8

Linear Mixed Effects Models

8.1 INTRODUCTION

In Chapters 5 and 6 we introduced models for longitudinal data where changes in the mean response, and their relation to covariates, can be expressed as

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

and where the primary goal is to make inferences about the population regression parameters, β. In Chapter 7 we described how the specification of this regression model for longitudinal data can be completed by making additional assumptions about the structure of $\text{Cov}(Y_i|X_i) = \text{Cov}(e_i) = \Sigma_i$. In this chapter we consider an alternative, but closely related, approach for analyzing longitudinal data using linear mixed effects models. The underlying premise of linear mixed effects models is that some subset of the regression parameters vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population. That is, individuals in the population are assumed to have their own subject-specific mean response trajectories over time and a subset of the regression parameters are now regarded as being random. The distinctive feature of linear mixed effects models is that the mean response is modeled as a combination of population characteristics, β, that are assumed to be shared by all individuals, and subject-specific effects that are unique to a particular individual. The former are referred to as *fixed effects*, while the latter are referred to as *random effects*. The term *mixed* is used in this context to denote that the model contains both fixed and random effects.

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

Although the linear mixed effects model assumes that the longitudinal responses depend on a combination of population (or fixed effects) parameters, β, and subject-specific effects, it nonetheless leads to a model for the marginal mean response (averaged over the distribution of the random effects) that can be expressed in the familiar form

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

However, the introduction of random effects induces covariance among the responses and $\text{Cov}(Y_i|X_i) = \Sigma_i$ has a distinctive random effects structure. With the inclusion of random effects, the covariances among the repeated measures can be expressed as functions of time. Unlike the covariance pattern models considered in Chapter 7, which do not distinguish the different sources of variability that have an impact on the covariance, linear mixed effects models explicitly distinguish between-subject and within-subject sources of variability. Moreover the induced random effects covariance structure can often be described with relatively few parameters, regardless of the number and timing of the measurement occasions.

Because linear mixed effects models explicitly distinguish between fixed and random effects, they allow the analysis of between-subject and within-subject sources of variation in the longitudinal responses. In addition it is not only possible to estimate parameters that describe how the mean response changes in the population of interest, it is also possible to predict how individual response trajectories change over time. For example, linear mixed effects models can be used to obtain predictions of individual growth trajectories over time. The latter will be of interest when the focus of inference is on the individual rather than the population of individuals. For example, in the physician–patient context, these predictions can be used to identify those patients who do not respond well to their assigned treatment in a clinical trial.

One very appealing aspect of linear mixed effects models is their flexibility in accommodating any degree of imbalance in longitudinal data, coupled with their ability to account for the covariance among the repeated measures in a relatively parsimonious way. That is, with linear mixed effects models we do not require the same number of observations on each subject nor that the measurements be taken at the same set of measurement occasions. As a result these models are particularly well suited for analyzing inherently unbalanced longitudinal data. While the regression models for the mean response described in Chapter 6 can also handle unbalanced longitudinal data, the class of covariance pattern models suitable for unbalanced data is very limited.

Example: Random Intercept Model

Recall that in earlier chapters we encountered the simplest possible case of a linear mixed effects model: the linear model with a randomly varying subject effect. In this model each subject is assumed to have an underlying level of response that persists over time. This subject effect is incorporated in the linear mixed effects model by regarding it as random, yielding the following model

$$Y_{ij} = X'_{ij}\beta + b_i + \epsilon_{ij}, \tag{8.1}$$

where b_i is the random subject effect and the ϵ_{ij} are regarded as measurement or sampling errors. Let us examine this simple model more closely. In this model, the response for the i^{th} subject at the j^{th} occasion is assumed to differ from the population mean, $X'_{ij}\beta$, by a subject effect, b_i, and a within-subject measurement error, ϵ_{ij}. Both the subject effect and the measurement error are assumed to be random, with mean zero, and with variances, Var $(b_i) = \sigma_b^2$ and Var $(\epsilon_{ij}) = \sigma^2$, respectively. In addition it is assumed that b_i and ϵ_{ij} are independent of one another. Note that this model describes the mean response trajectory over time for any individual,

$$E(Y_{ij}|b_i) = X'_{ij}\beta + b_i,$$

in addition to the mean response profile in the population,

$$E(Y_{ij}) = X'_{ij}\beta,$$

where the averaging is over all individuals in the population. We refer to the former as the *conditional* mean of Y_{ij}, given the subject-specific effect, and the latter as the *marginal* mean of Y_{ij} (averaged over the distribution of the subject-specific effects, b_i). There is potential for confusion in our use of this terminology, however, since in both cases the mean response is conditional also upon the covariates, X_{ij}; for notational convenience, we have suppressed the dependence on covariates. The alert reader will also have noticed a small change in notation from the previous chapters. The measurement or sampling errors in (8.1) are denoted by ϵ_{ij} (epsilon) not e_{ij}. This change in notation is intentional and reflects differences in interpretations of ϵ_{ij} and e_{ij}. In previous chapters, the error e_{ij} represents the deviation of Y_{ij} from the mean response in the population, $X'_{ij}\beta$. In this chapter, the *within-subject* error ϵ_{ij} represents the deviation of Y_{ij} from the subject-specific mean response, $X'_{ij}\beta + b_i$. Put another way, the random errors, e_{ij}, in previous chapters have now been decomposed into two random components, $e_{ij} = b_i + \epsilon_{ij}$, a between-subject component and a within-subject component.

Next consider the interpretation of the parameters in the model given by (8.1). The regression parameters β describe patterns of change in the mean response over time (and their relation to covariates) in the population of interest, while b_i describes how the trend over time for the i^{th} individual deviates from the population average. That is, b_i represents an individual's deviation from the population mean intercept, after the effects of the covariates have been accounted for. Thus, when combined with the fixed effects, b_i describes the mean response trajectory over time for any individual. This interpretation is often obscured by the use of vector and matrix notation, but is apparent if we express the model given by (8.1) as

$$\begin{aligned} Y_{ij} &= X'_{ij}\beta + b_i + \epsilon_{ij} \\ &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij} \\ &= \beta_1 + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij} \\ &= (\beta_1 + b_i) + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \epsilon_{ij}, \end{aligned}$$

where $X_{ij1} = 1$ for all i and j, and β_1 is then the fixed effect intercept term in the model. When expressed in this way, it can be seen that the intercept for the i^{th}

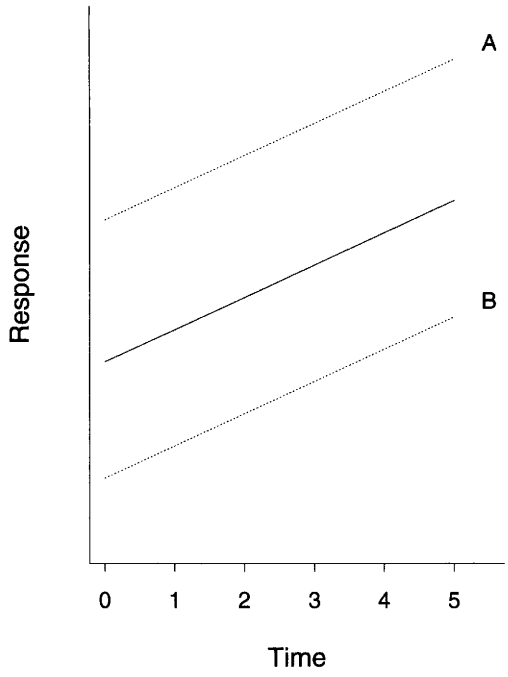

Fig. 8.1 Graphical representation of the marginal and conditional mean responses over time.

individual is $\beta_1 + b_i$ and varies randomly from one individual to another. Because the mean of the random effect b_i is assumed to be zero, b_i represents the deviation of the i^{th} individual's intercept ($\beta_1 + b_i$) from the population intercept, β_1.

For this simple example of a linear mixed effects model the fundamental ideas can be best understood by considering the graphical representation of the model equations. Figure 8.1 displays how the marginal mean response over time in the population changes linearly with time (denoted by the solid line), but also indicates how the conditional mean responses for two specific individuals, say subjects A and B, deviate from the population trend (denoted by the broken lines). In this simple illustration, individual A responds "higher" than the population average and thus has a positive b_i. On the other hand, individual B responds "lower" than the population average and has a negative b_i. Note that the mixed effects model with randomly varying intercepts does not posit that the repeated measures for individual A or B fall perfectly along these subject-specific response trajectories (represented by the broken lines in Figure 8.1). The inclusion of the measurement errors, ϵ_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories; this is illustrated in Figure 8.2.

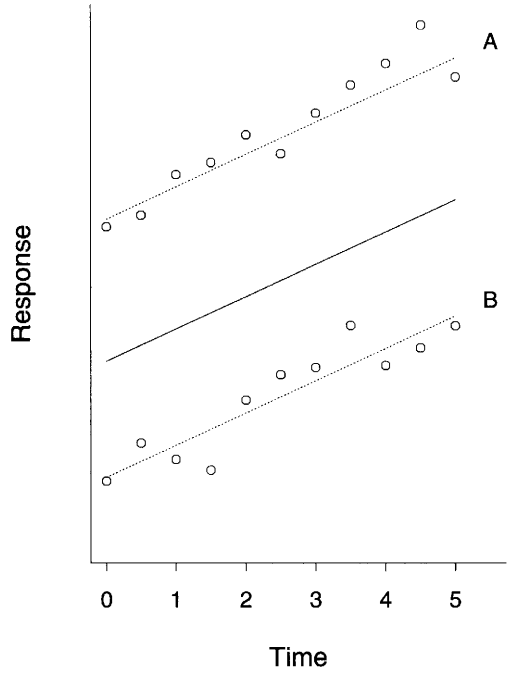

Fig. 8.2 Graphical representation of the marginal and conditional mean responses over time, plus measurement errors.

