate normal distribution). The model for the covariance must be correctly specified because conditional moments (e.g., conditional means) depend upon both the mean response vector and the covariance.

For example, if the responses are assumed to have a multivariate normal distribution, then predictions of the missing values are based on the conditional mean of Y_i^M, given Y_i^O. Using well-known properties of the multivariate normal distribution, the conditional mean of Y_i^M, given Y_i^O, can be expressed as

$$E(Y_i^M|Y_i^O) = \mu_i^M + \Sigma_i^{MO}\Sigma_i^{O^{-1}}(Y_i^O - \mu_i^O),$$

where μ_i^M and μ_i^O denote those components of the mean response vector corresponding to Y_i^M and Y_i^O, and Σ^O and Σ_i^{MO} denote those components of the covariance matrix corresponding to the covariance among the elements of Y_i^O and the covariance between Y_i^M and Y_i^O. The important aspect of the expression given above that we want to emphasize is the dependence of the prediction of Y_i^M on both the mean response vector

$$\mu_i = \begin{pmatrix} \mu_i^O \\ \mu_i^M \end{pmatrix},$$

and the covariance

$$\Sigma_i = \begin{pmatrix} \Sigma_i^O & \Sigma^{OM} \\ \Sigma_i^{MO} & \Sigma^M \end{pmatrix}.$$

When missing data are MAR, we must correctly model the entire joint distribution of Y_i, $f(Y_i|X_i)$ (e.g., both the mean and covariance when Y_i is assumed to have a multivariate normal distribution), to obtain valid estimates of β (and Σ_i).

With MAR, the missing values can be predicted using the observed data and a model for the joint distribution of Y_i. But, one does not need to use the model for $\Pr(R_i|Y_i^O, X_i)$ as a function of X_i and Y_i^O, only a model for Y_i given X_i. Since MCAR is a special case of MAR, the same is also true of MCAR, namely, one does not need to use the model for $\Pr(R_i|Y_i^O, X_i)$ to obtain valid likelihood-based inferences, only a model for $f(Y_i|X_i)$. Notice that not using $\Pr(R_i|Y_i, X_i)$ in the analysis has the important implication that we actually do not need to even posit a specific model for $\Pr(R_i|Y_i, X_i)$ other than to say it does not depend on the missing observations. Since it is common to use a model for $f(Y_i|X_i)$, valid likelihood-based analyses can be obtained with MAR or MCAR data with no extra assumptions, other than the general statement of MCAR or MAR. For this reason, MCAR and MAR are often referred to as *ignorable* mechanisms; the ignorability refers to the fact that once we establish that $\Pr(R_i|Y_i, X_i)$ does not depend on missing observations, we can ignore $\Pr(R_i|Y_i, X_i)$ and obtain a valid likelihood-based analysis provided we have a correct model for $f(Y_i|X_i)$. Alternatively, methods for analyzing longitudinal data that only require a model for the mean response (e.g., GEE methods), but do not specify the joint distribution of the response vector, can be adapted to provide a valid analysis by explicitly modelling $\Pr(R_i|Y_i, X_i)$; these methods are discussed in Section 14.5. A caveat, when data are MAR, it emphatically does not mean we can ignore the missing data problem and use any complete-case or available-data analysis we desire. Appropriate analyses are discussed in the next section.

The subtle distinction between MCAR and MAR is often not well understood. We find that statisticians and empirical researchers constantly confuse the definition of MAR with MCAR (and, admittedly, the choice of terminology has not helped matters). As we shall see in the next section, the distinction between MAR and MCAR has very important implications for the validity of different methods of analysis of longitudinal data. The MAR assumption is far less restrictive on $\Pr(R_i)$ than MCAR and may be considered to be a more plausible assumption about missing data in many applications. Of note, although the MAR assumption is less restrictive in the sense of restrictions on $\Pr(R_i)$, it can be considered more restrictive in terms of what methods of analyses are appropriate. In our view, the MAR assumption should be the default assumption for the analysis of partially missing longitudinal data unless there is a strong and compelling rationale to support the MCAR assumption.

Not Missing at Random

The third type of missing data mechanism is referred to as *not missing at random* (NMAR). Missing data are said to be NMAR when the probability that responses are missing is related to the specific values that should have been obtained. That is, the conditional distribution of R_i, given Y_i^O, is related to Y_i^M and

$$\Pr(R_i|Y_i^O, Y_i^M, X_i)$$

depends on at least some elements of Y_i^M. Let us return to the bivariate case where $Y_i = (Y_{i1}, Y_{i2})'$, Y_{i1} is fully observed, and Y_{i2} is sometimes missing. If missingness in Y_{i2} is NMAR, then

$$\Pr(R_{i2} = 1|Y_{i1}, Y_{i2}, X_i)$$

depends on the potentially unobserved value of Y_{i2}. That is, given Y_{i1}, the probability that Y_{i2} is missing depends on the value of Y_{i2} that should have been obtained. An NMAR mechanism is often referred to as *nonignorable* missingness. The term *nonignorable* refers to the fact that the missing data mechanism cannot be ignored when the goal is to make inferences about the distribution of the complete longitudinal responses.

An example where longitudinal data are NMAR arises when the outcome variable is a measure of "quality-of-life" and subjects fail to complete the instrument or questionnaire on occasions when their quality-of-life is compromised. Another example where missing data are NMAR is in the Muscatine Coronary Risk Factor (MCRF) study, when parents of children who were obese were either more or less likely to sign the consent form than parents of children who were not obese. In that case, missingness on weight and height is related to the obesity status of the child.

Sometimes the term *informative* is used to describe data that are NMAR; missingness is informative in the sense that the missingness (i.e., a component of R_i is equal to 0) informs us about the distribution of the missing observations. Specifically, the distribution of Y_i^M, conditional on Y_i^O, is not the same as that in the completers or in the target population, but rather, the distribution of Y_i^M depends upon Y_i^O and on $\Pr(R_i|Y_i, X_i)$. Thus, the model assumed for $\Pr(R_i)$ is crucial, it must be included in the analysis, and the specific model chosen can drive the results of the analysis.

14.3 IMPLICATIONS FOR LONGITUDINAL ANALYSIS

The statistical methods for analyzing longitudinal data described in Parts II and III can accommodate incomplete data. However, valid inferences from partially missing longitudinal data require assumptions about the missing data mechanism. In this section we summarize the key assumptions about missing data required for valid inferences when applying the techniques described in Parts II and III.

When the missing data mechanism is MCAR, individuals with missing data are a random subset of the sample. In this case the observed values of the responses are a random subsample of all values of the responses, and no bias will arise with

almost any method of analysis of the data (either the available data or the data on the completers only). In particular, all of the methods discussed in Part II and III will yield valid estimates of mean response trends (and within-subject associations) if the missing data can be assumed to be MCAR.

When the missing data mechanism is MAR, individuals with missing data are no longer a random subset of the sample. Only when stratified on their observed outcomes (i.e., conditional on Y_i^O) can they be considered a random subset of the sample belonging to that stratum. As a result the observed values are not necessarily a random subsample of the responses. In particular, the distribution of Y_i^O, the observed components of Y_i, differs from the distribution of the same components of Y_i in the target population. This implies that the sample means at each occasion (and the covariances) based on the available observations provide biased estimates of the means (and covariances) in the target population. Similarly, analyses restricted to the data from the completers also yield biased estimates of the means (and covariances). When missing data are MAR, but not MCAR, complete-case methods and standard GEE methods based on all of the available observations yield biased estimates of mean response trends. In contrast, likelihood-based methods that correctly specify the entire joint distribution of the responses yield valid estimates when missing data are MAR. However, there is a subtle, but important, proviso: the entire joint distribution must be correctly specified. In practice, this has the important implication that not only must the model for the mean response be correctly specified, but also the model for the within-subject association. Thus, when missing data are MAR, the likelihood-based methods discussed in Part II provide valid inferences about changes in the mean response over time provided that the covariance matrix has been correctly modelled. Similarly, the methods discussed in Chapter 12 provide valid estimates of the fixed effects provided that the random effects structure has been correctly specified. In summary, when missing data are MAR, but not MCAR, inferences about the mean response are sensitive to any form of misspecification of the joint distribution of the vector of responses. Accordingly, if longitudinal data are incomplete, somewhat greater care must be exercised when modelling the within-subject association.

The standard GEE approach requires that we have a model for the expected value of the observations given the covariates. With MAR, the model will generally not hold for the observed data, so the validity of the analysis is compromised. Methods have been devised for making adjustments to the analysis by using a simple weighted GEE. The weights have to be estimated using a model for $\Pr(R_i|Y_i, X_i)$, hence the non-response model must be explicitly specified and estimated, although the distribution of the error terms need not be. These weighting methods are discussed in Section 14.5.

Finally, when longitudinal responses are NMAR, almost all standard methods of longitudinal analysis are not valid. Both GEE methods and standard likelihood-based methods (that ignore the missing data mechanism) yield biased estimates of mean response trends. To obtain valid estimates, joint models for the response and the missing data mechanism are required. Indeed, the term *nonignorable* is used to emphasize that the missing data mechanism must be specified (i.e., cannot be ignored) for inferences about the complete responses. We must also stress that any assumptions made about

non-response being NMAR are completely unverifiable from the data at hand. That is, without external information about the reasons for missingness, the observed data provide no information that can either support or refute one NMAR mechanism over another. So, short of tracking down the missing data, any assumptions made about the missingness process are not verifiable. Therefore, when missingness is thought to be NMAR, it is important to carefully assess the sensitivity of inferences to a variety of plausible assumptions concerning the missingness process. However, sensitivity analysis under different assumptions about NMAR missingness is a topic that goes well beyond the scope of this chapter.

14.4 DROPOUT

As mentioned earlier, most longitudinal studies are designed to collect data on every individual in the sample at a planned sequence of occasions. However, longitudinal studies habitually suffer from the problem of attrition; that is, some individuals "drop out" of the study prematurely. The term *dropout* refers to the special case where if Y_{ik} is missing, then $Y_{ik+1}, ..., Y_{in}$ are also missing. Alternatively, when expressed in terms of the response indicators, dropout refers to the case where if $R_{ik} = 0$ then $R_{ik+1} = \cdots = R_{in} = 0$. This gives rise to the monotone missing data pattern displayed in Figure 14.1, in contrast to the non-monotone patterns that arise when data are missing intermittently. Note that intermittent missing data give rise to a considerably larger number of potential missing data patterns, but apart from that, do not raise any further technical considerations. As a result, the focus of the remainder of this chapter is on dropout.

When there is dropout in a longitudinal study, the key issue is whether those who "drop out" and those who remain in the study differ in any further relevant way. If they do not, then analyses restricted to those remaining in the study yield valid, albeit inefficient, inferences. If they do differ, then such "complete-case" analyses are potentially biased.

In the previous section three different types of missing data mechanisms were distinguished. The same taxonomy can be applied to dropout. That is, dropout can be *completely at random*, *at random*, or *not at random*. When dropout is completely at random the probability of dropout at each occasion is independent of all past, current, and future outcomes (given the covariates). With completely random dropout, an individual leaves the study in a process which is unrelated to that individual's outcomes. In contrast, when dropout is at random the probability of dropout at each occasion can depend on the previously observed outcomes up to, but not including, the current occasion. However, given the observed outcomes, dropout is assumed to be independent of the current and future outcomes. That is, with random dropout the process can depend on the outcomes that have been observed in the past, but given this information, it is unrelated to all future (unobserved) values of the outcome variable following dropout. Finally, when dropout is not at random, the probability of dropping out at each occasion can depend on current and future unobserved outcomes. That is, dropout is said to be not at random when the process depends on the unrecorded

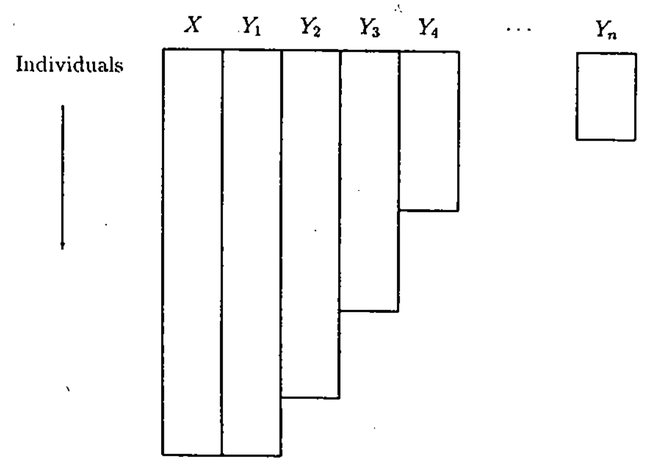

Fig. 14.1 Schematic representation of a monotone missing data pattern for dropout, with Y_j more observed than Y_{j+1} for $j = 1, ..., n-1$.

values of the outcome variable that would have been observed had the individual remained in the study. In the context of dropout in a longitudinal study, the term "informative" dropout often is used to refer to dropout that is NMAR (similarly, "non-informative" dropout often is used to refer to dropout that is either random or completely random). Here the fact of dropout is informative about the distribution of future observations. For example, consider two subjects with the same past history of responses (and covariates) up to time t. One drops out and the other does not. With MAR, the distribution of their future observations is the same. In contrast, dropout that is NMAR informs us that the distributions of the future observations will differ. In the most general case, nothing in the data can be used to determine the distribution of the future observations of the dropouts, hence the analysis depends strongly on the specification of $\Pr(R_i)$.

Illustration

To emphasize the main distinctions between the three types of dropout mechanism, and their potential impact on a longitudinal analysis, we consider the following simple illustration. Suppose repeated measurements, Y_{it} $(i = 1, ..., N; t = 1, ..., 5)$, are generated from a multivariate normal distribution with mean response

$$E(Y_{it}) = \mu_{it} = \beta_1 + \beta_2 t,$$

and covariance

$$\text{Cov}(Y_{is}, Y_{it}) = \rho^{|s-t|}; \quad \text{for } \rho \geq 0.$$

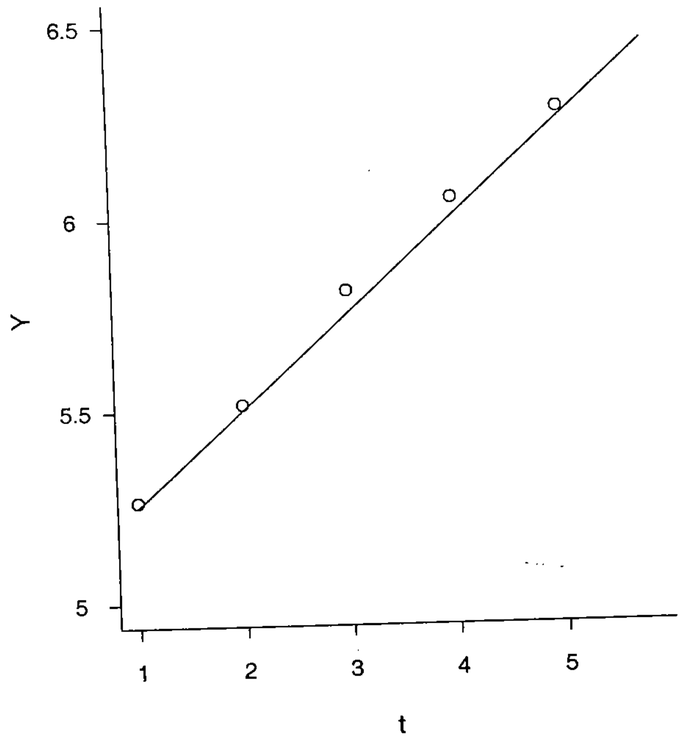

Fig. 14.2 Population regression line and empirical means at each occasion for simulated complete data.

That is, the variance at each occasion is 1 and assumed to be constant over time, while the correlations have a first-order autoregressive pattern. Figure 14.2 displays sample means of simulated data from this model, with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, and $\rho = 0.7$. The empirical means (i.e., the sample means at each occasion) show a clear increasing trend over time and virtually coincide with the population regression line (the solid line in Figure 14.2).

Next, suppose there is dropout. When there is dropout we can replace the vector of response indicators, R_{it} $(t = 1, ..., 5)$, with a simple dropout indicator variable, D_i, for each individual. The random variable D_i is recorded for all individuals and $D_i = k$ if an individual drops out between the $(k-1)^{th}$ and k^{th} occasion, that is, only the first $D_i - 1$ responses are observed. Assume that

$$\log\left\{\frac{\Pr(D_i = k | D_i \geq k, Y_{i1}, ..., Y_{ik})}{\Pr(D_i > k | D_i \geq k, Y_{i1}, ..., Y_{ik})}\right\} = \theta_1 + \theta_2 Y_{ik-1} + \theta_3 Y_{ik}.$$

This model specifies that the probability of dropout at any occasion, given dropout has not previously occurred, can depend on the current value and the prior value of

the response variable. We assume that the first response, Y_{i1}, is fully observed, that is, $\Pr(D_i = 1 | D_i \geq 1, Y_{i1}) = 0$. In terms of this model for dropout, consider the following three missing data mechanisms:

(a) Dropout is MCAR: $\theta_2 = \theta_3 = 0$;

(b) Dropout is MAR: $\theta_3 = 0$; and

(c) Dropout is NMAR: $\theta_3 \neq 0$.

Figure 14.3(a) displays simulated date from this model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is MCAR (with $\theta_1 = -0.5$, and $\theta_2 = \theta_3 = 0$). The conditional probability of dropout at the second through fifth occasions is 0.38 (or $\frac{e^{-0.5}}{1+e^{-0.5}}$). This results in approximately 38% of the responses missing at the second occasion, 61% missing at the third occasion, 76% missing at the fourth occasion, and 85% missing at the fifth occasion. Despite the large proportion of missing data, the empirical means at each occasion show a clear linearly increasing trend over time and almost coincide with the population regression line (solid line). Recall that when the missing data mechanism is MCAR, individuals with missing data can be considered a random subset of the sample, and no bias will arise with almost any method of analysis of the observed data (either the complete data or the available data). That is, all of the methods discussed in Part II and III will yield valid inferences when missing data are MCAR. To reinforce this point, the estimates of the regression parameters obtained using maximum likelihood (ML) estimation, with correctly specified covariance structure, and using a "working independence" GEE estimator are displayed at the top of Table 14.2. Recall that for a linear model with a "working independence" assumption for the covariance, the GEE estimator is identical to the ordinary least squares (OLS) estimator. As expected, both the ML and OLS (or "working independence" GEE) estimates of the intercept and slope are very close to the true values of the population parameters used to generate the data. The minor differences are simply due to sampling variability.

Figure 14.3(b) displays simulated date from the same model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is MAR (with $\theta_1 = -0.5$, $\theta_2 = 0.5$, and $\theta_3 = 0$). Here dropout at any occasion depends upon the previous response, but not the current response. Because the high responders progressively drop out ($\theta_2 > 0$), the empirical means at the second through fifth occasions are discernibly lower than the population regression line. As a result, available-data methods such as the GEE will yield biased estimates of mean response trends. In contrast, likelihood-based methods will yield valid estimates when missing data are MAR (or MCAR) and the model for the covariance has been correctly specified. The ML and GEE estimates of the intercept and slope are displayed in the middle of Table 14.2. The ML estimates are very close to the values of the population parameters and only differ due to sampling variability. On the other hand, the "working independence" GEE (or OLS) estimate of the slope shows very discernible bias and underestimates the rate of change over time ($\widehat{\beta}_2 = 0.18$ versus $\beta_2 = 0.25$).

Finally, Figure 14.3(c) displays simulated date from the same model (with $N = 1000$, $\beta_1 = 5$, $\beta_2 = 0.25$, $\rho = 0.7$) when dropout is NMAR (with $\theta_1 = -0.5$,

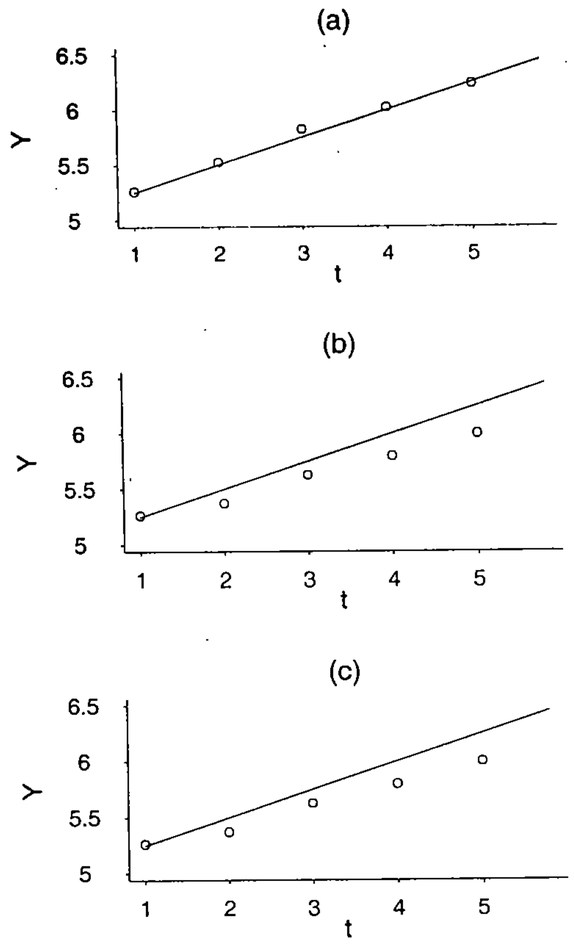

Fig. 14.3 Population regression line and observed data means at each occasion for simulated data when dropout is (a) completely at random (MCAR), (b) at random (MAR), and (c) not at random (NMAR).

$\theta_2 = 0$, and $\theta_3 = 0.5$). Here, dropout at any occasion depends upon the current value of the response. Because the high responders progressively drop out ($\theta_3 > 0$), the empirical means are discernibly lower than the population regression line. As a result, available-data methods such as the GEE yield biased estimates of mean response trends. Furthermore, likelihood-based methods that ignore the missing data mechanism also yield biased estimates of mean response trends. To reinforce this point, the ML and GEE estimates of the regression parameters are displayed at the bottom of Table 14.2. The ML and GEE estimates of the slope show large biases. Of note, the magnitude of the bias is somewhat smaller for ML; however, this cannot be

Table 14.2 Parameter estimates and standard errors for correctly specified likelihood analysis (ML) and "working independence" analysis (OLS/GEE) based on simulated data when dropout is (i) completely at random, (ii) at random, and (iii) not at random. The true regression parameters are $\beta_1 = 5.0$ and $\beta_2 = 0.25$.

		ML		OLS/GEE	
Dropout	Parameter	Estimate	SE	Estimate	SE†
MCAR	Intercept	5.015	0.031	5.022	0.032
	t	0.257	0.016	0.253	0.018
MAR	Intercept	5.003	0.041	5.062	0.043
	t	0.261	0.016	0.182	0.018
NMAR	Intercept	5.058	0.040	5.071	0.043
	t	0.201	0.016	0.162	0.018

† Standard errors for OLS/GEE are based on sandwich variance estimator.

expected in general unless the correlation among the responses is very high. That is, when the correlation among the responses is very high and dropout at any occasion depends only upon the current value of the response, the dropout mechanism can often be approximated by an ignorable dropout mechanism that conditions on all previously observed responses

$$\Pr(D_i = k | D_i \geq k, Y_{ik}) \approx \Pr(D_i = k | D_i \geq k, Y_{i1}, \ldots, Y_{ik-1}).$$

For example, when the data are simulated from the same model but with a correlation parameter $\rho = 0.9$ instead of $\rho = 0.7$, the ML estimate of the slope is $\widehat{\beta}_2 = 0.24$ and the magnitude of the bias is significantly reduced. In contrast, the OLS/GEE estimate of the slope is $\widehat{\beta}_2 = 0.11$ and remains highly biased under this NMAR dropout mechanism.

14.5 COMMON APPROACHES FOR HANDLING DROPOUT

In this section we present a short review of some of the most commonly used methods for handling dropout in longitudinal analysis. We also discuss the assumptions about dropout required for each of the methods to yield valid inferences. We note that many traditional methods for handling missing data (e.g., complete-case analysis,

imputation) became popular when the only approaches for analyzing data were ones based on complete and balanced data.

Complete-Case Analysis

One approach to handling dropout is to simply exclude all data from the analysis on any subject who drops out. That is, a so-called "complete-case" analysis can be performed by excluding any subjects that do not have data at all intended measurement occasions. We must stress that this method is very problematic and is rarely an acceptable approach to the analysis. It will yield unbiased estimates of mean response trends only when it can be assumed that dropout is MCAR. Recall that when dropout is MCAR, the study completers are a random subsample of the original sample from the population. However, even in cases where the MCAR assumption might be tenable, a complete-case analysis is very unappealing because of the reduction in the number of subjects contributing to the analysis. A complete-case analysis can be immensely inefficient, leading to an analysis with reduced statistical power.

Available-Data Analysis

Another approach for handling dropout is the "available-data" method. This is not a single method, but a very general term that refers to a wide collection of techniques that can readily incorporate vectors of repeated measures of unequal length in the analysis. For example, standard applications of generalized least squares (GLS) or the generalized estimating equations approach can be considered available-data methods, since these approaches base the analysis on all of the available observations. In general, available-data methods are more efficient than complete-case methods because they incorporate the partial information obtained from those who dropout. However, many available-data methods will yield valid analyses only if the conditional (i.e., conditional on X_i) means and covariances of the observed components of Y_i among those who dropout coincide with the corresponding conditional means and covariances of Y_i in the target population. As a result, available-data methods will yield biased estimates of mean response trends unless dropout is MCAR. In general, for available-data (and complete-case) methods to be valid we require that dropout is MCAR.

Imputation

A third approach, and one that is widely used in practice, is some form of imputation for the missing responses following dropout. The idea behind imputation is very simple: substitute or fill-in the values that were not recorded with imputed values. One of the chief attractions of imputation methods is that, once a filled-in data set has been constructed, standard methods for complete data can be applied. However, methods that rely on just a single imputation, creating only a single filled-in data set, fail to acknowledge the uncertainty inherent in the imputation of the unobserved

responses. Multiple imputation circumvents this difficulty. In multiple imputation the missing values are replaced by a set of m plausible values, thereby acknowledging the uncertainty about what values to impute for the missing responses. Typically, a small number of imputations, for instance, $5 \leq m \leq 10$, is sufficient to obtain realistic estimates of the sampling variability.

With multiple imputation, m filled-in data sets are created, producing m different sets of parameter estimates and their standard errors. These are then appropriately combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation of the unobserved responses. Specifically, a single estimate of the regression parameters is obtained by taking the arithmetic average of the estimates obtained from the m filled-in data sets. Letting $\widehat{\beta}^{(k)}$ and $\widehat{\mathrm{Cov}}(\widehat{\beta}^{(k)})$ denote the estimate of β and the estimated covariance of $\widehat{\beta}^{(k)}$ from the k^{th} filled-in data set (for $k = 1, ..., m$), a single estimate of β is given by

$$\overline{\beta} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}^{(k)},$$

and the estimated covariance of $\overline{\beta}$ is given by

$$\frac{1}{m}\sum_{k=1}^{m}\widehat{\mathrm{Cov}}(\widehat{\beta}^{(k)}) + \left(1+\frac{1}{m}\right)\frac{1}{m-1}\sum_{k=1}^{m}\left(\widehat{\beta}^{(k)}-\overline{\beta}\right)\left(\widehat{\beta}^{(k)}-\overline{\beta}\right)'.$$

Although the latter expression for calculating the standard errors appears somewhat complicated, it simply combines two inherent sources of variability: the within-imputation variance and the between-imputation variance. The main idea behind multiple imputation is very simple; what is less clear-cut is how to produce the imputed values for the missing responses. Next, we consider some of the commonly used methods for imputing missing data.

One widely used imputation method, especially in longitudinal clinical trials, is "last value carried forward" (LVCF), occasionally referred to as "last observation carried forward" (LOCF). This is a single imputation method that fills-in or imputes the missing values following dropout with the last observed value for that subject. Despite its widespread use, it should be recognized that LVCF makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value prior to dropout. Perhaps the only setting where this assumption might conceivably be appropriate is when dropout is due to recovery or cure. In the context of placebo-controlled longitudinal clinical trials, there appears to be some statistical folklore that LVCF yields a *conservative* estimate of the comparison of an active treatment versus the control. However, this is a gross misconception, and will only be true to the extent that the active treatment prior to dropout has carry-over effects following dropout. In many clinical trials, this is unlikely to be the case; instead, dropout from the active treatment (e.g., due to adverse side effects) might very well result in a deterioration of the response. Despite frequent and well-founded criticisms by statisticians, LVCF is still widely used to handle dropouts in clinical trials. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) seem

to encourage the continuing use of LVCF as a method for handling dropouts, despite all of its obvious shortcomings. Except in very rare cases (as mentioned above), we do not recommend the use of LVCF as a method for handling dropout.

Variations on the LVCF theme include baseline value carried forward and worst value carried forward. Worst value carried forward is most often used in comparisons of an active treatment to a placebo, since it is assumed to be conservative in that setting. However, both of these alternatives suffer the same difficulties as LVCF and cannot be counted on to give unbiased treatment estimates. In addition, all of the methods suffer from optimistic standard error estimates. It is easy to see that these analyses give smaller standard errors than complete-case, or even available-data estimates because they assume complete data on everyone. However they will generally give smaller standard errors than what we would expect if we had been fortunate enough to have complete data on everyone. This is because the variability of baseline measurements is usually smaller because of selection criteria into the study, and as we move out in time, the observations tend to become more variable. Hence substituting baseline or intermediate values for final values can be expected to give a less variable data set. It is also true if we use worst value, since worst values are often similar especially for responses based on a scale.

