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.

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

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. Al-

though 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 on 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) 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 on 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 and challenging problem in longitudinal studies. Indeed, missing data are the rule, not the exception, in longitudinal studies in the health sciences. For example, study participants do not always appear for a scheduled observation, or they may 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 Chapters 17 and 18 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 17.

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 uppercase 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 lowercase 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 on 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.

	Occasion				
Individual	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. We refer to μ_{ij} as the *conditional* mean response at the j^{th} occasion, where the term *conditional* is used to denote the dependence of the mean on covariates. 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); in Chapter 3 we introduce additional notation that makes the dependence of the mean on the covariates more transparent. A simple illustration of a model for the mean response that depends on time, and that allows

changes in the mean response to also depend on covariates, is presented in Section 2.4. In Chapters 5 and 6 we present a detailed discussion of two broad approaches for modeling 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. 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. 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 conditional expectation or mean of Y_{ij} by

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

then the conditional variance of Y_{ij} is defined as

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

While μ_{ij} provides a measure of the location of the center of the distribution of Y_{ij}, the conditional variance provides a measure of the spread or dispersion of the values of Y_{ij} around their conditional mean. The positive square-root of the conditional variance, σ_j, is known as the conditional *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 conditional *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}, given the covariates. 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 (given the covariates). 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 conditional correlation between Y_{ij} and Y_{ik} is denoted by

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

where σ_j and σ_k are the conditional 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 scatterplot (assuming the absence of covariates), 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. 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 conditional variances and covariances:

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

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

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 said to be *symmetric* in the sense that $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = \text{Corr}(Y_{ik}, Y_{ij})$. 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 to 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

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

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 modeling 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 changes in blood lead levels over time could be expressed as

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

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 will 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 semiparametric 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 scatterplot of each pair of repeated measures. Figure 2.1 displays scatterplots 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