Next consider the marginal covariance among the repeated measurements on the same individual. When averaged over the individual-specific effects, the marginal mean of Y_{ij} is given by

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

The marginal covariance among the Y_{ij} is defined in terms of deviations of Y_{ij} from the marginal mean, μ_{ij}. For example, in Figure 8.2 these deviations are positive at all measurement occasions for individual A and negative at all measurement occasions for individual B, indicating a strong positive correlation (marginally) among the responses over time. For the model with randomly varying intercepts, the marginal variance of each response is given by

$$\begin{aligned} \text{Var}\,(Y_{ij}) &= \text{Var}\,\left(X'_{ij}\beta + b_i + \epsilon_{ij}\right) \\ &= \text{Var}\,(b_i + \epsilon_{ij}) \\ &= \text{Var}\,(b_i) + \text{Var}\,(\epsilon_{ij}) \\ &= \sigma_b^2 + \sigma^2. \end{aligned}$$

Similarly the marginal covariance between any pair of responses, Y_{ij} and Y_{ik}, is given by

$$\begin{aligned}\text{Cov}\,(Y_{ij}, Y_{ik}) &= \text{Cov}\left(X'_{ij}\beta + b_i + \epsilon_{ij},\ X'_{ik}\beta + b_i + \epsilon_{ik}\right)\\ &= \text{Cov}\,(b_i + \epsilon_{ij},\ b_i + \epsilon_{ik})\\ &= \text{Cov}\,(b_i,\ b_i)\\ &= \text{Var}\,(b_i)\\ &= \sigma_b^2.\end{aligned}$$

Thus the marginal covariance matrix of the repeated measurements has the following compound symmetry pattern:

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

This is the only covariance model that arises in both the patterned (see Section 7.4) and random effects families.

Given that the covariance between any pair of repeated measurements is σ_b^2, the correlation is

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

This simple expression for the correlation emphasizes an important aspect of mixed effects models: the introduction of a random subject effect, b_i, can be seen to induce correlation among the repeated measurements. Although the randomly varying intercepts model is the simplest example of a linear mixed effects model, and the resulting covariance structure is not usually appropriate for longitudinal data, the basic ideas can be generalized to provide a very versatile model for analyzing longitudinal data.

8.2 LINEAR MIXED EFFECTS MODELS

In this section we consider generalizations of (8.1) by allowing additional regression coefficients to vary randomly. We also highlight some of the appealing aspects of the linear mixed effects model alluded to earlier. The underlying premise of the model is that some subset of the regression coefficients vary randomly from one individual to another. In the simplest case considered above, we assumed that the intercept varied randomly. The introduction of this single random effect induces covariance among the repeated measures, albeit with a somewhat restricted form. By allowing a subset of the regression coefficients to vary randomly, a very flexible, and yet quite parsimonious, class of random effects covariance structures becomes available.

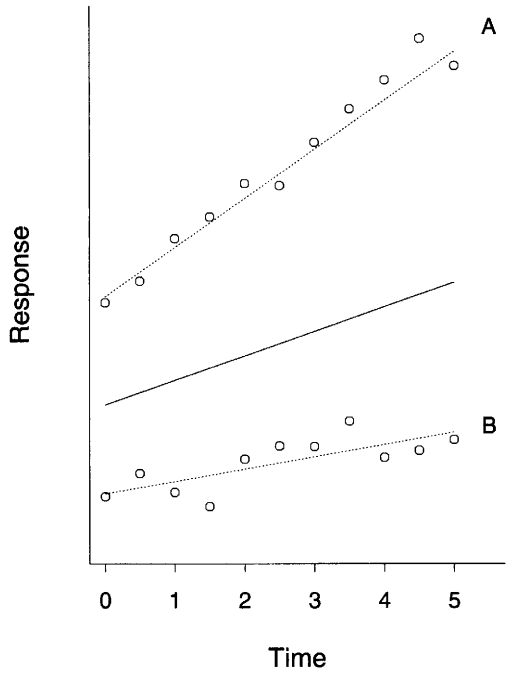

Fig. 8.3 Graphical representation of the marginal and conditional mean responses over time, plus measurement errors.

To fix ideas, consider the following example of a linear mixed effects model with intercepts and slopes that vary randomly among individuals. That is, for the i^{th} subject at the j^{th} measurement occasion,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}, \quad j = 1, ..., n_i.$$

In this model each subject varies not only in their baseline level of response (when $t_{i1} = 0$) but also in terms of changes in their responses over time. This can be best understood by considering the graphical representation of the model equations. Figure 8.3 displays how the marginal mean response in the population changes linearly with time (denoted by the solid line), but also indicates how the conditional mean responses for two specific individuals, say subjects A and B, deviate from the population trend (denoted by the broken lines). In this simple illustration, individual A has a "higher" baseline level of response ($\beta_1 + b_{1i}$) than the population average (β_1) and thus has a positive b_{1i}. On the other hand, individual B has a "lower" baseline level of response than the population average and thus has a negative b_{1i}. In addition individual A has a steeper rate of increase over time ($\beta_2 + b_{2i}$) than the population

average (β_2) and thus has a positive b_{2i}. Individual B has a less steep rate of increase over time than the population average and thus has a negative b_{2i}. Finally, the inclusion of the measurement errors, ϵ_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories. In this illustration there are randomly varying intercepts and slopes. However, the linear mixed effects model can be generalized to incorporate additional randomly varying regression coefficients and to allow the means of the random effects to depend on covariates.

In the following we assume there are N individuals on whom we have collected n_i repeated observations, with the response variable Y_{ij} measured at time t_{ij}. Thus the longitudinal data can be inherently unbalanced over time. In the most extreme case, each individual has a unique sequence of measurement occasions, $t_{i1}, ..., t_{in_i}$. Although no longitudinal study would ever be intentionally designed in this way, a change in the metameter for "time" may induce such a design. For example, a longitudinal design can be perfectly balanced when time is defined relative to the baseline measurement but become highly unbalanced if time is defined relative to some landmark event (e.g., puberty, menarche, or menopause).

Using vector and matrix notation, the linear mixed effects model can be expressed as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{8.2}$$

where β is a $(p \times 1)$ vector of fixed effects, b_i is a $(q \times 1)$ vector of random effects, X_i is a $(n_i \times p)$ matrix of covariates, and Z_i is a $(n_i \times q)$ matrix of covariates, with $q \leq p$. Here Z_i is a known design matrix linking the vector of random effects b_i to Y_i. In particular, for many models for longitudinal analysis the columns of Z_i are a subset of the columns of X_i. The reason for this restriction on the columns of Z_i will become evident in Section 8.4; in Chapter 19 we will encounter linear mixed effects models where the columns of Z_i are no longer a subset of the columns of X_i. In model (8.2) the particular subset of the regression parameters β that vary randomly is determined by the columns of X_i that comprise Z_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding column of X_i in Z_i, the design matrix for the random effects. The random effects, b_i, are assumed to be independent of the covariates, X_i, and to have a multivariate normal distribution with mean zero and covariance matrix G. That is, $E(b_i) = 0$ and $\text{Cov}(b_i) = G$. In principle, any multivariate distribution for b_i could be assumed; in practice, b_i are assumed to have a multivariate normal distribution.

If, in model (8.2), the vector of random effects, b_i, has mean zero, the random effects then have interpretation in terms of how the subset of regression parameters for the i^{th} individual deviate from those in the population. As mentioned previously, the particular subset of the regression parameters, β, that are assumed to vary randomly is determined by the columns of X_i that comprise Z_i. For example, in a model with only randomly varying intercepts, Z_i is a $(n_i \times 1)$ vector composed of 1's (since $X_{ij1} = 1$ for all i and j). Later we will consider the form of the design matrix Z_i for more general models.

An important distinction in the linear mixed effects model is that between the conditional and marginal means of Y_{ij}. The *conditional* or *subject-specific* mean of

Y_i, given b_i, is

$$E(Y_i|b_i) = X_i\beta + Z_i b_i,$$

while the *marginal* or population-averaged mean of Y_i, when averaged over the distribution of the random effects b_i, is

$$\begin{aligned} E(Y_i) &= \mu_i \\ &= E\{E(Y_i|b_i)\} \\ &= E(X_i\beta + Z_i b_i) \\ &= X_i\beta + Z_i E(b_i) \\ &= X_i\beta, \end{aligned}$$

since $E(b_i) = 0$. Thus, in the linear mixed effects model, the vector of regression parameters β (the *fixed effects*), are assumed to be the same for all individuals and have population-averaged interpretations, for example, in terms of changes in the mean response, averaged over all individuals in the population. In contrast to β, the vector b_i (when combined with the corresponding fixed effects) is comprised of subject-specific regression coefficients. These are the *random effects*, and when combined with the fixed effects, they describe the mean response profile of any *individual*. That is, the mean response profile for the i^{th} individual is given by

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

Finally, the $(n_i \times 1)$ vector of errors, ϵ_i, is assumed to be independent of b_i, and to also have a multivariate normal distribution with mean zero and covariance matrix R_i. Ordinarily it is further assumed that R_i is the diagonal matrix, $\sigma^2 I_{n_i}$, where I_{n_i} denotes an $n_i \times n_i$ identity matrix. In that case, ϵ_{ij} and ϵ_{ik} are uncorrelated, with equal variance, and the ϵ_{ij}'s can be thought of as sampling or measurement errors. In principle, we can allow correlation among the ϵ_{ij}'s by assuming R_i has a covariance pattern of the kind considered in Section 7.4. However, doing so would raise two potential complications. First, the ϵ_{ij}'s would no longer have a simple interpretation as measurement or sampling errors. This would alter the interpretation of the ϵ_{ij}'s, and hence b_i, implying that the ϵ_{ij}'s include a component of model misspecification at the individual level. Second, there can be subtle issues of model identification when R_i is assumed to have a non-diagonal covariance pattern since there may be insufficient information in the data at hand to support separate estimation of both G and a non-diagonal R_i. For example, it is not possible to estimate both G and an unstructured R_i. Throughout the remainder of this chapter we assume that the ϵ_{ij}'s are pure measurement or sampling errors and that $R_i = \sigma^2 I_{n_i}$.

Although we have assumed multivariate normality for both the random effects, b_i, and the measurement errors, ϵ_i, these distributional assumptions are not required for the model development. The form of the conditional and marginal means only requires that the measurement errors are independent of the random effects and that both have mean zero, $E(b_i) = 0$ and $E(\epsilon_i) = 0$. The multivariate normal assumption is required in subsequent sections where we consider estimation, testing, and prediction of random effects.

To clarify the vector and matrix notation introduced so far, consider the following linear mixed effects model with intercepts and slopes that vary randomly among individuals (see Figure 8.3). For the i^{th} subject at the j^{th} measurement occasion, assume that

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}, \quad j = 1, ..., n_i.$$

Using vector and matrix notation, this model can be expressed as

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i,$$

where

$$X_i = Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}.$$

Here $q = p = 2$ and Z_i is composed of the two columns of X_i. This model posits that individuals vary not only in their baseline level of response (when $t_{i1} = 0$), but also in terms of their changes in the mean response over time. The effects of covariates (e.g., due to treatments, exposures, or background characteristics of the individuals) can be incorporated by allowing the means of the intercepts and slopes to depend on these covariates (e.g., by allowing them to vary across the different treatment groups or levels of exposure).