There are other imputation methods that have a much firmer theoretical foundation by drawing values of Y_i^M from the conditional distribution of the missing responses given the observed responses, $f(Y_i^M|Y_i^O, X_i)$. With the monotone missing data patterns produced by dropouts, it is relatively straightforward to impute missing values by drawing values of Y_i^M from $f(Y_i^M|Y_i^O, X_i)$ in a sequential manner. A variety of imputation methods can be used to draw values from $f(Y_i^M|Y_i^O, X_i)$. One approach is known as the propensity score method. In propensity score methods, values to impute for the missing responses are obtained from observations on subjects who are equally likely to dropout (but who do not at that occasion). Propensity score methods require a model for the propensity or probability of dropout. For example, it might be assumed that

$$\log\left\{\frac{\Pr(D_i = k|D_i \geq k, Y_{i1}, ..., Y_{ik})}{\Pr(D_i > k|D_i \geq k, Y_{i1}, ..., Y_{ik})}\right\} = \theta_1 + \theta_2 Y_{ik-1}.$$

The estimated propensities (or probabilities of dropout) from this model are then used as follows. At each occasion the missing responses for subjects who dropout are imputed from the responses of subjects who do not dropout but have similar estimated propensities. The process is repeated sequentially until a filled-in data set is obtained; multiple imputations are obtained by repeating these steps m times. Another approach to imputation is known as predictive mean matching. In predictive mean matching, a series of regression models for Y_{ik}, given $Y_{i1} \ldots Y_{ik-1}$, are fit using the observed data. For example, when Y_{ik} is continuous, a series of linear regression models

$$E(Y_{ik}) = \gamma_1 + \gamma_2 Y_{i1} + \cdots + \gamma_k Y_{ik-1},$$

can be fit using the observed data on subjects who have not dropped out by the k^{th} occasion, yielding $\widehat{\gamma}$ and $\widehat{\sigma}^2$ (the latter is the residual variance from the linear

regression model). Parameters γ^* (and σ^*) are then drawn from the distribution of $\widehat{\gamma}$ (and $\widehat{\sigma}$), to account for the uncertainty in estimating γ (and σ). Missing values for Y_{ik} can then be imputed on the basis of the following predictions:

$$\gamma_1^* + \gamma_2^* Y_{i1} + \cdots + \gamma_k^* Y_{ik-1} + \sigma^* e_i,$$

where e_i is simulated from a standard normal distribution. Imputing values for missing Y_{ik} based on these predicted values represents a regression method for imputation; predictive mean matching imputes using the observed values from the data at hand that are closest to the predicted values. Multiple imputations are obtained by repeating these steps m times. The general approach can be extended to discrete responses by considering a series of generalized linear models

$$g\{E(Y_{ik})\} = \gamma_1 + \gamma_2 Y_{i1} + \cdots + \gamma_k Y_{ik-1},$$

where $g(\cdot)$ is some known link function.

When missing values are imputed from $f(Y_i^M|Y_i^O, X_i)$, regardless of the particular imputation method adopted, subsequent analyses of the observed and imputed data are valid when dropouts are MAR (or MCAR). Furthermore, multiple imputation ensures that the uncertainty is properly accounted for.

Finally, there is another related form of imputation where the missing responses are effectively imputed by modelling and estimating parameters for the joint distribution of Y_i, $f(Y_i|X_i)$. When dropout is MCAR or MAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data. If dropout is MCAR or MAR (and the parameters of the dropout and outcome processes are distinct, a technical requirement that can usually be assumed in practice), then ML estimates can be obtained by maximizing $f(Y_i^O|X_i)$, where $f(Y_i^O|X_i)$ denotes the ordinary marginal distribution of the particular subset of Y_i determined by Y_i^O. In a certain sense, the missing values are validly predicted by the observed data via the model for the conditional mean, $E(Y_i^M|Y_i^O)$. In an implementation of ML known as the *EM algorithm*, the two-step iterative algorithm alternates between filling-in missing values with their conditional means, given the observed responses and parameter estimates from the previous iteration (the expectation of E-step), and maximizing the likelihood for the resulting "complete data" (the maximization or M-step). For example, if the responses are assumed to have a multivariate normal distribution, then predictions of the missing values are based on the conditional mean of Y_i^M, given Y_i^O,

$$E(Y_i^M|Y_i^O) = \mu_i^M + \Sigma_i^{MO}\Sigma_i^{O^{-1}}(Y_i^O - \mu_i^O),$$

where μ_i^M and μ_i^O denote those components of the mean response vector corresponding to Y_i^M and Y_i^O, and Σ^O and Σ_i^{MO} denote those components of the covariance matrix corresponding to the covariance among the elements of Y_i^O and the covariance between Y_i^M and Y_i^O. Thus, when dropout is MAR, likelihood-based inference does not require specification of the dropout mechanism and the contribution of $\Pr(D_i|Y_i^O, X_i)$ to the likelihood can be ignored. However, recall that likelihood-based approaches do require full distributional assumptions about Y_i and the model for $f(Y_i|X_i)$ must be correctly specified (e.g., any misspecification of the covariance will, in general, yield biased estimates of the mean response trend).

Weighting Methods

An alternative approach for handling dropout is to weight the observed data in some appropriate way. In weighting methods, the under-representation of certain response profiles in the observed data is taken into account and corrected. A variety of different weighting methods that adjust for dropout have been proposed. These approaches are often called propensity weighted or inverse probability weighted methods. Here the underlying idea is to base estimation on the observed responses but weight them to account for the probability of remaining in the study. The propensities for dropout can be estimated as a function of the observed responses prior to dropout, and also as a function of the covariates and any additional variables or subject characteristics that are thought likely to predict dropout.

In the simplest version of this approach we need to estimate $w_i = \Pr(D_i = n+1)$ only for those who complete the study. As a result, w_i can be computed sequentially as the following product of the propensities for dropout:

$$w_i = (1 - \pi_{i1}) \times (1 - \pi_{i2}) \times \cdots \times (1 - \pi_{in}),$$

where $\pi_{ik} = \Pr(D_i = k | D_i \geq k)$ can be estimated from those remaining at the $(k-1)^{th}$ occasion, given the recorded history of all available data up to the $(k-1)^{th}$ occasion. Given $\widehat{w}_i$, the estimated probability that the i^{th} subject completes the study, a weighted complete-case analysis can be performed. For example, the GEE approach can be adapted to handle data that are MAR by making adjustments to the analysis for the propensities for dropout. One variant of this approach is to use a weighted GEE, with weights inversely proportional to the estimated propensities for dropout, to analyze the data from the "completers". In the weighted complete-case analysis each subject's contribution to the analysis is weighted inversely by $\widehat{w}_i$, thereby providing valid estimates of the mean response trends when dropout is MAR.

Inverse probability weighted methods were first proposed in the sample survey literature, where the weights are known and based on the survey design. The intuition behind the weighting methods is that each subject's contribution to the weighted complete-case analysis is replicated $\frac{1}{w_i}$ times, in order to count once for herself and $(\frac{1}{w_i} - 1)$ times for those subjects with the same history of responses and covariates, but who do not complete the study. For example, a subject with weight of $\frac{1}{4}$ (or $w_i = 0.25$) has a probability of completing the study of 0.25. As a result, in a complete-case analysis data from this subject should count once for herself and 3 times for those subjects who do not complete the study (recall that if the probability of completing the study is $\frac{1}{4}$, it means that 3 subjects are expected to dropout for every one that completes the study). In general, the weighting methods are valid provided that the model that produces the estimated w_i is correctly specified.

In contrast to sample surveys, note that w_i is not ordinarily known, but must be estimated from the observed data (e.g., using a repeated sequence of logistic regressions for the π_{ik}'s). Therefore, the variance of inverse probability weighted estimators must also account for estimation of w_i. Counter-intuitively, failure to account for the estimation of these weights will, in general, result in standard errors that are too large (i.e., estimation of the weights from the data at hand leads to

improvements in precision); however, this is a topic that goes beyond the scope of this chapter. Finally, we note that this general approach for handling dropout can also be made more efficient by conducting an appropriately weighted available-data analysis. If we let w_{ik} denote the probability that the i^{th} subject is still in the study at the k^{th} occasion, then

$$\begin{aligned} w_{i1} &= (1 - \pi_{i1}) \\ w_{i2} &= (1 - \pi_{i1}) \times (1 - \pi_{i2}) \\ &\vdots \quad \vdots \qquad\qquad \vdots \\ w_{in} &= (1 - \pi_{i1}) \times (1 - \pi_{i2}) \times \cdots \times (1 - \pi_{in}), \end{aligned}$$

and the available data at the k^{th} occasion are weighted by $\frac{1}{w_{ik}}$ in the analysis.

14.6 CASE STUDY

Next, we apply some of the methods described earlier for handling dropouts. The methods are applied to data from the longitudinal clinical trial of contracepting women (Machin *et al.*, 1988) discussed in Section 12.5. Recall that the goal of this trial was to compare the two treatments (100 mg or 150 mg of DMPA) in terms of how the rates of amenorrhea change over time *with continued use of the contraceptive method.* That is, the main interest is in an analysis that compares the rates of amenorrhea over time if those women who dropped out had remained on their assigned treatment. This is sometimes called an *explanatory* analysis (Schwartz and Lellouch, 1967). An "explanatory analysis", often referred to as an "as treated" analysis, focuses on what is thought to be the true underlying biological effects of the different treatments.

A total of 1151 women completed menstrual diaries and the diary data were used to generate a binary sequence for each woman, indicating whether or not she had experienced amenorrhea in four successive intervals. As indicated, a feature of this clinical trial is that there was substantial dropout. When the dropout rates are broken down by dosage group, the rates were marginally higher in the 150 mg dose group. For those women randomized to 100 mg (150 mg) of DMPA 37% (39%) dropped out before the completion of the trial; 17% (17%) dropped out after receiving only one injection of DMPA, 12% (15%) dropped out after receiving only two injections, and 8% (6%) dropped out after receiving three injections. For women who dropped out before the end of the 3-month interval between injections, a determination of whether or not they experienced amenorrhea was made, on a proportionate basis, using their existing menstrual diary data for that interval.

Letting $Y_{ij} = 1$ if the i^{th} woman experienced amenorrhea in the j^{th} injection interval, we considered the following logistic regression model for the marginal mean:

$$\text{logit}(\mu_{ij}) = 0 = \beta_1 + \beta_2 \mathtt{t_{ij}} + \beta_3 \mathtt{t_{ij}^2} + \beta_4 \mathtt{dose_i} + \beta_5 (\mathtt{t_{ij}} \times \mathtt{dose_i}) + \beta_6 (\mathtt{t_{ij}^2} \times \mathtt{dose_i}),$$

where $\mu_{ij} = \Pr(Y_{ij} = 1)$, $\mathtt{t} = 0, 1, 2, 3$ for the four consecutive 3-month injection intervals, $\mathtt{dose} = 1$ if randomized to 150 mg of DMPA, and $\mathtt{dose} = 0$ otherwise.

Table 14.3 Estimated marginal rates of amenorrhea for quadratic trend model using GEE under four different methods for handling dropouts: complete-case (CC), last value carried forward (LVCF), available-data (AD), and multiple imputation (MI) using propensity scores.

Method	Time	100 mg	150 mg	Difference	SE	Z
CC	3 months	0.176	0.155	−0.021	0.027	−0.79
	6 months	0.258	0.317	0.059	0.028	2.07
	9 months	0.369	0.463	0.094	0.033	2.83
	12 months	0.502	0.540	0.038	0.037	1.03
LVCF	3 months	0.184	0.201	0.017	0.023	0.75
	6 months	0.263	0.344	0.081	0.024	3.43
	9 months	0.350	0.453	0.103	0.027	3.78
	12 months	0.437	0.498	0.061	0.029	2.10
AD	3 months	0.184	0.201	0.017	0.023	0.73
	6 months	0.274	0.363	0.089	0.025	3.55
	9 months	0.388	0.499	0.111	0.030	3.68
	12 months	0.517	0.572	0.055	0.036	1.52
MI	3 months	0.183	0.198	0.015	0.026	0.57
	6 months	0.279	0.363	0.084	0.027	3.17
	9 months	0.394	0.500	0.106	0.030	3.48
	12 months	0.517	0.572	0.056	0.034	1.64

We first consider complete-case and available-data analyses of the data. If dropout is completely at random, then valid estimates of the marginal regression parameters can be obtained using a standard generalized estimating equations (GEE) approach. To account for the within-subject association among the repeated measures, we fit six separate pairwise log odds ratios. We note that the empirical and model-based standard errors are very similar in all of the analyses. For illustrative purposes, we also performed a GEE analysis using LVCF imputation of the missing data. The differences in results are more easily discerned by considering the dose-specific estimated rates of amenorrhea in each of the injection intervals for the quadratic trend model given above (see Table 14.3). Overall, the results of the complete-case and available-data GEE analyses suggest that the rates of amenorrhea in the second and third injection

intervals are significantly higher for those women who received the higher dose of DMPA, although these differences tend to decline by the end of the study. For example, during the third injection interval (6–9 months post-randomization) the predicted rates of amenorrhea from the available-data analysis are 0.499 in the 150 mg dose group and 0.388 in the 100 mg dose group. However, by the final follow-up visit there is no longer a discernible treatment difference, with predicted rates of amenorrhea of 0.572 in the 150 mg dose group and 0.517 in the 100 mg dose group.

The GEE analysis based on LVCF imputation produces discernibly lower estimated rates of amenorrhea during the third and fourth intervals, when compared to the available-data analysis, although the estimates of the treatment comparisons are not too dissimilar; however, the latter cannot be expected in general. Because LVCF uses a single imputation and does not reflect any uncertainty in the imputation, the standard errors for the estimated treatment comparisons are too small. Consequently, in contrast to the other methods, the analysis based on LVCF suggests that there are treatment differences in the estimated rates of amenorrhea at the end of the trial.

Note that if dropout is not completely at random, the complete-case and available-data GEE analyses of these data can yield biased estimates of the effects of treatment. Next, we consider handling dropout using propensity score multiple imputation methods. Recall that propensity score methods require a model for the probability of dropout. Within each dose group, we consider a sequence of logistic regression models that assume the log odds of dropout depends on all past observed responses. For example, the model for dropout at the k^{th} occasion is given by

$$\log\left\{\frac{\Pr(D_i = k | D_i \geq k, Y_{i1}, ..., Y_{ik-1})}{\Pr(D_i > k | D_i \geq k, Y_{i1}, ..., Y_{ik-1})}\right\} = \theta_1 + \theta_2 Y_{i1} + \cdots + \theta_k Y_{ik-1}.$$

Note that for each dose group, separate logistic regression models are fit to the data at the second through fifth occasions. Based on the estimated parameters from the logistic regression models, a propensity score is obtained for each observation at each occasion of dropout. Next, to draw imputed values from $f(Y_i^M | Y_i^O, X_i)$ the following procedure is used at each occasion of dropout:

(i) Observations are grouped on the basis of having similar propensity scores at that occasion. The observations are sorted into eight groups based on the propensity scores.

(ii) Within each group, let N^O denote the number of individuals with observed values for the response at that occasion, and N^M denote the number of individuals with missing values for the response. We randomly select N^O observations with replacement from the observed values for the response. As an aside, note that this step is analogous to the drawing of $\theta_1^*, \theta_2^*, ..., \theta_k^*$ from the joint distribution of $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_k$, to account for the uncertainty in estimating the propensity scores. This step ensure that the uncertainty in the estimation of the propensity scores is properly accounted for.

(iii) Finally, the N^M values for the missing responses are then randomly selected with replacement from the random sample of observed values drawn in (ii).

This three-step procedure is repeated sequentially at the second through fifth occasion to fill in all of the missing values for $Y_{i2}, \ldots, Y_{i5}$. To create 10 imputations, steps (ii) and (iii) are repeated 10 times and results from the 10 analyses of the filled-in data sets are appropriately combined.

The results from the analysis based on multiple imputation (see bottom of Table 14.3) are remarkably similar to those obtained from the available-data analysis. Both the point estimates of the rates of amenorrhea and their standard errors are similar. Under the assumption that dropout is at random, these results suggest that the rates of amenorrhea in the second and third injection intervals are significantly higher for those women who received the higher dose of DMPA, although these differences tend to decline by the end of the study.

14.7 FURTHER READING

A useful discussion of methods for handling dropout in longitudinal studies can be found in Heyting *et al.* (1992).

In longitudinal clinical trials, "last value carried forward" (LVCF) imputations are still widely used to handle dropouts; see Ware (2003) for a critique of this method.

Bibliographic Notes

Rubin (1976) developed the taxonomy for describing the assumptions concerning the dependence of the missingness process on observed and unobserved responses. Little and Rubin (2001) is the definitive textbook on missing data, providing a comprehensive description of the theory and application of methods for handling missing data; also, see Schafer (1997). Laird (1988) discusses missing data issues in longitudinal studies; also, see the review articles by Little (1995) and Kenward and Molenberghs (1999). The EM algorithm, a general technique for ML estimation with incomplete data, is discussed in the seminal paper by Dempster *et al.* (1977).

Inverse probability weighted methods were first proposed in the sample survey literature by Horvitz and Thompson (1952). Robins *et al.* (1995) developed an inverse probability weighted estimating equations approach for handling missing data in longitudinal studies. Propensity score methods are described in Rosenbaum and Rubin (1983). A comprehensive description of imputation methods can be found in Rubin (1987).

15

Some Aspects of the Design of Longitudinal Studies

15.1 INTRODUCTION

The focus in earlier chapters has been on methods for analyzing longitudinal data. In this chapter we focus on a number of issues concerning the design of a longitudinal study that also have important implications for the analysis. The first topic considers issues of sample size and power when designing a longitudinal study. We review sample size formulas for a univariate response and describe a simple, albeit approximate, method that allows direct application of standard sample size and power formulas in the longitudinal setting. The second topic is concerned with assumptions about time-varying covariates when they are not fixed by the study design but vary randomly (or stochastically). When a covariate is both time-varying and stochastic, subtle issues arise regarding the estimation and interpretation of regression parameters in models for longitudinal data. The final topic is concerned with longitudinal study designs that provide both longitudinal and cross-sectional sources of information. These studies provide opportunities to estimate both longitudinal and cross-sectional effects, and to distinguish, for example, aging effects from cohort effects in growth studies.

15.2 SAMPLE SIZE AND POWER

Questions about sample size and power arise in the earliest stages of the design of a study. Although the question can be posed in a variety of different ways, investigators typically need to know the answer to the following question: "How *large* should my study be?" The answer to this question is relatively straightforward when there is

only a single, univariate response: the *size* of a study is directly related to the number of subjects, that is, the sample size. However, for a longitudinal study the question of *size* is more complex. For example, in planning a longitudinal study to compare an active treatment to control, investigators need to determine not only how many subjects to enroll in the study, but also the duration of the study and the frequency and spacing of repeated measurements on the subjects.

When there is only a single, univariate response, statisticians have developed simple formulas for sample size and power calculations. Explicit formulas can be found in many introductory textbooks in statistics. In addition, some statistical software packages include procedures for sample size and power calculations and publicly available sample size and power calculators can be found on the Web. However, for the multivariate response obtained from a longitudinal study, accurate sample size (and power) determination is far more complicated and, in general, requires inversion of matrices and iterative solutions when no closed-form expressions can be obtained. The purpose of this section is not to derive complex sample size formulas for longitudinal studies; references to accurate, but also more complex, methods for calculating sample size can be found at the end of the chapter. Instead, we present a very simple, albeit approximate, method for sample size and power determination for longitudinal studies that reduces the problem to the univariate case. This allows direct application of standard sample size and power formulas.

In this section we begin with a review of sample size formulas for a single, univariate response. We emphasize the main considerations in determining how large the sample size needs to be to achieve a specified power to detect some effect of scientific interest; this section can be skimmed through for those already familiar with power and sample size calculations for a univariate response. For the case of longitudinal studies with a continuous response, we present a simple closed-form expression for sample size (and power) calculations that is based on the standard sample size (and power) formula for a univariate response. A similar closed-form expression for longitudinal binary responses is also presented; however, the latter expression is very approximate and provides conservative "ball park" estimates of sample size and power.

Sample Size for a Univariate Continuous Response

When planning a cross-sectional study, investigators must establish how many subjects they will need to achieve some specified power to detect an effect of subject-matter importance. For example, suppose that investigators are interested in comparing two treatments, an active drug and placebo. The investigators plan to randomize an equal number of subjects, say N, to receive either of the two treatments. At the completion of the study, the two treatment groups are to be compared in terms of the mean response. Let $\mu^{(A)}$ denote the mean response in the population of individuals treated with the active drug; similarly, let $\mu^{(B)}$ denote the mean response in the population of individuals treated with the placebo. The treatment effect can be expressed in a variety of different ways, but here we consider the simple difference in means, $\delta = \mu^{(A)} - \mu^{(B)}$. The null hypothesis of no treatment difference can be expressed

as H_0: $\delta = 0$. In this example, the investigators may be interested in establishing whether the active drug is superior to placebo, with the alternative hypothesis that $\delta > 0$.

Before we discuss sample size and power, we must consider the two types of errors that can arise when conducting a statistical test of H_0: $\delta = 0$. The first kind of error is called a type I error and is made if we reject the null hypothesis when in fact it is true. The probability of a type I error, also known as the significance level of the test, is usually denoted by α. Thus, for our example where H_0: $\delta = 0$,

$$\alpha = \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

Conventionally, α is chosen to be no greater than 0.05, that is, we are prepared to mistakenly reject the null hypothesis no more than 5% of the time. The second kind of error that can arise when conducting a statistical test is called a type II error. A type II error is made if we fail to reject the null hypothesis when in fact it is false. We denote the probability of a type II error by γ, with

$$\gamma = \Pr(\text{Fail to reject } H_0 \mid H_0 \text{ is false}).$$

(The usual convention is to denote the probability of a type II error by β; however, we have chosen to denote it by γ to avoid any potential confusion with our widespread use of β for the regression parameters in earlier chapters of the book.) Since γ is determined by considering the case where the null hypothesis is not true (i.e., $\delta \neq 0$), it necessarily depends upon the particular choice of value for $\delta \neq 0$ under the alternative hypothesis. Intuitively, the closer the true value of δ is to zero (the assumed value for δ under the null hypothesis) the more difficult it is to reject H_0: $\delta = 0$. Finally, the power of a statistical test is defined as $1 - \gamma$, that is,

$$\text{power} = 1 - \gamma = \Pr(\text{Reject } H_0 \mid H_0 \text{ is false}).$$

In simple terms, the power of a test is the probability that the study will determine that there is a treatment effect of some subject-matter importance when it truly exists. Since γ necessarily depends upon the particular choice of value for $\delta \neq 0$ under the alternative hypothesis, so too does the power of a test. Thus, with all other things being equal, the further the true value of δ is from zero the greater is the power of a test of H_0: $\delta = 0$.

Finally, by considering the two types of errors that can arise when conducting a statistical test, we can determine the sample size required to have some specified power to detect an effect, $\delta \neq 0$. For the special case of the two group comparison considered in our example, a formula for the approximate sample size in each group, N, is given by

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\sigma^2}{\delta^2}, \tag{15.1}$$

where σ^2 is the variance of the response (assumed to be common in the two groups), and $Z_{(1-\alpha/2)}$ and $Z_{(1-\gamma)}$ denote the $(1-\alpha/2)\times 100\%$ and $(1-\gamma)\times 100\%$ percentiles of a standard normal distribution (e.g., the 97.5th percentile of a standard normal

distribution is 1.96; or put in somewhat simpler terms, 97.5% of the area under the standard normal curve lies to the left of 1.96). Note that, in previous chapters, N denoted the total sample size in the study; here, N is used to denote the sample size in each of the two groups and the total sample size is $2N$.

How the sample size formula given by (15.1) was derived is not important. The main reason for displaying this formula is to highlight its main constituents. A closer examination of this formula reveals that the determination of sample size requires that all of the following be specified:

(i) significance level, α;

(ii) power, $1 - \gamma$;

(iii) effect size, δ; and

(iv) common variance, σ^2.

Ordinarily, the first two factors do not pose a great challenge for investigators. Conventionally, the significance level of a statistical test is fixed at the mythical 0.05 level (with $Z_{(1-\alpha/2)} = 1.96$ for a 2-tailed test). Similarly, the lower bound on what might be considered acceptable power is usually set at approximately 80% (with $Z_{(1-\gamma)} = 0.842$ for power = 0.8, or $Z_{(1-\gamma)} = 1.282$ for power = 0.9). This leaves only two key ingredients for which the investigators must provide information: the minimum effect size of scientific interest and an estimate of the variability in the data. Note that the former appears in the denominator of (15.1), while the latter appears in the numerator. As a result, for any fixed value of the variability, the required sample size decreases with increasing effect size, δ. Intuitively, fewer subjects (or less information) are needed when it is of interest to determine whether a true treatment effect is quite far from the null value. Similarly, for any fixed effect size, the required sample size decreases with decreasing variability. For example, the required sample size can be made smaller by using a more reliable measurement instrument.

Sample Size for a Longitudinal Continuous Response

Next, suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of changes in the mean response over time. The investigators plan to randomize an equal number of subjects (N) to receive either of the two treatments. They plan to take n repeated measurements of the response (not necessarily equally spaced measurements). At the completion of the study, the two treatment groups are to be compared in terms of changes in the mean response over the duration of the study. For simplicity, we assume that changes in the mean response can be expressed in terms of a linear trend and the treatment effect can be expressed in terms of the difference in slopes or rates of change, say δ. Under the null hypothesis of no treatment difference, that is, no treatment × linear trend interaction, H_0: $\delta = 0$.

In this section we show that sample size calculations for such a longitudinal study design can be simplified so that the standard sample size formula given by (15.1)

can be used. This is achieved by considering the two-stage model for longitudinal data described in Chapter 8 (see Section 8.4). Let us assume the following two-stage formulation. At the first stage, it is assumed that a simple parametric curve (e.g., linear trend in time) fits the observed responses for each subject. In the second stage, these individual-specific parameters are then related to covariates that describe the different groups from which the individuals have been drawn (e.g., treatment versus control).

Stage 1: In the first stage subjects are assumed to have their own unique individual-specific response trajectories. That is, in stage 1 we posit that the repeated measures on each individual follow a regression model having the same set of covariates, but with separate regression coefficients for each individual

$$Y_{ij} = \beta_{1i} + \beta_{2i}\, t_j + e_{ij},$$

where the errors, e_{ij}, are assumed to be independent and identically distributed, having a normal distribution with mean equal to zero and variance σ_e^2, that is, $e_{ij} \sim N(0, \sigma_e^2)$.

Stage 2: In the second stage we make the assumption that the individual-specific effects, $\beta_i = (\beta_{1i}, \beta_{2i})'$, are random. The mean and covariance of β_{1i} and β_{2i} are the population parameters that are modelled in the second stage. Specifically, variation in β_i is modelled as a function of between-individual covariates which we assume here includes only treatment group. Thus we can allow the mean of β_i (i.e., the mean intercept and slope) to depend upon treatment group,

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_2\, \texttt{Group}_\texttt{i}; \\ E(\beta_{2i}) &= \beta_3 + \beta_4\, \texttt{Group}_\texttt{i}, \end{aligned}$$

where $\texttt{Group}_\texttt{i} = 1$ if the i^{th} individual was assigned to the active treatment, and $\texttt{Group}_\texttt{i} = 0$ otherwise. In this model, β_3 is the mean slope, or constant rate of change in the mean response over time, in the control group, while $\beta_3 + \beta_4$ is the mean slope in the active treatment group. That is, β_4 has interpretation in terms of a treatment group difference in the mean slope or rate of change in the mean response and corresponds to the definition of δ given earlier. The residual between-individual variation in the β_i that cannot be explained by treatment group is expressed as

$$\text{Cov}(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \text{Var}(\beta_{1i})$, $g_{22} = \text{Var}(\beta_{2i})$, and $g_{12} = g_{21} = \text{Cov}(\beta_{1i}, \beta_{2i})$.

This two-stage formulation yields tractable forms for the component of variability required for sample size and power calculations. If each subject is measured at a common set of occasions, $t_1, \ldots, t_n$, and there are N subjects in each of the two treatment groups (for a total sample size of $2N$), we can derive simple expressions for sample size and power similar to the univariate setting. Letting $\widehat{\beta}_{2i}$ denote the

ordinary least squares (OLS) estimate of the slope for the i^{th} subject, the variability of $\widehat{\beta}_{2i}$ is given by

$$\sigma^2 = \text{Var}(\widehat{\beta}_{2i}) = \sigma_e^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

where

$$\bar{t} = \frac{1}{n} \sum_{j=1}^{n} t_j.$$

Thus the variability of $\widehat{\beta}_{2i}$ is composed of two components: the within-subject variance, $\sigma_e^2 \{\sum_{j=1}^{n} (t_j - \bar{t})^2\}^{-1}$, and the between-subject variance, $g_{22} = \text{Var}(\beta_{2i})$. To test whether the mean slopes are equal in the two treatment groups, we can construct the following z-test based on $\widehat{\beta}_{2i}$:

$$Z = \frac{\overline{\beta}_2^{(T)} - \overline{\beta}_2^{(C)}}{\sigma \sqrt{\frac{1}{N} + \frac{1}{N}}} = \frac{\overline{\beta}_2^{(T)} - \overline{\beta}_2^{(C)}}{\sigma \sqrt{\frac{2}{N}}},$$

where $\overline{\beta}_2^{(T)}$ and $\overline{\beta}_2^{(C)}$ are the sample averages of $\widehat{\beta}_{2i}$ in the treatment and control groups, respectively, and $\sigma^2 = \text{Var}(\widehat{\beta}_{2i})$.

Given estimates of g_{22}, the between-subject variability in slopes, and σ_e^2, the within-subject variability, the sample size can be determined from the standard formula given by (15.1),

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\sigma^2}{\delta^2}, \tag{15.2}$$

where now

$$\sigma^2 = \sigma_e^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

and δ is a treatment effect size of interest (i.e., δ is the treatment group difference in slopes or rates of change in the mean response). Notice that the sample size formula given by (15.2) is virtually identical to (15.1), except that σ^2 has two components: a within-subject variance component, $\sigma_e^2 \{\sum_{j=1}^{n} (t_j - \bar{t})^2\}^{-1}$, and a between-subject variance component, $g_{22} = \text{Var}(\beta_{2i})$.

Further examination of the simple formula for sample size given by (15.2) reveals how sample size (and power) is impacted by:

(i) the length of the study;

(ii) the number of repeated measures; and

(iii) the spacing of the repeated measures.