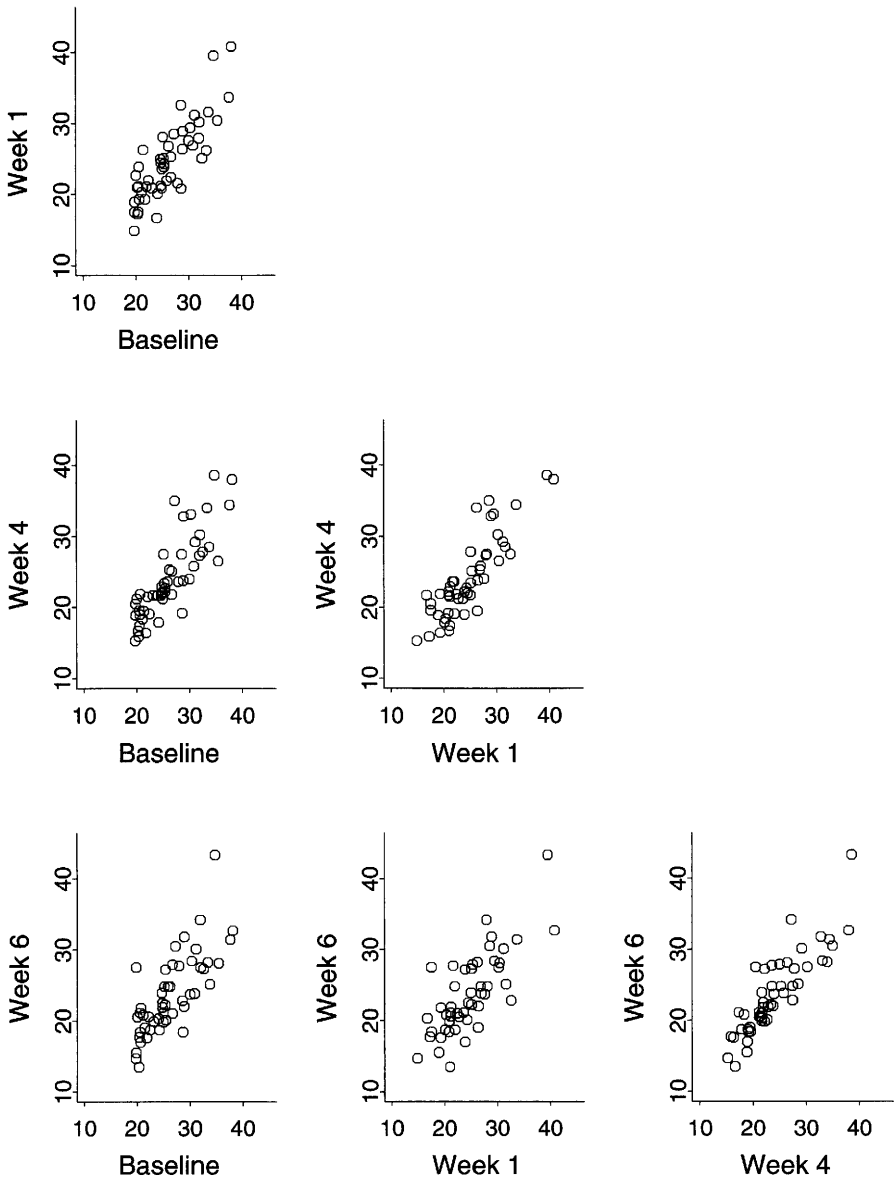

Fig. 2.1 Pairwise scatterplots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

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 scatterplots 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 scatterplot that plots the responses on the vertical axis and the times of measurements on the horizontal 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 scatterplot for only a few randomly selected individuals; it can be very difficult to discern the two distinct sources of variability in a scatterplot based on all of the individuals in the study. In Figure 2.2, based on four 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 segments

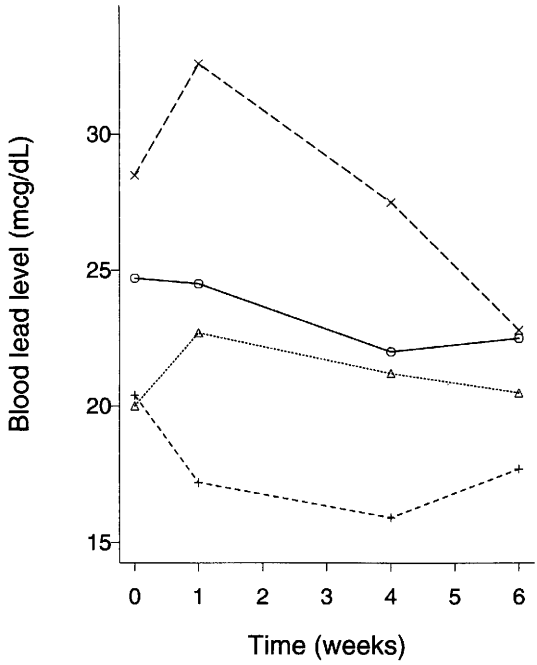

Fig. 2.2 Time plot of blood lead levels at baseline, week 1, week 4, and week 6 for four randomly selected children from the placebo group of the TLC trial.

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: (1) the correlations are positive, (2) the correlations often decrease with increasing time separation, (3) the correlations between repeated measures rarely ever approach zero, even in cases where they are taken many years apart, and (4) 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: (1) between-individual heterogeneity, (2) within-individual biological variation, and (3) 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, and heart rate.

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 that 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 on 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 among 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 that 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 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, because of 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.

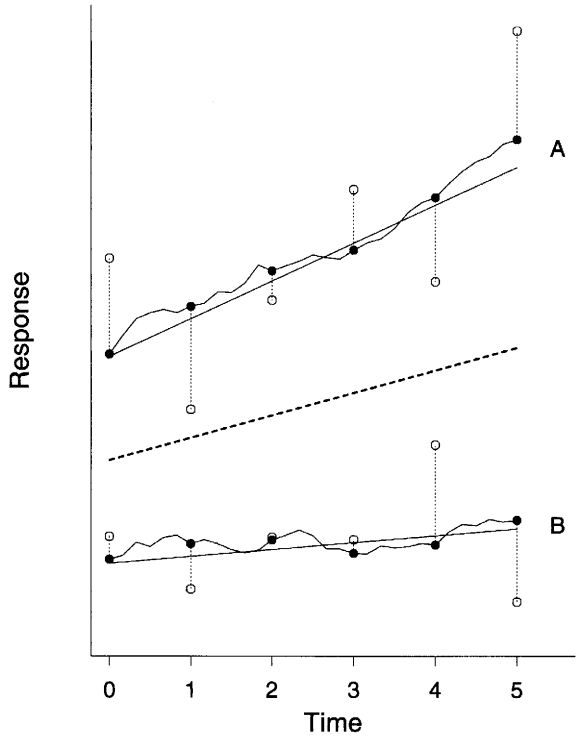

Fig. 2.3 Graphical representation of the three sources of variability in longitudinal data for two hypothetical individuals: • denotes repeated measure free of measurement error, ○ denotes observed repeated measure with measurement error.