For example, consider the hypothetical two-group study comparing a *treatment* and a *control* group discussed in Section 5.2. If the mean response changes in an approximately linear fashion over time, but with the means of the intercepts and slopes depending on group, the following linear mixed effects model can be adopted:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \, \text{Group}_i + \beta_4 t_{ij} \times \text{Group}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

where $\text{Group}_i = 1$ if the i^{th} individual was assigned to the treatment, and $\text{Group}_i = 0$ otherwise. In this model the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ 1 & t_{i2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,n_i} & 0 & 0 \end{pmatrix},$$

whereas for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & 1 & t_{i1} \\ 1 & t_{i2} & 1 & t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,n_i} & 1 & t_{i,n_i} \end{pmatrix}.$$

Note that the design matrix Z_i has the same form for both the treatment and control groups,

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{i,n_i} \end{pmatrix}.$$

Next consider the covariance among the components of Y_i in this linear mixed effects model with randomly varying intercepts and slopes. Let $\text{Var}\,(b_{1i}) = g_{11}$, $\text{Var}\,(b_{2i}) = g_{22}$, and $\text{Cov}\,(b_{1i}, b_{2i}) = g_{12}$. These are the three unique elements of the (2×2) covariance matrix $G = \text{Cov}(b_i)$. If we also assume that $R_i = \text{Cov}(\epsilon_i) = \sigma^2 I_{n_i}$, then it can be shown that

$$\begin{aligned} \text{Var}\,(Y_{ij}) &= \text{Var}\left(X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}\right) \\ &= \text{Var}\left(Z'_{ij}b_i + \epsilon_{ij}\right) \\ &= \text{Var}\,(b_{1i} + b_{2i}\,t_{ij} + \epsilon_{ij}) \\ &= \text{Var}(b_{1i}) + 2\,t_{ij}\,\text{Cov}(b_{1i}, b_{2i}) + t_{ij}^2\,\text{Var}(b_{2i}) + \text{Var}(\epsilon_{ij}) \\ &= g_{11} + 2\,t_{ij}\,g_{12} + t_{ij}^2\,g_{22} + \sigma^2. \end{aligned}$$

Similarly it can be shown that

$$\begin{aligned} \text{Cov}\,(Y_{ij}, Y_{ik}) &= \text{Cov}\left(X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij},\ X'_{ik}\beta + Z'_{ik}b_i + \epsilon_{ik}\right) \\ &= \text{Cov}\left(Z'_{ij}b_i + \epsilon_{ij},\ Z'_{ik}b_i + \epsilon_{ik}\right) \\ &= \text{Cov}\,(b_{1i} + b_{2i}\,t_{ij} + \epsilon_{ij},\ b_{1i} + b_{2i}\,t_{ik} + \epsilon_{ik}) \\ &= \text{Var}(b_{1i}) + (t_{ij} + t_{ik})\,\text{Cov}(b_{1i}, b_{2i}) + t_{ij}\,t_{ik}\,\text{Var}(b_{2i}) \\ &= g_{11} + (t_{ij} + t_{ik})\,g_{12} + t_{ij}\,t_{ik}\,g_{22}. \end{aligned}$$

Thus in this model for longitudinal data the covariance matrix, $\text{Cov}(Y_i)$, can be expressed as a function of time, t_{ij}. In particular, with the inclusion of random intercepts and slopes, the variance can increase or decrease over time as a quadratic function of the times of measurement. For example, the quadratic expression for $\text{Var}(Y_{ij})$ given above implies that the variance increases over time (for $t_{ij} \geq 0$) when $\text{Cov}(b_{1i}, b_{2i}) \geq 0$ but can decrease over time when $\text{Cov}(b_{1i}, b_{2i}) < 0$. Similarly the magnitude of the covariance (and correlation) between a pair of responses, say Y_{ij} and Y_{ik}, depends on the time separation between them (t_{ij} and t_{ik}). In Section 8.3 we consider the form of the induced random effects covariance structure in the more general case.

Note that the covariance matrix, G, for the vector of random effects is not invariant to a linear transformation of Z_i. Linear transformations of the columns of Z_i alter the interpretation of b_i and change the estimates of the variances and covariances of the random effects. For example, in the linear mixed effects model with randomly varying

intercepts and slopes, centering of the times of measurement (e.g., $t_{ij} - \tau$, for $\tau \neq 0$) alters the interpretation of the intercepts, and this leads to a change in the estimated variance of the random intercepts and their covariance with the random slopes. For example, in the model with untransformed times of measurement the variance of the "intercepts" is a measure of the between-subject variability in the response at time zero. However, in the model with transformed times of measurement, say centered at $\tau \neq 0$, the variance of the "intercepts" is a measure of the between-subject variability in the response at time τ. Centering not only changes the variance of the random intercepts but also changes the correlation between the random intercepts and slopes. Linear transformations of components of Z_i produce equivalent mixed effects models only when the covariance matrix, G, has been left unstructured. When G is unstructured the appropriate changes to the variances and covariances of the random effects can be produced. For this reason we strongly recommend that the covariance matrix, G, should always be left unstructured (unless there are compelling reasons, related to the specific analysis under consideration, that suggest this recommendation be relaxed).

Finally, an important issue in the linear mixed effects model concerns the "centering" of the times of measurement. In Chapter 6 we emphasized that "centering" can avoid problems of collinearity when the model for the mean includes linear, quadratic (and possibly higher-order polynomial) time trends. In the linear mixed effects model "centering" has implications for the proper interpretation of both the mean response and the variance of the random effects. In the illustration above, if t_{ij} represents time since baseline, then $\beta_1 + b_{1i}$ represents the subject-specific mean response at baseline (in the control group) and $\text{Var}(b_{1i}) = g_{11}$ is the between-subject variation in the mean response at baseline. On the other hand, if t_{ij} is an individual's age at the j^{th} measurement occasion, then $\beta_1 + b_{1i}$ does not have a useful interpretation since it represents the subject-specific mean response at age zero; similarly, $\text{Var}(b_{1i})$ does not have a useful interpretation. In that case there are two obvious choices for centering: (1) center the times of measurement for all subjects at some common fixed age within the age range of the study participants (i.e., $t_{ij} - a$, for some fixed value a), or (2) center at the mean age of each subject, when averaged over the subject's period of follow-up (i.e., $t_{ij} - \overline{a}_i$, where $\overline{a}_i$ is the average age, over the period of follow-up, for the i^{th} subject). The first option is preferable because $\beta_1 + b_{1i}$ represents the subject-specific mean response at the common age a and g_{11} is the between-subject variation in the mean response at that age. The second option should be avoided because $\beta_1 + b_{1i}$ then represents the subject-specific mean response at a specific subject's mean age over the period of follow-up. Since the mean age may vary considerably from one subject to another, g_{11} will be inflated and will not have a meaningful interpretation. In summary, with unbalanced longitudinal data, mean centering of the times of measurement should be avoided. Instead, we recommend that times of measurement should be centered at some common value of time (or age) in the center of the range of values for all individuals. By centering at a common value, the intercept is interpretable as the mean response at that common value for time (or age) and $\text{Var}(b_{1i})$ also has a meaningful interpretation.

8.3 RANDOM EFFECTS COVARIANCE STRUCTURE

Next we consider the form of the induced random effects covariance structure for longitudinal data in the more general case. In the linear mixed effects model

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

$R_i = \text{Cov}(\epsilon_i)$ describes the covariance among the longitudinal observations when focusing on the conditional mean response profile of a *specific* individual. That is, it is the covariance of the i^{th} individual's deviations from her mean response profile,

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

For example, in Figures 8.2 and 8.3 these deviations are positive and negative, and vary randomly about zero, for individuals A and B. As mentioned previously, it is usually assumed that R_i is a diagonal matrix, $\sigma^2 I_{n_i}$, where I_{n_i} denotes an $n_i \times n_i$ identity matrix. The latter is often referred to as a "conditional independence assumption"; that is, given the random effects b_i, the measurement errors are independently distributed with a common variance σ^2.

In the linear mixed effects model we can distinguish the conditional mean of Y_i, given b_i,

$$E(Y_i|b_i) = X_i\beta + Z_i b_i,$$

from the *marginal* or population-averaged mean of Y_i,

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

where averaging is over the distribution of the random effects, b_i. In a similar way we can distinguish between conditional and marginal covariances. The conditional covariance of Y_i, given b_i, is

$$\text{Cov}(Y_i|b_i) = \text{Cov}(\epsilon_i) = R_i,$$

while the marginal covariance of Y_i, averaged over the distribution of b_i, is

$$\begin{aligned} \text{Cov}(Y_i) &= \text{Cov}(Z_i b_i) + \text{Cov}(\epsilon_i) \\ &= Z_i \text{Cov}(b_i) Z_i' + \text{Cov}(\epsilon_i) \\ &= Z_i G Z_i' + R_i. \end{aligned}$$

This latter expression for the marginal covariance may be somewhat daunting at first glance. Even when $R_i = \text{Cov}(\epsilon_i) = \sigma^2 I_{n_i}$, a diagonal matrix (with all pairwise correlations equal to zero),

$$\text{Cov}(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i}$$

is emphatically not a diagonal matrix. That is, $\text{Cov}(Y_i)$ will, in general, have non-zero off-diagonal elements, thereby accounting for the correlation among the repeated

observations on the same individuals in a longitudinal study. Thus the introduction of random effects, b_i, induces correlation among the components of Y_i. An additional property of the linear mixed effects model is that Cov (Y_i) has been described in terms of a set of covariance parameters, some defining the matrix G and some defining the matrix R_i. That is, the linear mixed effects model allows for the explicit analysis of between-subject (G) and within-subject (R_i) sources of variation in the responses. Finally, the marginal covariance of Y_i is a function of the times of measurement. For example, in the model with randomly varying intercepts and slopes considered in Section 8.2, we saw that

$$\begin{aligned} \text{Var}\,(Y_{ij}) &= g_{11} + 2\,t_{ij}\,g_{12} + t_{ij}^2\,g_{22} + \sigma^2, \\ \text{Cov}\,(Y_{ij}, Y_{ik}) &= g_{11} + (t_{ij} + t_{ik})\,g_{12} + t_{ij}\,t_{ik}\,g_{22}, \end{aligned}$$

so that both Var (Y_{ij}) and Cov (Y_{ij}, Y_{ik}) depend on the measurement times.