Consider the design of a longitudinal study with n repeated measurements. Let $t_1 = 0$ denote the baseline measurement occasion and $t_n = \tau$ denote the time at the final measurement occasion (i.e., τ denotes the duration of the study). Study investigators can reduce the required sample size (or, correspondingly, increase the power for a fixed total sample size, $2N$) by reducing the magnitude of σ^2. Recall that σ^2 depends on both the between-subject and within-subject variability. In general, investigators have little control over the natural heterogeneity of the study population, denoted in this instance by $g_{22} = \text{Var}(\beta_{2i})$ and, more generally, by $G = \text{Var}(\beta_i)$. The between-subject variability can only be reduced by focusing on a more homogeneous population. However, doing so would alter the intended target of inference and could reduce the generalizability of the results. In principle, the within-subject variability, σ_e^2, can be reduced by using a more reliable measurement instrument; however, this will not always be possible or practical. Therefore, to reduce the magnitude of σ^2, we must focus on ways to increase the magnitude of

$$\sum_{j=1}^{n} (t_j - \bar{t})^2.$$

Since $\sum_{j=1}^{n} (t_j - \bar{t})^2$ is a divisor of σ_e^2, increasing its magnitude reduces the contribution of the within-subject variance to σ^2. Note that $\sum_{j=1}^{n} (t_j - \bar{t})^2$ is the sum of the squared deviations of the measurement times about their mean. It is a function of the duration of the study, τ, the number of repeated measurements, n, and the relative spacing of the repeated measurements. For a study of fixed length τ and fixed number of repeated measures n, $\sum_{j=1}^{n} (t_j - \bar{t})^2$ is maximized when $n/2$ measurements are taken at baseline and $n/2$ measurements are taken at the end of the study (when n is an even number). In general, such a study design would not be desirable because it relies too heavily on the assumption that changes in the response are linear over time and precludes examination of non-linear (e.g., quadratic) trends. Also, the notion of taking $n/2$ replicate measurements at the same occasion is not feasible or practical in many settings. So, for the remainder of this discussion, we assume that the measurement occasions will be equally spaced (at least approximately) throughout the duration of the study. That is, in a study of length τ, the n repeated measurements are to be taken at times $t_1 = 0, t_2 = \tau/(n-1), t_3 = 2\tau/(n-1), ..., t_n = \tau$. Then it can be shown that

$$\sum_{j=1}^{n} (t_j - \bar{t})^2 = \frac{\tau^2 \, n \, (n+1)}{12 \, (n-1)}.$$

(This expression can be derived by using the fact that the variance of the first n integers is $(n+1)(n-1)/12$.) Thus, for a fixed number of repeated measurements, doubling the length of the study decreases the impact of the within-subject variability by a factor of 4. Impressive as this may seem, there are a number of practical limitations that qualify this result. First, the length of a longitudinal study is usually determined by economic, logistical, and subject-matter factors that constrain the maximum length of follow-up. Second, changes in the mean response, as a function of exposure to some treatment or intervention, may be of limited duration and constrain the maximum

value of τ. As a result, many study investigators are restricted to a relatively narrow range of possible values for τ. Similarly, for a study of fixed length τ, there may be practical constraints on the number of repeated measurements. The simple formula for $\sum_{j=1}^{n}(t_j - \bar{t})^2$ given above indicates that for fixed τ the impact of the within-subject variability decreases non-linearly with increasing n. For example, increasing the number of repeated measurements from $n = 2$ (a simple pre/post longitudinal design) to $n = 4$, $n = 6$, $n = 8$, and $n = 10$, results in a 10%, 29%, 42% and 50% reduction in the impact of the within-subject variability. We must caution, however, that these results rely heavily on the assumption that changes in the mean response over the duration of the study are linear in time.

The sample size formula given by (15.2) can also be manipulated to determine the power of a test of H_0 for a given sample size, since (15.2) implies that

$$Z_{(1-\gamma)} = \sqrt{\frac{N\delta^2}{2\sigma^2}} - Z_{(1-\alpha/2)}. \tag{15.3}$$

This implies that the power, $1 - \gamma$, is given by $\Phi\{Z_{(1-\gamma)}\}$, where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. That is, the value of $Z_{(1-\gamma)}$ can be calculated from the formula given by (15.3) and the power is determined by the area under the standard normal curve that lies to the left of $Z_{(1-\gamma)}$.

The sample size formula can also be modified to allow groups of unequal size, say $N^{(T)}$ and $N^{(C)}$, where $N^{(T)} = 2N\pi$ and $N^{(C)} = 2N(1-\pi)$ (for $0 < \pi < 1$). Then, the following simple modification to the sample size formula is required:

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\,\sigma^2}{2\pi(1-\pi)\delta^2}, \tag{15.4}$$

and the total sample size required is $2N$, with $2N\pi$ subjects in one group and the remaining $2N(1-\pi)$ subjects in the other group.

Sample size formulas for two group comparisons of other coefficients in the model for the mean response can be derived by using the general formulation for the stage 1 model (see Section 8.4)

$$Y_i = Z_i\beta_i + e_i,$$

where the matrix Z_i specifies how an individual's responses change over time and β_i is a $q \times 1$ vector of individual-specific regression coefficients. Then, for any particular trend of interest, the variance, σ^2, in the sample size formula is simply obtained from the appropriate diagonal element of

$$\begin{aligned} \mathrm{Cov}(\widehat{\beta}_i) &= \sigma_e^2\,(Z_i'Z_i)^{-1} + \mathrm{Cov}(\beta_i) \\ &= \sigma_e^2\,(Z_i'Z_i)^{-1} + G. \end{aligned}$$

In our example, with random intercepts and slopes,

$$\begin{aligned} \mathrm{Cov}(\widehat{\beta}_i) &= \mathrm{Cov}\begin{pmatrix} \widehat{\beta}_{1i} \\ \widehat{\beta}_{2i} \end{pmatrix} \\ &= \sigma_e^2\left\{\begin{pmatrix} 1,\ldots,1 \\ t_1,\ldots,t_n \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}\right\}^{-1} + \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \end{aligned}$$

and $\mathrm{Var}(\widehat{\beta}_{2i})$ is the lower-diagonal element of this 2×2 matrix.

Finally, note that we have assumed no missing data or attrition. The impact of missing data is difficult to quantify precisely because it depends upon the patterns of missingness, and in this case simple formulas for $\mathrm{Cov}(\widehat{\beta}_i)$ no longer apply. An admittedly *ad hoc*, but conservative, approach for adjusting for attrition is to inflate the required sample size N in each group to account for the assumed rate of attrition (or proportion of subjects who drop out before the completion of the study). That is, if the rate of attrition is assumed to be 10% in each group, then the target sample size in each group should be $N/0.9$.

Example: Longitudinal Study with a Continuous Response

To illustrate the application of the sample size formula, let us return to the example introduced earlier. Suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of changes in the mean response over time. The investigators plan to randomize an equal number of subjects (N) to receive either of the two treatments. They plan to take 5 repeated measurements of the response, one at baseline, and the remainder at 6-month intervals until the completion of the study ($\tau = 2$ years). The response variable is assumed to have an approximate normal distribution. At the completion of the study, the two treatment groups are to be compared in terms of changes in the mean response over the duration of the study. For simplicity, we assume that changes in the mean response can be expressed in terms of a linear trend over time (in years) and the treatment effect can be expressed in terms of the difference in slopes, say δ.

Suppose that the investigators want to detect a minimum treatment effect of $\delta = 1.2$, that is, a difference in the annual rates of change in the treatment and control groups of no less than 1.2. Based on historical data from similar populations, the investigators posit that the between-subject variability in the rate of change, $\mathrm{Var}(\beta_{2i}) \approx 2$ and the within-subject variability, $\sigma_e^2 \approx 7$. Finally, the investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$). Given these specifications,

$$\sigma_e^2\left\{\sum_{j=1}^{n}(t_j - \bar{t})^2\right\}^{-1} = \frac{12\,(n-1)\,\sigma_e^2}{\tau^2\,n\,(n+1)} = \frac{12 \times 4 \times 7}{4 \times 5 \times 6} = 2.8,$$

Table 15.1 Power as a function of sample size and the number of equally spaced repeated measurements in a longitudinal study of fixed duration.

	Number of Repeated Measures (n)				
Sample Size (N)	2	4	6	8	10
20	0.37	0.39	0.43	0.47	0.50
40	0.63	0.66	0.72	0.76	0.79
60	0.80	0.83	0.87	0.90	0.93
80	0.90	0.92	0.95	0.97	0.98
100	0.95	0.96	0.98	0.99	0.99

Power when conducting a 2-sided test at the 5% significance level ($\alpha = 0.05$) when $\tau = 2$, $\delta = 1.2$, $\text{Var}(\beta_{2i}) = 2$, and $\sigma_e^2 = 7$.

and

$$\sigma^2 = \sigma_e^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + g_{22} = 2.8 + 2.0 = 4.8.$$

The projected sample size required in each of the two groups is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\sigma^2}{\delta^2} = \frac{(1.96 + 1.282)^2 \times 2 \times 4.8}{1.44} = 70.1.$$

Thus, to ensure that they have power of at least 90% the investigators will need to enroll a total of 142 subjects, randomizing an equal number (71) to each of the two treatment groups. Note that a study of the same duration ($\tau = 2$ years) with $n = 3$ repeated measurements, 12 months apart, would require a total of 182 subjects to achieve comparable power. Alternatively, if it were feasible to conduct the study over 3 years instead of 2 years (and have the same retention rate), with $n = 5$ repeated measurements taken 9 months apart, it would require a total of 96 subjects to achieve power of at least 90%.

For this example, it is of interest to study the relationship between power, sample size, and the number of repeated measurements (assuming a study of the same duration $\tau = 2$ years). Table 15.1 displays the power as a function of the sample size in each group, N, and the number of equally spaced repeated measurements, n. Table 15.1 is revealing about the tradeoffs of increasing the sample size versus increasing the number of repeated measurements. For example, doubling the sample size leads to a discernibly greater increase in power than doubling of the number of repeated measurements. This can be explained by the fact that increases in the number of repeated measurements only reduce the impact of the within-subject variance component in the formula for power. Recall that σ^2 depends on both the between-subject and within-subject variability. In contrast, increasing the sample size reduces the impact of both sources of variability.

Finally, it should be apparent from (15.2) and (15.3) that sample size and power calculations are sensitive to assumptions about the covariance among the repeated measures. Because σ^2 depends on assumptions about the magnitudes of the between-subject and within-subject variability, it is advisable to perform a sensitivity analysis to examine how sample size varies according to changes in the values of the between-subject and within-subject variances.

Sample Size for a Longitudinal Binary Response

Sample size determination for longitudinal studies with a binary response variable is somewhat more complicated. Complications arise from two main sources: (i) the non-linear link function (e.g., logit) usually adopted for the relationship between the mean response and covariates, and (ii) the dependence of the variance on the mean. Simple closed-form expressions for sample size (and power), comparable to those for a continuous response, cannot be derived. Instead, precise determination of sample size and power involves more complicated procedures that usually require iterative solutions.

In this section we derive a very approximate sample size formula for longitudinal studies with a binary response. The formula relies heavily on the assumption that the response probabilities are in the center of the range, say 0.2 to 0.8. When response probabilities lie between 0.2 to 0.8, linearity on the log odds scale (or the probit scale, for that matter) corresponds to approximate linearity on the probability scale. As a result, a generalized linear model for the response probabilities

$$g\{E(Y_{ij}|X_{ij})\} = X_{ij}'\beta$$

is well approximately by

$$E(Y_{ij}|X_{ij}) = X_{ij}'\beta^*$$

for some $\beta^* \neq \beta$. Here, the components of β^* have interpretation in terms of changes in the probabilities, while β has interpretation in terms of changes in the log odds if $g(\cdot)$ has a logistic form. Also, the variance of the binary response

$$\text{Var}(Y_{ij}) = E(Y_{ij})\{1 - E(Y_{ij})\}$$

changes somewhat more slowly when the response probabilities are in the range 0.2 to 0.8, with maximum value of $\frac{1}{4}$ when $E(Y_{ij}) = 0.5$. Therefore, a very approximate sample size formula can be based on (15.1) by expressing δ in terms of contrasts of the response probabilities, that is, in terms of a linear model for the response probabilities. Specifically,

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\sigma^2}{\delta^2}, \qquad (15.5)$$

where δ denotes a comparison of the two groups in terms of a linear contrast of the response probabilities over time, and

$$\sigma^2 = \text{Var}(Y_{ij})(1-\rho)\left\{\sum_{j=1}^{n}(t_j-\bar{t})^2\right\}^{-1} \approx \frac{1}{4}\times(1-\rho)\left\{\sum_{j=1}^{n}(t_j-\bar{t})^2\right\}^{-1},$$

where $\text{Var}(Y_{ij}) \approx \frac{1}{4}$, the maximum possible value for the variance of Y_{ij}, and $\rho = \text{Corr}(Y_{ij}, Y_{ik})$ for all $j \neq k$. In a study of length τ, with n repeated measurements assumed to be equally spaced (at least approximately),

$$\sigma^2 \approx \frac{1}{4}\times(1-\rho)\left\{\sum_{j=1}^{n}(t_j-\bar{t})^2\right\}^{-1} = \frac{3\,(n-1)(1-\rho)}{\tau^2\,n\,(n+1)}.$$

By assuming the maximum possible value for the variance, the sample size formula given above can be considered quite "conservative", in the sense of projecting a larger than necessary sample size for the desired power. We view this simple sample size formula as being very approximate and its main use is for producing "ball park" projections of the required sample size. More precise sample size (and power) formulas are available but require the use of matrix algebra functions and/or iterative solutions; some references to sample size formulas for longitudinal studies with binary responses are given at the end of this chapter.

Note that, in addition to assuming that the variances are constant, with $\text{Var}(Y_{ij}) \approx \frac{1}{4}$, the sample size formula given above also assumes that the correlation among any pair of repeated binary responses is constant, with $\rho = \text{Corr}(Y_{ij}, Y_{ik})$. This is awkward since making this assumption for linear models with a continuous response can dramatically underestimate the sample size. That is, for the case of a continuous response, it can be shown that the compound symmetry assumption (equal variances and equal correlations) always underestimates the sample size if a more complex random effects covariance structure (e.g., random intercepts and slopes) or an arbitrary covariance matrix actually holds; the impact of violations of this assumption for binary responses is not known, but is conjectured to be similar. The strong assumption about the correlation can, in principle, be relaxed by recognizing that σ^2 was derived from the following expression:

$$\sigma^2 \approx \frac{1}{4}\times\{(t_1,...,t_n)R^{-1}(t_1,...,t_n)'\}^{-1},$$

where R is the $n \times n$ matrix of pairwise correlations among the repeated binary responses (and R^{-1} denotes the inverse of the correlation matrix). Thus, for alternative assumptions about the correlation matrix, a different expression for σ^2 can be substituted into the sample size formula. However, for a general correlation matrix R, there is no simple expression for σ^2 and obtaining the inverse of R usually requires the use of computer software with matrix algebra functions.

Example: Longitudinal Study with a Binary Response

To illustrate the application of the sample size formula with binary responses, let us return to the example introduced earlier. Suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of changes in the probability of a binary response over time. The investigators plan to randomize an equal number of subjects (N) to receive either of the two treatments. They plan to take 5 repeated measurements of the response, one at baseline, and the remainder at 6-month intervals until the completion of the study ($\tau = 2$ years). At the completion of the study, the two treatment groups are to be compared in terms of changes in the probability of response over the duration of the study. For simplicity, we assume that changes in the response probabilities can be expressed (approximately) in terms of a linear trend and the treatment effect can be expressed in terms of the difference in slopes, say δ. We require this assumption to apply the approximate sample size formula given earlier.

The investigators assume that the baseline probability of response is approximately 0.3 (for both treatment groups). They want to be able to detect treatment group differences in the probability of response at 2 years of at least 0.15, that is, $\delta = 0.15/\tau = 0.15/2 = 0.075$. Based on historical data from similar populations the investigators posit that the correlation among pairs of responses is approximately 0.5. Finally, the investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$). Given these specifications,

$$\sigma^2 \approx \frac{3\,(n-1)(1-\rho)}{\tau^2\,n\,(n+1)} = \frac{3\times 4\times 0.5}{4\times 5\times 6} = 0.05.$$

The projected sample size required in each of the two groups is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\,2\,\sigma^2}{\delta^2} = \frac{(1.96+1.282)^2\times 2\times 0.05}{0.005625} = 186.9.$$

Thus, to ensure that they have power of at least 90%, the investigators will need to enroll a total of 374 subjects, randomizing an equal number (187) to each of the two treatment groups.

Summary

In this section we have shown that sample size determination for longitudinal studies can often be simplified so that well-established formulas for the univariate case can be applied. For the investigators, the main challenges are in the specification of the minimum effect size of subject-matter interest and in providing a realistic estimate of the anticipated variability in the measurements. The choice of an appropriate effect size must be made on purely subject-matter grounds. If the investigators expect a large effect, then it is likely to be detected with a relatively small sample size. In contrast, detection of small effects requires somewhat larger sample sizes. In planning a study investigators need to keep their optimism in check since gross overestimation of the

effect size will result in too few subjects and insufficient power to detect somewhat smaller, but nonetheless scientifically important, effects.

Perhaps the greatest challenge facing investigators is to provide a realistic estimate of the variability in the data. This will either require the provision of estimates of both between-subject and within-subject variability or, alternatively, an estimate of the covariance among the repeated measurements. Since scientific studies are rarely conducted in a vacuum, investigators can usually obtain some estimates of the variability based on historical data from related studies with similar populations. Alternatively, in the complete absence of any relevant historical data, it may be prudent to conduct a small pilot study. If there is much uncertainty regarding the anticipated variability in the data, a simple sensitivity analysis, examining the projected sample sizes across a range of plausible values for the variability, should be conducted.

Finally, as mentioned in Chapter 14, missing data are the rule, not the exception, in longitudinal studies. Therefore it is important to make some adjustment for the potential loss of information due to missing data when planning a longitudinal study, for example, by using relatively conservative estimates of sample size. In general, a consideration of the anticipated fraction of missing data (or the proportion of subjects who drop out before the completion of the study), say f, suggests that the sample size should be inflated by a factor of $\frac{1}{1-f}$. That is, if 15% of the observations are expected to be missing, then investigators should plan on increasing the sample size of the study by a factor of 1.18 (or $\frac{1}{1-0.15}$). Although this adjustment is crude, ignoring both the location of the missing observations and the correlation among repeated measurements, it will probably be adequate for most practical purposes. Failure to make any adjustment for missing data will result in an underestimation of the number of subjects required to attain the desired level of power.

15.3 INTERPRETATION OF STOCHASTIC TIME-VARYING COVARIATES

Next we consider aspects of interpretation of time-varying covariates. In earlier chapters we have used a common notation for the response variable and covariates in a longitudinal study. Specifically, we let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion ($i = 1, ..., N; j = 1, ..., n_i$). The response variables for the i^{th} subject can be grouped into an $n_i \times 1$ response vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, ..., N,$$

and associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \ \ j = 1, ..., n_i.$$

Note that X_{ij} is the vector of covariates associated with Y_{ij}, the response variable for the i^{th} subject at the j^{th} occasion, and may include two main types of covariates: covariates whose values do not change throughout the duration of the study and covariates whose values change over time. The former are referred to as time-stationary or between-subject covariates (e.g., gender and fixed experimental treatments), while the latter are referred to as time-varying or within-subject covariates (e.g., time since baseline, current smoking status, and environmental exposures). In the former case, the same values of the covariates are replicated in the corresponding rows of X_{ij}, for $j = 1, ..., n_i$. In the latter case, the values taken by the covariates can vary over time (for at least some individuals) and the values in the corresponding rows of X_{ij} can be different at each measurement occasion.

When considering time-varying covariates, we can distinguish covariates that vary systematically over time but are fixed by design of the study and covariates that vary randomly over time. An example of a time-varying covariate that is fixed by design is a treatment group indicator in a crossover trial. Another example, and one that is commonly encountered in a longitudinal study, is time since baseline (when the measurement occasions are fixed by the study design). Covariates that vary randomly over time are often referred to as *stochastic*, that is, values of the covariate at any occasion cannot be precisely predicted since they are governed by a random mechanism. An example of a time-varying covariate that is stochastic is current blood glucose level. In an observational study of diabetics, participants' blood sugar levels can vary randomly over the duration of the study. Additional examples include current smoking status or cumulative pack years, blood pressure, cholesterol level, fat intake, and exposure to environmental pollutants. As we shall see later, when a covariate is both time-varying and stochastic, new issues arise concerning the interpretation and estimation of regression parameters in models for longitudinal data.

Recall that many of the models for the mean response described in earlier chapters can be specified as

$$g\{E(Y_i|X_i)\} = X_i\beta \tag{15.6}$$

for some known link function $g(\cdot)$. This use of vector and matrix notation implies that the model for the mean at each occasion is given by

$$g\{E(Y_{ij}|X_i)\} = X'_{ij}\beta; \quad (j = 1, ..., n_i).$$

However, what is often overlooked is the implicit assumption that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends only on X_{ij}

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, \ ., X_{in_i}) = E(Y_{ij}|X_{ij}). \tag{15.7}$$

With time-stationary covariates, this assumption necessarily holds since $X_{ij} = X_{ik}$ for all occasions $k \neq j$. Also, with time-varying covariates that are fixed by design of the study (e.g., treatment group indicator in a crossover trial), the assumption also holds since values of the covariates at any occasion are determined *a priori* by study design and in a manner completely unrelated to the longitudinal response. However, when a covariate is time-varying and stochastic (15.7) may not necessarily hold. For

example, the assumption will be violated when the current value of Y_{ij}, given X_{ij}, predicts the subsequent value of X_{ij+1}. In that case

$$E(Y_{ij}|X_{ij}, X_{ij+1}) \neq E(Y_{ij}|X_{ij}),$$

and X_{ij+1} is said to confound the relationship between Y_{ij} and X_{ij}. In general, when (15.7) does not hold, then preceding and/or subsequent values of the time-varying covariate confound the relationship between Y_{ij} and X_{ij}; this can lead to biased estimates of β in (15.6).

To fix ideas, consider a longitudinal study designed to examine the effects of physical exercise on reducing blood glucose levels in patients with type 2 diabetes mellitus. We let X_{ij} denote the cumulative amount of physical activity at the j^{th} occasion and Y_{ij} denote a measure of blood glucose. The goal of the study is to determine the relationship between Y_{ij} and X_{ij}. Next, suppose that subjects with elevated blood glucose levels at the j^{th} occasion subsequently increase their level of physical activity, while subjects with the same cumulative amount of physical activity at the j^{th} occasion, but with normal blood glucose levels, continue to maintain their usual level of physical activity. Then, the assumption that

$$E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij})$$

does not hold and the relationship between Y_{ij} and X_{ij} is confounded by X_{ij+1}. In particular, the strength of the relationship between Y_{ij} and X_{ij} will be underestimated if subjects with elevated blood glucose levels subsequently increase their amount of physical activity.

In general, when a covariate is time-varying and stochastic, much greater care is needed in modelling its relationship to the response variable. It is important to assess the assumption made in (15.7), namely, that the conditional mean of Y_{ij}, given the entire time-varying covariate profile $X_{i1}, ..., X_{in_i}$, depends only on the covariate value at the j^{th} occasion, X_{ij}. Note, however, that X_{ij} can be defined in terms of functions of the explanatory variables measured at or preceding the j^{th} occasion (e.g., cumulative exposure at the j^{th} occasion). When (15.7) is violated the relationship between the mean of Y_{ij} and X_{ij}, expressed in terms of β, will be confounded by preceding and/or subsequent values of the covariate and misleading inferences about β can result.

Finally, even when (15.7) holds, there can be problems with the interpretation of regression parameters relating the mean response to stochastically time-varying covariates. In particular, the regression parameters β in (15.6) may not have the implied causal interpretation. For example, the model given by (15.6) may correctly specify the relationship between mean blood glucose level and physical activity at the last measurement occasion, since at the last occasion X_{in_i}, the cumulative amount of physical activity, is a function of the entire time-varying covariate profile. However, even though (15.7) holds, the regression parameters β may not have the implied causal interpretation without making additional assumptions. To see why, let us consider a simplified version of the example discussed earlier.

Suppose that a group of diabetics are measured at two occasions. Let Y_{i1} and Y_{i2} denote the blood glucose levels at baseline and follow-up, and X_{i1} and X_{i2}

denote measures of physical activity at the two occasions. Suppose it is of interest to determine the association between the cumulative amount of physical activity, $X_i^* = X_{i1} + X_{i2}$, and blood glucose level at the completion of the study, Y_{i2}. The following model is assumed:

$$E(Y_{i2}|X_i^*) = \beta_1 + \beta_2 X_i^*,$$

where, for ease of exposition, an identity link function is assumed. In this model β_2 appears to have interpretation as the effect of a unit increase in the cumulative amount of physical activity on the mean blood glucose level at follow-up, since

$$E(Y_{i2}|X_i^* = x + 1) - E(Y_{i2}|X_i^* = x) = \beta_2.$$

However, because X_{ij} is time-varying and stochastic, this interpretation of β_2 rests on the validity of *either* of the following two assumptions:

(i) Y_{i2} is not predicted by Y_{i1}, given X_{i1} and X_{i2};

or

(ii) X_{i2} is not predicted by Y_{i1}, given X_{i1}.

In particular, if neither of these assumptions hold, Y_{i1} "confounds" the relationship between Y_{i2} and X_i^* and β_2 does not have the desired causal interpretation. We loosely use the term "confounding" to emphasize that Y_{i1} obscures or distorts the association of real scientific interest between Y_{i2} and X_i^*. Strictly speaking, Y_{i1} can be considered both a "confounder" and a so-called "intermediate variable" on the causal path between Y_{i2} and X_i^*. When Y_{i1} is both a confounder and an intermediate variable, standard methods of adjustment for confounding no longer apply (since Y_{i1} is predicted by X_{i1}, and so should not be adjusted for, but also predicts X_{i2}, and so should be adjusted for in the analysis of the association between Y_{i2} and X_i^*). Instead, advanced statistical methods for causal inference are required when neither (i) or (ii) hold. However, a discussion of statistical methods for causal inference is beyond the scope of this chapter; some references to the statistical literature on this topic appear at the end of the chapter.

Let us consider these two assumptions in context. In a longitudinal study, it is unlikely that (i) would ever hold, since the repeated responses are usually positively correlated (given the covariates, X_{i1} and X_{i2}). Therefore, the causal interpretation of β_2 usually rests on the validity of (ii). For example, the assumption made in (ii) would be violated if subjects with elevated blood glucose levels at baseline subsequently increase their level of physical activity, while subjects with the same amount of physical activity at baseline, but with normal blood glucose levels, continue to maintain their usual level of physical activity. When assumption (ii) holds, the covariate is said to be *external* with respect to the response variable and β has the desired causal interpretation.

In summary, when a covariate is both time-varying and stochastic, we must consider the relationship between the response at any occasion, say Y_{ij}, and the subsequent value of the covariate, X_{ij+1}. A time-varying covariate is said to be *external*

when the current and preceding values of the response at the j^{th} occasion $(Y_{i1}, ..., Y_{ij})$, given the current and preceding values of the time-varying covariate $(X_{i1}, ..., X_{ij})$, do not predict the subsequent value of X_{ij+1}. More formally, a time-varying covariate is *external* (or sometimes referred to as *exogenous*) when

$$f(X_{ij+1}|X_{i1}, ..., X_{ij}, Y_{i1}, ..., Y_{ij}) = f(X_{ij+1}|X_{i1}, ..., X_{ij}); \qquad (15.8)$$

otherwise, the covariate is said to be *internal* (or *endogenous*). This generalizes the assumption made in (ii). Note that when a covariate is external,

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{i1}, ..., X_{ij}),$$

which is a weaker assumption than (15.7). An example of an external covariate is air pollution in studies of children's lung function growth. The outdoor levels of air pollutants (e.g., ozone, fine suspended particulate matter, and sulfur dioxide) are time-varying and stochastic, but conditional on past values, future values are not predicted by the lung function responses of the study participants and (15.8) holds. Note, however, that children's personal exposure to air pollution would not be considered an external covariate if children with poor lung function growth subsequently altered their daily behavior (e.g., spending less time outdoors) to avoid exposure to high levels of air pollution. In principle, it is possible to examine the assumption that a time-varying covariate is *external* by considering regression models for the dependence of X_{ij} on $Y_{i1}, ..., Y_{ij-1}$ (or some known function(s) of $Y_{i1}, ..., Y_{ij-1}$) and $X_{i1}, ..., X_{ij-1}$ (or some known function(s) of $X_{i1}, ..., X_{ij-1}$). The absence of any relationships between X_{ij} and $Y_{i1}, ..., Y_{ij-1}$, given the preceding covariate profile, $X_{i1}, ..., X_{ij-1}$, provides support for the validity of the assumption that the covariate process is *external*.

In conclusion, when covariates are time-varying and stochastic the regression parameters do not necessarily have the implied causal interpretation even when (15.7) holds. The regression parameters can be given a causal interpretation only when it can be further assumed that the time-varying covariates are external with respect to the response variable (i.e., when (15.8) holds).

15.4 LONGITUDINAL AND CROSS-SECTIONAL INFORMATION

In Chapter 1 we discussed the main distinctions between a longitudinal and cross-sectional study. In particular, we emphasized that the assessment of within-subject changes in the response over time can only be achieved within a longitudinal study design. In a cross-sectional study, where the response is measured at a single occasion, we cannot estimate the effect of growth or aging (an inherently within-subject effect); instead, we can only make comparisons among sub-populations that happen to differ in age. However, when the effect of aging is determined from a cross-sectional study, it is potentially confounded with cohort effects.

When an initial cross-sectional sample is measured repeatedly through time, it is then possible to make comparisons of longitudinal (or within-subject) and cross-sectional (or between-subject) estimates of changes in the response. For example,

the Muscatine Coronary Risk Factor (MCRF) study enrolled five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15 years. Repeated measurements of obesity were obtained biennially, from 1977 to 1981, with the objective of determining whether the risk of obesity increased with age. Note that the data from the MCRF study are unbalanced over time when the age of the child is used as the metameter for time. That is, baseline measurements are taken at the same calendar time (1977) for all subjects but age at entry to the study varies with subjects. As a result, there are two potential sources of information about changes in risk of obesity with age. First, there is cross-sectional or between-subject information about how the risk of obesity changes with age in the baseline observations obtained in 1977, since children enter the study at different ages. Second, there is longitudinal or within-subject information that arises because children are measured repeatedly over time, yielding measurements of obesity at different ages. It is possible that these two sources of information provide conflicting information about how the risk of obesity changes with age.

The main goal, indeed the *raison d'être*, of a longitudinal study is to characterize within-individual changes in the response over time. However, when a study provides both longitudinal and cross-sectional information, these two sources of information can be at odds. Therefore, somewhat greater care must be exercised in specifying models for the response to avoid confounding of longitudinal effects with cross-sectional effects. Specifically, it is important to consider a model for the data that includes separate parameters that represent the cross-sectional and longitudinal effects of age on the response. By doing so, it is possible to compare the cross-sectional and longitudinal effects, and report separate effects where necessary, or estimate a combined effect, based on both sources of information, if appropriate. In this section, we present a model for longitudinal data that allows simultaneous estimation of the cross-sectional and longitudinal effects of age on the response.

