Part II

Linear Models for Longitudinal Continuous Data

3

Overview of Linear Models for Longitudinal Data

3.1 INTRODUCTION

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

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

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

discuss statistical methods for constructing confidence intervals, testing hypotheses, and assessing the adequacy of models. In Chapter 13, where we discuss alternative methods of estimation, it will become more apparent that we do not require the assumption of multivariate normality.

3.2 NOTATION AND DISTRIBUTIONAL ASSUMPTIONS

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

Notation

In Chapter 2 we assumed that a sample of N subjects are measured repeatedly over time. We let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. As was mentioned in Chapter 2, either by design or happenstance, subjects may not have the same number of repeated measures and may not be measured at the same set of occasions. To accommodate both of these features, we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. For example, a study may be designed to take repeated measurements on all subjects at the same set of n occasions. However, missing data are a common problem in almost all longitudinal studies, and some subjects may not have observations at all n occasions (i.e., n_i denotes the number of *observed* responses on the i^{th} subject, where $n_i \leq n$). Missing data not only produce a varying number of repeated measurements of subjects in a longitudinal study but also have important consequences for the validity of any method of analysis. In Section 4.3 we outline some of the key issues and assumptions required for valid analyses when there are missing data; this topic is discussed in greater detail in Chapter 17. In addition to missing data, there may be mistimed measurements, in the sense that measurements are not obtained at the planned n occasions; instead, they are obtained some time before or after the intended measurement occasions. Thus both the number and the timing of the repeated measurements may not be common for all subjects. In later chapters the times of measurement, t_{ij}, are used to model trends in the mean response; they may also be required to appropriately account for the covariance among the repeated measurements.

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

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

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

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

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

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

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

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

baseline, current smoking status, and environmental exposures. In the former case, the same values of the covariates are replicated in the corresponding rows of X_{ij} for $j = 1, ..., n_i$. In the latter case, the values taken by the covariates can vary over time (for at least some individuals) and the values in the corresponding rows of X_{ij} can be different at each measurement occasion. The inclusion of time-varying covariates whose values at any occasion cannot be predicted (e.g., current smoking status) can raise subtle issues concerning the interpretation and estimation of the resulting models. A discussion of these issues is deferred until Chapter 13 (see Section 13.5).

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

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

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

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

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

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

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

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

where $\beta_1, ..., \beta_p$ are unknown regression coefficients relating the mean of Y_{ij} to its corresponding covariates. This regression model describes how the responses at

every occasion are related to the covariates. That is, there are n_i separate regression equations for the response variable at each of the n_i occasions

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

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

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

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

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

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

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

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

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

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

Illustration: Treatment of Lead-Exposed Children Trial

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

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

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

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

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

where β_1 represents the mean blood lead level at baseline (week = 0) and β_2 has interpretation as the change in mean blood level (in $\mu g/dL$) per week. For example, the expected change in mean blood level, from baseline to 6 weeks, is $\beta_2 \times 6$ for children in the placebo group. Similarly for children in the succimer group

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

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

hypothesis that the two treatments are equally effective in reducing blood lead levels corresponds to a hypothesis that $\beta_3 = 0$.

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

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

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

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

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

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

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

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

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

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

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

where

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

for children in the placebo group, and

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

for children in the succimer group.

Distributional Assumptions

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

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

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

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

refer to the distributions of Y_i and e_i and, more specifically, the covariance matrix of Y_i and e_i. Note that in discussing the distribution of Y_i, it should be understood that we are always referring to the *conditional* distribution of Y_i given the covariates, X_i.

Next Y_i, the vector of continuous responses, is assumed to have a conditional distribution that is multivariate normal, with mean response vector

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

and covariance matrix

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

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

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

Multivariate Normal Distribution

So far we have discussed the multivariate normal distribution in a very general way, noting how it can be seen as a very natural, multivariate extension of the univariate normal distribution. Next we present a more detailed description of the multivariate normal distribution, since it forms the basis for a general method of estimation that will be described in Chapter 4. However, we remind the reader that the statistical methods presented in later chapters use the assumption that the longitudinal responses have an approximate multivariate normal distribution to derive estimates and statistical

tests but do not require it when data are complete (i.e., no missing data), or when there is missingness but the observed data can be regarded as a random sample of the complete data.

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

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

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

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

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

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

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

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

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

In the context of a longitudinal study, with n_i repeated measures on the i^{th} individual, we have a vector of responses and need to consider their *joint* probability distribution. While a univariate probability density function describes the probability or relative frequency of occurrence of particular values of a single random variable, a joint probability density function describes the probability or relative frequency with

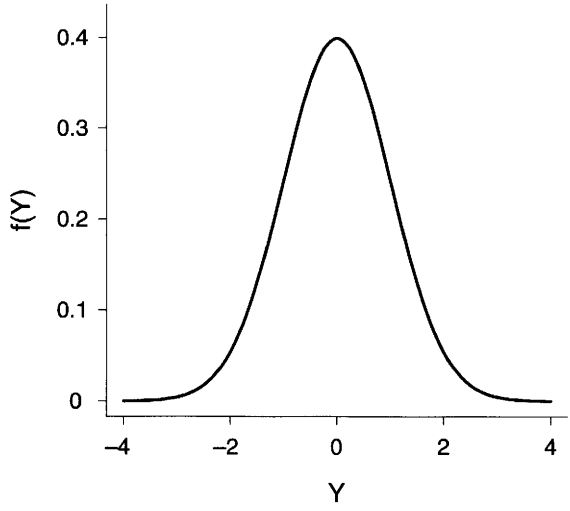

Fig. 3.1 Plot of univariate normal density function with zero mean and unit variance.

which the vector of responses take on a particular set of values. The multivariate normal distribution is a natural extension of the univariate normal distribution for a single response to a vector of responses. The multivariate normal joint probability density function for Y_i given X_i can be expressed as

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

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

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

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

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

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

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

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

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

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

where

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

Although this is a more complex expression, since it accounts for the correlation between Y_{i1} and Y_{i2}, it does nonetheless provide a single measure of distance that (1) adjusts for differences in the variances of Y_{i1} and Y_{i2}, by effectively down-weighting deviations from the mean when the variance is larger, and (2) adjusts for the magnitude of the correlation (or overlapping information) between Y_{i1} and Y_{i2}. When there is no correlation between Y_{i1} and Y_{i2} (and $\rho_{12} = 0$), the distance of y_i from μ_i is simply the sum of the component standardized distances. On the other hand, when there is strong positive correlation, the distance of y_i from μ_i also includes a component that factors in whether y_{i1} and y_{i2} are *both* larger (or smaller) than μ_{i1} and μ_{i2}, respectively. The latter adjustment is made because part of the standardized distance of y_{i2} from μ_{i2} is predictable from Y_{i2}'s correlation with Y_{i1}, and vice versa.