The induced random effects covariance structure,

$$\text{Cov}\,(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i},$$

can be contrasted with the covariance pattern models described in Chapter 7 (see Section 7.4). Recall that a defining feature of the covariance pattern models is that they take into account all sources of variability that have an impact on the covariance, but do not distinguish between the different sources of variability. In contrast, linear mixed effects models explicitly distinguish between-subject and within-subject sources of variability. For linear models for longitudinal continuous data, both approaches yield the same model for the *marginal* or population-averaged mean of Y_i,

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

and differ only in terms of the assumed model for the covariance. As we will see in Chapters 12 through 16, longitudinal models for discrete responses do not share this property; for discrete responses, different approaches for accounting for the covariance among the longitudinal responses can lead to models for the mean response having regression parameters with quite distinct interpretations.

The induced random effects covariance structure has certain features that are different from the covariance pattern models considered in Chapter 7. First, unlike many covariance pattern models, the random effects covariance structure does not require a balanced longitudinal design. Because the covariance is expressed as an explicit function of the times of measurement (when times of measurement, or functions of time, are included in Z_i), in principle, each individual can have a unique sequence of measurement times. This makes linear mixed effects models well suited for modeling data from inherently unbalanced longitudinal designs. In addition the number of covariance parameters is the same regardless of the number and timing of the measurements. Finally, unlike many of the covariance pattern models that make strong assumptions about homogeneity of variance over time, the random effects covariance structure allows the variance and covariance to increase or decrease as a function of the times of measurement (e.g., in the random intercepts and slopes model, the variance is a quadratic function of the times of measurement).

8.4 TWO-STAGE RANDOM EFFECTS FORMULATION

The linear mixed effects model given by (8.2) can be motivated by a two-stage random effects formulation of the model. Indeed, some of the main ideas behind the mixed effects model are often better understood by considering the model as arising from a two-stage specification. For purely pedagogical purposes, we find the two-stage specification to be quite helpful; however, we must caution the reader that the two-stage formulation of the linear mixed effects model does introduce some unnecessary restrictions on the model.

Stage 1

As the term implies, a two-stage random effects model can be conceived in two separate stages. In the first stage subjects are assumed to have their own unique individual-specific mean response trajectories. That is, in stage 1 we posit that the repeated measures on each individual follow a regression model having the same set of covariates, but with separate or distinct regression coefficients for each individual. This is expressed more formally as

$$Y_i = Z_i\beta_i + \epsilon_i,$$

where the vector of errors, ϵ_i, are assumed to have a normal distribution, with mean equal to zero and variance σ^2. That is, the ϵ_i can be thought of as measurement or sampling errors, with $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$. Note that the number of individual-specific regression coefficients is the same (i.e., the dimension of β_i is q), regardless of the number of longitudinal responses n_i. These individual-specific regression coefficients, β_i, can be interpreted as the i^{th} individual's "true" regression coefficients. Alternatively, $Z_i\beta_i$ can be thought of as the i^{th} individual's "true" underlying mean response trajectory. When viewed in this way, the longitudinal responses on the i^{th} individual are assumed to follow the individual-specific response trajectory given by $Z_i\beta_i$, but with the addition of measurement or sampling errors, ϵ_i.

Note that the matrix Z_i specifies how an individual's mean response changes over time and/or how the mean response changes with other time-varying covariates (e.g., height). For example, it might be assumed that the mean response trajectory is linear, quadratic, or a spline function of time. Consider a model that assumes the individual-specific trajectories are linear in time. Then, the first-stage model can be written as

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}.$$

The essential idea underlying the first-stage model is to fit separate linear regression models to the data for each individual, but with the proviso that these regressions

involve the same set of covariates, Z_i. This is an important observation since it implies that, in principle (and given a sufficient number of repeated measures on each individual), it should be possible to estimate β_i (and σ^2) using data from only the i^{th} individual.

Recall that a particular feature of this first-stage formulation is that the matrix of covariates Z_i is restricted to contain only within-individual or time-varying covariates (with the exception of the column of 1's for the intercept). Time-invariant or between-individual covariates (e.g., gender, treatment group, exposure group) cannot be included in Z_i since their effects would simply be absorbed into the intercept term. Instead, between-individual covariates are introduced in the second stage of the model formulation.

Stage 2

In the second stage we make the assumption that the individual-specific effects, β_i, are random. Given that the β_i are random variables, they have some probability distribution, with a mean and covariance. The mean and covariance of the β_i are the population parameters that are modeled in the second stage. Specifically, variation in β_i from one individual to another is modeled as a function of a set of between-individual (or time-invariant) covariates (e.g., gender, treatment group). In particular, the mean of the β_i can be expressed as a linear function of a set of between-individual covariates, A_i,

$$E(\beta_i) = A_i\beta,$$

where A_i is a $q \times p$ matrix. The remaining residual between-individual variation in the β_i that cannot be explained by A_i is expressed as

$$\text{Cov}(\beta_i) = G.$$

Specification of a model for the mean and covariance of the β_i completes the second stage of the model formulation.[1]

For example, consider the hypothetical two-group study comparing a *treatment* and a *control* discussed earlier. If we assume that individual-specific changes in the mean response over time are linear, the first stage model is given by

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}.$$

In the second stage, we can allow the mean of β_i (i.e., the mean intercept and slope) to depend on group. For example, a model that allows both the mean intercept and

[1]Note that, in this section, in a slight abuse of notation, we are using β to denote a fixed parameter and β_i to denote a random variable.

slope to depend on group is given by

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_3 \,\mathrm{Group}_i \\ E(\beta_{2i}) &= \beta_2 + \beta_4 \,\mathrm{Group}_i \end{aligned}$$

where $\mathrm{Group}_i = 1$ if the i^{th} individual was assigned to the treatment, and $\mathrm{Group}_i = 0$ otherwise. In this model, β_1 is the mean intercept in the control group, while $\beta_1 + \beta_3$ is the mean intercept in the treatment group. That is, β_3 represents the treatment group difference in the mean intercept. When t_{ij} is the time since baseline, β_3 has a useful interpretation in terms of a treatment group difference in the mean response at baseline. Similarly β_2 is the mean slope, or rate of change in the mean response over time, in the control group, while $\beta_2 + \beta_4$ is the mean slope in the treatment group. That is, β_4 has interpretation in terms of a treatment group difference in the mean slope or rate of change in the mean response over time. In this model the design matrix A_i of between-individual covariates has the following form:

$$A_i = \begin{pmatrix} 1 & 0 & \mathrm{Group}_i & 0 \\ 0 & 1 & 0 & \mathrm{Group}_i \end{pmatrix}.$$

Thus, for the control group, the model for the mean is

$$E\begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix};$$

similarly, for the treatment group, the model for the mean is

$$E\begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_3 \\ \beta_2 + \beta_4 \end{pmatrix}.$$

It is also assumed that the remaining residual variation in β_i, which cannot be explained by the effect of group, is

$$\mathrm{Cov}(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \mathrm{Var}(\beta_{1i})$, $g_{22} = \mathrm{Var}(\beta_{2i})$, and $g_{12} = g_{21} = \mathrm{Cov}(\beta_{1i}, \beta_{2i})$. Thus g_{11} is the variance of β_{1i}, after adjusting for the effect of treatment group, and so on.

The two components of the two-stage model can be combined to yield a linear mixed effects model for Y_i, albeit one that has some restrictions. To see how this can be achieved, let us rewrite the subject-specific effects, β_i, as

$$\beta_i = A_i\beta + b_i,$$

where b_i has a multivariate normal distribution with mean zero and covariance matrix, G. Here the b_i yield the regression coefficients from an individual's *residual* trajectory over time, after the covariate effects have been accounted for. Put another way, the b_i represent the i^{th} individual's deviation from the population mean response. Next, by combining the two components of the two-stage model, we obtain

$$\begin{aligned} Y_i &= Z_i\beta_i + \epsilon_i \\ &= Z_i(A_i\beta + b_i) + \epsilon_i \\ &= (Z_iA_i)\beta + Z_ib_i + \epsilon_i \\ &= X_i\beta + Z_ib_i + \epsilon_i, \end{aligned}$$

where $X_i = Z_iA_i$. When averaged over the random effects, b_i,

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

and

$$\text{Cov}(Y_i) = Z_iGZ_i' + \sigma^2 I_{n_i}.$$

Note that in the two-stage formulation, Z_i appears in both the models for the marginal mean and covariance.

While this model is remarkably similar to the linear mixed effects model introduced in the previous section, there is one important difference. The two-stage model places a constraint on the choice of the design matrix for the fixed effects. That is, the two-stage formulation requires that the design matrix for the fixed effects has the special structure $X_i = Z_iA_i$, where A_i contains only between-subject (or time-invariant) covariates and Z_i contains only within-subject (or time-varying) covariates. This form for the design matrix for the fixed effects implies that any time-varying covariates must be specified as random effects to ensure their inclusion in the model for the population mean response.[2] This constraint is unnecessary and, in many settings, it can be somewhat inconvenient. In some applications this constraint forces us to consider rather more complex models than may be necessary. For example, in order to allow a sufficiently complex model for the mean response over time (specified in terms of $Z_iA_i\beta$), it may be necessary to include many covariates in Z_i. However, in the two-stage model formulation, this can only be achieved by also introducing an equally complex model for the covariance, since

$$\text{Cov}(Y_i) = Z_iGZ_i' + \sigma^2 I_{n_i}.$$

An example of this arises in developing a model for FEV_1 in children. Previous studies have shown that both age (as a linear spline) and log height are important

[2] Alternatively, this constraint can be relaxed by assuming that some components of β_i are constant, not random. However, when some components of β_i are assumed to be constant, then the repeated measures on each individual no longer follow a regression model with distinct regression coefficients for each individual in stage 1.

covariates. Thus four subject-specific regression coefficients (an intercept, two coefficients for age, and one coefficient for log height) are needed to model the mean. But a 4×4 covariance matrix for G is very unwieldy and difficult to fit without very large samples.