To accommodate unbalanced data, we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at age t_{ij}. Here, in a slight departure from the notation used in previous chapters, t_{ij} denotes the age of the i^{th} subject at the j^{th} measurement occasion. Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates. The vector of covariates can be partitioned into two main types of covariates: covariates whose values do not change throughout the duration of the study and covariates whose values change over time. The former are referred to as time-stationary or between-subject covariates (e.g., gender and fixed experimental treatments), while the latter are referred to as time-varying or within-subject covariates (e.g., time since baseline, current smoking status, and environmental exposures). In another departure from the notation used in previous chapters, we let X_{ij} denote the $q \times 1$ vector of time-varying covariates and Z_{ij} denote the $(p - q) \times 1$ vector of time-stationary covariates. For the latter, the same values of the covariates are replicated in the corresponding rows of Z_{ij}, for $j = 1, ..., n_i$; so we can drop the second subscript and denote the time-stationary covariates by Z_i.

A method for simultaneously modelling cross-sectional and longitudinal effects can be motivated by the following linear model for Y_{ij}:

$$Y_{ij} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij},$$

where X_{ij}' is the row vector of q time-varying covariates for the j^{th} response on the i^{th} individual. This representation allows both cross-sectional effects, $\beta^{(C)}$, and longitudinal effects, $\beta^{(L)}$, to be modelled simultaneously. The interpretation of the model parameters, $\beta^{(C)}$ and $\beta^{(L)}$, becomes more transparent when the implied models for the initial response and subsequent within-subject changes are considered. First, consider the model for the initial response, Y_{i1},

$$Y_{i1} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1},$$

since $(X_{i1}' - X_{i1}') = 0$. In the model for the initial response, $\beta^{(C)}$ represents a vector of regression parameters for cross-sectional effects. For example, certain components of $\beta^{(C)}$ represent cross-sectional effects of age since they describe how the mean response at baseline changes with age at baseline. The regression parameters β_0 represent the effects of the time-stationary covariates. Next, consider the model for the within-subject changes from the initial response, $Y_{ij} - Y_{i1}$,

$$\begin{aligned}(Y_{ij} - Y_{i1}) &= Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij} \\ &\quad - (Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1}) \\ &= (X_{ij}' - X_{i1}')\beta^{(L)} + (e_{ij} - e_{i1}).\end{aligned}$$

In the model for the within-subject changes, $Y_{ij} - Y_{i1}$, $\beta^{(L)}$ represents a vector of regression parameters for longitudinal effects. For example, certain components of $\beta^{(L)}$ represent longitudinal effects of age since they describe how within-subject changes in the response are related to within-subject changes in age.

One advantage of simultaneously modelling cross-sectional and longitudinal effect. is that formal comparisons can be made by testing H_0: $\beta^{(C)} = \beta^{(L)}$ (or by comparing certain components of $\beta^{(C)}$ with the corresponding components of $\beta^{(L)}$). Differences between $\beta^{(C)}$ and $\beta^{(L)}$ can arise when there are cohort or period effects. Cohort effects will introduce bias in the cross-sectional estimates but not the longitudinal estimates. Period effects will introduce bias in the longitudinal estimates but not the cross-sectional estimates. Alternatively, differences between $\beta^{(C)}$ and $\beta^{(L)}$ can be due to the biasing effects of selective dropouts. Note that w: en $\beta^{(C)} = \beta^{(L)} = \beta$, the model simplifies to

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta + e_{ij}.$$

On the other hand, when $\beta^{(C)} \neq \beta^{(L)}$ but the model for the data does not allow for separate estimation of the cross-sectional and longitudinal effects on the response, that is,

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta + e_{ij},$$

then β cannot be interpreted as pure longitudinal effects. Instead, the parameters of β are some weighted combination of $\beta^{(C)}$ and $\beta^{(L)}$ and may not reflect effects of subject-matter interest. That is, failure to distinguish cross-sectional and longitudinal effects can result in a distorted estimate of the effects of age that reflects neither the cross-sectional nor the longitudinal effects of age on the response.

Finally, this method for simultaneously modelling cross-sectional and longitudinal effects posits the following linear model for the mean response:

$$E(Y_{ij}) = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)},$$

since the errors, e_{ij}, have mean equal to zero. This motivates how the method can be extended to different types of responses (e.g., binary responses and counts) by considering the following generalized linear model for the mean response:

$$g\{E(Y_{ij})\} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)},$$

for any suitable link function $g(\cdot)$.

Illustration

The main distinction between cross-sectional and longitudinal effects is highlighted in the following simple illustration. Suppose that three age-cohorts of children, initially aged 5, 6, and 7 years, are measured at baseline and followed annually for three years. Suppose the cross-sectional effect of age on the baseline response is linear, with

$$E(Y_{i1}) = \beta^{(C)}\text{Age}_{i1},$$

(for simplicity, a model with intercept equal to zero is assumed) and the mean response increases linearly with changes in age in each cohort

$$E(Y_{ij} - Y_{i1}) = \beta^{(L)}(\text{Age}_{ij} - \text{Age}_{i1}),$$

but with slope $\beta^{(L)} \neq \beta^{(C)}$. A graphical representation of the model for the mean response versus age, when $\beta^{(C)} = 0.75$ and $\beta^{(L)} = 0.25$, is given in Figure 15.1. In this illustration there is a discernible difference between the longitudinal (solid line) and cross-sectional (dotted line) effects of aging on the mean response. When these differences between the longitudinal and cross-sectional effects of aging are ignored (see Figure 15.1), changes in the mean response (averaged over the three age-cohorts) with age of measurement (dashed line) reflect a combination of $\beta^{(C)}$ and $\beta^{(L)}$. For example, the change in mean response from age 5 to age 6 is $(\beta^{(L)} + \beta^{(C)})/2 = 0.5$. This is the average of the longitudinal effect of aging in the cohort of children initially aged 5 and the cross-sectional effect of aging obtained from comparing the baseline mean response in the cohorts of children initially aged 5 and 6. In general, when differences between the longitudinal and cross-sectional effects are ignored, the naive analysis assuming $\beta^{(C)} = \beta^{(L)}$ estimates a weighted average of $\beta^{(C)}$ and $\beta^{(L)}$.

This simple illustration highlights why the parameters β in standard regression models for longitudinal data that do not incorporate separate parameters for cross-sectional and longitudinal effects of aging can yield misleading inferences. That

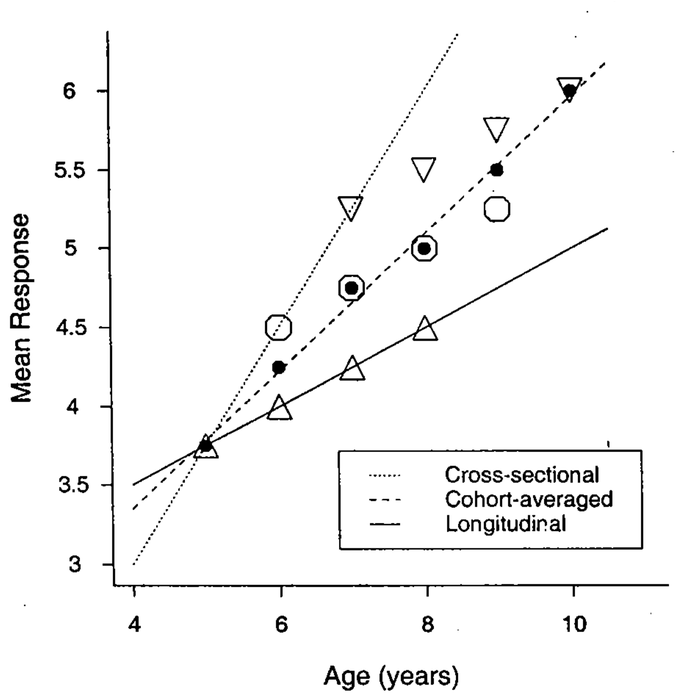

Fig. 15.1 Plot of the longitudinal, cross-sectional, and cohort-averaged regression lines for the three age-cohorts: △ denotes mean response of the age-cohort of children initially aged 5 years; ○ denotes mean response of the age-cohort of children initially aged 6 years; and ▽ denotes mean response of the age-cohort of children initially aged 7 years. (● denotes averages over cohorts).

is, failure to acknowledge that the cross-sectional effect of aging differs from the longitudinal effect of aging can lead to a conclusion about the effect of aging that confounds one effect with the other.

15.5 FURTHER READING

Schlesselman (1973a,b), in two companion papers on the design of longitudinal studies, discusses sample size calculations and issues concerning study duration and the frequency of measurement; also see Raudenbush and Liu (2001). Hedeker *et al.* (1999) discuss sample size estimation for longitudinal study designs with a continuous outcome and allow for attrition and a variety of covariance structures for the repeated measurements.

Ware *et al.* (1990) present an accessible discussion of regression models for longitudinal data that incorporate separate parameters for cross-sectional and longitudinal

effects of aging; also see discussion of discrepancies between longitudinal and cross-sectional effects in Louis *et al.* (1986).

Bibliographic Notes

Lipsitz and Fitzmaurice (1994) describe a method, based on generalized least squares, for calculating sample size for longitudinal studies with binary responses; also see Pan (2001). A more general method for computing sample size and statistical power for longitudinal studies, based on the generalized estimating equations approach, has been developed by Liu and Liang (1997); this method does not, in general, yield closed-form expressions for sample size and power, but the method can be implemented numerically.

The implicit assumption in marginal regression models with time-varying covariates given by (15.7) is discussed in Fitzmaurice *et al.* (1993), Pepe and Anderson (1994), Robins *et al.* (1999), and Pan *et al.* (2000).

A general discussion of methods for estimating the causal effect of time-varying covariates in marginal models for longitudinal data can be found in Robins *et al.* (1999) and the references therein. Chapter 12 of Diggle *et al.* (2002) presents a useful summary of the key ideas.

16

Repeated Measures and Related Designs

16.1 INTRODUCTION

In this chapter we discuss the application of methods for longitudinal data to closely related study designs. In these settings, individuals have multiple commensurate measurements made under different circumstances and possibly also at different times. However, the major interest of the analysis is not in changes in the response over time, but in the effect of different circumstances of measurement and/or the effects of covariates on the responses.

The first design that we will consider is the classical repeated measures design. In this setting, each subject is measured under a fixed number of different conditions, often corresponding to different treatments. Interest centers on comparing the effects of the different experimental conditions on the outcome. Similar to time in a longitudinal study, experimental condition is a within-subject factor and the conditions are compared using within-subject contrasts. Such designs are popular because subject-to-subject differences in outcome are accounted for in the design. Since each subject acts as his or her own control, comparisons of the outcome under different experimental conditions are estimated free of any between-subject variation in the outcome. As a result, the design is potentially very efficient relative to comparing the different experimental conditions on different groups of subjects.

The second design is one that produces what we refer to as "multiple source" data. In this setting, the primary outcome of interest is measured by more than one instrument or rater. This frequently happens when the outcome is difficult to measure and is thus determined under multiple different circumstances. In Section 16.3 we describe an example in the context of measuring psychopathology in children. In

this context, there may be some interest in comparing the different measures, but ordinarily the main focus of the analysis is the effects of subject-specific covariates on the outcome. Hence, unlike a typical longitudinal study and also unlike a classical repeated measures design, the main interest centers on the effects of subject-specific variables on response, and possibly also their interaction with the multiple sources. In many settings, the fact that there are multiple sources could be regarded as a "nuisance" feature of the study design.

The distinction we make between repeated measures and multiple source data is based on what is of primary interest in the analysis. Both share the same analytical methods for regression models with correlated data. Sometimes, however, this distinction is blurred. For example, Hernández *et al.* (2000) report on a validity study of a new questionnaire designed to measure physical activity and inactivity in school children in Mexico. A self-reported questionnaire was administered both to the mother and the child on two different occasions; in addition, a 24-hour recall (considered the best measure, but only limited to a single day) was administered on each occasion. The average of the two 24-hour recalls was considered the "gold standard" for the analysis. Here the objective of the analysis was two-fold: to compare mean responses of the two child and two mother assessments to the average of the two 24-hour recalls, and to look at the correlations between these measures. This is clearly a multiple source data set, since the mother and child questionnaires were both intended to measure the child's average activity levels over the period. However, the primary focus of the analysis was comparing the multiple reports to each other and to the "gold standard" and examining the correlations; here the subject-specific variables, age, gender, and socio-economic level, were used to adjust the means and the correlations for between-subject differences. Thus, in this example, the analytic goal of comparing means is closer to a repeated measures analysis than a multiple source analysis, although comparing correlations is more typical in the multiple source analysis.

We will first describe the main features of these two designs in greater detail, and then provide some examples to illustrate the application of regression methods for correlated data to repeated measures and multiple source data.

16.2 REPEATED MEASURES DESIGNS

Repeated measures designs are frequently encountered in applications. In the experimental context, the repeated measures design is also sometime called a randomized block design. In the simplest setting, subjects each receive n treatments or experimental conditions, and the outcome is recorded for each condition. Thus each subject has a vector of n measurements, $Y_i = (Y_{i1}, \ldots, Y_{in})'$. The treatments may be given sequentially in a randomly assigned order, but in some settings they can be given simultaneously. An example of the latter is a classic study of topical treatments for leprosy, where each patient was given four different treatments simultaneously at four different locations on their body. After a number of days, the skin lesions were recorded at each of the four locations. Letting Y_{ij} denote the response to the j^{th} treatment ($j = 1, ..., 4$), the primary interest is in comparing the four treatments. However, the analysis must account for the fact that the observations on different treatments, $Y_i = (Y_{i1}, ..., Y_{i4})'$, are correlated.

As mentioned earlier, repeated measures designs are popular because subject differences in outcome are accounted for in the design. By removing between-subject variation in the outcome from the comparisons of different experimental conditions, the repeated measures design can be very efficient relative to comparing the different experimental conditions on different groups of subjects. To illustrate this point, consider a simple repeated measures design with N subjects receiving $n = 2$ treatment conditions (producing a total of $2N$ observations). Let Y_{i1} denote the response for the i^{th} subject under the first condition and Y_{i2} denote the response under the second condition. A natural estimate of the effect of treatment on the mean response is

$$\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_2,$$

where

$$\widehat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}.$$

The variability of this estimator of the effect of treatment is given by

$$\mathrm{Var}(\hat{\delta}) = \mathrm{Var}\left\{\frac{1}{N}\sum_{i=1}^{N}(Y_{i1} - Y_{i2})\right\} = \frac{1}{N}\left(\sigma_1^2 + \sigma_2^2 - 2\,\sigma_{12}\right),$$

where $\sigma_j^2 = \mathrm{Var}(Y_{ij})$, $\sigma_{12} = \mathrm{Cov}(Y_{i1}, Y_{i2}) = \rho\,\sigma_1\sigma_2$, and $\rho = \mathrm{Corr}(Y_{i1}, Y_{i2})$. To simplify, we assume that treatment may have an impact on the mean response, but not on the variance; this assumption can be justified when treatments are allocated within subjects randomly by time. Then, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and the variability of this estimator simplifies to

$$\mathrm{Var}(\hat{\delta}) = \frac{2}{N}\left\{\sigma^2(1-\rho)\right\}.$$

Next, consider a two-group study design, where the treatment conditions are randomized to $2N$ subjects drawn from the population, with N subjects allocated to one condition, and the remaining N subjects allocated to the other condition (producing a total of $2N$ observations). The effect of treatment on the mean response can be estimated by comparing the mean response in the two groups of subjects. The natural estimate of treatment effect is the same as before:

$$\widehat{\gamma} = \widehat{\mu}_1 - \widehat{\mu}_2,$$

where

$$\widehat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}$$

is the sample mean response in the group of subjects that were randomized to the j^{th} treatment condition. The variability of this estimator of the effect of treatment is

given by

$$\mathrm{Var}(\hat{\gamma}) = \frac{1}{N}\left(\sigma_1^2 + \sigma_2^2\right),$$

where $\sigma_j^2 = \mathrm{Var}(Y_{ij})$. Again, if we assume that treatment may have an impact on the mean response, but not on the variance, then $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and

$$\mathrm{Var}(\hat{\gamma}) = \frac{2\sigma^2}{N}.$$

The potential gain in efficiency of the repeated measures design is quantified by taking the ratio of the variances of the two estimators of the effect of treatment:

$$\frac{\mathrm{Var}(\hat{\delta})}{\mathrm{Var}(\hat{\gamma})} = (1 - \rho).$$

Provided that the correlation among the repeated measures is positive, the repeated measures design provides a more precise estimate of the effect of treatment. For example, when $\rho = 0.75$, it is four times more efficient than the two-group design or, put another way, the repeated measures design only requires a quarter of the number of observations to attain the same level of precision as the two-group design.

There are many variations on the repeated measures design. If, for example, treatments are given sequentially in a random order, it is common to use a restricted randomization scheme to balance the treatments over time. This would allow equal numbers of each treatment to be assigned at each time point, for example, with two treatments denoted A and B, half the subjects would be assigned to AB, and half to BA. Thus, systematic time or sequence differences will be eliminated, and the design is potentially more efficient. In this case the number of subjects is a multiple of n, the number of treatment conditions. A special case of these designs is the crossover design, where each subject receives each treatment in a random order.

A closely related design is the split-plot design. Here there are two treatment factors, a so-called main plot factor and a sub-plot factor. When subjects are the main plots, the main plot factor is a between-subject factor and each subject receives only one level of this factor. The sub-plot factor is a within-subject factor and each subject receives all levels of the within-subject factor. Thus, in the split-plot design, one factor is randomized within subjects, just as in the repeated measures design, and the other factor is randomized between subjects. For example, in a study comparing the effects of different antibiotics (drugs) and topical gels (gels) for the treatment of eye infections, subjects (main plots) are randomized to receive one of three different oral antibiotics. The main plot factor is drug (antibiotics). In addition, each subject is required to apply two different topical gels directly to the left and right eye; the two different gels are randomized to the left and right eyes. Here, the sub-plot factor is gels. In the split-plot design, both the main effects and the interaction of the two factors are usually of interest, although this design provides more precise information about the within-subject factor than the between-subject factor.

Much of the literature on repeated measures is in the context of designed experiments, where treatments are allocated within subjects randomly by time or location.

This has several ramifications for our discussion. First, because the factors in a randomized design are typically categorical, the analysis of these designs is ordinarily presented in the context of analysis of variance (ANOVA) rather than a general regression model with correlated data. However, the analysis of variance model can be viewed as a special case of the general linear regression model for correlated data presented in Part II. Hence the regression models for correlated data apply quite straightforwardly to the classical repeated measures design.

Second, with randomization, arguments can be made that allow one to simplify the analysis of repeated measures data, especially with balanced and complete designs. There are two main approaches to the analysis of repeated measures data (see Section 3.6), repeated measures analysis by ANOVA and repeated measures analysis by multivariate ANOVA (MANOVA). With the former, one assumes that any contrast between any two repeated measures on the same subject, say $Y_{ij} - Y_{ik}$, has the same variance, that is, $\mathrm{Var}(Y_{ij} - Y_{ik})$ is constant for all choices of i and $j \neq k$. For example, if the covariance matrix of the vector of repeated observations, Y_i, takes on a compound symmetry form, then the requirement for the repeated measures analysis by ANOVA is satisfied. In contrast, the model for repeated measures analysis by MANOVA allows the vector of repeated observations to have an arbitrary covariance structure, but the standard analysis is usually limited to balanced and complete designs. In addition, the analysis of repeated measures by ANOVA can be considerably more powerful than the analysis by MANOVA (when the assumptions for the former hold).

If the within-subject factor is randomly allocated to subjects, then randomization arguments can be made to show that the constant variance condition on the contrast does hold. More generally, one can show that $\mathrm{Var}(Y_{ij})$ is constant for all i and j and $\mathrm{Cov}(Y_{ij}, Y_{ik})$ is constant for all i and $j \neq k$. The general approach to a randomization argument involves treating the randomization indicators as random variables, and the observed outcomes as fixed. While the approach provides an attractive justification for using the repeated measures ANOVA in the randomized experiment, it may not be justifiable in the more general setting. In addition, the justification relies on a linear model. Note that with constant variance and covariance, we can formulate a repeated measures analysis using mixed effects models with a random subject effect b_i (where $\mathrm{Var}(b_i) = \sigma_b^2$) and independent errors e_{ij} (where $\mathrm{Var}(e_{ij}) = \sigma_e^2$). Then $\mathrm{Var}(\widehat{\delta}) = 2\,\sigma_e^2/N$ and $\mathrm{Var}(\widehat{\gamma}) = 2(\sigma_e^2 + \sigma_b^2)/N$, because b_i drops out of any within-subject contrast.

Finally, there are many variations on repeated measures designs which are difficult to handle with the classical analytic approaches. For example, missing data lead to unbalanced designs with a variable number of observations on subjects. In addition, the outcomes may be count or binary data, neither of which can be handled by classical ANOVA techniques. Using generalized linear models for correlated outcomes enables us to handle these variations in a unified way.

16.3 MULTIPLE SOURCE DATA

Multiple source data usually arise in the context of epidemiological studies, where outcomes and/or risk factors may be difficult to measure. For example, Fitzmaurice *et al.* (1996) discuss a study involving child psychopathology, which used standardized questionnaires administered to both the child's mother and teacher. Responses to these questionnaires were used to construct, for each informant, dichotomous indicators of whether psychopathology was present in the child. Other informants sometimes used in this setting include the child, the father, and peers. Because informants interact in different settings with the child, the information from different informants reflects different perspectives. However, the primary interest of the study was not to compare informants, but rather to study the effects of covariates on child psychopathology, broadly defined. By including informant as a variable in the correlated data model, and its interaction with other covariates of interest, one can also examine how informants differ overall, and whether or not covariate effects differ by informant.

In this section we use the term "multiple source" to encompass data that are simultaneously obtained from multiple informants or raters (e.g., self-reports, family members, health care providers, administrators) or via different/parallel instruments or methods (e.g., symptom rating scales, standardized diagnostic interviews, or clinical diagnoses). For example, in psychiatric studies of children, the child's parent is routinely used as a proxy data source; other sources (e.g., self-report, peers, teachers, clinicians, or trained observers) may also be employed, depending on the child's age and the nature of psychopathology under study. Multiple source data have become increasingly common in hospital-based and outpatient-based assessments of the effectiveness of treatments. For example, evaluations of managed care programs for the United States Medicaid population require analysis of multiple sources of information, including patient satisfaction with health care, treatment utilization, and appropriate care. Other areas where multiple reports arise include studies of severe mental illness, such as schizophrenia or Alzheimer's disease, where the affected subject is often unable to provide self-report data, family history studies, where many relatives are interviewed about the status of the proband and other family members, and behavioral studies of alcohol/drug use or of eating disorders, where information is obtained from the subject, as well as family members or other sources.

Historically there has been little consensus as to how to analyze multiple source data. Sometimes investigators conduct completely separate univariate analyses for each source. Alternately the multiple source measures may be combined to make a single outcome. For measured responses, investigators may take a mean and for dichotomous responses the "and" or the "or" rules are often used. In the "and" rule binary source data are considered to be positive if *all* of the source data are positive, and negative otherwise; in the "or" rule binary source data are considered to be positive if *any* of the source data are positive, and negative otherwise. All of these *ad hoc* strategies usually require discarding data when any sources are missing; the separate analysis strategy makes it difficult to compare the results for the two (or more) sources, while the analysis using a combined response can obscure interesting differences. Using correlated data models similar to those discussed in previous chapters allows one to directly compare source effects and to handle missing data in a unified framework.

16.4 CASE STUDY 1: REPEATED MEASURES EXPERIMENT

In this section we analyze data from a randomized crossover design to illustrate the use of mixed effects models in the repeated measures setting. The study was designed to compare two active drugs and placebo for relief of tension headache. The two analgesic drugs were identical in their active ingredients except that one contained caffeine. The primary comparison of interest was between the two active drugs; the placebo was included for regulatory purposes. What makes the analysis non-standard is that there were three treatments, but only two periods; that is, each subject received only two of the three treatments in a random order. With three treatments and two periods, there are six possible treatment sequences, AB, BA, AP, PA, BP, PB, where A, B and P denote the two active drugs and placebo. If each sequence is assigned an equal number of subjects, then we have what is known as a balanced incomplete block design. It is balanced because all possible sequences are equally represented, but incomplete because each subject is "missing" a third treatment. In our example, the AB and BA sequences were assigned three times as many subjects as the remaining four because of the interest in the A versus B comparison. The descriptive statistics for one measure of pain relief used in the crossover study of analgesics are given in Table 16.1. There were actually two headaches treated within each period, but that feature of the data is ignored here and we use the average of the two measures of pain relief. A few observations were missing, but we have access only to subjects with complete data.

An important issue in crossover designs is the possibility of carryover effects. The presence of a carryover effect means that the treatment taken in the first period may influence the treatment effect in the second period, that is, the drug in the first period carries over. In our analysis we will show the results of fitting two different models, one with only treatment and period effects included in the model and one with the carryover effects included. We display the model which includes carryover effects by giving the expected value of the response for each sequence and each period. Letting Y_{ij} denote the treatment response for the i^{th} subject in the j^{th} period, we write:

$$E(Y_{ij}) = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_6 X_{6ij}, \qquad (16.1)$$

where $X_{1ij} = 1$ for all i and j, X_{2ij} and X_{3ij} are dummy variable indicators of the main effects of treatment A and B (P is the reference condition), X_{4ij} is a dummy variable indicator of period 2 (so for all i, $X_{4i1} = 0$ and $X_{4i2} = 1$) and X_{5ij} and X_{6ij} are dummy variable indicators of the carryover effects for A and B, respectively. For subjects assigned to treatment A in the j^{th} period, $X_{2ij} = 1$, and 0 otherwise; for those assigned B in the j^{th} period, $X_{3ij} = 1$, and 0 otherwise ($X_{2ij} = X_{3ij} = 0$ for subjects assigned to P in the j^{th} period). For subjects in sequences where A is taken first $X_{5i2} = 1$, and $X_{6i2} = 1$ if the first treatment taken is B; for all other values of i, X_{5i2} and X_{6i2} are zero. Likewise, for all i, $X_{5i1} = X_{6i1} = 0$ since by definition

Table 16.1 Descriptive statistics (means, standard deviations, and correlation), by sequence, for total pain relief for headache in periods 1 and 2.

		Period 1		Period 2		
Sequence	N	Mean	SD	Mean	SD	$\rho^\dagger$
AB	126	10.196	3.347	9.153	3.429	0.20
BA	127	9.581	3.881	10.791	3.530	0.30
AP	43	10.477	3.546	7.273	4.451	0.31
PA	43	8.366	3.777	10.855	3.204	0.47
BP	42	10.333	3.306	8.357	3.944	0.72
PB	42	7.464	4.265	9.911	4.183	0.67

†*Note*: ρ is the correlation between pain relief scores in periods 1 and 2. The estimate of the common correlation (ignoring sequence) is $\widehat{\rho} = 0.38$.

there are no carryover effects in the first period. In Table 16.2 we summarize the interpretations of $\beta_1, ..., \beta_6$ in terms of the mean response for each sequence during each period.

This model includes effects of treatment (β_2 and β_3), period (β_4), and carryover (β_5 and β_6). The active drug comparison is given by $\beta_2 - \beta_3$. In our model, the carryover effects are assumed to depend only on the treatments taken in period one. That is, the model assumes that the carryover of one active treatment to the other active treatment is the same as the carryover of that active treatment to the placebo. More complicated models can be used (see, e.g., Laird, Skinner and Kenward, 1992), but are not considered here.

Since subjects are randomized, we will assume $\text{Var}(Y_{i1}) = \text{Var}(Y_{i2})$. A more general covariance structure could easily be accommodated by allowing the variance to depend upon period, but we do not do so here. Using the compound symmetry assumption allows us to analyze the data using mixed effects models with a random subject effect. Handling missing responses simply involves removing rows from the design matrix and the response vector (as discussed in Section 5.3).

The balanced incomplete block design has two types of information about treatment. One source of information comes from within-subject contrasts; that is, in the absence of carryover effects, $Y_{i2} - Y_{i1}$ is a treatment contrast with variance $2\sigma^2(1-\rho)$. There is also information about treatment contrasts from between-subject comparisons, that is, in the absence of carryover effects, $Y_{ij} - Y_{i'j}$ estimates a treatment contrast for two subjects with different treatments in period j. Now, however, the variance of the contrast is $2\sigma^2$. When there are carryover effects in this design, information about treatment effects can be found in both within-subject and between-subject contrasts. If ρ is modest ($\rho < 0.4$), it is often suggested that the

Table 16.2 Regression parameters for the mean total pain relief for headache, by sequence and period.

Sequence	Period 1	Period 2
AB	$\beta_1 + \beta_2$	$\beta_1 + \beta_3 + \beta_4 + \beta_5$
BA	$\beta_1 + \beta_3$	$\beta_1 + \beta_2 + \beta_4 + \beta_6$
AP	$\beta_1 + \beta_2$	$\beta_1 + \beta_4 + \beta_5$
PA	β_1	$\beta_1 + \beta_2 + \beta_4$
BP	$\beta_1 + \beta_3$	$\beta_1 + \beta_4 + \beta_6$
PB	β_1	$\beta_1 + \beta_3 + \beta_4$

between-subject information be ignored, and the analysis use only the within-subject contrasts. The rationale behind this approach is that the within-subject contrast yields a simple structure with no need to estimate variance and covariance components for Y_{i1} and Y_{i2}, while using all the information requires estimating the between- and within-subject error variance. Of note, $\widehat{\rho} \approx 0.4$ for the pain relief data in our example (see Table 16.1).

The within-subject analysis is especially easy to do with just two periods by subtracting Y_{i1} from Y_{i2}, and analyzing the difference $d_i = Y_{i2} - Y_{i1}$:

$$E(d_i) = \beta_2(X_{2i2} - X_{2i1}) + \beta_3(X_{3i2} - X_{3i1}) + \beta_4 + \beta_5 X_{5i2} + \beta_6 X_{6i2}. \quad (16.2)$$

Note that β_1 has vanished from the model, and β_4 acts as the constant (or intercept) in the model, since $X_{4i1} = 0$ and $X_{4i2} = 1$ for all i. The parameters β_2 and β_3 remain the main effects of treatment and β_5 and β_6 remain the carryover effects. In our analysis the main focus is on the active drug comparison, $\beta_2 - \beta_3$. It is a special feature of this design that carryover effects can be estimated from within-subject contrasts (Koch *et al.*, 1989), but as we will show, most of the information about the carryover effects comes from the between-subject information. This issue is similar to that raised in Chapter 15 in the discussion of cross-sectional and longitudinal information. In the context of a balanced incomplete randomized trial, or a complete randomized trial with carryover effects, the issue of bias does not arise, only that of efficiency.