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

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

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

Perhaps the most useful assessment of the validity of the assumption of multivariate normality is through the use of graphical displays. For example, histograms and

box and whisker plots of the residuals at each occasion can be used to detect gross departures of e_{ij} from univariate normality (we discuss residual diagnostics in greater detail in Chapter 10). These simple graphical displays can also be used to determine an appropriate transformation of the response variable so that the marginal distributions of the e_{ij}'s more closely approximate normal distributions. However, a caveat of the use of this technique is that although a multivariate normal distribution for e_i implies that each of the separate e_{ij} has a univariate normal distribution, univariate normal distributions for the e_{ij} do not necessarily imply that e_i has a multivariate normal distribution. Therefore the assumption of multivariate normality cannot be formally verified by examination of each of the component variables separately. However, gross departures from univariate normality can be taken to indicate that the multivariate normal assumption is not tenable.

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

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

In summary, in longitudinal studies the repeated measurements on the same individual are inherently dependent or correlated. This lack of independence can be accounted for by considering the multivariate distribution of the entire vector of repeated measurements (given the covariates). Note that while the repeated measurements are correlated, we implicitly assume that the vectors of observations from different individuals are independent of one another. In Part II of this book we are primarily concerned with longitudinal data that are continuous, and we make the assumption that their joint distribution is multivariate normal for the purpose of deriving estimates and statistical tests. However, the methods that are discussed do not require the assumption that the responses have a multivariate normal distribution. In practice, longitudinal data are not anticipated to have a joint distribution that is *exactly* multivariate normal. The multivariate normal distribution is adopted as an approximation, but one that has many convenient statistical properties. In Chapter 4 we present a method for estimating β and Σ_i, and for making inferences about β and Σ_i, which is derived from the multivariate normal assumption for the longitudinal responses. In later chapters we discuss how the assumption that Y_i has a multivariate normal distribution can be avoided altogether.

One final comment on notation and terminology. For the remainder of the book, for simplicity of notation, we often replace $E(Y_i|X_i)$ in many equations by $E(Y_i)$ when it should be clear from the context that it denotes the *conditional* mean of the responses given the covariates. Likewise we often replace $\text{Cov}(Y_i|X_i)$ by $\text{Cov}(Y_i)$ when it should also be clear from the context that it denotes the *conditional* covariance of the responses given the covariates. In a similar vein, in discussing the distribution of Y_i, it should be understood that we are always referring to the *conditional* distribution of Y_i given the covariates.

3.3 SIMPLE DESCRIPTIVE METHODS OF ANALYSIS

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

Also note that the time plot does not indicate which data points represent repeated measurements on the same individual. To circumvent the latter problem, the time plot can be supplemented by joining or connecting successive repeated measures on

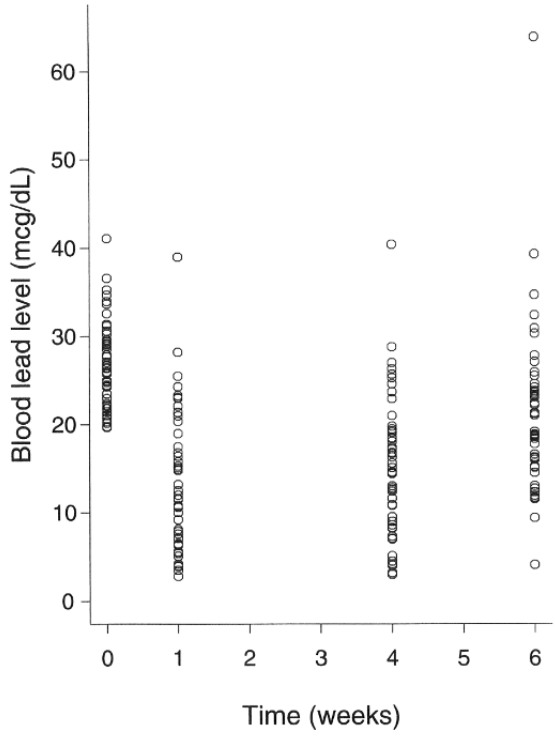

Fig. 3.2 Time plot of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

the same individual with straight lines. However, the resulting line segments do not necessarily enhance the time plot; more often than not, the result is a "spaghetti" plot that is not very informative about trends in the mean response over time. Perhaps the only useful source of information provided by the simple time plot concerns the presence of extreme outliers in the data and whether the variability in the data changes discernibly with time.

Some of the problems with the time plot of longitudinal data can be illustrated using data from the *Treatment of Lead-Exposed Children Trial*. Figure 3.2 displays a time plot of the blood lead level data for the group treated with succimer. Recall that in this study repeated measurements were taken at the same set of occasions for all subjects in the trial. Because of the resulting overlap of data points at the four measurement occasions, it is difficult to discern any pattern in the mean response trend over time. As noted earlier, the most extreme case of this problem arises when the response variable is binary; then it is impossible to discern any information about time trends from the resulting time plot due to the completely overlapping data points.