Alternatively, in the two-stage formulation a very simple structure for the covariance imposes an often unrealistically simple structure on the mean response. The most extreme example of this is the two-stage model, which induces a compound symmetry covariance. In that case a compound symmetry covariance is obtained from a two-stage model with randomly varying intercepts,

$$Y_i = Z_i\beta_i + \epsilon_i,$$

where Z_i is a $(n_i \times 1)$ vector of 1's. While marginally (or averaged over the random effects) this model induces a simple compound symmetry covariance structure

$$\text{Cov}(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i} = g_{11} J_{n_i} + \sigma^2 I_{n_i},$$

where J_{n_i} denotes a $(n_i \times n_i)$ matrix of 1's, this model precludes any dependence of the mean response on time. That is,

$$E(Y_i) = (Z_i A_i)\beta$$

cannot depend on time since time, a within-subject covariate, has not been included in Z_i in the first stage. Thus, when formulated as a two-stage model, the randomly varying intercepts model excludes the most salient within-subject covariate in a longitudinal study (i.e., time), and thereby does not allow for estimation of changes in the mean response over time, the primary goal in a longitudinal analysis!

In summary, we view the two-stage formulation as being most useful for motivating the main ideas and concepts underlying linear mixed effects models. The inherent restrictions in the two-stage formulations can be circumvented by considering linear mixed effects models with an arbitrary design matrix, X_i, for the fixed effects, and by allowing the dimension of Z_i to be arbitrary. For many models for longitudinal analysis the only restrictions placed on X_i and Z_i is that Z_i is composed of a subset of the columns of X_i (in Chapter 19 we will encounter linear mixed effects models where the columns of Z_i are no longer a subset of the columns of X_i). The latter constraint ensures that $Z_i b_i$ can be interpreted as a zero mean between-subject residual trajectory (or, put another way, the discrepancy between the i^{th} individual's conditional mean response trajectory and the mean response trajectory in the population). Thus in the linear mixed effects model,

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

the restriction that the columns of Z_i are a subset of the columns of X_i allows us to partition the columns of X_i into a set of columns corresponding to the effects that are fixed and a complementary set of columns corresponding to the effects that are random. If we denote the former by $X_i^{(F)}$ and the latter by $X_i^{(R)}$, the model for Y_i can then be rewritten as

$$Y_i = X_i^{(F)}\beta^{(F)} + X_i^{(R)}\beta_i^{(R)} + \epsilon_i,$$

where β has been similarly partitioned into effects that are considered to be fixed, $\beta^{(F)}$, and effects that are considered to be random, $\beta_i^{(R)}$.

Finally, on an historical note, a version of the two-stage formulation was popularized by biostatisticians working at the U.S. National Institutes of Health (NIH). They proposed a method for analyzing repeated measures data where in the first stage subject-specific regression coefficients are estimated using ordinary least-squares regression (based only on the observations for each subject). In the second stage, the estimated regression coefficients are then analyzed as summary measures using standard parametric (or nonparametric) methods. This method for analyzing repeated measures data became known as the "NIH method"[3] and is a variant of the summary measure analyses considered in Section 3.6.

8.5 CHOICE AMONG RANDOM EFFECTS COVARIANCE MODELS

Although the linear mixed effects model assumes that the longitudinal responses depend on a combination of population and subject-specific effects, when averaged over the distribution of the random effects,

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

and the covariance among the responses has the distinctive random effects structure,

$$\text{Cov}\,(Y_i) = Z_iGZ_i' + \sigma^2 I_{n_i}.$$

From the perspective of modeling the covariance, the random effects structure is appealing because the number of covariance parameters, $q \times (q+1)/2 + 1$, is the same regardless of the number and timing of the measurement occasions. In many applications, it will be sufficient to include only random intercepts and slopes for time (a total of $2 \times (2+1)/2 + 1 = 4$ covariance parameters), thereby allowing for heterogeneity in the variances and correlations that can be expressed as functions of time. In other applications, a more complex random effects structure may be required.

In choosing a model for the covariance, it will often be of interest to compare two nested models, one with q correlated random effects, the other with $q+1$ correlated random effects. The difference in the number of covariance parameters between these two models is $q+1$, since there is one additional variance and q additional covariances in the "full" model. As mentioned in Section 7.5, in general, the likelihood ratio test provides a valid method for comparing nested models for the covariance. However, in certain cases, the usual null distribution for the likelihood ratio test is no longer valid. In particular, the comparison of random effects models for the covariance is such a non-standard problem.

[3]It is difficult to attribute the popularization of the so-called NIH method to any single biostatistician at NIH. During their time at NIH, Sam Greenhouse, Max Halperin, and Jerry Cornfield introduced many biostatisticians to this technique.

In general, when testing a null hypothesis that is on the boundary of the parameter space (e.g., the variance of a random effect equals zero), the usual null distribution for the likelihood ratio test is no longer a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models; instead, the null distribution is a mixture of chi-squared distributions. For example, when comparing two nested models, one with q correlated random effects, the other with $q + 1$ correlated random effects, the null distribution of the likelihood ratio test is a 50:50 mixture of chi-squared distributions with q and $q + 1$ degrees of freedom. In Section 7.5 we considered the special case where $q = 0$. A table of critical values, when the null distribution is a known 50:50 mixture of chi-squared distributions with q and $q + 1$ degrees of freedom, is provided in Table C.1 in Appendix C. The critical values in Table C.1 can be used for making inferences about the complexity of the random effects covariance structure. For example, when comparing a model with two correlated random effects (e.g., random intercepts and slopes) versus a model with one random effect (e.g., random intercepts only), the critical value for the 0.05 significance level can be found in the second row ($q = 1$) of Table C.1. This yields a critical value of 5.14, which is somewhat smaller than the critical value of 5.99 from a standard chi-squared distribution with 2 degrees of freedom.

Alternatively, and especially for more complex comparisons among nested random effects models for the covariance where the null distribution of the likelihood ratio test is not well understood (e.g., comparisons of nested models with q correlated random effects and $q + k$ correlated random effects, where $k > 1$), we recommend the use of $\alpha = 0.1$, instead of $\alpha = 0.05$, when judging the statistical significance of the likelihood ratio test. The latter procedure is somewhat ad hoc but will protect against selection of a model that is too parsimonious. In conclusion, for simple comparisons among nested random effects models, the likelihood ratio test statistic can be compared with the critical values in Table C.1. For more complex comparisons, we recommend the use of the $\alpha = 0.1$, instead of $\alpha = 0.05$, significance level.

8.6 PREDICTION OF RANDOM EFFECTS

In this section we provide a non-technical discussion on the prediction of random effects. A good grasp of the material in this section is all that is required for an understanding of the notion of predicting random effects. In Section 8.7 we present a more detailed and technical discussion of the same topic. Many of our readers may find the level of mathematical difficulty of the material in Section 8.7 too challenging. While we encourage all of our readers to tackle Section 8.7, we note that it can be omitted at first reading without loss of continuity.

In many applications where longitudinal data arise, inference is focused on the fixed effects, β. These regression parameters have interpretation in terms of changes in the mean response over time, and their relation to covariates. However, in some longitudinal studies, we may want to predict subject-specific response profiles. For example, in studies of growth it may be of interest to obtain subject-specific growth

trajectories. In other types of longitudinal studies, it may be of interest to identify those individuals who showed the greatest increase or decrease in the response over time. Because the linear mixed effects model explicitly distinguishes between fixed and random effects, we can also estimate (or predict) individual-specific response trajectories over time. That is, it is possible to obtain predictions of the subject-specific effects, b_i, or of the subject-specific response trajectories, $X_i\beta + Z_i b_i$. Technically, because the b_i are random variables, and not fixed population parameters, we customarily refer to "predicting" the random effects rather than "estimating" them.

In general, the problem of predicting a random variable can be shown to be that of predicting its conditional mean, given the available data. Thus the best predictor of b_i is the conditional mean of b_i, given the vector of responses Y_i (and $\widehat{\beta}$),

$$E(b_i|Y_i) = GZ_i'\Sigma_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\Sigma_i = \text{Cov}(Y_i) = Z_iGZ_i' + R_i$. This is known as the "best linear unbiased predictor" (or BLUP). This predictor of b_i depends on the unknown covariance among the Y_i. When the unknown covariance parameters are replaced by their REML (or ML) estimates, the resulting predictor

$$\widehat{b}_i = \widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}),$$

is referred to as the "empirical BLUP." Given the "empirical BLUP," $\widehat{b}_i$, we can also obtain the i^{th} subject's predicted response profile as follows:

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i.$$

Interestingly, the i^{th} subject's predicted response profile can also be expressed as a weighted average of the estimated population-averaged mean response profile, $X_i\widehat{\beta}$, and the i^{th} subject's observed response profile Y_i. Specifically,

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i = (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I_{n_i} - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i.$$

This expression helps to explain why it is often said that the empirical BLUP estimator "shrinks" the i^{th} subject's predicted response profile towards the population-averaged mean response profile.

The amount of "shrinkage" toward the population mean depends on the relative magnitude of R_i and Σ_i. Recall that R_i characterizes the within-subject variability, while Σ_i incorporates both within-subject and between-subject sources of variability. As a result, when R_i is relatively "large," and the within-subject variability is greater than the between-subject variability, more weight is assigned to $X_i\widehat{\beta}$, the estimated population-averaged mean response profile, than to the i^{th} individual's observed responses. On the other hand, when the between-subject variability is large relative to the within-subject variability, more weight is given to the i^{th} subject's observed responses, Y_i. Intuitively, this weighting scheme is quite sensible since greater weight should be given to the i^{th} individual's observed responses when any within-subject variability in the longitudinal responses (e.g., due to measurement error) is relatively

small when compared to the natural heterogeneity in the individual-specific longitudinal response trajectories. On the other hand, less weight should be given to the i^{th} individual's observed responses when the within-subject variability is relatively large and the population is relatively homogeneous. Finally, the amount of "shrinkage" toward the population mean depends also on n_i, the number of observation on the i^{th} subject. In general, there is more shrinkage toward the population mean curve when n_i is small. Intuitively, this is also quite sensible since less weight should be given to the i^{th} individual's observed responses when fewer data points are available.

8.7 PREDICTION AND SHRINKAGE*

In this section† we present a more detailed and technical discussion on prediction of random effects in the linear mixed effects model. In doing so, we provide some motivation for, and expressions that support, the main results outlined in Section 8.6.