We will illustrate the analysis using both the simple regression on the differences (d_i) and a mixed effects model analysis on the full data (Y_{i1} and Y_{i2}), with subject as a random effect (b_i). Table 16.3 illustrates the conventional wisdom that the within-subject (differences) analysis is highly efficient for treatment effects when carryover is absent. (Compare the two rows of Table 16.3 for the treatment effect estimates assuming no carryover effects.) However, if carryover effects are present the within-subject analysis is very inefficient for both the main effect of treatment and for carryover effect. (Compare the two rows of Table 16.3 for the treatment and carryover effects estimates.) In this case investigators basing their conclusion on the within-subject (differences)

Table 16.3 Results of comparison of the two active treatments (estimate ± SE) based on analysis of differences and mixed effects model analysis, with and without carryover effects.

	No Carryover	With Carryover	
	Treatment	Treatment	Carryover
Method of Analysis	$(\beta_2 - \beta_3)$	$(\beta_2 - \beta_3)$	$(\beta_5 - \beta_6)$
Differences	1.06 ± 0.24	0.46 ± 0.67	-1.19 ± 1.26
Mixed Effects Model†	1.02 ± 0.23	0.55 ± 0.32	-1.06 ± 0.52

†Linear mixed effects model with random subject effect.

analysis would be led to conclude erroneously that the active drug comparison is not confounded with carryover ($Z = -1.19/1.26 = -0.94$, $p > 0.30$) and to use the results of the model without carryover to obtain a significant difference between the active treatments ($Z = 1.06/0.24 = 4.42$, $p < 0.0001$). That is, the inefficient analysis of carryover effects based on differences leads to the erroneous conclusion that the treatment comparison does not requite any adjustment for carryover effects. Using the mixed effects model analysis, however, we come to quite a different conclusion; there is a substantial carryover effect ($Z = -1.06/0.52 = -2.05$, $p < 0.05$), but no evidence of a statistically significant treatment effect ($Z = 0.55/0.32 = 1.71$, $p > 0.05$) when the carryover effects are included in the model and adjusted for in the analysis. The widespread availability of software to implement a mixed effects model analysis makes it relatively easy to capture the "between-subject information" even in complex repeated measures designs.

16.5 CASE STUDY 2: MULTIPLE SOURCE DATA

Data for this example come from two surveys of children's mental health (Zahner *et al.*, 1992; Zahner *et al.* 1993). A standardized measure of childhood psychopathology was used both by parents (Child Behavior Checklist, CBC) and teachers (Teacher Report Forms, TRF) to assess children in the study. We use here the externalizing scale, which assesses delinquent and aggressive behavior. The scale has been dichotomized at the cut point for borderline/clinical psychopathology. The cut points are normed separately for males and females; thus we expect to see small gender effects in these data. Because of the multiple levels of permissions and reporting, a substantial number of children were missing the TRF. Our analysis is based on 1428 children who had both parent and teacher responses, and an additional 1073 children with only a parent response; a total of 2501 children participated in the study. In this example the two sources or respondents are the children's parents and teachers; in the psychiatric literature, these sources are often referred to as "informants".

Table 16.4 Results of fitting separate logistic regressions to data on externalizing behavior from each source.

Informant	N	Intercept†	Single Parent†	Child Health†
Parent	2501	-2.156 ± 0.092	0.620 ± 0.124	0.600 ± 0.113
Teacher	1428	-1.694 ± 0.105	0.655 ± 0.157	0.175 ± 0.135

†Estimated coefficient ± standard error.

The objective of the analysis is to study the influence of several explanatory variables on the prevalence of externalizing behavior in these children. For simplicity, we limit our analysis to single parent status (coded 1: single, 0: otherwise) and child's physical health problems (coded 1: fair to bad health, 0: good health). In addition, we are interested in describing the level of association between the two respondents, and determining whether the effects of the covariates depend on informant. We will use standard regression models to address these issues, but because we have two different measures of the outcome, we will use correlated data models, one for each outcome. Since externalizing behavior is dichotomous, we use logistic models for the regressions.

The basic approach uses two separate regression models, one with the CBC as an outcome and one with the TRF as an outcome. Both models have the same set of covariates (here single parent status and physical health problems), but the coefficients may differ for the different sources. In addition, we have an "informant" indicator variable which identifies the source of the response. To motivate the approach, we first fit two completely separate logistic regressions each with the full complement of covariates, one for each informant outcome. Let μ^P and μ^T denote the probability of a positive response on externalizing behavior as measured by parents and teachers, respectively. Then the two regression models are:

$$\text{logit}(\mu_i^P) = X_i'\beta^P$$

and

$$\text{logit}(\mu_i^T) = X_i'\beta^T,$$

where X_i is a $p \times 1$ vector of covariates for the i^{th} subject, and

$$E(Y_i^P) = \mu_i^P \text{ and } E(Y_i^T) = \mu_i^T.$$

The first logistic regression model was fit with the parent response as outcome using all 2501 children, and the second was fit with the teacher response using only the 1428 children with a teacher response. The results are displayed in Table 16.4.

The estimated coefficients for single parent status are similar and statistically significant (at the 0.05 level) for each informant report; the estimated coefficients for

child health problems are rather different, and significant only for the parent report. In both cases, standard errors for the parent response are smaller, reflecting the larger sample size. Fitting these two logistic regression models separately does not allow us to formally quantify the differences in $\widehat{\beta}^P$ and $\widehat{\beta}^T$ because the estimated regression parameters are correlated.

We now show how to fit these two regression models simultaneously using multivariate methods that take the association between the responses into account. To begin, we rewrite the two separate models as a set of bivariate models with a common regression coefficient β, which will have dimension six (or $2p$ in general). First, we simply change notation as follows. Let $Y_i^P = Y_{i1}, Y_i^T = Y_{i2}, \mu_i^P = \mu_{i1}, \mu_i^T = \mu_{i2}, \beta' = (\beta^{P\prime}, \beta^{T\prime})$, and let Z_i be an $n_i \times 6$ matrix where n_i is the number of informants available for the i^{th} child. For a child with both informants ($n_i = 2$), the first row of Z_i is given by

$$Z_{1i}' = (X_i', 0, 0, 0)$$

and the second row of Z_i simply interchanges X_i' with (0,0,0),

$$Z_{2i}' = (0, 0, 0, X_i').$$

If an informant is missing ($n_i = 1$), we delete the row corresponding to that informant (i.e., delete Z_{2i} if the teacher does not give a report for the i^{th} child). We now write

$$E(Y_{ij}) = \mu_{ij}, \quad i = 1, ..., N \text{ and } j = 1, ..., n_i,$$

and specify the following bivariate model,

$$\text{logit}(\mu_{ij}) = Z_{ij}'\beta, \quad j = 1, 2. \tag{16.3}$$

This is exactly the mean model for a marginal model (see Chapter 11), with a special structure for the design matrix, here labelled Z_i instead of the usual X_i. To complete the marginal model, all we need is a specification of $\text{Cov}(Y_{i1}, Y_{i2})$, using one of the approaches discussed in Chapter 11. For the analysis here we use the odds ratio to measure the association. Given the model for the mean and $\text{Cov}(Y_{i1}, Y_{i2})$, GEE methods can be used to estimate β. If we specify the entire joint distribution for the Y_{ij}'s we can use maximum likelihood estimation (Fitzmaurice *et al.*, 1995). We choose to present a GEE analysis since it is easier to implement. Note that, because of the way we defined Z_i and β, the first three components of β correspond to β^P and the second three correspond to β^T. The variance-covariance matrix of $\widehat{\beta}$ is now 6×6; the 3×3 diagonal blocks are the variance-covariance matrices of $\widehat{\beta}^P$ and $\widehat{\beta}^T$, while the off-diagonal 3×3 block gives the covariance between the two.

Table 16.5 shows the results of fitting the regression model, using the log odds ratio to model association between parent and teacher response. The estimates and standard errors for β^P are nearly unchanged, as we might expect, but those for β^T are different, reflecting the fact that many children were missing the teacher response. The parent response provides some information about teacher response because of the relatively high association between parent and teacher response (estimated odds ratio is 4.75, with a 95% confidence interval of 3.52 to 6.39). For both $\widehat{\beta}^T$ and $\widehat{\beta}^P$ the

Table 16.5 Results of fitting two regression models simultaneously to externalizing behavior data on 2501 children using GEE method.

Informant	Intercept†	Single Parent†	Child Health†
Parent	-2.154 ± 0.091	0.616 ± 0.124	0.598 ± 0.113
Teacher	-1.683 ± 0.104	0.602 ± 0.155	0.146 ± 0.135

†Estimated coefficient ± empirical standard error.

standard errors are slightly smaller than with separate logistic regressions, but only very slightly so for the $\widehat{\beta}^P$.

If we use a "working independence" model for $\text{Cov}(Y_{i1}, Y_{i2})$, setting the log odds ratio for the association between parent and teacher response to zero, then the results (not shown) of a GEE analysis yield exactly the same result as separate regressions for $\widehat{\beta}$ (i.e., GEE estimates of β are identical to those reported in Table 16.4). The model-based standard errors are also identical to those reported in Table 16.4 and the 3×3 off-diagonal block of the covariance matrix for $\widehat{\beta}$ is zero. The empirical (or "sandwich") standard errors are very similar to the model-based standard errors. However, they differ slightly because the 3×3 off-diagonal block of the covariance matrix for $\widehat{\beta}$ is estimated as zero by the model-based variance estimator, whereas the empirical variance estimator correctly estimates the covariance between $\widehat{\beta}^P$ and $\widehat{\beta}^T$. That is, the empirical standard errors account for the correlations (ranging from approximately 0.15 to 0.25) between components of $\widehat{\beta}^P$ and $\widehat{\beta}^T$.

The model given by (16.3) is a very general model; its advantages over the separate regressions are that:

(i) We can test whether $\beta_k^P = \beta_k^T$ for the k^{th} covariate (or for the whole vector, test $\beta^P = \beta^T$, using $\text{Cov}(\widehat{\beta})$ provided by the GEE analysis);

(ii) We can use all available data; and

(iii) It provides a measure of association between the two informants.

With a large number of covariates, we will usually want to fit simpler models. The way we have defined β and Z_i in model (16.3) means that the first p components of β correspond to β^P and the second p correspond to β^T. To formulate simpler models, we need to create a dichotomous indicator variable of informant.

To illustrate, we introduce a dichotomous variable (X_2) which is 1 if the informant is the parent, and 0 if the informant is the teacher. Denoting single-parent status and child health problems by X_3 and X_4, consider a model of the form

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij}, \tag{16.4}$$

Table 16.6 Results of fitting regression models, with common or shared parameters, simultaneously to data on externalizing behavior using GEE method.

Variable	Estimate	SE	Z
Intercept	−1.685	0.100	−16.85
Informant	−0.467	0.118	−3.96
Single Parent	0.611	0.108	5.68
Child Health	0.146	0.135	1.08
Informant × Child Health	0.452	0.157	2.87

†Estimated coefficient ± empirical standard error.

where $X_{1ij} = 1$ for all i and j. This model specifies that the effects of single-parent status (measured by β_3) and child's physical health problems (measured by β_4) are the same regardless of the source of the information, but the mean level may be higher or lower (measured by β_2) depending on informant. A positive β_2 suggests that a positive rating ($Y_{ij} = 1$), here denoting externalizing behaviors in the borderline/clinical range, is more likely from a parent's report than from a teacher's, holding single parent status and physical health problems constant. Notice that only informant is a within-subject variable, that is, $X_{2i1} \neq X_{2i2}$ while X_3 and X_4 are both between-subject variables. Forcing the coefficients of single parent status to be equal seems reasonable in view of the results presented in Tables 16.4 and 16.5, but not for child health problems.

We construct a model which allows the effect of physical health problems to depend on informant by simply adding the interaction:

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij}, \qquad (16.5)$$

where $X_{5ij} = X_{2ij} X_{4ij}$. Fitting a model which includes the interaction allows the probability of a positive rating to depend on informant, single-parent status, and child health problems, and allows the effect of child health problems to differ by informant. The results of this model are displayed in Table 16.6. The estimated odds ratios change very little from the model given by (16.3). The interpretation of the results in Table 16.6 becomes more transparent if the model given by (16.5) is written separately for the parent and teacher informant:

$$\begin{aligned} \text{logit}(\mu_{i1}) &= \text{logit}(\mu_i^P) = (\beta_1 + \beta_2) + \beta_3 X_{3ij} + (\beta_4 + \beta_5) X_{4ij}; \\ \text{logit}(\mu_{i2}) &= \text{logit}(\mu_i^T) = \beta_1 + \beta_3 X_{3ij} + \beta_4 X_{4ij}. \end{aligned}$$

Because we code informant as 1 for parent response, $\widehat{\beta}_1$ and $\widehat{\beta}_4$ can be regarded as estimated teacher parameters for the intercept and child physical health problems.

For single-parent status, $\widehat{\beta}_3$ is the common coefficient for both informants; notice that its standard error is considerably smaller than the two corresponding standard errors for model (16.3) reported in Table 16.5. Finally, $\widehat{\beta}_5$ estimates the difference in effects of child health problems as estimated by parent and teacher evaluations. Comparison of $\widehat{\beta}_5$ to its standard error provides a formal test of the null hypothesis that informant does not matter in evaluating the effect of child health problems. The rejection of this hypothesis may reflect the fact that the rating of child health problems was given by the parent, and the teacher may lack information about child's physical health.

As a general rule, if informant interactions are included for all the covariates, then the model is basically equivalent to fitting separate regressions. However, the estimated coefficients obtained from using GEE with a non-zero correlation will differ from those obtained by fitting separate regressions because of the non-linearity of the logistic regression model, even with complete data on each informant. If there are missing data, the coefficients may differ substantially. In our example, the estimated effects for the parent respondents are very similar to those obtained from a separate regression on Y_{i1}, with all 2501 observations. The differences for the teacher respondents were more pronounced because of the substantial missing data, and because the two informant responses are highly associated. When we use GEE with a non-zero correlation, the coefficients for the teacher responses use the information in the parent response to provide some information for the missing teacher respondents.

When simpler models are fit (i.e., not all interactions with informant are present) we can expect to gain efficiency in the analysis for the common coefficients. This point is illustrated by comparing the standard errors of the coefficient for single-parent status in Tables 16.4 or 16.5 with Table 16.6.

16.6 SUMMARY

This chapter has illustrated how two other types of studies, repeated measures and multiple sources, can be analyzed using methods for correlated data that are comparable to those used for longitudinal data analysis. This is certainly not an exhaustive list, as many study designs produce repeated measures and multiple source data and their proper analysis requites linear or generalized linear models for correlated data.

Our repeated measures example is also an example of a crossover design, but there are many examples of simpler repeated measures designs with no period or carryover effects; in these cases there are often between-subject variables and the interaction of those with the within-subject variables will be of interest. For our example, the response variable is continuous and a linear mixed model is appropriate, but in other settings one may wish to use a more general covariance structure, or, when the response variable is discrete, a generalized linear model for correlated data.

Our multiple informant example used dichotomous reports and logistic regression models, but often multiple source data will be continuous responses. The general approach to constructing multivariate correlated regression models remains the same,

but now maximum likelihood and GEE approaches are viable alternatives for the analysis. Likewise, there may be more than two sources ($n > 2$). The general approach to analyzing multiple source data can handle any number of informants or sources using $n - 1$ (one less than the number of informants) dichotomous indicators of informant and their interactions with the covariates of interest.

16.7 FURTHER READING

Repeated measures designs are frequently encountered in applications and there is a very large literature on their design and analysis. Accessible descriptions of methods for analyzing repeated measures data can be found in the review articles by Keselman and Keselman (1984), Everitt (1995), and Omar *et al.* (1999).

Methods for analyzing crossover trials are discussed in the review articles by Matthews (1994) and Jones and Donev (1996). Finally, Fitzmaurice *et al.* (1995) and Goldwasser and Fitzmaurice (2001) discuss the use of regression models for analyzing multiple source data and present a substantive analysis of the multiple informant data from the Connecticut Child Surveys.

Bibliographic Notes

A comprehensive description of the multitude of techniques available for analyzing repeated measures data can be found in the books by Crowder and Hand (1990), Lindsey (1999), and Davis (2002), and the references therein.

A discussion of split-plot designs can be found in Chapter 7 of Cochran and Cox (1957). A comprehensive description of the design and analysis of crossover trials can be found in the books by Jones and Kenward (1989) and Senn (2002), and the references therein.

17

Multilevel Models

17.1 INTRODUCTION

In Parts I–III of the book, the major focus has been on the analysis of longitudinal data. A distinctive feature of longitudinal data is that they are *clustered*. In longitudinal studies the cluster is composed of the repeated measurements obtained from a single individual at different occasions. There are, however, many studies in the health sciences that are not longitudinal but which give rise to data that are clustered or cluster-correlated. For example, clustered data commonly arise when intact groups are randomized to health interventions in so-called cluster-randomized trials or when naturally occurring groups in the population are randomly sampled. In addition, there can be more than a single level of clustering in the data. The term *multilevel data* (or *hierarchical data*) encompasses all of these cases. The distinctive feature of multilevel data is that measurements on units within a cluster are more similar than measurements on units in different clusters. The clustering can be expressed in terms of correlation among the measurements on units within the same cluster and this correlation must be appropriately accounted for in the analysis.

Because longitudinal data are a special case of multilevel data, with only a single level of clustering and a natural ordering of the measurements within a cluster, this chapter provides a description of regression models for multilevel data, more broadly defined. One of our goals is to demonstrate that many of the methods for the analysis of longitudinal data considered in earlier chapters are, more or less, special cases of more general regression methods for multilevel data. The overview of multilevel models presented here provides a basic introduction to a general methodology for

analyzing the wide range of clustered data that commonly arise in studies in the biomedical and health sciences.

17.2 MULTILEVEL DATA

Multilevel data arise when there is a hierarchical or clustered structure to the data. Data of this kind frequently arise in the health sciences since individuals can be grouped in so many different ways. For example, in studies of health services and outcomes, assessments of quality of care are often obtained from patients who are nested within different clinics. Such data can be regarded as multilevel, with patients referred to as the level 1 units and clinics the level 2 units. In this example there are two levels in the data hierarchy and, by convention, the lowest level of the hierarchy is referred to as level 1. The term "level", as used in this context, signifies the position of a unit of observation within a hierarchy.

Broadly speaking, the clustering in multilevel data can be a consequence of the study design or due to a naturally occurring hierarchy in the target population, or sometimes due to both. An example of a naturally occurring two-level data hierarchy arises in developmental toxicity studies. In a typical developmental toxicity experiment, pregnant mice or rats (dams) are assigned to increasing doses of a chemical or a test substance over the period of major organogenesis (when organ systems are developing in a growing fetus). Following sacrifice, each fetus in the litter is weighed (a continuous response) and examined for evidence of malformations (a binary response, present or absent). Data collected in developmental toxicity experiments are clustered (i.e., the litter is the cluster), with observations on the fetuses (level 1 units) nested within dams/litters (level 2 units). Two-level data also arise in family studies designed to assess the association or "aggregation" of disease (or markers of disease development) among relatives. In family studies, the goal is to determine whether the presence of disease in a family member is associated with increased risk of disease for relatives. The associations among siblings and between parent-child pairs are of main interest because their relative magnitudes can be used to provide indirect evidence of genetic risk due to the sharing of the same genes. Data from studies of nuclear families are clustered, with observations on the mother, father, and children (level 1 units) nested within families (level 2 units).

Other common examples of naturally occurring clusters in the population are households, hospital wards, clinics, medical practices, neighborhoods, and schools. Furthermore, naturally occurring hierarchical data structures can have more than two levels. For example, observations may be obtained on patients nested within clinics, that, in turn, are nested within different geographical regions of the country. Another example of a naturally occurring data hierarchy is when observations are obtained on children nested within classrooms, nested within schools. In both of these examples there are three levels in the data hierarchy. In principle, there can be many levels in the data hierarchy.

Alternatively, the hierarchical data structure can be a consequence of the study design. For example, the United States National Health and Nutrition Examination Survey (NHANES) uses a multi-stage sampling design to produce information on nutrition and health status. The target population is the total U.S. civilian non-institutionalized population, 2 months of age or over. Because it is not practical to obtain a simple random sample of the U.S. population, complex sampling methods are commonly used. For example, NHANES III, conducted in 1988–1994, used the following multi-stage sampling design (National Center for Health Statistics, 1992, 1994). In the first stage, so-called "primary sampling units" (PSUs) were defined based on counties or combined counties in the United States. A first-stage random sample of PSUs were selected from these geographical regions. In the second-stage sampling, within each of the selected PSUs, a random sample of area segments consisting of census blocks were selected. In the third stage, within each of the selected area segments, a random sample of households were selected. Finally, in the fourth stage, eligible persons were randomly selected within households. The resulting data can be regarded as hierarchical, with individuals being the level 1 units, households the level 2 units, area segments the level 3 units, and counties the level 4 units.

Additional examples of study designs that produce multilevel data structures include cluster-randomized clinical trials, repeated measures experiments, and longitudinal studies. In a cluster-randomized trial, groups of individuals, rather than the individuals themselves, are randomized to different treatments or health interventions. For example, the Promotion of Breastfeeding Intervention Trial (PROBIT) was designed to determine whether efforts to promote breastfeeding have any impact on the duration and exclusivity of breastfeeding (Kramer *et al.*, 2001). In this trial, maternity clinics, rather than the mothers themselves, were randomized to either the intervention or control (standard care). The mothers were followed-up for one year after the birth of their infants and the effectiveness of the health intervention was assessed by the responses of mothers in each treatment group. When regarded as multilevel data, the level 1 units are the mothers and the level 2 units are the maternity clinics. Of note, the main covariate of interest, denoting the assignment to intervention or control, is defined at level 2. Longitudinal studies are another common example where the study design produces data with a two-level structure. In a longitudinal study the clusters are composed of the repeated measurements obtained from a single individual at different occasions. When longitudinal data are regarded as multilevel data, the level 1 units are the repeated occasions of measurement and the level 2 units are the subjects.

Finally, the clustering in multilevel data can be due to both the design of the study and naturally occurring hierarchies in the target population. For example, clinical trials are often conducted in many centers to ensure sufficient numbers of patients and/or to assess the effectiveness of the treatment in different settings. These studies are referred to as multi-center trials. Observations from a multi-center longitudinal clinical trial can be regarded as multilevel data having 3 levels, with repeated measurement occasions (level 1 units) nested within subjects (level 2 units) nested within clinics (level 3 units).

Although we have distinguished between clustering that occurs naturally and clustering due to study design, the consequence of clustering at different levels is the same: units that are grouped at any level are likely to respond more similarly. For example,

two patients selected at random from the same clinic are expected to respond more similarly than two patients randomly selected from different clinics. In general, the degree of clustering can be expressed in terms of correlation among the observations on units within the same level. Statistical models for multilevel data must account for the intra-cluster correlation at each level; failure to do so can result in misleading inferences.

17.3 MULTILEVEL LINEAR MODELS

In this section we discuss linear models for multilevel data. The dominant approaches to multilevel modelling have the same basis: clustering in the data is accounted for via the introduction of random effects at different levels in the hierarchy. Multilevel linear models can be regarded as extensions of the linear mixed effects models described in Chapter 8, which allow random effects to be incorporated at more than one level. In addition to accounting for clustering in the data, multilevel models permit estimation of the effects of covariates, measured at any of the levels of the hierarchy, on the outcome.

In a multilevel model, the response is obtained on the lowest level (or level 1) units, but covariate information can by measured at any level. Combining covariates measured at different levels of the hierarchy within a single regression model is central to multilevel modelling. For example, multilevel models can determine and disentangle the relative importance of patient-level, clinic-level and regional-level factors on quality of care. In general, multilevel models can be used to make inferences about the population of units at any level of the hierarchy and to discern how variation in the outcome at different levels depends on covariates. In this section we present an overview of multilevel linear models for a continuous outcome. We begin with a discussion of models for two-level data. The models generalize in a natural way when there is additional clustering in the data (e.g., three- and higher-level data). The major focus of this section is on the specification of multilevel models; estimation is mentioned but not emphasized.

17.3.1 Two-Level Linear Models

Before describing models for two-level data we need to introduce some notation. For a two-level data structure, let i index level 1 units and j index level 2 units. We assume that there are n_2 units at level 2 in the sample. Each of these clusters (for $j = 1, ..., n_2$) is composed of n_{1j} level 1 units. For example, consider a multi-center clinical trial comparing two treatments (active drug versus placebo) conducted in 20 medical clinics. Patients are enrolled from each clinic and randomly assigned to one of the two treatment conditions. In this example, clinics are the level 2 units ($j = 1, ..., 20$) and patients are the level 1 units ($i = 1, ..., n_{1j}$), where n_{1j} is the number of patients enrolled in the study from the j^{th} clinic (and $n_2 = 20$ is the number of clinics). Alternatively, consider two-level data arising from a longitudinal

study where 150 subjects are measured at four occasions. In this example, subjects are the level 2 units ($j = 1, ..., 150$) and measurement occasions are the level 1 units ($i = 1, ..., 4$), with $n_2 = 150$ level 2 units, and $n_{1j} = 4$ level 1 units (within each level 2 unit). For the latter example, the alert reader will have noticed that the indices i and j have now been reversed from their use in earlier chapters; here, we have adopted the usual convention in much of the multilevel modelling literature of letting i denote level 1 units, j denote level 2 units, and so on. We must caution the reader that some of the literature on multilevel modelling reverses this notation (and/or occasionally reverses the ordering of the levels).

Let Y_{ij} denote the response on the i^{th} level 1 unit within the j^{th} level 2 cluster. For example, Y_{ij} might denote the primary outcome for the i^{th} patient in the j^{th} clinic. Associated with each Y_{ij} is a $1 \times p$ (row) vector of covariates, X_{ij}. These can include covariates defined at each of the two levels and can also include so-called "compositional" covariates, formed by aggregating values over lower-level units. For example, severity of disease defines a patient-level (or level 1) covariate. However, a "compositional" covariate at the clinic level can be formed by taking the average disease severity for all patients within each clinic.

Consider the following linear model relating the mean response to the covariates:

$$E(Y_{ij}) = X_{ij}\beta. \tag{17.1}$$

For example, in a multi-center clinical trial, a simple model for the mean response is given by

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \texttt{Group}_{ij},$$

where $\texttt{Group}_{ij}$ denotes the treatment assignment for the i^{th} patient in the j^{th} clinic, with $\texttt{Group}_{ij} = 1$ for active drug and $\texttt{Group}_{ij} = 0$ for placebo. The model given by (17.1) specifies how the mean response depends on covariates, where the covariates can be defined at level 2 and/or level 1. A multilevel model accounts for the variability in Y_{ij}, around its mean, by allowing for random variation across both level 1 and level 2 units. In particular, a multilevel model for Y_{ij} assumes there is random variation across level 1 units and random variation in a subset of the regression parameters across level 2 units. The two-level linear model for Y_{ij} is given by

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + e_{ij}, \tag{17.2}$$

where Z_{ij} is a design matrix for the random effects at level 2, formed from a subset of the appropriate columns of X_{ij}. The random effects, b_j, vary across level 2 units but, for a given level 2 unit, are constant for all level 1 units. These random effects are assumed to be independent across level 2 units, with mean zero and covariance, $\text{Cov}(b_j) = G$. The level 1 random components, e_{ij}, are also assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(e_{ij}) = \sigma^2$. In addition, the e_{ij}'s are assumed to be independent of the b_j's, with $\text{Cov}(e_{ij}, b_j) = 0$. That is, the level 1 units are assumed to be conditionally independent given the level 2 random effects (and the covariates).

The regression parameters, β, are the fixed effects and describe the effects of covariates on the mean response

$$E(Y_{ij}) = X_{ij}\beta,$$

where tl.e mean response is averaged over both level 1 and level 2 units. The two-level model given by (17.2) also describes the effects of covariates on the conditional mean response

$$E(Y_{ij}|b_j) = X_{ij}\beta + Z_{ij}b_j,$$

where the response is averaged over level 1 units only.

Let us return to the multi-center clinical trial example introduced earlier. A simple two-level model for the data is given by

$$Y_{ij} = \beta_1 + \beta_2\,\mathtt{Group_{ij}} + b_{1j} + e_{ij},$$

where b_{1j} is a random clinic effect. The random effect b_{1j} varies across clinics but, for a given clinic, is constant and shared by all patients belonging to that clinic. The inclusion of b_{1j} accounts for the clustering of patients within clinics, due perhaps to similarities in severity of illness and/or quality of care. The model explicitly accounts for the fact that some clinics have patients that respond higher (or lower) than patients in other clinics. However, the model assumes that the effect of treatment is the same across all clinics. This assumption can be relaxed by allowing the effect of treatment to vary among clinics

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2\,\mathtt{Group_{ij}} + b_{1j} + b_{2j}\,\mathtt{Group_{ij}} + e_{ij} \\ &= (\beta_1 + b_{1j}) + (\beta_2 + b_{2j})\,\mathtt{Group_{ij}} + e_{ij}. \end{aligned}$$

In this model, the magnitude of the effect of treatment varies randomly across the different clinics. The average effect of treatment, when averaged over the population of clinics (and not simply those included in the trial), is β_2.

The example just presented involves randomizing patients (level 1) within clinics (level 2). In the language of experimental design, patients (level 1) are *nested* within clinics (level 2), but treatment is *crossed* with clinics because patients within each clinic are randomized to each treatment. Another very different type of design is one where patients (level 1) are nested within a clinic (level 2), but clinics are randomized to treatments, so that all patients from any given clinic receive the same treatment. In this case, clinics are nested within treatment, and not crossed. Formally, the same model just presented can be used for the analyses of these data, except that the effect of treatment can no longer vary randomly across the different clinics, because each clinic is assigned to only one treatment group. Note also that the treatment group variable, $\mathtt{Group_{ij}}$, does not vary over i for fixed j, hence can be replaced by $\mathtt{Group_j}$. However, it is important to note that the nesting of clinics within treatment has a negative impact on efficiency of the treatment effect estimate, relative to a design with no nesting. This general principle will be illustrated later with analyses of the *Television, School and Family Smoking Prevention and Cessation Project.* In this study, schools were randomized to treatments, and in the analysis both classroom and

student variability were accounted for. The first design, where clinics are crossed with treatment, is generally more efficient than a design which does not stratify on clinic. The principle behind this is the same as that of a longitudinal study, where we can generally measure change more efficiently by using repeated measures on the same subject than using a cross-sectional design.