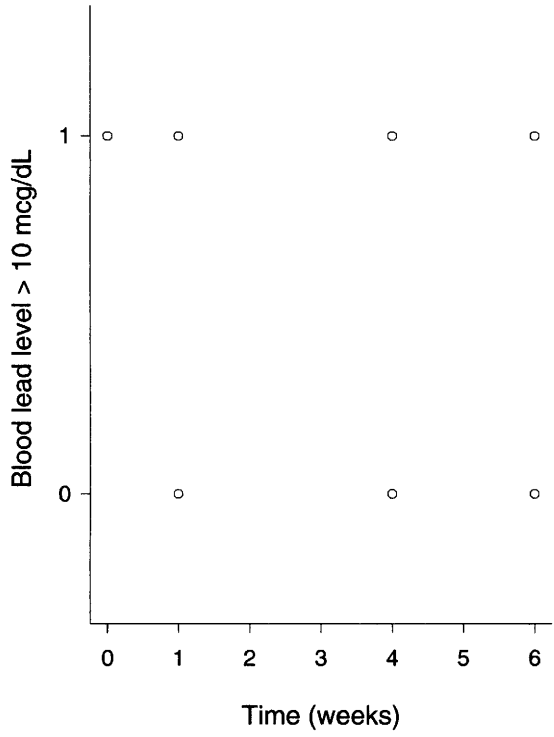

Fig. 3.3 Time plot of blood lead levels > 10 mcg/dL at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

Figure 3.3 displays a time plot of the repeated binary response, indicating whether each child has a blood lead level below 10 $\mu g/dL$. (The U.S. Centers for Disease Control defines 10 $\mu g/dL$ as the threshold for concern about exposure to lead.) Here, the binary response $Y_{ij} = 1$ if the i^{th} child's blood lead level is above 10 $\mu g/dL$ at the j^{th} occasion, and $Y_{ij} = 0$ otherwise. Due to the overlapping 0's and 1's, the time plot provides no information about the trend in the mean response (or probability that a blood lead level is above 10 $\mu g/dL$) over time.

In Figure 3.4 the time plot of blood lead levels is supplemented with line segments joining successive measures on the same individual. However, Figure 3.4 is only a little more informative about trends in the mean response over time than Figure 3.2. Although, in principle, Figure 3.4 distinguishes two sources of variability in the data, between-subject variability and within-subject variability, in practice, it is difficult to assess their relative magnitude from the time plot. However, Figure 3.4 does reveal an observation at week 6 that is a potential outlier, given the previous measurements

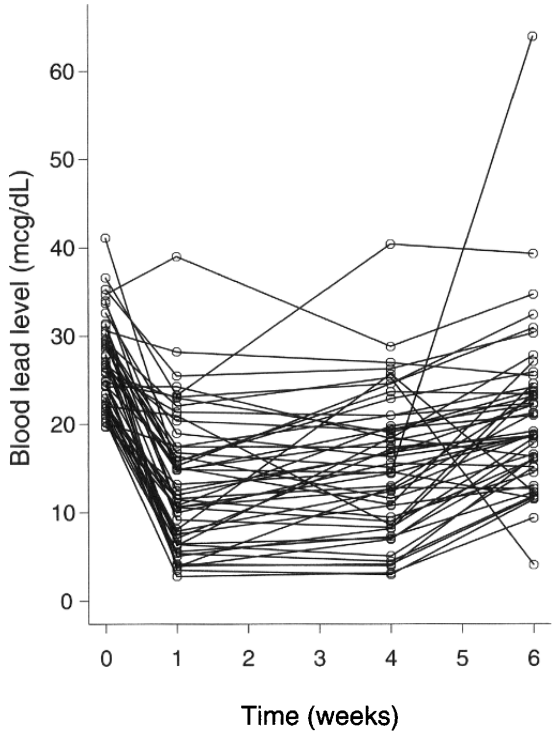

Fig. 3.4 Time plot, with joined line segments, of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial.

of blood lead levels for that child. In summary, time plots of the raw data, with or without joined line segments for successive repeated measurements on the same individual, can reveal important features of the data. However, time plots of the raw data are not always the most informative displays of longitudinal data, especially when the data are balanced over time. With time plots of balanced longitudinal data it can be difficult to discern the "signal" (i.e., the trend in the mean response over time) from the "noise" in the data and the between-subject and within-subject sources of variability are often almost completely obscured. With highly unbalanced data, time plots of the raw data, with joined line segments, are easier to interpret.

In general, it is usually more informative to display a time plot of the average or mean response, with successive points on the graph joined by straight lines. In addition time plots of the mean response for different levels of discrete covariates (e.g., different treatment or exposure groups) can be overlayed on the same graph. The construction of these plots is relatively straightforward when the timing of the

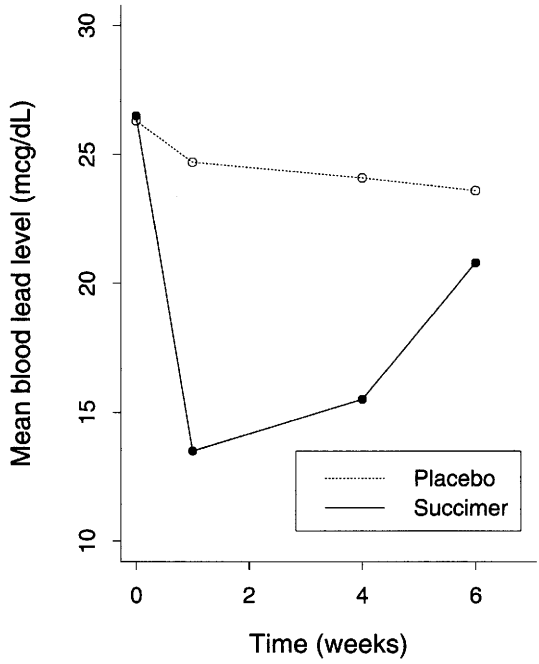

Fig. 3.5 Time plot, with joined line segments, of the mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups.

repeated measures is the same for all individuals. The time plots can also be enhanced by including standard error bars for the mean response at each occasion. For example, Figure 3.5 displays the mean blood lead levels in the succimer and placebo groups at weeks 0, 1, 4, and 6. From this simple display it is readily apparent that the effect of succimer is greater after one week of treatment and that there appears to be a rebound effect thereafter.

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

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

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

Ideally at least two or three reference categories for a quantitative covariate should be chosen, and in such a way that investigators in the field can readily appreciate the substantive importance of going from one level to the other. For example, the change in going from "low" to "medium" to "high," or vice versa, should have some subject-matter meaning. For some quantitative covariates, there may be natural choices for the reference levels (e.g., corresponding to intervals that represent "normal" and "abnormal" ranges). For other quantitative covariates, especially those that are less well established or unfamiliar to investigators in the field, there may not be an obvious choice for the reference categories. In the latter case the choice can be made on the basis of the data at hand. For example, one possible choice is to group the covariate at the 25th and 75th percentiles. This will produce "low" (or lowest quartile), "medium" (2nd or 3rd quartiles), and "high" (or highest quartile) reference categories. It must be acknowledged, though, that the number and choices of reference groups is, to some extent, arbitrary; reference groups that are more or less extreme than those suggested here could equally be chosen (e.g., tertiles or quintiles).