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

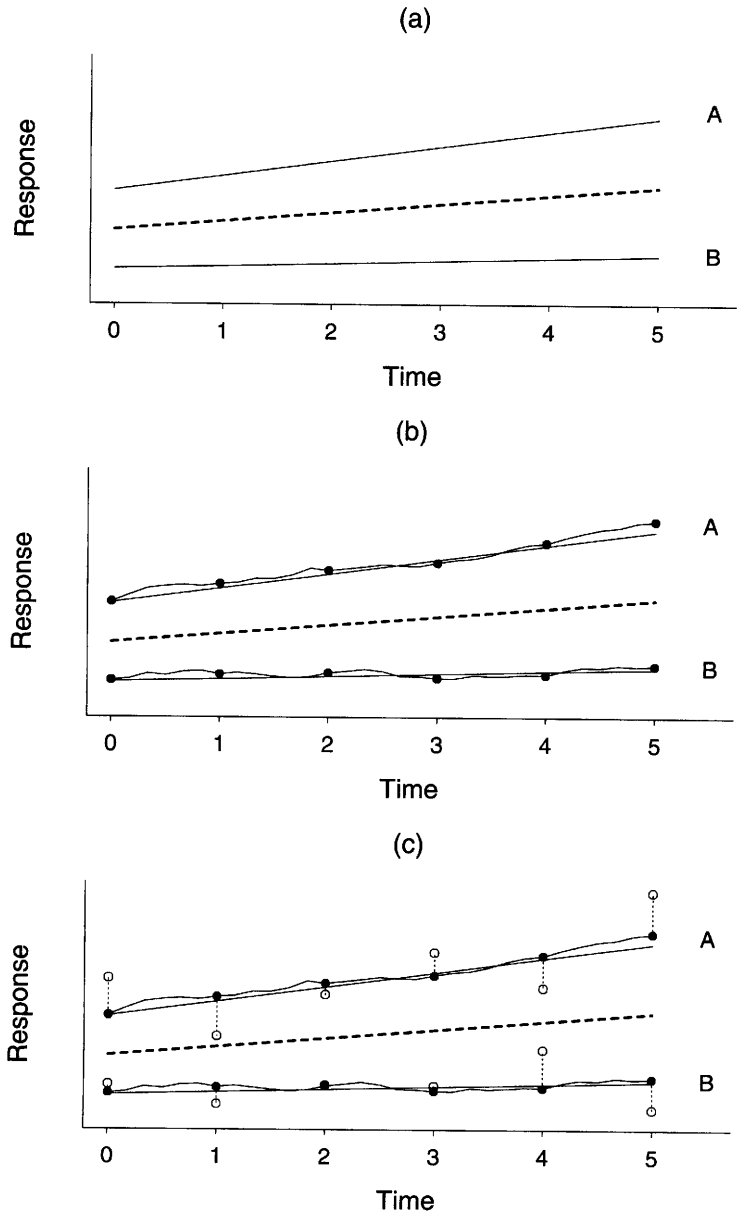

Fig. 2.4 Graphical representation of the cumulative impact of three sources of variability in longitudinal data: (a) between-individual heterogeneity, (b) within-individual biological variation (where • denotes repeated measure free of measurement error), and (c) measurement error (where ○ denotes observed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

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. 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 conditional variance and covariance are measures defined in terms of a particular model for the conditional 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 that 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 (see Table 1.2), 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 variances ($\widehat{\sigma}_1^2 = 25.2$ and $\widehat{\sigma}_2^2 = 58.9$) and covariance ($\widehat{\sigma}_{12} = 15.5$) for the succimer group 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, which can lead to quite misleading inferences. This topic will be discussed at much greater length in later chapters.

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 to 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.