So far, our discussion of two-level models has very closely paralleled the description of the linear mixed effects model given in Chapter 8. In Chapter 8 we focused on models for two level data where measurement occasions are the level 1 units and subjects are the level 2 units. However, it should be recognized that longitudinal data are simply a special case of two-level data and the linear mixed effects model given by (17.2) can be applied more broadly.

Finally, the two-level model given by (17.2) can also be written in terms of two models, one for each level of the hierarchy, using the two-stage formulation described in Section 8.4. That is. the two-level model can be expressed in terms of a level 1 model,

$$Y_{ij} = Z_{ij}\beta_j + e_{ij},$$

where e_{ij} are assumed to be independent across level 1 units, with mean zero and variance, $\mathrm{Var}(e_{ij}) = \sigma^2$, and a level 2 model,

$$\beta_j = A_j\beta + b_j,$$

where b_j are assumed to vary independently across level 2 units, with mean zero and covariance, $\mathrm{Cov}(b_j) = G$. Substituting the second model equation into the first yields (17.2)

$$\begin{aligned} Y_{ij} &= Z_{ij}(A_j\beta + b_j) + e_{ij} \\ &= (Z_{ij}A_j)\beta + Z_{ij}b_j + e_{ij} \\ &= X_{ij}\beta + Z_{ij}b_j + e_{ij}, \end{aligned}$$

where $X_{ij} = Z_{ij}A_j$. An advantage of specifying a multilevel model in terms of a series of models for each level of the hierarchy, rather than as a combined model, is that it becomes more transparent which covariates are operating at which level of the model. However, this does introduce some unnecessary restrictions on the model of the kind discussed in Section 8.4.

In summary, the two-level linear model given by (17.2) accounts for the clustering of the level 1 units by incorporating random effects at level 2. The model explicitly distinguishes two main sources of variation in the response: variation across level 2 units and variation across level 1 units (within level 2 units). The relative magnitude of these two sources of variability determines the degree of clustering in the data. The larger the variance of the level 2 random effects, relative to the level 1 (within level 2) variability, the greater the degree of clustering. Next, we describe how the linear mixed effects model can be generalized to three-level data structures; the extensions to four or more levels follows directly.

17.3.2 Three-Level Linear Models

As mentioned earlier, there can be many levels in the data hierarchy. The extension of the two-level model given by (17.2) to three or more levels is very natural. With three-level data, there is clustering in the data that is assumed to be due to variation in the response across level 1, level 2, and level 3 units. Although the extension from 2 to 3 levels is conceptually straightforward, the description of the three-level model does require the introduction of additional notation that often obfuscates the salient features of the model. The basis of a three-level model is that variability in the response is accounted for by the introduction of random effects at all higher levels in the hierarchy (e.g., by allowing random variation in a subset of the regression parameters at both levels 2 and 3). The model explicitly distinguishes three sources of variation in the response: (1) variation across level 3 units, (2) variation across level 2 units (within level 3 units), and (3) variation across level 1 units (within level 2 units nested within level 3 units).

For a three-level data structure, let i index level 1 units, j index level 2 units, and k index level 3 units. We assume that there are n_3 units at level 3 in the sample. Each of these clusters (for $k = 1, ..., n_3$) is composed of n_{2k} level 2 clusters, and each of these, in turn, is composed of n_{1jk} level 1 units. For example, consider a multi-center *longitudinal* clinical trial comparing two treatments (active drug versus placebo) conducted in 20 different centers or clinics. Patients in each clinic are measured at baseline and at three post-treatment occasions. In this example, clinics are the level 3 units ($k = 1, ..., 20$), patients are the level 2 units ($j = 1, ..., n_{2k}$), and measurement occasions are the level 1 units ($i = 1, ..., 4$), where n_{2k} is the number of patients in the k^{th} clinic ($n_3 = 20$ and $n_{1jk} = 4$, for all j and k).

Let Y_{ijk} denote the response of the i^{th} level 1 unit within the j^{th} level 2 cluster within the k^{th} level 3 cluster. For example, in a multi-center longitudinal clinical trial, Y_{ijk} denotes the outcome at the i^{th} occasion for the j^{th} patient in the k^{th} clinic. Associated with each Y_{ijk} is a $1 \times p$ (row) vector of covariates, X_{ijk}, with the covariates defined at different levels. A three-level model for Y_{ijk} is given by

$$Y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk}\, b^{(3)}_k + Z^{(2)}_{ijk}\, b^{(2)}_{jk} + e_{ijk}, \qquad (17.3)$$

where $Z^{(3)}_{ijk}$ is a design matrix for the random effects at level 3, formed from a subset of the appropriate columns of X_{ijk}, and $Z^{(2)}_{ijk}$ is a design matrix for the random effects at level 2 (also formed from a subset of the appropriate columns of X_{ijk}). In this notation the superscripts attached to $b^{(3)}_k$ and $b^{(2)}_{jk}$ denote the levels at which the random effects vary. In general, the design matrices for the random effects contain covariates that vary at lower levels than that of the corresponding random effects. That is, $Z^{(3)}_{ijk}$, the design matrix for $b^{(3)}_k$, will, in general, contain covariates that vary across level 2 and level 1 units.

To fix ideas, consider the example of a multi-center longitudinal clinical trial introduced earlier. A three-level model for the outcome is given by

$$Y_{ijk} = \beta_1 + \beta_2\, t_{ijk} + \beta_3\, (\texttt{Group}_{ij} \times t_{ijk}) + b^{(3)}_k + b^{(2)}_{jk} + e_{ijk},$$

where t_{ijk} denotes the time since baseline for the i^{th} observation on the j^{th} patient in the k^{th} clinic. In this model, $b^{(3)}_k$ is a random clinic effect and $b^{(2)}_{jk}$ is a random patient effect. The inclusion of the former accounts for the clustering of patients within clinics, while the inclusion of the latter accounts for the positive correlation among the repeated measures on the same patient. Additional random effects, at both levels 2 and 3, can easily be incorporated in the model (e.g., random slopes for time, and possibly random effects for the $\texttt{Group}_{ij} \times t_{ijk}$ interaction).

The three-level model makes the following two assumptions about the different sources of variability:

(i) The random effects $b^{(3)}_k$ are assumed to be independent across level 3 units, with mean zero and covariance, $\text{Cov}(b^{(3)}_k) = G^{(3)}$; similarly, the random effects $b^{(2)}_{jk}$ are assumed to be independent across level 2 units, with mean zero and covariance, $\text{Cov}(b^{(2)}_{jk}) = G^{(2)}$. Random effects may be correlated within a given level, but not between levels.

(ii) The level 1 random components, e_{ijk}, are assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(e_{ijk}) = \sigma^2$. In addition, the e_{ijk}'s are assumed to be independent of the random effects, $b^{(3)}_k$ and $b^{(2)}_{jk}$.

That is, the random effects at the same level are, in general, correlated within units at that level but not between units; random effects at different levels are assumed to be independent of each other and of the level 1 random components, e_{ijk}. In principle, we can replace e_{ijk} in (17.3) with $Z^{(1)}_{ijk}\, b^{(1)}_{ijk}$, where $b^{(1)}_{ijk}$ has mean zero and $\text{Cov}(b^{(1)}_{ijk}) = G^{(1)}$. This would allow for heterogeneity in the level 1 variability, with possible dependence of the level 1 variance on certain covariates. However, for the remainder of this discussion, we assume the simpler variance structure for the e_{ijk}, with $\text{Var}(e_{ijk}) = \sigma^2$ (i.e., we assume $Z^{(1)}_{ijk} = 1$ for all i, j, and k).

In model (17.3) the regression parameters, β, are the fixed effects and describe the effects of covariates on the mean response (averaged over level 1, level 2, and level 3 units),

$$E\,(Y_{ijk}) = X_{ijk}\beta.$$

The three-level model given by (17.3) also describes the effects of covariates on the conditional mean response

$$E\left(Y_{ijk} \mid b^{(3)}_k\right) = X_{ijk}\beta + Z^{(3)}_{ijk}\, b^{(3)}_k,$$

where the response is averaged over level 1 and level 2 units only, and the effects of covariates on the conditional mean response

$$E\left(Y_{ijk} \mid b^{(3)}_k, b^{(2)}_{jk}\right) = X_{ijk}\beta + Z^{(3)}_{ijk}\, b^{(3)}_k + Z^{(2)}_{ijk}\, b^{(2)}_{jk},$$

where the response is averaged over level 1 units only.

17.3.3 Estimation

So far, we have described a general specification of two- and three-level models that can be readily extended to more levels. The parameters of multilevel models are the fixed effects regression parameters, β, and the covariance (or variance) of the random effects at each level. Given estimates of the latter, predictions (empirical BLUPs) of the random effects at any level can also be obtained. For multilevel linear models, it is common to assume that the random components have multivariate normal distributions. For example, in the three-level model, it is usually assumed that $b_k^{(3)} \sim N(0, G^{(3)})$, $b_{jk}^{(2)} \sim N(0, G^{(2)})$, and $e_{ijk} \sim N(0, \sigma^2)$. Given these distributional assumptions, maximum likelihood (ML) estimation of the multilevel model parameters is relatively straightforward.

The ML estimate of β is obtained from the generalized least squares (GLS) estimator. For the two-level model, the GLS estimator has the same form as that given in Chapter 4 (albeit, with the indices i and j reversed). For the three-level model, the GLS estimator of β also has a closed-form expression and is given by

$$\widehat{\beta} = \left\{ \sum_{k=1}^{n_3} \left(X_k' V_k^{-1} X_k \right) \right\}^{-1} \sum_{k=1}^{n_3} \left(X_k' V_k^{-1} Y_k \right),$$

where Y_k is a column vector, of length $\sum_{j=1}^{n_{2k}} n_{1jk}$, formed by stacking the responses for all second- and first-level units within the k^{th} cluster. Similarly, X_k is an $(\sum_{j=1}^{n_{2k}} n_{1jk}) \times p$ matrix formed by stacking the covariates for all second- and first-level units within the k^{th} cluster. Finally, V_k is the covariance among observations on first- and second-level units within the k^{th} cluster and has a random effects covariance structure, expressed as a function of $G^{(3)}, G^{(2)}$, and σ^2 (and the corresponding design matrices for the random effects).

Restricted maximum likelihood (REML) estimation of $V_k(G^{(3)}, G^{(2)}, \sigma^2)$ proceeds in the same way as with estimation of β. That is, the REML (or ML) estimates of $G^{(3)}, G^{(2)}$, and σ^2 are obtained by maximizing the restricted log-likelihood with respect to $G^{(3)}, G^{(2)}$, and σ^2. In general, it is not possible to write down simple, closed-form expressions for the REML (or ML) estimators of $G^{(3)}, G^{(2)}$, and σ^2; instead, estimates must be obtained using iterative techniques. Once the REML (or ML) estimates of $G^{(3)}, G^{(2)}$, and σ^2 have been obtained, the estimate of $V_k(G^{(3)}, G^{(2)}, \sigma^2)$, say $V_k(\widehat{G}^{(3)}, \widehat{G}^{(2)}, \widehat{\sigma}^2)$, is substituted into the generalized least squares estimator of β to obtain the REML (or ML) estimate of β. REML estimation for multilevel linear models has been implemented in many major statistical software packages (e.g., PROC MIXED in SAS and the *lme* function in S-PLUS) and in stand-alone programs that have been specifically tailored for multilevel modelling (e.g., MLwiN and HLM).

17.3.4 Case Studies

Next we illustrate the main ideas by conducting analyses of two- and three-level data. The first example analyzes two-level data on fetal weight from a developmental toxicity study of laboratory mice exposed to ethylene glycol (EG). The data on

Table 17.1 Descriptive statistics on fetal weight from the ethylene glycol (EG) experiment.

Dose (mg/kg)	$\sqrt{\text{Dose}/750}$	Dams	Fetuses	Weight (gm) Mean	Weight (gm) St. Deviation†
0	0	25	297	0.972	0.098
750	1	24	276	0.877	0.104
1500	1.4	22	229	0.764	0.107
3000	2	23	226	0.704	0.124

†Calculated ignoring clustering.

the weights of live fetuses, nested within litters, are from an experiment conducted through the National Toxicology Program (NTP) (Price *et al.*, 1985). The second example analyzes three-level data from a cluster-randomized trial to determine the efficacy of school-based interventions to prevent tobacco use. The data on 7th grade children, nested within classrooms, nested within schools are from the *Television, School, and Family Smoking Prevention and Cessation Project* (TVSFP) (Flay *et al.*, 1995, Hedeker *et al.*, 1994).

Developmental Toxicity Study of Ethylene Glycol

Developmental toxicity studies of laboratory animals play a crucial role in the testing and regulation of chemicals and pharmaceutical compounds. Exposure to developmental toxicants typically causes a variety of adverse effects, such as fetal malformations and reduced fetal weight at term. In a typical developmental toxicity experiment, laboratory animals are assigned to increasing doses of a chemical or test substance. In this section we describe an analysis of data from a development toxicity study of ethylene glycol (EG). Ethylene glycol is a high-volume industrial chemical used in many applications. It is used as an antifreeze, as a solvent in the paint and plastics industries, and in the formulation of various types of inks. In a study of laboratory mice conducted through the National Toxicology Program (NTP), EG was administered at doses of 0, 750, 1500, or 3000 mg/kg/day to 94 pregnant mice (dams) over the period of major organogenesis, beginning just after implantation. (See Price *et al.*, 1985, for additional details concerning the study design.) Following sacrifice, fetal weight and evidence of malformations were recorded for each live fetus. In our analysis of the data, we focus on the effects of dose on fetal weight; in Section 17.4 we present a complementary analysis that examines the effect of dose on fetal malformations.

Summary statistics (ignoring clustering in the data) for fetal weight for the 94 litters (composed of a total of 1028 live fetuses) are presented in Table 17.1. Fetal weight decreases monotonically with increasing dose, with the average weight ranging from

Table 17.2 Fixed and random effects estimates for the fetal weight data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	0.984	0.016	61.32
$\sqrt{\text{Dose}/750}$	−0.134	0.012	−10.85
Level 2 Variance:			
$\sigma_2^2(\times 100)$	0.726	0.119	6.11
Level 1 Variance:			
$\sigma_1^2(\times 100)$	0.556	0.026	21.55

0.972 (gm) in the control group to 0.704 (gm) in the group administered the highest dose. The decrease in fetal weight is not linear in increasing dose, but is approximately linear in increasing $\sqrt{\text{dose}}$.

The data on fetal weight from this experiment are clustered, with observations on the fetuses (level 1 units) nested within dams/litters (level 2 units). The litter sizes range from 1 to 16. Letting Y_{ij} denote the fetal weight of the i^{th} live fetus from the j^{th} litter, we considered the following model relating the fetal weight outcome to dose:

$$Y_{ij} = \beta_1 + \beta_2 \, \mathrm{d_j} + b_j + e_{ij},$$

where $\mathrm{d_j} = \sqrt{\text{Dose}_j/750}$ is the square-root transformed dose administered to the j^{th} dam. The random effect b_j is assumed to vary independently across litters, with $b_j \sim N(0, \sigma_2^2)$. The errors, e_{ij}, are assumed to vary independently across fetuses (within a litter), with $e_{ij} \sim N(0, \sigma_1^2)$. Note that, in a slight departure from the notation introduced previously, the first- and second-level variances are denoted by σ_1^2 and σ_2^2, respectively. This model assumes that fetuses within a cluster are exchangeable and the positive correlation among the fetal weights is accounted for by their sharing a common random effect, b_j. The degree of clustering in the data can be expressed in terms of the intra-cluster (or intra-litter) correlation

$$\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

In Table 17.2 the results of fitting the model to the fetal weight data are presented. The REML estimate of the regression parameter for (transformed) dose indicates that the mean fetal weight decreases with increasing dose. The estimated decrease in weight, comparing the highest dose group to the control group, is 0.27 (or 2×-0.134, with 95% confidence interval: -0.316 to -0.220). Of note, we calculated both model-based and empirical (or "sandwich") standard errors and they were very similar,

suggesting that the simple random effect structure for the clustering of fetal weights is adequate. The estimate of the intra-cluster correlation, $\hat{\rho} = 0.57$, indicates that there are moderate litter effects.

Finally, to assess the adequacy of the linear dose–response trend, we considered a model that included a quadratic effect of (transformed) dose. Both Wald and likelihood ratio tests of the quadratic effect of dose indicated that the linear trend is adequate for these data (Wald $W^2 = 1.38$, with 1 df, $p > 0.20$; likelihood ratio $G^2 = 1.37$, with 1 df, $p > 0.20$).

Television School and Family Smoking Prevention and Cessation Project

Though smoking prevalence has declined among adults in recent decades, substantial numbers of young people begin to smoke and become addicted to tobacco. The *Television, School and Family Smoking Prevention and Cessation Project* (TVSFP) was a study designed to determine the efficacy of a school-based smoking prevention curriculum in conjunction with a television-based prevention program, in terms of preventing smoking onset and increasing smoking cessation (Flay *et al.*, 1995). The study used a 2×2 factorial design, with four intervention conditions determined by the cross-classification of a school-based social-resistance curriculum (CC: coded 1 = yes, 0 = no) with a television-based prevention program (TV: coded 1 = yes. 0 = no). Randomization to one of the four intervention conditions was at the school level, while much of the intervention was delivered at the classroom level. The TVSFP study is described in greater detail in Flay *et al.* (1995).

The original study involved 6695 students in 47 schools in Southern California. Our analysis focuses on a subset of 1600 seventh-grade students from 135 classes in 28 schools in Los Angeles. The response variable, a tobacco and health knowledge scale (THKS), was administered before and after randomization of schools to one of the four intervention conditions. The scale assessed a student's knowledge of tobacco and health.

We considered a linear model for the post-intervention THKS score, with the baseline or pre-intervention THKS score as a covariate. This model for the adjusted change in THKS scores included the main effects of CC and TV and the CC × TV interaction. School and classroom effects were modelled by incorporating random effects at levels 3 and 2, respectively. Letting Y_{ijk} denote the post-intervention THKS score of the i^{th} student within the j^{th} classroom within the k^{th} school, our model is given by

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + b_k^{(3)} + b_{jk}^{(2)} + e_{ijk},$$

where $e_{ijk} \sim N(0, \sigma_1^2)$, $b_{jk}^{(2)} \sim N(0, \sigma_2^2)$, and $b_k^{(3)} \sim N(0, \sigma_3^2)$. Once again, in a slight departure from the notation introduced previously, the first-, second-, and third-level variances are denoted by σ_1^2, σ_2^2, and σ_3^2, respectively.

The results of fitting this model to the data are presented in Table 17.3. The REML estimates of the three sources of variability indicate that there is variability at both classroom and school levels, with almost twice as much variability among

Table 17.3 Fixed and random effects estimates for the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.702	0.1254	13.57
Pre-Intervention THKS	0.305	0.0259	11.79
CC	0.641	0.1609	3.99
TV	0.182	0.1572	1.16
CC × TV	−0.331	0.2245	−1.47
Level 3 Variance:			
σ_3^2	0.039	0.0253	1.52
Level 2 Variance:			
σ_2^2	0.065	0.0286	2.26
Level 1 Variance:			
σ_1^2	1.602	0.0591	27.10

classrooms within a school as among schools themselves. The correlation among the THKS scores for classmates (or children within the same classroom within the same school) is approximately 0.061 (or $\frac{0.039+0.065}{0.039+0.06+1.602}$), while the correlation among the THKS scores for children from different classrooms within the same school is approximately 0.023 (or $\frac{0.039}{0.039+0.06+1.602}$). The estimates of the fixed effects for the intervention conditions, when compared to their standard errors, indicate that neither the mass-media intervention (TV) nor its interaction with the social-resistance classroom curriculum (CC) have an impact on adjusted changes in the THKS scores from baseline. There is a significant effect of the social-resistance classroom curriculum, with children assigned to the social-resistance curriculum showing increased knowledge about tobacco and health. The estimate of the main effect of CC, in the model that excludes the CC × TV interaction, is 0.47 (SE = 0.113, $p < 0.0001$).

The intra-cluster correlations at both the school and classroom levels are relatively small. The reader might be tempted to regard this as an indication that the clustering in these data is inconsequential. However, such a conclusion would be erroneous. Although the intra-cluster correlations are relatively small, they have a substantial impact on inferences concerning the effects of the intervention conditions. To illustrate this, consider the following model for the adjusted changes in THKS scores:

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + e_{ijk},$$

where $e_{ijk} \sim N(0, \sigma^2)$. This model ignores clustering in the data at the classroom and school levels; it is a standard linear regression model and assumes independent

Table 17.4 Fixed effects estimates from analysis that ignores clustering in the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.661	0.0844	19.69
Pre-Intervention THKS	0.325	0.0258	12.58
CC	0.641	0.0921	6.95
TV	0.199	0.0900	2.21
CC × TV	−0.322	0.1302	−2.47

observations and homogeneous variance. The results of fitting this model to the THKS scores are presented in Table 17.4. The estimates of the fixed effects are similar to those reported in Table 17.3. However, the model-based standard errors (assuming no clustering) are misleadingly small for the randomized intervention effects and lead to substantively different conclusions about the effects of the intervention conditions. This highlights an important lesson: the impact of clustering depends on both the magnitude of the intra-cluster correlation and the cluster size. For the data from the TVSFP, the cluster sizes vary from 1–13 classrooms within a school and from 2–28 students within a classroom. With relatively large cluster sizes, even very modest intra-cluster correlation can have a discernible impact on inferences.

17.4 MULTILEVEL GENERALIZED LINEAR MODELS

The discussion of multilevel models in the previous section focused on linear models for a continuous response, where clustering was accounted for through the introduction of random effects at different levels. In this section we briefly describe how multilevel modelling can be extended to discrete response data. These models can be thought of as multilevel generalized linear models and they extend in a natural way the conceptual approach described in Section 17.3. They differ from the models described in the previous sections primarily in terms of assumptions concerning the distribution of observations at level 1. The level 1 observations are no longer required to have a normal distribution; instead, they are assumed to have a distribution belonging to the exponential family (e.g., Bernoulli or Poisson). We focus on models for two- and three-level data; the generalizations to more levels follow directly.

17.4.1 Two-Level Generalized Linear Models

The basic premise of multilevel generalized linear models is that clustering among units can be thought of as arising from their sharing a set of random effects. For

example, with two-level binary data, it is assumed that the clustering of level 1 units (with n level 2 units) can be accounted for by heterogeneity across level 2 clusters in a subset of the regression coefficients from a generalized linear model (e.g., a logistic regression model with randomly varying intercepts). Conditional on the random effects, the level 1 observations are assumed to be independent and with a distribution belonging to the exponential family (e.g., Bernoulli).

In our description of two-level generalized linear models we adopt the notation used earlier. Let Y_{ij} denote the response on the i^{th} level 1 unit in the j^{th} level 2 cluster; the response can be continuous, binary, or a count. Associated with each Y_{ij} is a $1 \times p$ (row) vector of covariates, X_{ij}. We can formulate two-level models for discrete (and continuous) outcomes, Y_{ij}, using the familiar three-part specification of generalized linear mixed effects models outlined in Chapter 12:

1. We assume that the conditional distribution of each Y_{ij}, given a vector of random effects b_j (and the covariates), belongs to the exponential family of distributions and that $\text{Var}(Y_{ij}|b_j) = v\{E(Y_{ij}|b_j)\}\phi$, where $v(\cdot)$ is a known variance function, a function of the conditional mean, $E(Y_{ij}|b_j)$, and ϕ is a scale or dispersion parameter. In addition, given the random effects b_j, it is assumed that the Y_{ij} are independent of one another.

2. The conditional mean of Y_{ij} is assumed to depend upon fixed and random effects via the following linear predictor,

$$\eta_{ij} = X_{ij}\beta + Z_{ij}b_j,$$

with

$$g\{E(Y_{ij}|b_j)\} = \eta_{ij} = X_{ij}\beta + Z_{ij}b_j$$

for some known link function, $g(\cdot)$.

3. Finally, the random effects are assumed to have some probability distribution. In principle, any multivariate distribution can be assumed for the b_j; in practice, for computational convenience, the random effects are usually assumed to have a multivariate normal distribution, with zero mean and covariance matrix, G.

These three components completely specify a broad class of two-level generalized linear models for different types of responses. Next, to clarify the main ideas, we consider two examples of multilevel generalized linear models in greater detail.

Example 1: Two-Level Generalized Linear Model for Counts

Consider a study comparing cross-national rates of skin cancer and the factors (e.g., climate, economic and social factors, regional differences in diagnostic procedures) that influence variability in the rates of disease. Suppose we have counts of the number of cases of skin cancer in a set of well-defined regions, indexed by i, within countries, indexed by j. Let Y_{ij} be a count of the number of individuals who develop skin cancer within the i^{th} region of the j^{th} country during a given period of time (say, 5 years). The resulting counts have a two-level structure with regional units at the lower level (level 1 units) nested within countries (level 2 units). Usually, the analysis of count data requires knowledge of the denominator, the population at risk. That is, the *rate* at which the disease occurs is of more direct interest than the corresponding count.

Counts are often modelled as Poisson random variables using a log link function. This motivates the following illustration of a two-level generalized linear model for Y_{ij} given by the three-part specification:

1. Conditional on a vector of random effects b_j, the Y_{ij} are assumed to be independent observations from a Poisson distribution, with $\text{Var}(Y_{ij}|b_j) = E(Y_{ij}|b_j)$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends upon fixed and random effects via the following log link function,

$$\log\{E(Y_{ij}|b_j)\} = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_j,$$

where T_{ij} is the population at risk in the i^{th} region of the j^{th} country and $\log(T_{ij})$ is an *offset*.

3. The random effects are assumed to have a multivariate normal distribution, with zero mean and covariance matrix G.

This is an example of a two-level log-linear model that assumes a linear relationship between the log rate of disease occurrence and the covariates.

Example 2: Two-Level Generalized Linear Model for Binary Responses

Consider a study of men with newly diagnosed prostate cancer. The study is designed to evaluate the factors that determine physician recommendations for surgery (radical prostatectomy) versus radiation therapy. In particular, it is of interest to determine the relative importance of patient factors (e.g., patient's age, level of prostate specific antigen) and physician factors (e.g., specialty training, years of experience) on physician recommendations for treatment. Many patients in the study seek the recommendation of the same physician. As a result, patients (level 1 units) are nested within physicians (level 2 units). For each patient we have a binary outcome denoting the physician recommendation (surgery versus radiation therapy).

Let Y_{ij} be the binary response, taking values 0 and 1 (e.g., denoting surgery or radiation therapy) for the i^{th} patient of the j^{th} physician. An illustrative example of a two-level logistic model for Y_{ij} is given by the following three-part specification:

1. Conditional on a single random effect b_j, the Y_{ij} are independent and have a Bernoulli distribution, with $\text{Var}(Y_{ij}|b_j) = E(Y_{ij}|b_j)\{1 - E(Y_{ij}|b_j)\}$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends upon fixed and random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}\beta + b_j,$$

with

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_j)}{\Pr(Y_{ij}=0|b_j)}\right\} = \eta_{ij} = X_{ij}\beta + b_j.$$

That is, the conditional mean of Y_{ij} is related to the linear predictor by a logit link function.

3. The single random effect b_j is assumed to have a univariate normal distribution, with zero mean and variance g_{11}.

In this example, the model is a simple two-level logistic regression model with randomly varying intercepts. In principle, the linear predictor can include additional random effects. However, some caution must be exercised because there is usually not much information in binary data to estimate more than a single variance component unless the number of level 1 units is relatively large.

In Section 10.3 we discussed how the logistic regression model for binary data can also be developed through the notion of an underlying latent variable distribution. The same notion can be applied to multilevel models for binary responses. Suppose that U_{ij} is a latent (i.e., unobserved) continuous variable and that a positive response is observed only when U_{ij} exceeds some threshold. Consider the following two-level linear model for U_{ij},

$$U_{ij} = X_{ij}\beta + Z_{ij}b_j + e_{ij},$$

where the random effects are assumed to have a multivariate normal distribution, with mean zero and covariance matrix, G, and the e_{ij} are assumed to have a standard logistic distribution, with mean zero and variance $\pi^2/3$. Without loss of generality, we can assume the threshold for categorizing U_{ij} is zero, with

$$Y_{ij} = 1 \quad \text{if} \quad U_{ij} > 0,$$

$$Y_{ij} = 0 \quad \text{if} \quad U_{ij} \leq 0.$$

Then the relationship between Y_{ij} and U_{ij} results in a logistic regression model for $\Pr(Y_{ij} = 1|b_j)$. That is, the two-level linear model for U_{ij} with standard logistic errors,

$$U_{ij} = X_{ij}\beta + Z_{ij}b_j + e_{ij},$$

implies the two-level logistic regression model for Y_i,

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_j)}{\Pr(Y_{ij}=0|b_j)}\right\} = X_{ij}\beta + Z_{ij}b_j.$$

Using the notion of an underlying latent variable distribution, we can then compare the magnitudes of the between-subject and within-subject sources of variability of the U_{ij}. For example, in a two-level logistic regression model with a single random effect b_j (with variance g_{11}), the relative magnitudes of the between-subject and within-subject sources of variability can be summarized in terms of the intra-cluster correlation

$$\rho = \text{Corr}(U_{ij}, U_{ik}) = \frac{g_{11}}{g_{11} + \pi^2/3}.$$

Note that ρ is the marginal correlation among the unobserved U_{ij}; it is not the marginal correlation among the Y_{ij}.

Although in both of the examples of two-level generalized linear models we have chosen canonical link functions to relate the conditional mean of Y_{ij} to η_{ij}, in principle, any suitable link function can be selected. These two examples are intended to be purely illustrative. They demonstrate how the choices of the three components might differ according to the type of response variable.

So far, our discussion of two-level models has closely paralleled the description of the generalized linear mixed effects model given in Chapter 12 (albeit, with the indices i and j reversed). In Chapter 12 we focused on two-level models where measurement occasions are the level 1 units and subjects are the level 2 units; this is a special case of two-level data. However, the methods in Chapter 12 can be applied more broadly to different types of two-level data and also extend naturally to more than two levels.