So far our discussion has assumed that many, if not all, individuals are measured at the same set of occasions. When the times of measurement are not the same for all individuals, construction of time plots of the mean response can pose difficulties due to sparseness of data at any particular occasion. For example, Figure 3.6 displays a time plot of longitudinal data on lung function growth in children and adolescents from the Six Cities Study of Air Pollution and Health. The data are from a cohort of 300 school-age girls living in Topeka, Kansas, who, in most cases, were enrolled in the first or second grade (between the ages of six and seven). The girls were measured annually until high school graduation (approximately at age eighteen) or loss to follow-up, and each girl provided a minimum of one and a maximum of 12 observations. At each examination, pulmonary function measurements were obtained from simple spirometry. The basic maneuver in simple spirometry is maximal inspiration followed by forced exhalation as rapidly as possible into a closed chamber. A widely used measure computed from simple spirometry is the volume of air exhaled in the first second of the maneuver, FEV_1. Figure 3.6 displays a time plot of $\log(FEV_1/\text{height})$ versus age for the 300 girls. Although children were measured approximately annually, the data are highly unbalanced when age, rather than chronological time, is used as the metameter for lung function growth. Figure 3.7 displays a time plot, with joined line segments, of $\log(FEV_1/\text{height})$ versus age for 50 randomly selected girls. Because each girl is not measured at the same age, construction of plots of the mean response versus age can pose difficulties due to sparseness of data at any particular age.

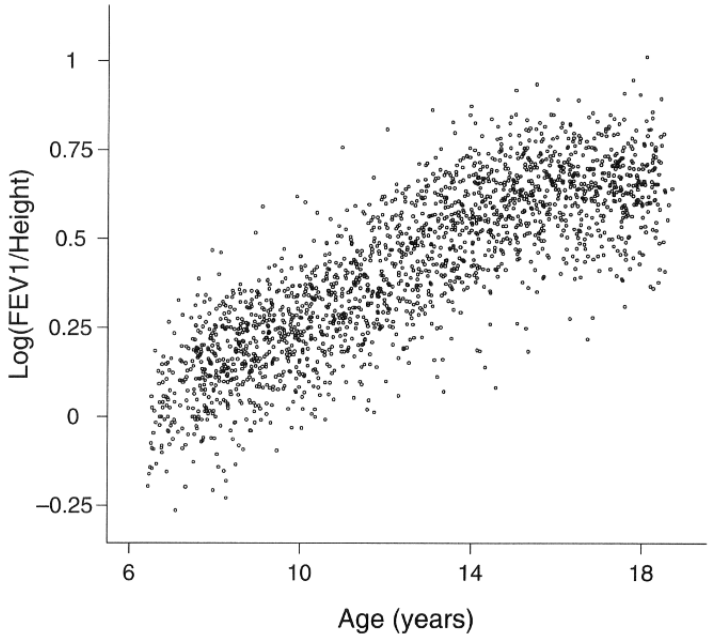

Fig. 3.6 Time plot of log(FEV_1/height) versus age in years for girls from Topeka.

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

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

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

where k is some positive integer (e.g., $k = 1$ or $k = 2$) and we refer to $2k+1$ as being the *order* of the moving average. This expression for the moving average assumes that all N individuals are measured at the same set of occasions. With highly unbalanced

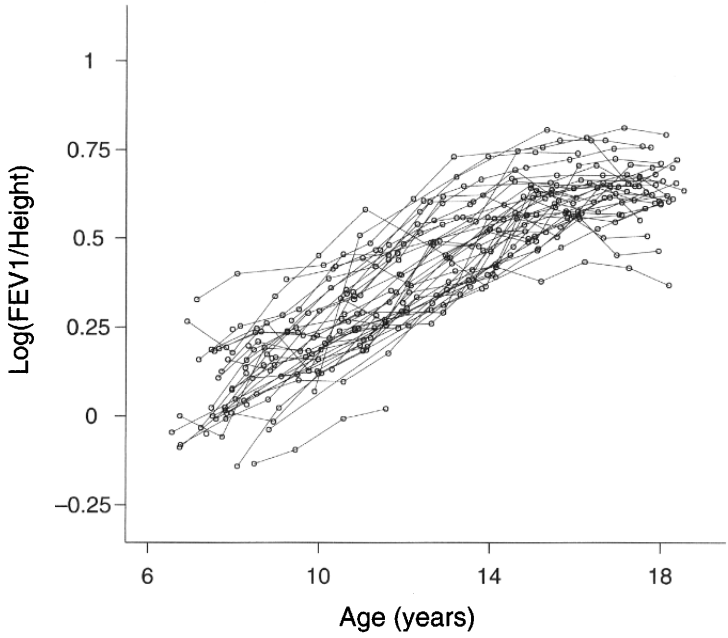

Fig. 3.7 Time plot, with joined line segments, of log(FEV_1/height) versus age in years for 50 randomly selected girls from Topeka.

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

A simple example of a "moving average" is

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

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

Moving averages are best suited to smoothing observations that are approximately equally separated in time. They are not ideal for handling completely irregularly spaced observations. When longitudinal data are irregularly spaced and unbalanced over time, other nonparametric regression methods can be used to estimate the mean response trend over time. One popular method available in most standard statistical software packages is locally weighted regression or *lowess*. The basic idea behind most of the nonparametric regression methods is very similar. They attempt to trace the salient features of the mean response as a function of time while making only minimal assumptions about the form of the relationship. For example, the lowess estimate at time t is best understood by imagining that there is a "window" centered at time t. One estimate of the mean at time t is obtained by taking some weighted average of all the observations that fall within the window. The lowess estimate, however, is not based on a simple weighted average of the observations within the window. Instead, it is determined by fitting a straight line to the data within the window using a robust regression technique that gives more weight to observations close to the center of the window and that also down-weights potential outliers. The lowess estimate of the mean at time t is simply the predicted value at time t from the fitted regression line. The entire lowess curve is obtained by moving a window of fixed width from the first measurement occasion to the last, and repeating the process at every time. Figure 3.8 displays a lowess curve for the lung function data described earlier. Unlike the time plot of the raw data in Figure 3.6, the lowess curve is informative about changes in lung function as the children grow older. The smooth curve produced by the lowess procedure indicates ages where lung function appears to develop more rapidly.