Because the linear mixed effects model explicitly distinguishes between fixed and random effects, we can estimate both types of effects. As noted in the previous section, the prediction of random effects translates into the problem of predicting the conditional mean of b_i, given the vector of responses Y_i, $E(b_i|Y_i)$. Under the assumptions of the linear mixed effects model, Y_i and b_i have a joint multivariate normal distribution. Using well-known properties of the multivariate normal distribution, it can be shown that the conditional mean of b_i given Y_i (and $\widehat{\beta}$) is

$$E(b_i|Y_i) = GZ_i'\Sigma_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\Sigma_i = \text{Cov}(Y_i) = Z_iGZ_i' + R_i$. This is known as the "best linear unbiased predictor" (or BLUP). From a practical standpoint, this predictor of b_i is unusable because it depends on the unknown covariance parameters. When the unknown covariance parameters are replaced by their REML estimates, the resulting predictor

$$\widehat{b}_i = \widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}),$$

is referred to as the "empirical BLUP" or the "empirical Bayes" (EB) estimator (since $\widehat{b}_i$ can also be derived from a fully Bayesian formulation). In addition to obtaining a prediction of b_i, we can obtain standard errors for the prediction based on the expression

$$\text{Cov}(\widehat{b}_i - b_i) = G - GZ_i'\Sigma_i^{-1}Z_iG + GZ_i'\Sigma_i^{-1}X_i\left(\sum_{i=1}^{N} X_i'\Sigma_i^{-1}X_i\right)^{-1}X_i'\Sigma_i^{-1}Z_iG.$$

Note that $\text{Cov}(\widehat{b}_i - b_i)$ is used to assess the precision of the prediction of b_i, rather than $\text{Cov}(\widehat{b}_i)$, because the latter would fail to recognize the random variation of b_i.

†This section provides a more technical presentation of prediction of random effects and the notion of shrinkage; it can be omitted at first reading without loss of continuity.

Standard errors for the prediction are obtained by simply substituting $\widehat{\Sigma}_i$ and $\widehat{G}$, the REML estimates of the covariance parameters, in the previous expression for $\text{Cov}(\widehat{b}_i - b_i)$.

Given the prediction of b_i, the i^{th} subject's predicted response profile is given by

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i.$$

This expression for the predicted response profile can be re-expressed as follows:

$$\begin{aligned}
\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\
&= X_i\widehat{\beta} + Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}) \\
&= (I_{n_i} - Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1}Y_i \\
&= (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I_{n_i} - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i,
\end{aligned}$$

since

$$\widehat{\Sigma}_i\widehat{\Sigma}_i^{-1} = I_{n_i} = (Z_i\widehat{G}Z_i' + \widehat{R}_i)\widehat{\Sigma}_i^{-1} = Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1} + \widehat{R}_i\widehat{\Sigma}_i^{-1}.$$

The latter expression for $\widehat{Y}_i$ shows how the empirical Bayes estimator "shrinks" the i^{th} subject's predicted response profile toward the population-averaged mean response profile. As noted in Section 8.6, the amount of "shrinkage" depends on the relative magnitude of R_i and Σ_i. When the within-subject variability, R_i, is large relative to the between-subject variability, more weight is assigned to $X_i\widehat{\beta}$ than to the i^{th} individual's observed responses. Conversely, when the between-subject variability is large relative to the within-subject variability, more weight is given to the i^{th} subject's observed responses, Y_i.

Similarly it can be shown that the prediction of individual-specific regression coefficients is a weighted average of the REML estimate of the fixed effects and the corresponding OLS estimate based only on the individual's observations. Specifically, when the linear mixed effects model has a two-stage representation, with $X_i = Z_iA_i$ and $R_i = \sigma^2 I_{n_i}$, the "empirical BLUP" of $\beta_i = A_i\beta + b_i$ is a weighted average of $A_i\widehat{\beta}$ and $\widehat{\beta}_i^{\text{OLS}}$, where $\widehat{\beta}$ is the usual REML estimate of β obtained from the available data on all subjects and $\widehat{\beta}_i^{\text{OLS}}$ is the ordinary least squares estimate of β_i based only on the n_i observations for the i^{th} subject. More formally, the "empirical BLUP" of β_i can be expressed as

$$\widehat{\beta}_i = A_i\widehat{\beta} + \widehat{b}_i = W_i\widehat{\beta}_i^{\text{OLS}} + (I_q - W_i)A_i\widehat{\beta},$$

where the "weight," W_i, is a ratio of the between-subject variability to the sum of the between- and within-subject variability,

$$W_i = G\{G + \sigma^2(Z_i'Z_i)^{-1}\}^{-1},$$

and I_q denotes a $q \times q$ identity matrix. Although this expression for the "weight", W_i, appears somewhat daunting, note that when there is very little within-subject variability (and $\sigma^2 \approx 0$), $W_i \approx I_d$ and then $\widehat{\beta}_i \approx \widehat{\beta}_i^{\text{OLS}}$. That is, when there is very little

within-subject variability, we have almost perfect information about b_i from Y_i alone. On the other hand, when there is very little between-subject variability (and $G \approx 0$), $W_i \approx 0$ and then $\widehat{\beta}_i \approx A_i \widehat{\beta}$. Thus, when there is very little between-subject variability in the individual-specific trajectories, it is quite sensible to base our "estimate" or prediction of b_i on data from all of the individuals in the study. The expression for the weight W_i also highlights how the number of repeated measurements influences the compromise between $\widehat{\beta}_i^{\text{OLS}}$ and $A_i \widehat{\beta}$. For example, consider the special case of the model with randomly varying intercepts, where Z_i is an $n_i \times 1$ vector of 1's. It can be shown that

$$W_i = G\{G + \sigma^2 (Z_i' Z_i)^{-1}\}^{-1} = \frac{n_i g_{11}}{n_i g_{11} + \sigma^2},$$

where $g_{11} = \text{Var}(b_{1i})$ is the variance of the random intercept. Thus, for fixed values of the within- and between-subject variability, the more observations that are available on the i^{th} individual the more the "empirical BLUP" of β_i relies on that individual's data to "estimate" the random effect.

8.8 CASE STUDIES

Next we illustrate the main ideas presented in this chapter by considering linear mixed effects models for analyzing data from three different studies. The first illustration uses lung function growth data in a sample of children and adolescents from the Six Cities Study of Air Pollution and Health. The second illustration uses data on body fat accretion from a prospective study of the development of obesity in a cohort of girls. The third illustration uses data on CD4 counts from a randomized clinical trial of AIDS patients with advanced immune suppression.

Six Cities Study of Air Pollution and Health

The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth (Dockery et al., 1983). A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the United States: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of 6 and 7) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian. The basic maneuver in simple spirometry is maximal inspiration (or breathing in) followed by forced exhalation as rapidly as possible into a closed container. Many different measures can be derived from the spirometric curve of volume exhaled versus time.

Table 8.1 Data on age, height, and FEV_1 for four girls selected from the Topeka data set.

Subject ID	Age	Height	Time	FEV_1
001	9.34	1.20	0.00	1.24
001	10.39	1.28	1.05	1.45
001	11.45	1.33	2.11	1.63
001	12.46	1.42	3.12	2.12
001	13.42	1.48	4.08	2.30
001	15.47	1.50	6.13	2.44
001	16.37	1.52	7.03	2.39
002	6.58	1.13	0.00	1.36
002	7.65	1.19	1.06	1.42
002	12.74	1.49	6.15	2.13
002	13.77	1.53	7.19	2.38
002	14.69	1.55	8.11	2.85
002	15.82	1.56	9.23	3.17
002	16.67	1.57	10.08	2.52
002	17.63	1.57	11.04	3.11
003	6.91	1.18	0.00	1.54
003	7.97	1.23	1.06	1.47
003	8.97	1.30	2.05	1.82
003	9.99	1.35	3.08	2.12
003	11.08	1.47	4.16	2.63
003	13.07	1.57	6.16	2.45
003	14.10	1.59	7.19	2.77
003	15.08	1.60	8.17	3.02
003	16.02	1.60	9.10	2.96
007	6.43	1.18	0.00	0.97
007	7.50	1.25	1.06	1.10
007	13.63	1.64	7.19	2.62
007	14.56	1.67	8.12	2.53
007	15.64	1.68	9.21	2.76
007	16.50	1.69	10.06	2.80
007	17.49	1.69	11.06	2.67

Note: Time represents time since entry to study.

One widely used measure is the total volume of air exhaled in the first second of the maneuver (FEV_1).

In this section we present an analysis of a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV_1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time; data for four selected girls are presented in Table 8.1. Note that although girls with only a single observation do not directly provide information about longitudinal or intra-individual change over time, their observations at a single occasion do contribute to the analysis (e.g., these observations

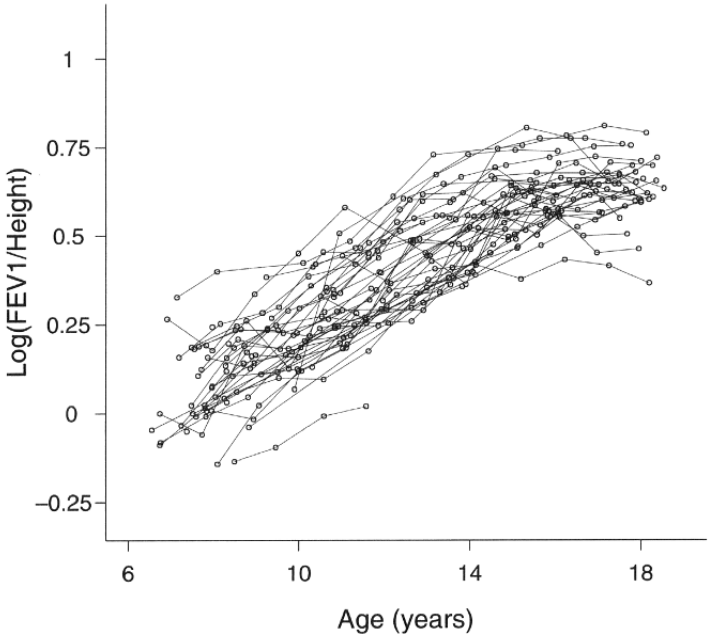

Fig. 8.4 Time plot, with joined line segments, of log(FEV_1/height) versus age in years for 50 randomly selected girls from the Topeka data set.

contribute information to the estimation of variances and between-subject effects). Examination of Table 8.1 reveals that these data are inherently unbalanced over time, and the degree of imbalance is even more marked when the age of the child is used as the metameter for time. That is, in this data set children enter the study at different ages and also have different occasions of measurement. Figure 8.4 displays a time plot, with joined line segments, of log(FEV_1/height) versus age for 50 randomly selected girls.