17.4.2 Three-Level Generalized Linear Models

The extension of two-level generalized linear models to three or more levels is straightforward and follows from the previous sections. With three-level data the variability of the response is accounted for by the introduction of random effects at both levels 2 and 3. For a three-level data structure, we adopt the same notation as in Section 17.3 except that the response can be continuous, binary, or a count. Let Y_{ijk} denote the response on the i^{th} level 1 unit within the j^{th} level 2 cluster within the k^{th} level 3 cluster. Associated with each Y_{ijk} is a $1 \times p$ (row) vector of covariates, X_{ijk}, with the covariates defined at different levels. A three-level generalized linear model for Y_{ijk} is given by the following three part specification:

1. We assume that the conditional distribution of each Y_{ijk}, given vectors of random effects, $b_k^{(3)}$ and $b_{jk}^{(2)}$ (defined at levels 3 and 2, respectively), belongs to the exponential family of distributions and that $\text{Var}(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)}) = v\{E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})\}\,\phi$, where $v(\cdot)$ is a known variance function, a function of the conditional mean, $E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})$, and ϕ is a scale or dispersion parameter. In addition, given $b_k^{(3)}$ and $b_{jk}^{(2)}$, it is assumed that the Y_{ijk} are independent of one another.

2. The conditional mean of Y_{ijk} is assumed to depend upon fixed and random effects via the following linear predictor,

$$\eta_{ijk} = X_{ijk}\beta + Z_{ijk}^{(3)}b_k^{(3)} + Z_{ijk}^{(2)}b_{jk}^{(2)},$$

with

$$g\{E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})\} = \eta_{ijk} = X_{ijk}\beta + Z_{ijk}^{(3)}b_k^{(3)} + Z_{ijk}^{(2)}b_{jk}^{(2)},$$

for some known link function, $g(\cdot)$.

3. Finally, the random effects are assumed to have multivariate normal distributions, with mean zero and covariance matrices, $\text{Cov}(b_k^{(3)}) = G^{(3)}$ and $\text{Cov}(b_{jk}^{(2)}) = G^{(2)}$. Although random effects may be correlated within a given level, random effects at different levels are assumed to be independent of each other.

These three components completely specify a broad class of three-level generalized linear models.

17.4.3 Estimation

The multilevel generalized linear models described in the previous section fully specify the joint distribution of the responses at level 1 and the random effects at all higher levels. As a result, we can base estimation and inference on the likelihood function. However, unlike the case with a continuous response assumed to have a normal distribution, maximum likelihood (ML) estimation for multilevel generalized linear models is not straightforward and will, in general, require numerical quadrature.

For example, for three-level data, inference about β, $G^{(2)}$ and $G^{(3)}$ is based on the marginal likelihood function. The marginal likelihood can be expressed as the product of the probability density functions, $f(y_k)$, for the level 3 units. Specifically, the marginal log-likelihood function is given by the following sum:

$$\sum_{k=1}^{n_3} \log f(y_k),$$

where $f(y_k)$ can be obtained by recognizing that observations on level 2 units (within level 3 units) are conditionally independent of one another given the level 3 random effects, $b^{(3)}$; similarly, observations on level 1 units (within level 2 units) are conditionally independent of one another given the level 3 and level 2 random effects, $b^{(3)}$ and $b^{(2)}$. The k^{th} level 3 unit's contribution to the likelihood function is

$$f(y_k) = \int \prod_{j=1}^{n_{2k}} \left\{ \int \prod_{i=1}^{n_{1jk}} f(y_{ijk}|b_k^{(3)}, b_{jk}^{(2)},) f(b_{jk}^{(2)})\, db_{jk}^{(2)} \right\} f(b_k^{(3)})\, db_k^{(3)},$$

where $f(b_{jk}^{(2)})$ and $f(b_k^{(3)})$ denote the multivariate normal distributions for the random effects at levels 2 and 3, respectively. The ML estimates of β, $G^{(2)}$ and $G^{(3)}$ are simply those values of β, $G^{(2)}$ and $G^{(3)}$ that maximize the marginal log-likelihood function.

The primary reason for displaying the expression given above is to highlight that multivariate integrals must be evaluated to compute the marginal log-likelihood. That is, the log-likelihood function is obtained by integrating out or averaging over the distribution of the random effects, $b_{jk}^{(2)}$ and $b_k^{(3)}$. Because integrals (denoting averaging over the distribution of the random effects) appear in the log-likelihood function, there are no simple, closed-form solutions. Instead, numerical integration techniques, for instance, Gaussian quadrature, are required for maximizing the log-likelihood function. ML estimation, using Gaussian quadrature, for two-level generalized linear models is implemented in some of the major statistical software packages (e.g., PROC NLMIXED in SAS). Various alternative approximations to ML estimation for the extensions to three or more levels are implemented in more specialized, stand-alone programs that have been specifically developed for multilevel modelling (e.g., MLwiN and HLM).

17.4.4 Case Studies

Next we illustrate the main ideas by conducting analyses of two-level data where the observations at level 1 are counts and binary outcomes. The first example analyzes two-level count data from a study of malignant melanoma mortality and ultraviolet (UV) radiation exposure. The second example analyzes two-level data on fetal malformations, a binary outcome, from the developmental toxicity study of ethylene glycol (EG) described in Section 17.3. For the latter, we present a traditional multilevel analysis of the fetal malformation data and contrast the results with those obtained from a marginal model that accounts for clustering in the data in a different way.

Malignant Melanoma Mortality and Ultraviolet Light Exposure

In a study of the effects of ultraviolet (UV) light exposure on malignant melanoma mortality (Langford *et al.*, 1998), counts of the number of deaths due to malignant melanoma were recorded for males of all ages in the United Kingdom (UK). The counts of the number of deaths between 1975 and 1980 were aggregated over areas that correspond to counties or shires; hereafter referred to as counties. Data were collected on 70 counties nested within 11 regions of the United Kingdom. The resulting data structure is multilevel, with counties at level 1 (indexed by i) nested within regions at level 2 (indexed by j). The main predictor of interest is exposure to ultraviolet light in the B band (UVB). An index of UVB dose reaching the earth's surface was calculated for each county. The mean UVB index in the United Kingdom was 10.9, with a standard deviation of 1.5.

Let Y_{ij} denote the count of the number of deaths in the i^{th} county in the j^{th} region due to malignant melanoma. Within a given region, we assume Y_{ij} has a Poisson distribution to account for level 1 variation in the counts. Variation in the counts across regions is accounted for by the inclusion of a random region effect, b_j. That is, conditional on a random region effect b_j, the counts are assumed to have a Poisson distribution with conditional mean related to UVB dose via a log link function,

$$\log\{E(Y_{ij}|b_j)\} = \log(T_{ij}) + \beta_1 + \beta_2 \text{UVB}_{ij} + b_j,$$

where UVB_{ij} is the UVB index (centered at the mean UVB index in the United Kingdom) in the i^{th} county of the j^{th} region. For each county, T_{ij} is the number of deaths that would be expected were U.K. national age- and gender-specific death rates to apply to the population of the county. Note that T_{ij} is known and $\log(T_{ij})$ is

Table 17.5 Fixed and random effects estimates for the malignant melanoma mortality data for males in the United Kingdom.

Variable	Estimate	SE	Z
Intercept	−0.0365	0.0352	−1.04
UVB	0.1301	0.0279	4.67
Level 2 Variance:			
$\sigma^2 \times 100$	0.6222	0.5087	1.22

ML estimation based on 50-point adaptive Gaussian quadrature.

an *offset* in this model. The ratio of the observed number of deaths to the expected number of deaths, Y/T, is referred to as the *standardized mortality ratio* (SMR) in each county. Our model assumes a linear relationship between the log SMR due to malignant melanoma and county level UV radiation exposure. Finally, we assumed the random region effect has a normal distribution, $b_j \sim N(0, \sigma^2)$.

The results of fitting this model to the U.K. malignant melanoma mortality data, using maximum likelihood estimation, are presented in Table 17.5. There is a significant positive relationship between the SMRs and exposure to UVB. Recall that the standard deviation of UVB in the United Kingdom is 1.5. Therefore the estimated effect of UVB dose indicates that the SMR is approximately 1.5 times larger (or $e^{3\times 0.13}$, with 95% confidence interval: 1.25 to 1.74) when comparing a county with UVB index 1 standard deviation above the U.K. average to a county with UVB index 1 standard deviation below. Finally, in interpreting these results it should be remembered that the UVB covariate used here is simply an index of exposure for each county; it is the potential, not the actual, UVB dose experienced by the population of residents in each county.

Developmental Toxicity Study of Ethylene Glycol

Next we consider a two-level logistic regression model for binary data on fetal malformations from the developmental toxicity study of ethylene glycol (EG). Recall that in this study, EG was administered (0, 750, 1500, or 3000 mg/kg/day) to 94 pregnant mice (dams) over the period of major organogenesis, beginning just after implantation (Price *et al.*, 1985). Following sacrifice, each live fetus was examined for evidence of malformations, recorded as present or absent. The primary question of scientific interest is the effect of dose on fetal malformations.

Summary statistics (ignoring clustering in the data) for fetal malformations for the 94 litters (composed of a total of 1028 live fetuses) are presented in Table 17.6. The percentage of fetal malformations increases monotonically with increasing dose,

Table 17.6 Descriptive statistics on fetal malformations from the ethylene glycol (EG) experiment.

Dose (mg/kg)	Dams	Fetuses	Fetal Malformations	
			Number	Percentage
0	25	297	1	0.34
750	24	276	26	9.42
1500	22	229	89	38.86
3000	23	226	129	57.08

Table 17.7 Fixed and random effects estimates for the fetal malformation data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	−4.360	0.440	−9.92
Dose	1.336	0.166	8.06
Level 2 Variance:			
σ_b^2	2.517	0.685	3.68

ML estimation based on 50-point adaptive Gaussian quadrature.

with less than 1% in the control group and almost 60% in the group administered the highest dose.

Letting $Y_{ij} = 1$ denote the presence of fetal malformations in the i^{th} live fetus from the j^{th} litter (and $Y_{ij} = 0$, otherwise), we considered the following logistic model relating the log odds of fetal malformations to a linear effect of dose:

$$\text{logit}\{E(Y_{ij}|b_j)\} = \beta_1 + \beta_2\, \mathrm{d_j} + b_j,$$

where $\mathrm{d_j} = \mathrm{Dose_j}/750$ denotes the dose (in units of 750 mg/kg) administered to the j^{th} dam (cluster). The random effect b_j is assumed to vary independently across litters, with $b_j \sim N(0, \sigma_b^2)$. This model assumes that fetuses within a litter are exchangeable and the positive association among the fetal malformation outcomes is accounted for by their sharing a common random effect, b_j.

The results of fitting the model to the fetal malformation data, using maximum likelihood estimation, are presented in Table 17.7. The estimated regression parameter

Table 17.8 Estimates of regression parameters from marginal model for the fetal malformation data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	−3.190	0.220	−14.53
Dose	0.960	0.099	9.66
Log Odds Ratio	1.447	0.221	6.56

for dose indicates that the log odds of malformation increases with increasing dose. In particular, the odds ratio for malformation, comparing the highest dose group to the control group, is 209.2 (or $e^{4\times1.336}$; with 95% confidence interval: 56.1 to 779.9). This provides overwhelming evidence of the increased risk of malformations at the highest dose of EG. The odds ratio for malformations, comparing the lowest dose group to the control group, is 3.80 (or $e^{1.336}$; with 95% confidence interval: 2.75 to 5.26). Finally, the estimate of σ_b^2 indicates that there are moderate litter effects, with heterogeneity across dams in the underlying risk of producing fetuses with malformations. For example, in the control group, 95% of dams have a risk of producing fetuses with malformations between 0 and 22%. Alternatively, if we appeal to the notion of a latent variable distribution and assume an underlying two-level linear model for the latent variable with standard logistic errors, the estimated intra-cluster correlation is

$$\widehat{\rho} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \pi^2/3} = \frac{2.517}{2.517 + 3.290} = 0.43.$$

For illustrative purposes, we also consider a marginal logistic regression model relating the log odds of fetal malformations to a linear effect of dose

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2 \, \text{d}_\text{j},$$

and account for the intra-litter association by a common log odds ratio,

$$\log\{\text{OR}(Y_{ij}, Y_{ik})\} = \log\left\{\frac{\Pr(Y_{ij}=1, Y_{ik}=1)\Pr(Y_{ij}=0, Y_{ik}=0)}{\Pr(Y_{ij}=1, Y_{ik}=0)\Pr(Y_{ij}=0, Y_{ik}=1)}\right\}.$$

This can be thought of as a marginal model analogue of the random intercepts model for the within-litter association. The results of fitting this marginal model, using GEE methods, are presented in Table 17.8. The estimate of the effect of dose indicates that the odds ratio for malformations, comparing the highest dose group to the control group, is 46.4 (or $e^{4\times0.960}$; with 95% confidence interval: 21.3 to 101.2). These results also provide strong evidence of the increased risk of malformations at the highest dose of EG. The within-litter odds ratio of 4.26 (or $e^{1.447}$) indicates that there is clustering in the fetal malformations data.

Note that the estimated effect of dose in the marginal model is discernibly smaller than that reported in Table 17.7. This should not be too surprising given the important distinctions between the regression parameters in marginal models and generalized linear mixed effects models that were highlighted in Chapter 13. The regression parameters for dose in the two models have quite different interpretations. In the logistic regression model with a random litter effect, β_2 describes the change in the risk of producing fetuses with malformations for *any given dam*; this change in the risk for a single-unit change in dose depends on b_j, a specific dam's random effect or underlying propensity for producing fetuses with malformations. In the marginal model, β_2 describes changes in the prevalence of fetal malformations when sub-populations of dams exposed to different doses of EG are compared. Although both models account for clustering in the data, the targets of inference are different and the two analyses address distinct scientific questions.

17.5 SUMMARY

In previous sections we described models for data with a hierarchical structure, where lower-level units are nested within higher-level units. The dominant approach for modelling such data is regression models where random effects are introduced at different levels. A central feature of multilevel modelling is the incorporation of covariates that can be measured at any level of the hierarchy, thereby allowing the effects at each level to be disentangled. By combining covariates that have been measured at different levels within a single regression model, their relative importance can be determined. For example, multilevel models can address questions about the effects of individuals' characteristics (e.g., disease severity) while adjusting for their context (e.g., being treated at a large university teaching hospital versus a rural clinic).

Multilevel data can be challenging to analyze for at least two main reasons. First, the covariates can be measured at different levels, and the same covariate can operate at many different levels. As a result somewhat greater care is required in the interpretation of regression parameters in multilevel models. It is not always transparent how best to combine covariates measured at different levels within a single model so that the regression parameters have useful interpretations. In our brief overview of multilevel models, we have not touched upon this important topic; for more information, the interested reader is directed to the references at the end of this chapter.

The second challenge in the analysis of multilevel data is how best to account for the clustering that can arise at different levels of the hierarchy. In the multilevel modelling literature, the dominant approach for accounting for the intra-cluster correlations at different levels is via the introduction of random effects at different levels. This gives rise to mixed effects models that can be extended in a very natural way to any number of levels of clustering in the data. For linear models, this is certainly a very natural way to account for clustering. However, for generalized linear models for discrete data, it does raise subtle issues concerning the interpretation of the fixed effect regression parameters and questions about what is the relevant target of inference. The same issues that were given an airing in Chapter 13 apply equally to multilevel

models for discrete data. These issues were highlighted in our analyses of the two-level clustered data on fetal malformations in Section 17.4, where the estimated effect of dose was discernibly different depending upon how the clustering was accounted for. The estimates of the effect of dose from the marginal and random effects logistic regression models differed because the corresponding regression parameters have distinct interpretations and address somewhat different scientific questions. In general, the fixed effects parameters in a two-level model for discrete data represent changes in the (transformed) mean response, for a single-unit change in the corresponding covariate, *for any given level 2 unit*. In Chapters 12 and 13, in the context of longitudinal data, these regression coefficients were referred to as "subject-specific"; here, they are "cluster-specific" and describe covariate effects for an individual cluster. In contrast, the regression parameters in a marginal model represent changes in the (transformed) mean response when sub-populations defined by different values of the corresponding covariate are compared. The regression parameters in marginal models address the dependence of the population-averaged response (where averaging is over all possible units in the hierarchy) on the covariates. These regression parameters do not have any direct interpretation for an individual cluster when there is heterogeneity across clusters.

Although much of the multilevel literature on the analysis of discrete data is dominated by the use of generalized linear mixed effects models, we note that marginal models can also be used to account for clustering at different levels. All of the issues discussed in Chapter 13 for two-level longitudinal data apply equally to two- and higher-level data more broadly defined. In general, the choice between the two classes of models should not be driven by the availability of software for multilevel modelling but on the basis of careful thought about the questions of scientific interest.

17.6 FURTHER READING

There is an extensive literature on multilevel models that appears in the statistical, psychometric, and educational literature. A comprehensive description of multilevel models, and their application to a wide range of problems, can be found in the books by Raudenbush and Bryk (2002), Longford (1993), and Goldstein (2003). For readers who find the level of mathematical difficulty in these books too challenging, the books by Hox (2002), Kreft and De Leeuw (1998), and Snijders and Bosker (1999) provide a more introductory and accessible presentation of similar topics targeted at empirical researchers.

For illustrations of the application of multilevel models in the biomedical and health sciences, we recommend the edited volume of articles in Leyland and Goldstein (2001) and the review articles by Sullivan *et al.* (1999), Goldstein *et al.* (2002), and Subramanian *et al.* (2003).

Bibliographic Notes

Although most of the statistical literature on marginal models has focused on two-level data, Qaqish and Liang (1992) discuss the use of marginal models for multilevel binary data, with multiple levels of nesting.

Daniels and Gatsonis (1999) describe multilevel modelling in a Bayesian framework; also see Lindley and Smith (1972), Zeger and Karim (1991), Browne *et al.* (2002), Carlin and Louis (2000), and Chapters 15 and 16 of Gelman *et al.* (2003). Bayesian methods for multilevel modelling can be implemented using the publicly available software WinBUGS (Spiegelhalter *et al.*, 1999).

Appendix A
Gentle Introduction to Vectors and Matrices

We present a very brief introduction to vectors and matrices, intended for readers with no prior exposure to matrix algebra. Specifically, we cover basic definitions and summarize some of the main properties of vectors and matrices. Vectors and matrices allow us to perform common mathematical operations (e.g., addition, subtraction, and multiplication) on a collection of numbers; they also facilitate the description of statistical methods for multivariate data. Our primary motivation for using them is the conciseness and compactness with which statistical techniques for analyzing longitudinal data can be presented when expressed in terms of vectors and matrices.

Mastery of the material presented in this section is a prerequisite for understanding the statistical methods for longitudinal data described in the book. Although we do not assume a profound understanding of matrix algebra, vectors and matrices are used extensively throughout the book to simplify notation and the reader is required to have some basic facility with the addition and multiplication of vectors and matrices.

Basic Concepts and Definitions

A *matrix* is a rectangular array of elements (e.g., numbers), arranged in rows and columns. For example,

$$A = \begin{pmatrix} 2 & 7 & 11 & 5 \\ 4 & 9 & 3 & 1 \\ 13 & 8 & 2 & 6 \end{pmatrix}$$

is a matrix with three rows and four columns. The *element* or *entry* in the i^{th} row and the j^{th} column of the matrix is referred to as the $(i, j)^{th}$ element of the matrix. For example, the entry in the 2^{nd} row and 3^{rd} column of A is 3. If we let a_{ij} denote the element in the i^{th} row and the j^{th} column of the matrix A, then

$$\begin{array}{llll} a_{11} = 2, & a_{12} = 7, & a_{13} = 11, & a_{14} = 5; \\ a_{21} = 4, & a_{22} = 9, & a_{23} = 3, & a_{24} = 1; \\ a_{31} = 13, & a_{32} = 8, & a_{33} = 2, & a_{34} = 6. \end{array}$$

The subscripts on the element a_{ij} denote its position in the i^{th} row and the j^{th} column of the matrix A.

The *dimension* of a matrix is the number of rows and columns in the matrix. By convention, the number of rows is listed first, and then the number of columns. Thus, we refer to the matrix A above as being a 3×4, or a "3 by 4", matrix.

A *vector* is a special kind of matrix, having either a single row or a single column. For example,

$$V = \begin{pmatrix} 2 \\ 4 \\ 9 \\ 7 \\ 3 \end{pmatrix}$$

is a 5×1 (column) vector. Since the dimension of a vector corresponds to the number of elements in the vector, the dimension of a vector is often loosely referred to as its *length*[1].

Finally, a *scalar* is a single element (e.g., a single number), and hence can be treated either as a single element vector or as a 1×1 matrix.

[1] In matrix algebra, vectors have a geometric meaning, denoting the coordinates of a point in Euclidean space. The geometric concept of the "length" (or magnitude) of a vector in Euclidean space has a very precise definition and technical meaning that is quite different from our informal use of the term here.

Transpose

The *transpose* is a function that interchanges the rows and columns of a matrix. That is, the first row becomes the first column, the second row becomes the second column, and so on. By convention, the transpose of a matrix A is denoted A' (or "A prime"). (Note that in some texts, a superscript 'T', instead of a prime, is used to denote the transpose of a matrix, e.g., A^T.)

For example, consider the 3×4 matrix A,

$$A = \begin{pmatrix} 2 & 7 & 11 & 5 \\ 4 & 9 & 3 & 1 \\ 13 & 8 & 2 & 6 \end{pmatrix}.$$

The transpose of A,

$$A' = \begin{pmatrix} 2 & 4 & 13 \\ 7 & 9 & 8 \\ 11 & 3 & 2 \\ 5 & 1 & 6 \end{pmatrix},$$

is the 4×3 matrix with rows and columns interchanged. Similarly, since a vector is a matrix with either a single row or column, if

$$V = \begin{pmatrix} 2 \\ 4 \\ 9 \\ 7 \\ 3 \end{pmatrix}, \quad \text{then} \quad V' = \begin{pmatrix} 2 & 4 & 9 & 7 & 3 \end{pmatrix}.$$

Examples of vectors and matrices that play key roles in the analysis of longitudinal data are the *response vector*, often denoted Y, and the *covariate matrix*, often denoted X. For example, consider the following data from a subject participating in a longitudinal clinical trial. In this trial, the subject was assigned to the placebo group (Group = 0 if assigned to placebo, Group = 1 if assigned to active treatment) and four repeated measures of blood lead levels were obtained at baseline (or week 0), week 1, week 4, and week 6:

Blood Lead	Treatment Group	Week
30.8	0	0
26.9	0	1
25.8	0	4
23.8	0	6

If we let Y denote the vector of repeated measurements of the response variable, then

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} 30.8 \\ 26.9 \\ 25.8 \\ 23.8 \end{pmatrix}.$$

Similarly, we can let X denote a matrix of covariates associated with the vector of repeated measurements, with

$$X = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & X_{43} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 4 \\ 1 & 0 & 6 \end{pmatrix}.$$

The first column of X contains only 1's, while the second column of X contains a variable denoting the treatment group assignment and the third column contains the times of the repeated measurements.

Square and Symmetric Matrices

A matrix is said to be *square* if it has the same number of rows and columns. A square matrix is *symmetric* if it equals its transpose. For example,

$$S = \begin{pmatrix} 2 & 3 & 7 & 11 \\ 3 & 9 & 1 & 2 \\ 7 & 1 & 5 & 8 \\ 11 & 2 & 8 & 4 \end{pmatrix}$$

is a symmetric matrix since it equals its transpose

$$S' = \begin{pmatrix} 2 & 3 & 7 & 11 \\ 3 & 9 & 1 & 2 \\ 7 & 1 & 5 & 8 \\ 11 & 2 & 8 & 4 \end{pmatrix}.$$

Examples of symmetric matrices that play an important role in the analysis of longitudinal data are the covariance and correlation matrices for the repeated measures on the same individuals.

Finally, a *diagonal* matrix is a special case of a symmetric square matrix that has non-zero elements only in the main diagonal positions, and zeros elsewhere. The main diagonal elements are those in the same row and column, from the upper left to

the lower right corners of the matrix. For example,

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

is a diagonal matrix. The diagonal matrix having all ones along the main diagonal is known as the *identity* matrix and is often denoted by I or I_n, where the subscript n denotes the dimension of the identity matrix. Thus

$$I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Arithmetic Operations

Addition and subtraction of matrices are defined only for matrices of the same dimension. That is, the matrices must share the same number of rows and the same number of columns. The sum of two matrices is obtained by adding their corresponding elements. For example,

$$\begin{pmatrix} 2 & 7 & 11 \\ 4 & 9 & 3 \\ 13 & 8 & 2 \end{pmatrix} + \begin{pmatrix} 3 & 2 & 14 \\ 7 & 8 & 4 \\ 6 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 2+3 & 7+2 & 11+14 \\ 4+7 & 9+8 & 3+4 \\ 13+6 & 8+5 & 2+9 \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 9 & 25 \\ 11 & 17 & 7 \\ 19 & 13 & 11 \end{pmatrix}.$$

Subtraction of matrices is defined in a similar way. For example,

$$\begin{pmatrix} 2 & 7 & 11 \\ 4 & 9 & 3 \\ 13 & 8 & 2 \end{pmatrix} - \begin{pmatrix} 3 & 2 & 14 \\ 7 & 8 & 4 \\ 6 & 5 & 9 \end{pmatrix} = \begin{pmatrix} 2-3 & 7-2 & 11-14 \\ 4-7 & 9-8 & 3-4 \\ 13-6 & 8-5 & 2-9 \end{pmatrix}$$

$$= \begin{pmatrix} -1 & 5 & -3 \\ -3 & 1 & -1 \\ 7 & 3 & -7 \end{pmatrix}.$$

Scalar Multiplication of a Matrix

A scalar is a single number, as opposed to a vector or matrix of numbers. The scalar multiple of a matrix is formed by multiplying each element of the matrix by the scalar. For example, if

$$A = \begin{pmatrix} 2 & 7 & 11 & 5 \\ 4 & 9 & 3 & 1 \\ 13 & 8 & 2 & 6 \end{pmatrix}, \quad \text{then} \quad 2A = \begin{pmatrix} 4 & 14 & 22 & 10 \\ 8 & 18 & 6 & 2 \\ 26 & 16 & 4 & 12 \end{pmatrix}.$$

Multiplication of Matrices

The multiplication of two matrices is somewhat more involved. The multiplication of two matrices A and B, denoted AB, is defined only if the number of columns of A is equal to the number of rows of B. For example, if A is a $p \times q$ matrix and B is a $q \times r$ matrix, then the product of the two matrices AB is a $p \times r$ matrix. Letting C be the product of A and B,

$$C = AB,$$

the $(i, j)^{th}$ element of C is the sum of the products of the corresponding elements in the i^{th} row of A and the j^{th} column of B. Specifically, if c_{ij} is the element in the i^{th} row and the j^{th} column of the matrix $C = AB$, then

$$c_{ij} = \sum_{k=1}^{q} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj}, \quad i = 1, ..., p; \quad j = 1, ..., r;$$

where q is the number of columns in A or the number of rows in B. Matrix multiplication is best understood by considering a simple example. Suppose

$$A = \begin{pmatrix} 2 & 7 & 11 \\ 4 & 9 & 3 \\ 13 & 8 & 2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 4 \end{pmatrix}$$

then

$$AB = \begin{pmatrix} (2 \times 1) + (7 \times 3) + (11 \times 2) & (2 \times 2) + (7 \times 1) + (11 \times 4) \\ (4 \times 1) + (9 \times 3) + (3 \times 2) & (4 \times 2) + (9 \times 1) + (3 \times 4) \\ (13 \times 1) + (8 \times 3) + (2 \times 2) & (13 \times 2) + (8 \times 1) + (2 \times 4) \end{pmatrix}$$

$$= \begin{pmatrix} 45 & 55 \\ 37 & 29 \\ 41 & 42 \end{pmatrix}.$$

Note that the order of multiplication is very important. For example, if A and B are both square matrices of the same dimension, then AB is usually not equal to BA.

The multiplication of a vector by a matrix is a particularly important operation that plays a key role in longitudinal analysis. Let B be a $p \times 1$ vector and X be a $n \times p$ matrix. Then the product,

$$C = XB,$$

is a $n \times 1$ vector with

$$c_i = \sum_{k=1}^{p} x_{ik}b_k, \quad i = 1, ..., n;$$

where x_{ij} is the element in the i^{th} row and the j^{th} column of the matrix X and b_j is the element in the j^{th} row of the vector B. That is,

$$c_1 = x_{11}b_1 + x_{12}b_2 + \cdots + x_{1p}b_p,$$

$$c_2 = x_{21}b_1 + x_{22}b_2 + \cdots + x_{2p}b_p,$$

$$c_3 = x_{31}b_1 + x_{32}b_2 + \cdots + x_{3p}b_p,$$

and so on.

Let us return to the example introduced earlier, with repeated measures of blood lead levels obtained on four occasions. Letting Y denote the vector of repeated measurements of the response variable,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix},$$

and X a matrix of covariates associated with the vector of repeated measurements,

$$X = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & X_{43} \end{pmatrix},$$

a linear regression model for the mean of each response can be expressed in vector and matrix notation as

$$E(Y) = X\beta,$$

where $E(Y)$ denotes the expected value or mean of Y (see *Properties of Expectations and Variances* in Appendix B) and β is a 3×1 vector of regression coefficients,

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Specifically, the product

$$E(Y) = X\beta,$$

is a 4×1 vector

$$\begin{pmatrix} E(Y_1) \\ E(Y_2) \\ E(Y_3) \\ E(Y_4) \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & X_{43} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} \\ \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} \\ \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33} \\ \beta_1 X_{41} + \beta_2 X_{42} + \beta_3 X_{43} \end{pmatrix}.$$

That is,

$$E(Y) = X\beta,$$

is simply a shorthand representation for the following series of linear regression equations

$$E(Y_1) = \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13},$$

$$E(Y_2) = \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23},$$

$$E(Y_3) = \beta_1 X_{31} + \beta_2 X_{32} + \beta_3 X_{33},$$

$$E(Y_4) = \beta_1 X_{41} + \beta_2 X_{42} + \beta_3 X_{43}.$$

Inverse

The *inverse* of a square matrix A, denoted A^{-1}, is defined as a square matrix whose elements are such that

$$AA^{-1} = A^{-1}A = I,$$

where I is the identity matrix, a diagonal matrix having all ones along the main diagonal. That is, the product of A by its inverse is equal to the identity matrix. The inverse of a square matrix does not always exist. The inverse of a matrix only exists if the matrix is *non-singular*.