All smoothing techniques require that a smoothing parameter, often referred to as the *bandwidth* parameter, be specified. This parameter controls the amount of smoothing. For example, the width of the window in the lowess procedure determines how jagged or smooth the resulting plot appears; the wider the window, the smoother the resulting curve will be. The choice of smoothing or bandwidth parameter involves the classical trade-off between bias and precision. Excessive smoothing decreases the variance of the estimate of the mean trend but at the risk of introducing bias. Insufficient smoothing is unlikely to introduce bias but will result is a quite variable estimate of the mean response trend. All smoothing techniques must compromise in some way, and the goal is to find an appropriate trade-off between these two competing forces: increased bias versus decreased variance of the estimated mean response trend over time.

Finally, we note that standard applications of nonparametric smoothing techniques to longitudinal data ignore the correlation among repeated measures on the same individual. On the whole, the correlation among repeated measures is not likely to grossly

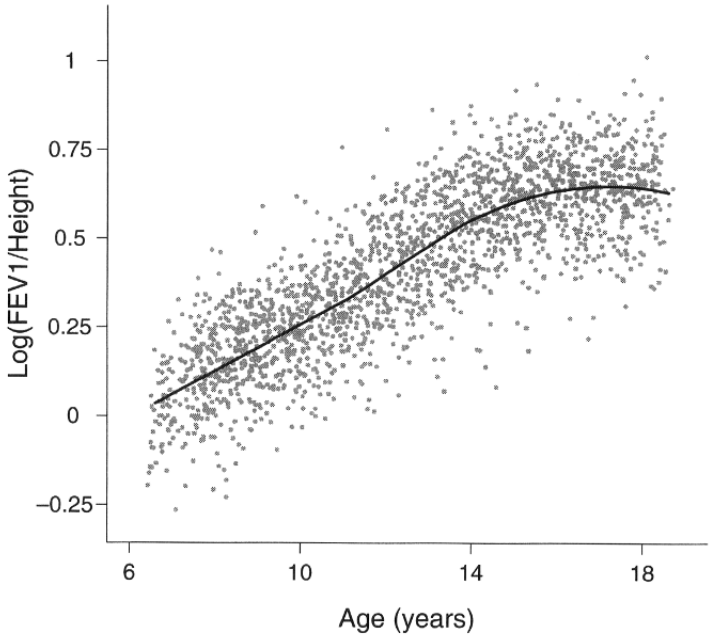

Fig. 3.8 Time plot of log(FEV_1/height) versus age in years, with *lowess* smoothed curve superimposed, for girls from Topeka.

distort the estimated mean response trend when standard smoothing techniques are applied to longitudinal data. Correlation is likely to have a much greater impact on the construction of confidence bands for the smoothed curve. As a result we caution the reader that the confidence bands produced by standard statistical software for lowess, and other nonparametric smoothing techniques, are likely to be optimistically biased. That is, confidence bands constructed under the assumption that the correlation among repeated measures is zero will be too narrow and could potentially lead to misleading inferences. In summary, the routine application of nonparametric smoothing methods to longitudinal data can be useful for exposing trends in the mean response over time, with the caveat that confidence bands produced by standard statistical software packages should be ignored. A final note concerns attrition over time. If attrition or dropout occurs in a substantial number ($> 5\%$) of subjects, then the end of the estimated mean curve can be distorted if subjects who leave the study differ from those who remain; the impact of dropout, and of missingness more generally, is discussed in greater detail in Section 4.3 and Chapter 17.

3.4 MODELING THE MEAN

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

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

In Section 2.4 we mentioned that regression models for longitudinal data can usually be formulated to encapsulate the main research questions of interest in terms of a set of regression parameters. That is, certain regression parameters will have interpretations that bear directly on the scientific question or questions of interest. For example, in a regression model for the longitudinal blood lead level data from the TLC trial, treatment group interaction effects have direct interpretation in terms of how the underlying rate of change in mean blood lead levels differs between the two treatment groups. Before discussing approaches for modeling the mean response over time, it is important to clarify the distinction between *substantive* and *nuisance* parameters in the context of a longitudinal study.

Substantive and Nuisance Parameters for Longitudinal Data

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

By making this distinction between substantive and nuisance parameters, the covariance among longitudinal responses is, in a certain sense, regarded as a secondary aspect of the data (relative to the mean response over time). However, we must emphasize that this distinction does not imply that the covariance can be disregarded or simply ignored. Indeed, the covariance among repeated measures must be properly acknowledged to assure an appropriate method of analysis. The distinction between

substantive and nuisance parameters has some important ramifications for the types of statistical methods that are adopted. For example, there will be a high premium attached to methods that yield valid estimates of the substantive parameters across a broad range of different assumptions about the nuisance parameters. In the context of longitudinal data, this implies that there will be a premium attached to methods that yield unbiased estimates of change in the mean response over time under a broad range of assumptions about the structure of the covariance among longitudinal responses.

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

Modeling the Mean Response over Time

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

A second approach is to assume a parametric curve (e.g., linear or quadratic trend) for the mean response over time. This approach can dramatically reduce the number of model parameters. By their very nature, parametric curves provide a very parsimonious description of trends in the mean response over time, and of covariate effects on the mean response over time. For example, a linear trend in the mean response can be characterized by a single regression parameter that has interpretation in terms

of the constant rate of change in the mean response over time. In addition parametric curves describe the mean response as an explicit function of time. As a result, and in contrast to profile analysis, there is no necessity to require that all individuals in the study have the same set of measurement times, nor even the same number of repeated measurements.

