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. During these "golden years" of NIH expansion, the entire NIH budget grew from $8 million in 1947 to more than $1 billion in 1966. 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 modeled 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.

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware.

Although 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 30 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 lon-

gitudinal 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, leveling off approximately 4 years after menarche. Suppose that 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 each individual alone, 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 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. Although the analysis of multilevel data is not the primary focus of this book, multilevel data are discussed in Chapter 22.

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, Stata, SPSS, R, 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 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 U.S. 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 to 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 to 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.

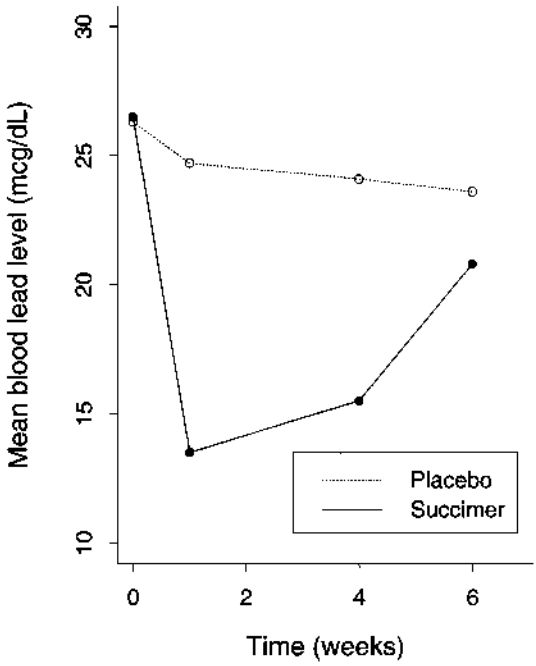

Fig. 1.1 Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

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 that 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				
None 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. 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)

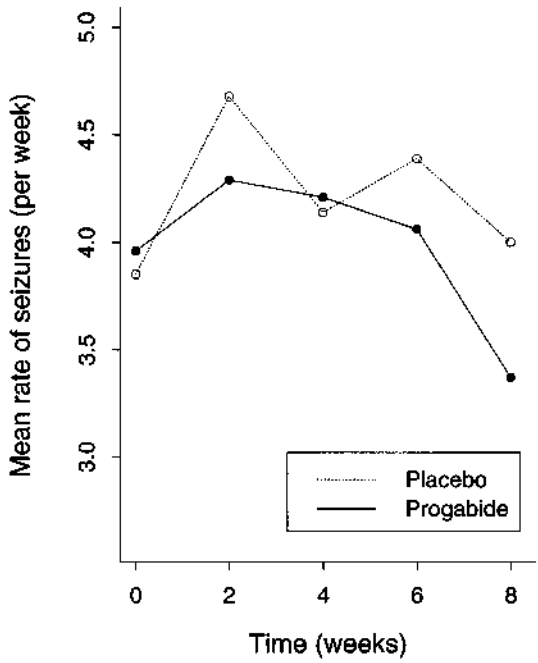

Fig. 1.2 Mean rate of seizures (per week) at baseline, week 2, week 4, week 6, and week 8 in the progabide and placebo groups.

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, 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 and 1987 in New Haven, Connecticut, a predominantly minority metropolitan center. The second survey, the Eastern Connecticut Child Survey (ECCS) was conducted in 1988 and 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

In the last 30 years we have seen remarkable advances in methods for analyzing longitudinal and clustered data. In particular, we now have a broad and flexible class of models for correlated data based on a regression paradigm. Indeed, all 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

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

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

and

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

are all cases where the mean is linear in the regression parameters (where $E(Y|X)$ denotes the *conditional* mean or expectation of Y given X). 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

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

and

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

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 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 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 modeling 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 that 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 that

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 that 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 THE BOOK

The book is organized into five 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 through 10, 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 modeling 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 modeling: the mean response over time and the covariance among repeated measures on the same individuals. In Chapters 5 and 6, the emphasis is on modeling the mean response. Two main approaches are distinguished: the analysis of response profiles (Chapter 5) and parametric or semiparametric curves (Chapter 6). In Chapter 7, we discuss models for the covariance in longitudinal data and develop an overall modeling 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 data 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 modeled as a combination of fixed effects that are assumed to be shared by all individuals, and random effects that are unique or specific to a particular individual. In Chapter 9, we discuss an alternative, but closely related, class of regression models for longitudinal data known as linear "fixed effects" models. These models treat the subject-specific effects as fixed rather than random. We review the main features of linear fixed effects models for longitudinal data and discuss their potential advantages and disadvantages relative to linear mixed effects models. In Chapter 10, 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 11 through 16, 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 11, 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 12 through 15, 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 16, we compare and contrast these two classes of models. One of the underlying themes emphasized in Chapters 12 through 16 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 fourth part of the book, consisting of Chapters 17 and 18, addresses the issue of missing data in longitudinal studies. In Chapter 17, we review the assumptions about missing data required to ensure that the methods discussed in earlier chapters provide valid inferences. Two methods for handling missing data, multiple imputation and inverse probability weighted methods, are discussed in detail in Chapter 18.

The final part of the book, consisting of Chapters 19 through 22, focuses on a number of advanced topics. In Chapter 19, we discuss smoothing methods for longitudinal analysis that allow greater flexibility for the form of the relationship between the mean response and the covariates. This chapter focuses on the connection between penalized splines and linear mixed effects models. Chapter 20 considers the design of a longitudinal study, focusing on the determination of sample size and power. In Chapter 21, 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 22, we present an overview of methods for analyzing multilevel data. Chapters 21 and 22 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) and Gelman and Hill (2007); a more advanced presentation of similar topics can be found in Neter et al. (1996).