When age is used as the metameter for time, there are two sources of information about the relationship between FEV_1 and age. First, there is "cross-sectional" or between-subject information that arises because children enter the study at different ages. For example, there is information about how FEV_1 changes with age in the baseline (or time = 0) observations. Second, there is "longitudinal" or within-subject information that arises because children are measured repeatedly over time, yielding measurements of FEV_1 at different ages. Because there are two potentially conflicting sources of information about the relationship between FEV_1 and age, it is important to model these data in a way that allows for separate estimation of the "cross-sectional" and "longitudinal" effects of age of FEV_1. In doing so, it is then possible to test whether there are differences between the cross-sectional and longitudinal effects of age on FEV_1, and report separate effects where necessary or estimate a combined

effect, based on both sources of information, if appropriate. Note that the same issues arise in examining the relationship between FEV_1 and height. A more detailed discussion of the main issues surrounding the analysis of longitudinal designs that provide both longitudinal and cross-sectional sources of information can be found in Chapter 9 (see Sections 9.4 and 9.5).

The Six Cities Study was designed to characterize lung function growth between the ages of 6 and 18. The goal of the following analyses is to determine how changes in lung function (as determined by FEV_1) over time are related to the age and height of the child. Previous research has indicated that $\log(FEV_1)$ has an approximately linear relationship with age and log(height) in children and adolescents. To distinguish between the cross-sectional and longitudinal effects of age and log(height) on $\log(FEV_1)$, baseline and current values of these covariates were included in the model for the mean. Because these data are inherently unbalanced, accounting for the covariance among the repeated observation on the same child via a random effects structure is very appealing. Here we allow the intercept and slope for age to varying randomly from one child to another. Specifically, we consider the following model for $\log(FEV_1)$:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 \text{Age}_{ij} + \beta_3 \log(\text{Ht})_{ij} + \beta_4 \text{Age}_{i1} + \beta_5 \log(\text{Ht})_{i1} + b_{1i} + b_{2i}\text{Age}_{ij},$$

where Y_{ij} is the $\log(FEV_1)$ for the i^{th} child at the j^{th} occasion, and Age_{i1} and $\log(\text{Ht})_{i1}$ are the initial or baseline age and log(height) for the i^{th} child. In this model, β_2 and β_3 are the longitudinal effects of age and log(height), respectively, while $(\beta_2 + \beta_4)$ and $(\beta_3 + \beta_5)$ are the corresponding cross-sectional effects. That is, β_4 and β_5 represent the differences between the longitudinal and cross-sectional effects of age and log(height), respectively. (See Sections 9.4 and 9.5 for a more detailed discussion of, and an alternative representation of, models that allow for estimation of both longitudinal and cross-sectional effects.)

Preliminary analysis of the data revealed a measurement of FEV_1 that was clearly outlying. This observation was from a girl who had only a baseline measurement available. The observation was removed, and all subsequent analyses are based on the data from 299 girls (with a total of 1993 measurements). The REML estimates of the fixed effects are displayed in Table 8.2. Based on the magnitude of the estimates of β_4 and β_5, relative to their standard errors, there is evidence of a significant difference between the longitudinal and cross-sectional effects of age, but not of log(height). From a subject-matter point of view, the magnitudes of the longitudinal and cross-sectional effects of log(height) are quite similar (2.24 versus 2.46), whereas the magnitudes of the longitudinal and cross-sectional effects of age are strikingly different (0.024 versus 0.007). That is, the longitudinal and cross-sectional effects of age on changes in FEV_1 ($e^{0.024} \approx 1.025$ versus $e^{0.007} \approx 1.007$) are discernibly different. This may be due, in part, to the relatively small amount of variability in ages at baseline (relative to the variability in ages throughout the duration of the study), resulting in the cross-sectional effect of age being poorly estimated from the data at baseline alone; in Sections 9.4 and 9.5 we present an alternative model where estimation of the cross-sectional effect of age is based on measurements at all occa-

Table 8.2 Estimated regression coefficients (fixed effects) and standard errors for the log(FEV_1) data from the Six Cities Study.

Variable	Estimate	SE	Z
Intercept	−0.2883	0.0387	−7.45
Age	0.0235	0.0014	16.86
Log(Height)	2.2372	0.0435	51.39
Initial Age	−0.0165	0.0075	−2.21
Initial Log(Height)	0.2182	0.1455	1.50

sions. From the longitudinal effects of age and log(height), there is clear evidence that changes in log(FEV_1) are related to both age and height.

Next we consider the interpretation of the fixed effects estimates. The model for the mean, averaged over the distribution of the subject-specific random effects, is given by

$$E(Y_{ij}) = \beta_1 + \beta_2 \,\mathrm{Age}_{ij} + \beta_3 \log(\mathrm{Ht})_{ij} + \beta_4 \,\mathrm{Age}_{i1} + \beta_5 \log(\mathrm{Ht})_{i1}.$$

Furthermore this model can be re-expressed in terms of two models, a cross-sectional model and a longitudinal model. The former is given by

$$\begin{aligned} E(Y_{i1}) &= \beta_1 + \beta_2 \,\mathrm{Age}_{i1} + \beta_3 \log(\mathrm{Ht})_{i1} + \beta_4 \,\mathrm{Age}_{i1} + \beta_5 \log(\mathrm{Ht})_{i1} \\ &= \beta_1 + (\beta_2 + \beta_4) \,\mathrm{Age}_{i1} + (\beta_3 + \beta_5) \log(\mathrm{Ht})_{i1}, \end{aligned}$$

while the latter is given by

$$\begin{aligned} E(Y_{ij} - Y_{i1}) &= \beta_1 + \beta_2 \,\mathrm{Age}_{ij} + \beta_3 \log(\mathrm{Ht})_{ij} + \beta_4 \,\mathrm{Age}_{i1} + \beta_5 \log(\mathrm{Ht})_{i1} \\ &\quad - \{\beta_1 + \beta_2 \,\mathrm{Age}_{i1} + \beta_3 \log(\mathrm{Ht})_{i1} + \beta_4 \,\mathrm{Age}_{i1} + \beta_5 \log(\mathrm{Ht})_{i1}\} \\ &= \beta_2 \,(\mathrm{Age}_{ij} - \mathrm{Age}_{i1}) + \beta_3 \left\{\log(\mathrm{Ht})_{ij} - \log(\mathrm{Ht})_{i1}\right\}. \end{aligned}$$

We note that there are alternative ways to decompose the cross-sectional and longitudinal effects; this topic is explored in greater detail in Chapter 9 (see Sections 9.4 and 9.5).

The longitudinal effect of log(height), β_3, has interpretation in terms of the changes in mean log(FEV_1) for a single-unit increase in log(height), for any given change in age (e.g., during a two-year interval). Similarly the longitudinal effect of age, β_2, has interpretation in terms of the changes in mean log(FEV_1) for a one-year increase in age, for any given change in log(height). The coefficient for log(height), 2.24, is not directly interpretable because a single-unit change in log(height) corresponds to an almost threefold (or $e^{1.0} \approx 2.7$) increase in height. Instead, it is probably more

Table 8.3 Estimated covariance of the random effects and standard errors ($\times 100$) for the log(FEV_1) data from the Six Cities Study.

Parameter	Estimate	SE
$\text{Var}(b_{1i}) = g_{11}$	1.2207	0.1924
$\text{Cov}(b_{1i}, b_{2i}) = g_{12}$	−0.0435	0.0122
$\text{Var}(b_{2i}) = g_{22}$	0.0050	0.0010
$\text{Var}(\epsilon_i) = \sigma^2$	0.3629	0.0133

meaningful to consider the effect of a 10% increase in height. On the logarithmic scale this corresponds to a 0.1 increase in log(height), since $e^{0.1} \approx 1.1$. Therefore a 10% increase in height (corresponding to an approximate 0.1 increase in log(height)) is associated with a 0.224 increase in log(FEV_1). Note that a 0.224 increase in log(FEV_1) corresponds to a 25% increase in FEV_1 (since $e^{0.224} = 1.25$). On the other hand, the coefficient for age, 0.024, is more directly interpretable. The estimate of the longitudinal effect of age indicates that a single year increase in age is associated with a 0.024 increase in log(FEV_1) or an approximate 2.5% ($e^{0.024} \approx 1.025$) increase in FEV_1, for any fixed change in height.

Next consider the estimates[4] of the variances and covariances of the random effects (see Table 8.3). The marginal covariance among the repeated observations is a function of these variance and covariance parameters (and σ^2) and the ages of the child when the observations were obtained. The estimated correlations for annual measurements from ages 7 to 18 are displayed in Table 8.4, and these results indicate that there is strong positive correlation among measurements of log(FEV_1) that declines by a modest amount over the 11 years of follow-up. This pattern of correlation reinforces an observation that we made in earlier chapters of the book: the correlation among repeated measurements of many health outcomes rarely decays to zero, even when they are separated by many years.

Finally, we note that the correlation among repeated measurements has been accounted for by the introduction of random intercepts and slopes for age. Alternatively, we could have considered a random effects model with randomly varying slopes for log(height). By assuming that the slope for log(height) varies randomly for individuals, we can also induce covariance among the repeated observations but with correlations that are a function not of age, but of the height of the child. For illustra-

[4] Z statistics are intentionally omitted from Table 8.3 because, in general, we do not recommend testing hypotheses about the covariance parameters using Wald tests. In particular, the sampling distribution of a variance parameter estimate does not have an approximate normal distribution when the sample size is relatively small and the population variance is close to zero (see Chapter 4, Section 4.4).

Table 8.4 Estimated marginal correlations among repeated measures of log(FEV_1) between the ages of 7 and 18.