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

3.5 MODELING THE COVARIANCE

The defining feature of longitudinal data is that repeated responses are obtained on the same individuals over time and the resulting responses on the same individual are correlated. Although the correlation, or more generally, the covariance among the repeated responses, is not usually of intrinsic interest, it cannot simply be ignored. Instead, the covariance among repeated measures is an important aspect of the data that must be properly accounted for to yield valid inferences about the regression parameters of primary interest. Accounting for the correlation among repeated measures completes the specification of any regression model for longitudinal data and usually increases efficiency or the precision with which the regression parameters can be estimated. When an appropriate model for the covariance has been adopted, correct standard errors are obtained and valid inferences about the regression parameters can be made. In addition, when there are missing data, correct modeling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In Chapter 13 we also consider a method for analyzing longitudinal data that ignores the correlation among the repeated measures, for the purposes of estimation of the regression parameters, but makes an appropriate adjustment to the standard errors for the purposes of inference.

Three broad approaches to modeling the covariance among repeated measures can be distinguished: (1) unstructured covariance, (2) covariance pattern models, and (3) random effects covariance structures. The first is to allow any arbitrary pattern of covariance among the repeated measures. This results in what is ordinarily referred to as an "unstructured" covariance. That is, no explicit structure is assumed for the covariance among the repeated measures (other than the homogeneity of covariance across different individuals). Thus, when there are n repeated measures, the n variances at each occasion and the $n \times (n-1)/2$ pairwise covariances (or correlations) are estimated. Historically the unstructured covariance matrix has been the model of choice for the covariance in the analysis of response profiles. That is, the analysis of response profiles assumes arbitrary patterns for both the mean response

over time (and their relation to covariates) and for the variances and covariances. This approach to modeling the covariance, however, is not limited to the analysis of response profiles and could equally be adopted when the mean response is modeled with parametric or semiparametric curves. There are two potential drawbacks with this approach. The first is that the number of covariance parameters can be quite large. If there are n measurement occasions, the $n \times n$ covariance matrix has $n \times (n+1)/2$ unique parameters. Thus, in a longitudinal study with 10 measurement occasions, an unstructured covariance has 55 parameters (10 variances and 45 covariances). When the number of covariance parameters to be estimated is large relative to the sample size, then estimates are likely to be unstable. The second drawback of this approach is that it is only applicable when all individuals are measured at the same set of occasions. That is, it cannot accommodate mistimed measurements or, more generally, irregularly timed measurements.

Alternative approaches to modeling the covariance place structure on the covariance matrix. There are two main strategies. The first approach borrows ideas from the statistical literature on time series analysis. Time series data, in contrast to longitudinal data, arise from studies with a small number of replications or individuals (often only a single replication) and a large number of repeated measures. That is, in times series data N, the number of replications is small (often $N = 1$) relative to the number of repeated measures, n. With longitudinal data, it is the reverse situation, with N being large relative to the number of repeated measures, n. Thus time series data consist of a small number of long sequences of repeated measurements, whereas longitudinal data consist of a large number of relatively short sequences of repeated measurements. Although times series data and longitudinal data are dissimilar in structure, and the data analytic goals are usually different, they do share one common feature: the repeated measures are correlated. Most of the models for the covariance in the time series literature incorporate at least one important aspect of longitudinal data: repeated measures taken closer together in time are expected to be more highly correlated than repeated measures further apart in time. This implies that the correlations decay as the time separation increases. Quite often the correlation among repeated measures is expressed as an explicit function of the time separation. In the latter case these models can be used with unequally spaced observations. In addition many of the models for the variance assume *stationarity*, namely that the variance does not change as a function of time. Much of the statistical literature on the analysis of time series data has focused on parametric models that can adequately describe the covariance structure among the repeated measures with only a few parameters. These parsimonious models for the covariance can also be adopted for longitudinal data and are discussed in Chapter 7.

An alternative, and somewhat indirect, strategy for imposing structure on the covariance is through the introduction of *random effects*. Historically simple random effects models were one of the earliest approaches for analyzing repeated measures data. In the so-called univariate repeated measures ANOVA model, the correlation among repeated measurements is accounted for by the inclusion of a single individual-specific random effect. This effect can be thought of as a randomly varying intercept, representing an aggregation of all the unobserved or unmeasured factors that make

some individuals "high responders" and some individuals "low responders." The consequence of adding a single individual-specific random effect to every measurement on any given individual is that the resulting repeated measurements will be positively correlated. Thus the inclusion of random effects imposes structure on the covariance.

The univariate repeated measures ANOVA model has a very long history and has enjoyed widespread use in many fields of application; this model is discussed in greater detail in Section 3.6. Although the introduction of a single individual-specific random effect induces correlation among repeated measures, a feature of the model is that the resulting positive correlation is constant, and does not vary as a function of the time between any pair of repeated measurements. In addition the variance is constant over time. These constraints on the covariance structure are somewhat unappealing for longitudinal data. However, this problem can be easily remedied by the inclusion of more than one random effect. That is, the constraints on the covariance induced by the repeated measures ANOVA model can be relaxed by assuming that a subset of the regression parameters (e.g., intercepts and slopes) vary randomly across individuals. If the inclusion of a single individual-specific random effect induces positive correlation among repeated measures, albeit with a somewhat unappealing structure on the correlations, it should not come as a surprise that the inclusion of additional randomly varying coefficients induces patterns of correlation among the repeated measures that are somewhat less restrictive. In addition these models permit the variance to change over time in a smooth fashion. Indeed, random effects models provide both very flexible and parsimonious models for the covariance and are particularly well suited to handling longitudinal data that are irregularly timed. These models are discussed at length in Chapter 8.

3.6 HISTORICAL APPROACHES

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

From a historical perspective, three methods for the analysis of repeated measures data can be distinguished: (1) univariate repeated measures analysis of variance (ANOVA), (2) multivariate repeated measures analysis of variance (MANOVA), and (3) methods based on summary measures. All three of these approaches have had varying degrees of popularity, and some are still in widespread use, in different areas of application. Many of these approaches are unnecessarily restrictive in their assumptions and their analytic goals. For example, ANOVA and MANOVA focus

on comparing groups in terms of their mean response trend over time but provide little information about how individuals change over time. Also, as we will see later, ANOVA and MANOVA have numerous features that limit their usefulness for the analysis of longitudinal data. In contrast, the regression models that are discussed throughout the remainder of this book make more realistic assumptions and can address the major scientific questions of interest in a longitudinal study. For all the reasons that were outlined in Section 1.4, we view the regression paradigm as being the most useful, general, and versatile approach for analyzing longitudinal data arising from the health sciences.