In matrix algebra, the inverse plays the role of the reciprocal, and thus multiplication by an inverse, A^{-1}, can loosely be thought of as "division" by the matrix A. Methods for calculating the inverse of a matrix will not be discussed here. In practice, the inverse of a matrix is usually obtained with the aid of a computer.

Finally, the *determinant* of a square matrix is a unique scalar (or single number) function of its elements and is denoted by $|A|$. For example, if A is a 2×2 matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then the determinant of A is the following function of its elements

$$|A| = a_{11}\,a_{22} - a_{12}\,a_{21}.$$

The corresponding expression for the determinant of a 3×3 square matrix, and of matrices of higher dimensions, is more involved and the details are not important. A useful property of the determinant is that it provides a test of whether the inverse of a matrix exists. In particular, if $|A| \neq 0$, then the inverse of A exists; if $|A| = 0$, then the matrix is said to be *singular* and the inverse of A does not exist.

The determinant also plays a role in the definition of the multivariate normal distribution (see Section 3.2). The multivariate normal density includes a term involving the determinant of the covariance matrix. The determinant of the covariance matrix is often referred to as the *generalized variance* and characterizes the salient features of the variation expressed by the covariance matrix in a single number summary.

Appendix B
Properties of Expectations and Variances

Let Y denote a random variable that takes on values according to some probability density function if Y is continuous or some probability mass function if Y is discrete. The *expected value*, or *expectation*, of Y is simply its *mean* or average value and is usually denoted by

$$\mu = E(Y).$$

It is often referred to as the first *moment* of Y, since it describes the location of the center of the distribution. The precise definition of the expectation of Y is that it is a *weighted* average of all the possible values of Y, with weights determined by the probabilities associated with each possible value.

The *variance* of Y, often denoted by $\sigma^2 = \text{Var}(Y)$, is a measure of the dispersion or variability around the mean or expected value of Y. The variance is often referred to as the second *central moment* of Y and is defined as

$$\sigma^2 = \text{Var}(Y) = E\{Y - E(Y)\}^2.$$

The variance is a weighted average of the squared deviations of Y around its mean. Because the variance is expressed in squared units of Y, a measure of variability in

the original units of Y is given by the *standard deviation*

$$\sigma = \sqrt{\text{Var}(Y)}.$$

Finally, the covariance between two random variables, X and Y, is defined as

$$\text{Cov}(X, Y) = E\left[\{X - E(X)\}\{Y - E(Y)\}\right],$$

and is a measure of the *linear dependence* between X and Y. If X and Y are *independent*, then $\text{Cov}(X, Y) = 0$. Note that the covariance of a variable with itself is the variance, $\text{Cov}(Y, Y) = \text{Var}(Y)$.

Properties of Expectations and Variances

Next, we consider some properties of expectations and variances. Let X and Y be two (possibly dependent) random variables and let a and b denote non-random constants. Then the expectation operator, $E(\cdot)$, has the following five important properties:

1. $E(a) = a$
2. $E(b\,X) = b\,E(X)$
3. $E(a + b\,X) = a + b\,E(X)$
4. $E(a\,X + b\,Y) = a\,E(X) + b\,E(Y)$
5. $E(XY) \neq E(X)\,E(Y)$ (unless X and Y are *independent*.)

Thus expectation is a linear operator in the sense that it respects or preserves the arithmetic operations of addition and multiplication by a constant. As a result, the expected value of a linear function of Y (e.g., $a + b\,Y$) is simply the same linear function of the expected value of Y (e.g., $a + b\,E(Y)$).

The variance operator, $\text{Var}(\cdot)$, has the following five important properties:

1. $\text{Var}(a) = 0$
2. $\text{Var}(b\,Y) = b^2\,\text{Var}(Y)$
3. $\text{Var}(a + b\,Y) = b^2\,\text{Var}(Y)$
4. $\text{Var}(a\,X + b\,Y) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2\,a\,b\,\text{Cov}(X, Y)$
5. $\text{Var}(a\,X - b\,Y) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) - 2\,a\,b\,\text{Cov}(X, Y)$.

In particular, if X and Y are dependent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y).$$

Finally, we note that the expectation and variance operators can also be applied to vectors of random variable. For example, let Y be a $n \times 1$ (column) response vector (e.g., repeated measurements at n different occasions),

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

then

$$E(Y) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix},$$

and

$$\text{Cov}(Y) = \begin{pmatrix} \text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \dots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_2, Y_1) & \text{Var}(Y_2) & \dots & \text{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Y_1) & \text{Cov}(Y_n, Y_2) & \dots & \text{Var}(Y_n) \end{pmatrix}.$$

Appendix C
Critical Points for a 50:50 Mixture of Chi-Squared Distributions

Table C.1 Critical points[a] for a 50:50 mixture of chi-squared distributions with q and $q+1$ degrees of freedom; right-hand tail probabilities. Adapted from Monette *et al.* (2002).

	Significance Level									
q	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
0	0.28	0.71	1.64	2.71	3.84	5.41	6.63	7.88	9.55	10.83
1	1.76	2.50	3.81	5.14	6.48	8.27	9.63	11.00	12.81	14.18
2	3.06	3.98	5.53	7.05	8.54	10.50	11.97	13.43	15.36	16.80
3	4.29	5.36	7.09	8.76	10.38	12.48	14.04	15.59	17.61	19.13
4	5.49	6.68	8.57	10.37	12.10	14.32	15.97	17.59	19.69	21.27
5	6.66	7.96	10.00	11.91	13.7·	16.07	17.79	19.47	21.66	23.29
6	7.82	9.21	11.38	13.40	15.32	17.76	19.54	21.29	23.55	25.23
7	8.97	10.44	12.74	14.85	16.86	19.38	21.23	23.04	25.37	27.10
8	10.10	11.66	14.07	16.27	18.35	20.97	22.88	24.74	27.13	28.91
9	11.23	12.87	15.38	17.67	19.82	22.52	24.49	26.40	28.86	30.68
10	12.35	14.06	16.67	19.04	21.27	24.05	26.07	28.02	30.54	32.40

[a] Critical value c such that right-hand tail probability equals $0.5 \times \Pr(\chi^2_q > c) + 0.5 \times \Pr(\chi^2_{q+1} > c)$, where χ^2_q and χ^2_{q+1} denote chi-squared distributions with q and $q+1$ degrees of freedom, respectively.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B*, **46**, 118–119.

Altman, D.G. (1990). *Practical Statistics for Medical Research*. New York: Chapman and Hall/CRC Press.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Anderson, D.A. and Aitkin, M. (1985). Variance components models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203–210.

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research*. New York: McGraw-Hill.

Bandini, L.G., Must, A., Spadano, J.L. and Dietz, W.H. (2002). Relation of body composition, parental overweight, pubertal stage, and race-ethnicity to energy expenditure among premenarcheal girls. *American Journal of Clinical Nutrition*, **76**, 1040–1047.

Bock, R.D. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

Box, G.E.P. (1950). Problems in the analysis of growth and wear data. *Biometrics*, **6**, 362–389.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Brown, H. and Prescott, R. (1999). *Applied Mixed Models in Medicine*. Chichester, UK: Wiley.

Browne, W.J., Draper, D., Goldstein, H. and Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis*, **39**, 203–225.

Burton, P., Gurrin, L. and Sly, P. (1998). Tutorial in biostatistics: Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level modelling. *Statistics in Medicine*, **17**, 1261–1291.

Byar, D. and Blackard, C. (1977). Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage I bladder cancer. *Urology*, **10**, 556–561.

Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. New York: Chapman and Hall/CRC Press.

Chi, E.M. and Reinsel, G.C. (1989). Models of longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, **84**, 452–459.

Cnaan, A., Laird, N.M. and Slasor, P. (1997). Tutorial in biostatistics: Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, **16**, 2349–2380.

Cochran, W.G. and Cox, G.H. (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

Cole, J.W.L. and Grizzle, J.E. (1966). Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, **22**, 810–828.

Collett, D. (1991). *Modelling Binary Data*. London: Chapman and Hall/CRC Press.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall/CRC Press.

Cox, D.R. (1970). *Analysis of Binary Data*, 1st ed. New York: Chapman and Hall/CRC Press.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D.R. and Wermuth, N. (1996). *Multivariate Dependencies. Models, Analysis and Interpretation*. New York: Chapman and Hall/CRC Press.

Crowder, M.J. and Hand, D.J. (1990). *Analysis of Repeated Measures*. London: Chapman and Hall/CRC Press.

Danford, M.B., Hughes, H.M. and McNee, R.C. (1960). On the analysis of repeated measurements experiments. *Biometrics*, **16**, 547–565.

Daniels, M. and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, **94**, 29–42.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall/CRC Press.

Davis, C.S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, **10**, 1959–1980.

Davis, C.S. (2002) *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Dawber, T.R. (1980). *The Framingham Study: The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press.

Dawber, T.R., Meadors, G.F. and Moore, F.E.J. (1951). Epidemiological approaches to heart disease: The Framingham Study. *American Journal of Public Health*, **41**, 279–286.

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: A double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. *Journal of the American Academy of Dermatology*, **38**, 57–63.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Diggle, P.J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Oxford, UK: Oxford University Press.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall/CRC Press.

Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E. and Ferris, B.G. (1983). Distribution of FVC and FEV_1 in children 6 to 11 years old. *American Review of Respiratory Disease*, **128**, 405–412.

Drum, M. and McCullagh, P. (1993). Comment on "Regression models for discrete longitudinal responses". *Statistical Science*, **8**, 300–301.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Biometrika*, **47**, 139–153.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, **236**, 119–127.

Emrich, L.J. and Piedmonte, M.R (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, **41**, 19–29.

Everitt, B.S. (1995). The analysis of repeated measures: A practical review with examples. *Statistician*, **44**, 113–135.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. New York: Springer.

Fay, M.P. and Graubard, B.I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, **57**, 1198–1206.

Fay, M.P., Graubard, B.I., Freedman, L.S. and Midthune, D.N. (1998). Conditional logistic regression with sandwich estimators: Application to a meta-analysis. *Biometrics*, **54**, 195–208.

Feldman, H.A. (1988). Families of lines: Random effects in linear regression analysis. *Journal of Applied Physiology*, **64**, 1721–1732.

Firth, D. (1991). Generalized linear models. In Hinkley, D.V., Reid, N. and Snell, E.J. (editors), *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*. London: Chapman and Hall/CRC Press.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd.

Fitzmaurice, G.M. (2001). A conundrum in the analysis of change. *Nutrition*, **17**, 360–361.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Fitzmaurice, G.M., Laird, N.M. and Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statistical Science*, **8**, 248–309.

Fitzmaurice, G.M., Laird, N.M., Zahner, G.E.P. and Daskalakis, C. (1995). Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *American Journal of Epidemiology*, **142**, 1194–1203.

Flay, B.R., Miller, T.Q., Hedeker, D., Siddiqui, O., Brannon, B.R., Johnson, C.A., Hansen, W.B., Sussman, S. and Dent, C. (1995). The television, school and family smoking prevention and cessation project: VIII. Student outcomes and mediating variables. *Preventive Medicine*, **24**, 29–40.

Freund, R.J., Littell, R.C. and Spector, P.C. (1986). *SAS System for Linear Models*. Cary, NC: SAS Institute Inc.

Friedman, G.D., Cutter, G.R., Donahue, R., Hughes, G.H., Hulley, S., Jacobs, D.R., Liu, K. and Savage, P.J. (1988). CARDIA: Study design, recruitment,

and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, **41**, 1105–1116.

Geisser, S. (1963). Multivariate analysis of variance for a special covariance case. *Journal of the American Statistical Association*, **58**, 660–669.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd ed. New York: Chapman and Hall/CRC Press.

Gibbons, R.D., Hedeker, D., Waternaux C. and Davis, J.M. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacological Bulletin*, **24**, 438–443.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach*. Thousand Oaks, CA: Sage Publications.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by generalized linear mixed models. *Biometrika*, **72**, 593–599.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1212.

Goldstein, H. (2003). *Multilevel Statistical Methods*, 3rd ed. London: Edward Arnold.

Goldstein, H., Browne, W. and Rasbach, J. (2002). Multilevel modelling of medical data. *Statistics in Medicine*, **21**, 3291–3315.

Goldwasser, M. and Fitzmaurice, G.M. (2001). Multivariate linear regression analysis of childhood psychopathology using multiple informant data. *International Journal of Methods in Psychiatric Research*, **10**, 1–10.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo-maximum likelihood methods: Theory. *Econometrica*, **52**, 681–700.

Graubard, B.I. and Korn, E.L. (1994). Regression analysis with clustered data. *Statistics in Medicine*, **13**, 509–522.

Greenberg, E.R., Baron, J.A., Stevens, M.M., Stukel, T.A., Mandel, J.S., Spencer, S.K., Elias, P.M., Lowe, N., Nierenberg, D.W., Bayrd, G. and Vance, J.C. (1989). The Skin Cancer Prevention Study: Design of a clinical trial of beta-carotene among persons at high risk for nonmelanoma skin cancer. *Controlled Clinical Trials*, **10**, 153–166.

Greenberg, E.R., Baron, J.A., Stukel, T.A., Stevens, M.M., Mandel, J.S., Spencer, S.K., Elias, P.M., Lowe, N., Nierenberg, D.W., Bayrd, G., Vance, J.C., Freeman, D.H., Clendenning, W.E., Kwan, T. and the Skin Cancer Prevention Study Group (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, **323**, 789–795.

Greenhouse, S.W. and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, **32**, 95–112.

Grizzle, J.E. and Allen, D.W. (1969). Analysis of growth and dose response curves. *Biometrics*, **25**, 357–381.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, **15**, 489–504.

Gunsolley, J.C., Getchell, C. and Chinchilli, V.M. (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics, Simulation and Computation*, **24**, 869–878.

Hand, D.J. and Crowder, M.J. (1996). *Practical Longitudinal Data Analysis*. New York: Chapman and Hall/CRC Press.

Hand, D.J. and Taylor, C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures*. New York: Chapman and Hall/CRC Press.

Hartley, H.O. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338.

Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–944.

Hedeker, D. and Gibbons, R.D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, **49**, 157–176.

Hedeker, D., Gibbons, R.D. and Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, **24**, 70–93.

Henderson, C.R. (1963). Selection index and expected genetic advance. In Hanson, W.D. and Robinson, H.F. (editors), *Statistical Genetics and Plant Breeding*. Washington, D.C.: National Academy of Sciences-National Research Council.

Henry, K., Erice, A., Tierney, C., Balfour, H.H. Jr, Fischl, M.A., Kmack, A., Liou, S.H., Kenton, A., Hirsch, M.S., Phair, J., Martinez, A. and Kahn J.O. for the AIDS Clinical Trial Group 193A Study Team (1998). A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **19**, 339–349.

Hernández, B., Gortmaker, S.L., Laird, N.M., Colditz, G.A., Parra-Cabrera, S. and Peterson, K.E. (2000). Validity and reproducibility of a physical activity and inactivity questionnaire for Mexico City's schoolchildren. *Salud Publica de Mexico*, **42**, 315–323.

Heyting, A., Tolboom, J. and Essers, J. (1992). Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine*, **11**, 2043–2061.

Hinkley, D.V. and Wang, S. (1991). Efficiency of robust standard errors for regression coefficients. *Communications in Statistics, Theory and Methods*, **20**, 1–11.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Hosmer, D.W. Jr. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. London: Lawrence Erlbaum Associates.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol 1, pp. 221–233. Berkeley, CA: University of California Press.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**, 805–820.

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, 5th ed. Englewood Cliff, NJ: Prentice-Hall.

Jones, B. and Donev, A.N. (1996). Modelling and design of cross-over trials. *Statistics in Medicine*, **15**, 1435–1446.

Jones, B. and Kenward, M.G. (1989). *Design and Analysis of Cross-over Trials*. London: Chapman and Hall/CRC Press.

Kackar, R.N. and Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics, Theory and Methods*, **10**, 1249–1261.

Kakwani, N.C. (1967). The unbiasedness of Zellner's seemingly unrelated regression equations estimators. *Journal of the American Statistical Association*, **62**, 141–142.

Kauermann, G. and Carroll, R.J. (2001). The sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals. *Journal of the American Statistical Association*, **96**, 1387–1396.

Kenward, M.G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, **8**, 51–83.

Kenward, M.G. and Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.

Keselman, H.J. and Keselman, J.C. (1984). The analysis of repeated measures designs in medical research. *Statistics in Medicine*, **3**, 185–195.

Khatri, C.G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, **18**, 75–86.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. and Nizam, A. (1998). *Applied Regression Analysis and Multivariable Methods*, 3rd ed. Pacific Grove, CA: Duxbury Press.

Koch, G.G., Carr, G.J., Amara, I.A., Stokes, M.E. and Uryniak, T.J. (1990). Categorical data analysis. In Berry, D.A. (editor), *Statistical Methodology in the Pharmaceutical Sciences*. New York: Marcel Dekker.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133–158.

Kramer, M.S., Chalmers, B., Hodnett, E.D., Sevkovskaya, Z., Dzikovich, I. and Shapiro, S. for the PROBIT Study Group (2001). Promotion of breastfeeding intervention trial (PROBIT): A randomized trial in the Republic of Belarus. *Journal of the American Medical Association*, **285**, 413–420.

Kreft, I.I. and de Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications.

Laird, N.M. (1983). Further comparative analyses of pretest-posttest research designs. *American Statistician*, **37**, 329–330.

Laird, N.M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.

Laird, N.M., Donnelly, C. and Ware, J.H. (1992). Longitudinal models with continuous responses. *Statistical Methods in Medical Research*, **1**, 225–247.

Laird, N.M., Lange, N. and Stram, D.O. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, **82**, 97–105.

Laird, N.M., Skinner, J. and Kenward, M.G. (1992). An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine*, **11**, 1967–1979.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Langford, I.H., Bentham, G. and McDonald, A. (1998). Multilevel modelling of geographically aggregated health data: A case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine*, **17**, 41–58.

Lauer, R.M., Clarke, W.R. and Burns, T.L. (1997). Obesity in childhood: The Muscatine Study. *Acta Pædiatrica Scandinavica*, **38**, 432–437.

Leppik, I., Dreifuss, F.E., Porter, R., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stackman J., Sutula, T.P., Graves, N., Welty, T., Vickery, T., Bundage, R., Gates, J., Gummit, R. and Gutierres, A. (1987). A controlled study of progabide in partial seizure: Methodology and results. *Neurology*, **37**, 963–968.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, **50**, 325–335.

Leyland, A.H. and Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. Chichester, UK: Wiley.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K.-Y. and Zeger, S.L. (1995). Inference based on estimating functions in the presence of nuisance parameters (with discussion). *Statistical Science*, **10**, 158–199.

Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, **84**, 1074–1078.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1017.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Lindsey, J.K. (1999). *Models for Repeated Measurements*, 2nd ed. Oxford, UK: Clarendon Press.

Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association*, **83**, 1014–1022.

Lipsitz, S.R. and Fitzmaurice, G.M. (1994). Sample size for repeated measures studies with binary responses. *Statistics in Medicine*, **13**, 1233–1239.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

Littell, R.C., Pendergast, J. and Natarajan, R. (2000). Tutorial in biostatistics: Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793–1819.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

Little, R.J.A. and Rubin, D.B. (2001). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.

Liu, G. and Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics*, **53**, 937–947.

Longford, N. (1993). *Random Coefficient Models*. Oxford, UK: Oxford University Press.

Lord, F. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, **68**, 304–305.

Louis, T.A., Robins, J., Dockery, D.W., Spiro, A. and Ware, J.H. (1986). Explaining discrepancies between longitudinal and cross-sectional models. *Journal of Chronic Diseases*, **39**, 831–839.

Machin, D., Farley, T., Busca, B., Campbell, M. and d'Arcangues, C. (1988). Assessing changes in vaginal bleeding patterns in contracepting women. *Contraception*, **38**, 165–179.

Mancl, L.A. and DeRouen, T.A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, **57**, 126–134.

Matthews, J.N.S. (1994). Multi-period crossover trials. *Statistical Methods in Medical Research*, **3**, 383–405.

Matthews, J.N.S., Altman, D.G., Campbell, M.J. and Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, **300**, 230–235.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall/CRC Press.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics*, **5**, 746–762.

Monette, G., Kwan, E., Rivilis, A. and Shao, Q. (2002). A first look at multilevel models. Unpublished manuscript. Department of Mathematics and Statistics, York University.

Morrison, D.F. (1972). The analysis of a single sample of repeated measurements. *Biometrics*, **28**, 55–71.

Morrison, D.F. (1990). *Multivariate Statistical Methods*, 3rd ed. New York: McGraw-Hill.

Myers, R.H., Montgomery, D.C. and Vining, G.G. (2001). *Generalized Linear Models: With Applications in Engineering and the Sciences*. New York: Wiley.

National Center for Health Statistics (1992). Sample design: Third National Health and Nutrition Examination Survey. *Vital and Health Statistics*, Series 2, No. 113.

National Center for Health Statistics (1994). Plan and operation of the third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics*, Series 1, No. 32.

Naumova, E.N., Must, A. and Laird, N.M. (2001). Evaluating the impact of "critical periods" in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology*, **30**, 1332–1341.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Regression Models*, 3rd ed. Homewood, IL: Richard D. Irwin.

Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 25–35.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R.F. and McFadden, D. (editors). *Handbook of Econometrics Vol 4*. Amsterdam: North-Holland.

Omar, R.Z., Wright, E.M., Turner, R.M. and Thompson, S.G. (1999). Analyzing repeated measures data: A practical comparison of methods. *Statistics in Medicine*. **18**, 1587–1603.

Pagano, M. and Gauvreau, K. (2000). *Principles of Biostatistics*, 2nd ed. Pacific Grove, CA: Duxbury Press.

Pan, W. (2001). Sample size and power calculations with correlated binary data. *Controlled Clinical Trials*, **22**, 211–227.

Pan, W., Louis, T.A. and Connett, J.E. (2000). A note on marginal linear regression with correlated response data. *American Statistician*, **54**, 191–195.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pepe, M.S. and Anderson, G.A. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Simulation and Computation*, **23**, 939–951.

Phillips, S.M., Bandini, L.G., Compton, D.V., Naumova, E.N. and Must, A. (2003). A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls. *Journal of Nutrition*, **133**, 1419–1425.

Pierce, D.A. and Sands, B.R. (1975). Extra-Bernoulli variation in binary data. Technical Report 46, Department of Statistics, Oregon State University.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Potthoff, R.F. and Roy, S.W. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterization. *Biometrika*, **86**, 677–690.

Price, C.J., Kimmel, C.A., Tyl, R.W. and Marr, M.C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicological Applications in Pharmacology*, **81**, 113–127.

Qaqish, B.F. and Liang, K.-Y. (1992). Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics*, **48**, 939–950.

Rao, C.R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1–17.

Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 447–458.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Newbury Park, CA: Sage Publications.

Raudenbush, S.W. and Liu, X.F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, **6**, 387–401.

Rencher, A.C. (2002). *Methods of Multivariate Analysis*, 2nd ed. New York: Wiley.

Rijcken, B., Schouten, J.P., Weiss, S.T., Speizer, F.E. and van der Lende, R. (1987). The relationship of nonspecific bronchial responsiveness to respiratory symptoms in a random population sample. *American Review of Respiratory Disease*, **136**, 62–68.

Robins, J.M., Greenland, S. and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**, 687–712.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6**, 15–51.

Rogan, W.J., Dietrich, K.N., Ware, J.H., Dockery, D.W., Salganik, M., Radcliffe, J., Jones, R.L., Ragan, N.B., Chisolm, J.J. and Rhoads, G.G. (2001). The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *New England Journal of Medicine*, **344**, 1421–1426.

Rosenbaum, P.R. and Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, **45**, 212–218.

Rosenman, R.H, Brand, R.J., Jenkins, C.D., Friedman, M., Straus, R. and Wurm, M. (1975). Coronary heart disease in the Western Collaborative Study: Final follow-up experience of $8\frac{1}{2}$ years. *Journal of the American Medical Association*, **233**, 872–877.

Rowell, J.G. and Walters, D.E. (1976). Analysing data with repeated observations on each experimental unit. *Journal of Agricultural Science*, **87**, 423–432.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Russell, T.S. and Bradley, R.A. (1958). One-way variances in a two-way classification. *Biometrika*, **45**, 111–129.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110–114.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall/CRC Press.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 40, 719–727.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Schiesselman, J.J. (1973a). Planning a longitudinal study: I. Sample size determination. *Journal of Chronic Diseases*, **26**, 553–560.

Schlesselman, J.J. (1973b). Planning a longitudinal study: II. Frequency of measurement and study duration. *Journal of Chronic Diseases*, **26**, 561–570.

Schoenfield, L.J., Lachin, J.M., Baum, R.A., Habig, R.L., Hanson, R.F., Hersh, T., Hightower, N.C., Hofmann, A.F., Lasser E.C., Marks, J.W., Mekhjian, H., Okun, R., Schaefer, R.A., Shaw, L., Soloway, R.D., Thistle, J.L., Thomas, F.B., Tyor, M.P. (1981). National Cooperative Gallstone Study: A controlled trial of the efficacy and safety of chenodeoxycholic acid for dissolution of gallstones. *Annals of Internal Medicine*, **95**, 257–282.

Schwartz, D. and Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, **20**, 637–648.

Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Selvin, S. (1995). *Practical Biostatistical Methods*, 1st ed. Belmont, CA: Duxbury Press.

Senn, S. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. New York: Wiley.

Silvapulle, M.J. (1996). A test in the presence of nuisance parameters. *Journal of the American Statistical Association*, **91**, 1690–1693.

Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, **90**, 342–349.

Singer, J.D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, **25**, 323–355.

Singer, J.D. and Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Snedecor, G.W. and Cochran, W.G. (1967). *Statistical Methods*, 6th ed. Ames, IA: Iowa State Press.

Snijders, T.A.B. and Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit, Cambridge, UK.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 197–206. Berkeley, CA: University of California Press.

Stiratelli, R., Laird, N.M. and Ware, J.H. (1984). Random effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.

Stokes, M.E., Davis, C.S. and Koch, G.G. (1995). *Categorical Data Analysis using the SAS System*. Cary, NC: SAS Institute, Inc.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

Stram, D.O. and Lee, J.W. (1995). Correction to "Variance components testing in the longitudinal mixed effects model". *Biometrics*, **51**, 1196.

Stukel, T.A. (1993). Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine*, **12**, 1339–1351.

Subramanian, S.V., Jones, K. and Duncan, C. (2003). Multilevel methods for public health research. In Kawachi, I. and Berkman, L.F. (editors), *Neighborhoods and Health*. New York: Oxford University Press.

Sullivan, L.M., Dukes, K.A. and Losina, E. (1999). Tutorial in biostatistics: An introduction to hierarchical linear modeling. *Statistics in Medicine*, **18**, 855–888.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.

Thompson, W.A. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, **33**, 273–289.

Treatment of Lead-Exposed Children (TLC) Trial Group (2000). Safety and efficacy of succimer in toddlers with blood leads of 20-44 μg/dL. *Pediatric Research*, **48**, 593–599.

van der Lende, R., Kok, T.J., Peset, R., Quanjer, P.H., Schouten, J.P. and Orie, N.G.M. (1981). Decreases in VC and FEV_1 with time: Indicators for effects of smoking and air pollution. *Bulletin of European Physiopathology and Respiration*, **17**, 775–792.

Van Marter, L.J., Leviton, A., Kuban, K.C.K., Pagano, M. and Allred, E.N. (1990). Maternal glucocorticoid therapy and reduced risk of bronchopulmonary dysplasia. *Pediatrics*, **86**, 331–336.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.

Ware, J.H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician*, **39**, 95–101.

Ware, J.H. (2003). Interpreting incomplete data in studies of diet and weight loss. *New England Journal of Medicine*, **348**, 2136–2137.

Ware, J.H., Dockery, D., Louis, T.A., Xu, X., Ferris, B.G. and Speizer, F.E. (1990). Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *American Journal of Epidemiology*, **132**, 685–700.

Ware, J.H. and Liang, K.-Y. (1996). The design and analysis of longitudinal studies: A historical perspective. In Armitage, P. and David, H.A. (editors), *Advances in Biometry*. New York: Wiley.

Waternaux, C., Laird, N.M. and Ware, J.H. (1989). Methods for analysis of longitudinal data: Blood-lead concentration and cognitive development. *Journal of the American Statistical Association*, **84**, 33–41.

Waternaux, C. and Ware, J.H. (1991). Unconditional linear models for analysis of longitudinal data. In Dwyer, J.H., Feinleib, M., Lippert, P. and Hoffmeister, H. (editors), *Statistical Models for Longitudinal Studies of Health*. New York: Oxford University Press.

Wei, L.J. and Lachin, J.M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, **79**, 653–661.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–1073.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144–148.

Winer, B.J. (1971). *Statistical Principles in Experimental Design*, 2nd ed. New York: McGraw-Hill.

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, **30**, 16–28.

Wong, G.Y. and Mason, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, **80**, 513–524.

Woolson, R.F. and Clarke, W.R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society, Series A*, **147**, 87–99.

Yates, F. (1935). Complex experiments (with discussion). *Supplement to the Journal of the Royal Statistical Society*, **2**, 181–247.

Zahner, G.E.P., Jacobs, J.H., Freeman, D.H. and Trainor, K.F. (1993). Rural-urban child psychopathology in a northeastern U.S. state: 1986–1989. *Journal of the American Academy of Child and Adolescent Psychiatry*, **32**, 378–387.

Zahner, G.E.P., Pawelkiewicz, W., DeFrancesco, J.J. and Adnopoz, J. (1992). Children s mental health service needs and utilization patterns in an urban community: An epidemiological assessment. *Journal of the American Academy of Child and Adolescent Psychiatry*, **31**, 951–960.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zhao, L.P., Prentice, R. and Self, S. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B*, **54**, 805–811.

Zimmerman, D.L. and Nunez-Anton, V. (2001). Parametric modelling of growth curve data: An overview (with discussion). *Test*, **10**, 1–73.

Index

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Third Edition*
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Statistical Modelling
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods

*Now available in a lower priced paperback edition in the Wiley Classics Library.

*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
*DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

*Now available in a lower priced paperback edition in the Wiley Classics Library.

HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HELLER · MACSYMA for Statisticians
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KASS and VOS · Geometrical Foundations of Asymptotic Inference

*Now available in a lower priced paperback edition in the Wiley Classics Library.

KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
McFADDEN · Management of Data in Clinical Trials

*Now available in a lower priced paperback edition in the Wiley Classics Library.