Repeated Measures Analysis by ANOVA

One of the earliest proposals for analyzing correlated responses was the repeated measures analysis of variance (ANOVA), sometimes referred to as the "univariate" or "mixed-model" analysis of variance. The analysis of variance paradigm was developed in the early part of the twentieth century by R. A. Fisher. Although many of the early applications of ANOVA were to designed experiments in agriculture, since then it has found widespread application in many other disciplines. In the repeated measures ANOVA model, the correlation among repeated measurements is assumed to arise from the additive contribution of an individual-specific random effect to each measurement on any given individual. Thus the model assumes the correlation between repeated measurements arises because each subject has an underlying (or latent) level of response that persists over time and influences all repeated measurements on that subject. This individual-specific effect is regarded as a random variable.

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

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij},$$

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

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

$$E\,(Y_{ij}|X_{ij}) = \mu_{ij} = X'_{ij}\beta.$$

Thus, in the repeated measures ANOVA model, the response for the i^{th} individual is assumed to differ from the population mean, μ_{ij}, by an individual-specific random effect, b_i, that persists throughout all measurement occasions, and a within-subject

measurement error, ϵ_{ij}. That is, the repeated measures ANOVA model distinguishes two main sources of variation in the data: between-subject variation, σ_b^2, and within-subject variation, σ_ϵ^2. The between-subject variation acknowledges the simple fact that subjects respond differently; some are "high" responders, some are "low" responders, and some are "medium" responders. The within-subject variation acknowledges that there are random fluctuations that arise from the process of measurement, for example, due to measurement error and/or sampling variability.

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

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

The derivation of the variances and covariances is not important; a more detailed account will be given in Chapter 8 (see Section 8.1). What is important to note is that the variances at every occasion are equal, $(\sigma_b^2 + \sigma_\epsilon^2)$, as are the covariances, σ_b^2. Consequently, the correlation among any pair of repeated measures,

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2},$$

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

This particular covariance structure is also known as *compound symmetry* and has a randomization justification in certain repeated measures designs (e.g., split-plot designs). In an experiment where the within-subject factor is randomly allocated to subjects, randomization arguments can be made to show that the constant variance and constant correlation conditions hold (see Chapter 21 for a more detailed discussion of the randomization argument). Historically this provided an attractive justification for using the repeated measures analysis by ANOVA in randomized experiments. The randomization argument is simply not justifiable in the longitudinal data setting; measurement occasions cannot be randomly allocated to subjects. As a result the compound symmetry assumption for the covariance is often inappropriate for longitudinal data. That is, the constraint on the correlation among repeated measurements is somewhat unappealing for longitudinal data, where the correlations are expected to decay with increasing separation in time. Also the assumption of constant variance across time is often unrealistic. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. Finally, as originally conceived, the repeated measures ANOVA model was developed for the analysis of data from designed experiments, where the repeated measures are obtained at a set of occasions common to all individuals, the covariates are discrete factors (e.g., treatment group and time), and the data

are complete. Thus the repeated measures ANOVA could not be readily applied to longitudinal data that were irregularly spaced, incomplete, or when it was of interest to include quantitative covariates in the analysis.

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

Repeated Measures Analysis by MANOVA

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

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

Since MANOVA was originally developed for the analysis of a multivariate vector of response variables, it is worth emphasizing some of the distinctions between longitudinal responses and more general cases of multivariate responses. Recall that longitudinal data give rise to a vector of responses. Thus the responses in a longitudinal study are inherently multivariate. On the other hand, the multivariate responses arising from a longitudinal study are commensurate, being repeated measures over time of the same response variable. With longitudinal data, the repeated measures represent selected observations of the main features of some underlying continuous process that is potentially changing over time. This is in contrast to having a single measure of multiple, but substantively different or distinct, response variables (e.g., blood pressure, blood glucose, and LDL cholesterol levels). With more general multivariate data where we have single measurements of multiple, but distinct, response

variables, there is no notion of an underlying continuum. Finally, with longitudinal data, the covariance among the repeated measures can be expected to have certain features or patterns; with more general multivariate data, there is rarely any indication of structure to the covariance matrix.

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

The repeated measures analysis by MANOVA is a special case of a more general approach known as *profile analysis*. The analysis of response profiles will be discussed in much greater detail in Chapter 5. Here we describe the basic idea underlying the repeated measures analysis by MANOVA, but without much technical detail. In Chapter 5 we will illustrate how the repeated measures analysis by MANOVA relates to profile analysis and highlight the potential limitations of this approach for analyzing longitudinal data.

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

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

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

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

A simple example will help to motivate the main ideas. Suppose that in a longitudinal clinical trial, designed to compare a new treatment to placebo, repeated measures

of the response variable are obtained on three occasions ($n = 3$). A repeated measures analysis by MANOVA proceeds by constructing three derived variables, say V_{i1}, V_{i2}, and V_{i3}. The first derived variable is simply the sum (or average) of the responses. That is, for each individual we can construct

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

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

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

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

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

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

to the original vector of responses. That is,

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

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

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

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

Given the set of derived variables, the repeated measures analysis by MANOVA proceeds as follows. First, the two derived variables representing contrasts of time, V_{i2} and V_{i3}, are analyzed using MANOVA. For example, the first question can be addressed by comparing the groups in terms of these two derived variables. Specifically, in our simple example with two groups, this is achieved by using a multivariate extension of the two-sample t-test, which is known as Hotelling's T^2 test. A test of no differences between groups on these two derived variables is equivalent to a test of the "group $\times$ time interaction." Next, and assuming that there is no "group $\times$ time interaction," the second question can be addressed. The second question is concerned with the shape of the overall (i.e., averaged over groups) trend in the mean response over time. If the mean response trend over time is flat, then the two derived variables have expectation zero. As a result the second question can be addressed by using a multivariate extension of the single sample t-test, the single sample Hotelling's T^2 test, of the hypothesis that V_{i2} and V_{i3} have mean zero. Finally, the third question can be addressed by focusing on the first derived variable, V_{i1}. This variable is proportional to the mean of the repeated measures and can be used to assess whether there is a "group" effect. Specifically, group difference in the overall mean response, averaged over measurement occasions (or time), can be examined using a simple two-sample t-test, or, in the case of more than two groups, using ANOVA. A standard ANOVA of the first derived variable (the sum or average of the repeated measures) is performed and provides a test of the group or between-subject factor. Note that there is nothing intrinsically multivariate in this last part of the analysis since the analysis is based on a single derived variable. It is the first part of the analysis that is intrinsically multivariate.

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

The repeated measures analysis by MANOVA has a number of features that make it unappealing for the analysis of longitudinal data. In particular, the MANOVA

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

Summary Measure Analysis

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

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

When the covariates are discrete, and the repeated measures for each individual are obtained at the same set of occasions, the AUC analysis can also be based on the results of an analysis of response profiles (that assumes arbitrary patterns for the mean responses over time). Given the covariates, the AUC for the mean response over time is the same as the average (or mean) of the individual-specific AUCs. That is, for the case of linear models for continuous longitudinal responses, the AUC for

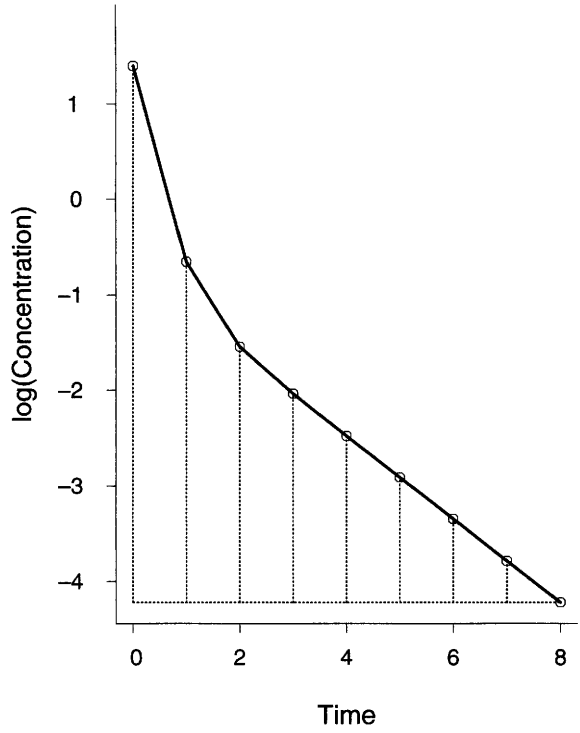

Fig. 3.9 Time plot of log(concentration) versus time, illustrating how the area under the curve (AUC) can by calculated using the trapezoidal rule.

the mean response over time coincides with the mean of the AUCs for the individuals in the population of interest. In a limited context, reducing the sequence of repeated measures for each individual to an AUC can provide a useful basis for the analysis of longitudinal data. However, the analysis of AUCs is problematic with unbalanced longitudinal data.

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

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

where b_{1i} and b_{2i} are regression parameters specific to the i^{th} individual and the errors, ϵ_{ij}, are implicitly assumed to be independent within any individual. Estimates of the individual-specific slopes (and intercepts) can then be obtained from a linear regression line fit to each individual's repeated measures. The resulting slopes can

then be related to the covariates using standard parametric or nonparametric methods for the analysis of a univariate response. This approach does not require that the repeated measures for each individual be obtained at the same set of occasions. Finally, we note that extensions of this particular summary measure analysis approach lead naturally to a class of models referred to as "growth curve models". Studies of growth and aging are classic examples of observational longitudinal studies. In these studies the goal is to describe naturally occurring changes in the response over time, due to developmental or aging processes, and to compare these growth curve profiles in different groups (e.g., males and females). To meet this goal, growth curve models have been developed to summarize the pattern of response over time and allow for the possibility that individuals may belong to or be drawn from different groups. Growth curve models can be motivated in terms of a two-stage model. Indeed, growth curve models are sometimes referred to as "two-stage" growth models. At the first stage, it is assumed that a parametric curve (e.g., linear or quadratic trend in time) fits the observed responses for each subject. In the second stage, these individual-specific growth parameters are then related to covariates that describe the different groups from which the individuals have been drawn. Growth curve models will be discussed in greater detail in Chapter 8.

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

In those cases where the summary measures are well defined when individuals have missing data or different numbers of repeated measures, the analysis becomes more complicated because the derived summary measures no longer have the same variance. Similarly, if the repeated measures are taken at irregular times for different

individuals, the resulting summary measures may also have different variances. In all these cases the variance of the derived summary measures is not constant, violating a fundamental assumption made by many standard statistical methods for univariate responses. Thus, in general, the standard parametric methods for the analysis of a univariate response (e.g., *t*-tests, ANOVA, linear regression) cannot be validly applied to the summary measures when the design is unbalanced over time due to missing data, different numbers of repeated measures, or sequences of repeated measures taken at irregular times for different individuals.

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

3.7 FURTHER READING

Winer (1971) provides a very accessible discussion of the application of repeated measures analysis by ANOVA; also see McCulloch (2005) for a comparison of repeated measures ANOVA with more modern methods of analyses. A comprehensive description of repeated measures analysis by MANOVA, targeted at applied researchers, can be found in the book by Hand and Taylor (1987). Finally, for a non-technical discussion of the analysis of summary measures, readers are referred to the review articles by Matthews et al. (1990) and Everitt (1995).

Bibliographic Notes

An excellent discussion of scatterplot smoothing techniques can be found in Chapter 3 of Ruppert et al. (2003). Ware and Liang (1996) provide an interesting historical perspective on the development of statistical methods for the analysis of longitudinal data, with emphasis on the contributions that have been made in the biostatistical literature; also, see Chapter 1 of Fitzmaurice et al. (2009).

The foundations for the repeated measures analysis of variance can be found in the seminal monograph by Fisher (1925) and in the method for analyzing split-plot experiments proposed by Yates (1935); also see Scheffé (1959). Greenhouse and Geisser (1959) described an adjustment to the repeated measures analysis by ANOVA when the required assumption about the covariance matrix (compound symmetry)

does not hold. The repeated measures analysis by MANOVA was introduced in the statistical literature by Box (1950); also see Danford et al. (1960), Geisser (1963), Cole and Grizzle (1966), and Morrison (1972). A discussion of repeated measures analyses by ANOVA and MANOVA, and the relationship between the two methods, can be found in Chapters 2, 3, and 11 of Hand and Crowder (1996).

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