Age (years)											
7	8	9	10	11	12	13	14	15	16	17	18
1.00	0.70	0.69	0.68	0.67	0.66	0.64	0.62	0.60	0.58	0.56	0.54
0.70	1.00	0.70	0.69	0.68	0.67	0.66	0.65	0.63	0.61	0.60	0.58
0.69	0.70	1.00	0.70	0.70	0.69	0.68	0.67	0.66	0.64	0.63	0.61
0.68	0.69	0.70	1.00	0.70	0.70	0.70	0.69	0.68	0.67	0.66	0.64
0.67	0.68	0.70	0.70	1.00	0.71	0.71	0.70	0.70	0.69	0.68	0.67
0.66	0.67	0.69	0.70	0.71	1.00	0.72	0.72	0.71	0.71	0.70	0.70
0.64	0.66	0.68	0.70	0.71	0.72	1.00	0.73	0.73	0.73	0.72	0.72
0.62	0.65	0.67	0.69	0.70	0.72	0.73	1.00	0.74	0.74	0.74	0.74
0.60	0.63	0.66	0.68	0.70	0.71	0.73	0.74	1.00	0.75	0.75	0.75
0.58	0.61	0.64	0.67	0.69	0.71	0.73	0.74	0.75	1.00	0.76	0.76
0.56	0.60	0.63	0.66	0.68	0.70	0.72	0.74	0.75	0.76	1.00	0.77
0.54	0.58	0.61	0.64	0.67	0.70	0.72	0.74	0.75	0.76	0.77	1.00

tive purposes we considered the following model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 \text{Age}_{ij} + \beta_3 \log(\text{Ht})_{ij} + \beta_4 \text{Age}_{i1} + \beta_5 \log(\text{Ht})_{i1} + b_{1i} + b_{2i} \log(\text{Ht})_{ij}.$$

The REML estimates of the fixed effects are displayed in Table 8.5 and are qualitatively very similar to those presented in Table 8.2. The reader might then ask which model is more appropriate for the data at hand, a model with randomly varying slopes for age or randomly varying slopes for log(height)? Fortunately, since both models have the same number of covariance parameters, we can make that judgement based on a direct comparison of their maximized REML log-likelihoods. For the model with randomly varying slopes for age the maximized REML log-likelihood is 2283.9, while for the model with randomly varying slopes for log(height) the maximized REML log-likelihood is 2294.7. The comparison of the maximized log-likelihoods indicates that the model with randomly varying slopes for log(height) is to be preferred. For illustrative purposes we also considered a random effects model with randomly varying slopes for both age and log(height). By assuming that the slopes for age and log(height) vary randomly this would induce covariances among the repeated observations that are functions of both the age and height of the child. For the latter model the maximized REML log-likelihood is 2294.9 and does not lead to a discernible improvement in fit over the model with randomly varying slopes for log(height) only.

Table 8.5 Estimated regression coefficients (fixed effects) and standard errors for the log(FEV_1) data from the Six Cities Study.

Variable	Estimate	SE	Z
Intercept	−0.2846	0.0390	−7.30
Age	0.0233	0.0012	18.65
Log(Height)	2.2523	0.0461	48.82
Initial Age	−0.0163	0.0074	−2.19
Initial Log(Height)	0.1808	0.1455	1.24

Formally, the likelihood ratio test statistic, $G^2 = 0.4$, can be compared to the critical values in the third row ($q = 2$) of Table C.1 in Appendix C.

Study of Influence of Menarche on Changes in Body Fat Accretion

The second illustration uses longitudinal data from a prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study (Bandini et al., 2002; Phillips et al., 2003). The data represent a subset of the study materials and should not be used to draw substantive conclusions.

It is known that increases in body fatness in girls begin just before or around menarche. Although it has been presumed that the increase in body fatness levels off approximately four years after menarche, these changes in body fat accretion had not been studied in population-based samples. Naumova et al. (2001) examined changes in body fat before and after menarche. At the start of the study, all of the girls were pre-menarcheal and non-obese, as determined by a triceps skinfold thickness less than the 85th percentile. All girls were followed over time according to a schedule of annual measurements until four years after menarche. The final measurement was scheduled on the fourth anniversary of their reported date of menarche.

At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis. Percent body fat (%BF) was derived from three basic measurements of body weight (Wt. in kg), height (Ht. in cm), and bioelectric impedance resistance (R). Percent body fat is calculated using the equation:

$$\%\mathrm{BF} = \left(1 - \frac{\mathrm{TBW}}{0.73}\mathrm{Wt}\right) \times 100\%,$$

where total body water, $\mathrm{TBW} = (0.7\mathrm{Ht}^2/\mathrm{R}) - 0.32$.

In this section we present an analysis of the changes in percent body fat before and after menarche. For the purposes of these analyses, "time" is coded as time since menarche and can be positive or negative. Although the measurement protocol is the

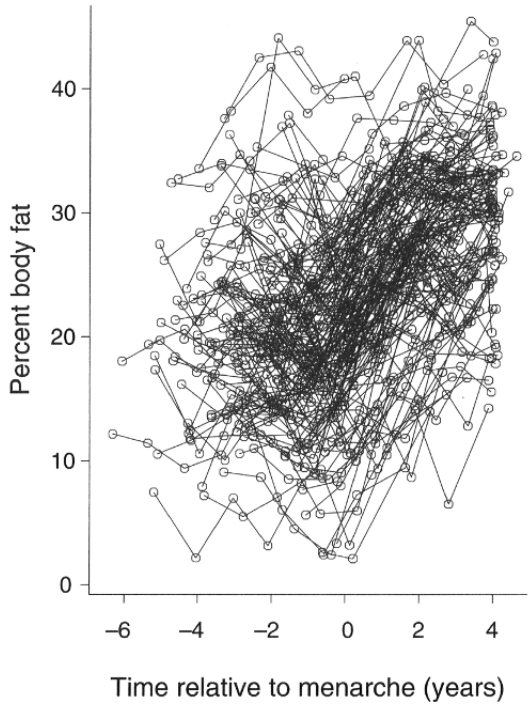

Fig. 8.5 Time plot of percent body fat against time, relative to age of menarche (in years).

same for all girls and the study design is balanced if the timing of measurement is defined as the time since the baseline measurement, it is inherently unbalanced when the timing of measurements is defined as the time since a girl experienced menarche.

In this data set there are a total of 1049 percent body fat measurements, with an average of 6.4 measurements per subject. The numbers of measurements per subject pre- and post-menarche are approximately equal, with 497 measurements for the pre-menarcheal period (producing an average of 3.1 measurements per subject) and 552 measurements for the post-menarcheal period (producing an average of 3.5 measurements per subject). In this sample the average age at menarche was 12.8 years.

Figure 8.5 shows a time plot of the individual response profiles (where time is relative to the individual age at menarche). This graph reveals some information about the greater variability of measurement times before menarche. However, it is difficult to discern whether the changes in percent body fat in the pre-menarcheal period are similar to the changes in the post-menarcheal period. In Figure 8.6 the trend in the mean response is assessed using a *lowess* smoothed curve. Recall that *lowess* is a nonparametric, robust regression method that traces the salient features

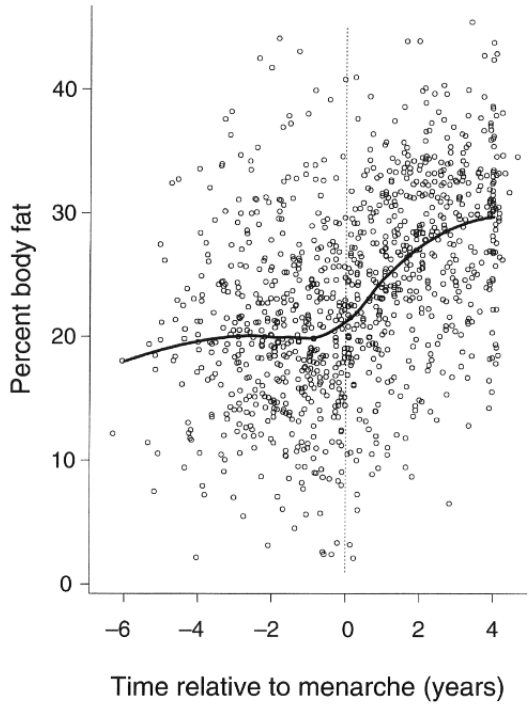

Fig. 8.6 Time plot of percent body fat against time, relative to age of menarche (in years), with *lowess* smoothed curve.

of the mean response as a function of time while making only minimal assumptions about the form of the relationship. The *lowess* curve reveals that the mean response remains relatively flat during the pre-menarcheal period and then rises sharply during the post-menarcheal period.

In the following analysis we consider the hypothesis that percent body fat accretion increases linearly with age, but with different slopes before and after menarche. Specifically, we assume that each girl has a piecewise linear spline growth curve with a knot at the time of menarche. That is, each girl's growth curve can be described with an intercept and two slopes, one slope for the changes in response before menarche, another slope for the changes in response after menarche. Note that unlike the piecewise linear splines considered in Section 6.3, the knot is not the same age for all subjects.

Let t_{ij} denote the time of the j^{th} measurement on the i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche). We fit the following linear mixed effects model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

Table 8.6 Estimated regression coefficients (fixed effects) and standard errors for the percent body fat data.

Variable	Estimate	SE	Z
Intercept	21.3614	0.5646	37.84
Time	0.4171	0.1572	2.65
$(\text{Time})_+$	2.0471	0.2280	8.98

Table 8.7 Estimated covariance of the random effects and standard errors for the percent body fat data.

Parameter	Estimate	SE
$\text{Var}(b_{1i}) = g_{11}$	45.9413	5.7393
$\text{Var}(b_{2i}) = g_{22}$	1.6311	0.4331
$\text{Var}(b_{3i}) = g_{33}$	2.7497	0.9635
$\text{Cov}(b_{1i}, b_{2i}) = g_{12}$	2.5263	1.2185
$\text{Cov}(b_{1i}, b_{3i}) = g_{13}$	−6.1096	1.8730
$\text{Cov}(b_{2i}, b_{3i}) = g_{23}$	−1.7505	0.5980
$\text{Var}(\epsilon_i) = \sigma^2$	9.4732	0.5443

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$. In this model, $(\beta_1 + b_{1i})$ is the intercept for the i^{th} subject and has interpretation as the true percent body fat at menarche (when $t_{ij} = 0$). Of note, the actual percent body fat at menarche is not observed and cannot be directly estimated from the data at hand. As a result we use the term "true" percent body fat at menarche to remind the reader that this is a parameter in the model. Similarly $(\beta_2 + b_{2i})$ is the i^{th} subject's slope, or rate of change in percent body fat during the pre-menarcheal period. Finally, the i^{th} subject's slope during the post-menarcheal period is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$. Since the overall goal of the analysis is to assess whether the population slopes for fat accretion differ before and after menarche, this can be translated into the null hypothesis, H_0: $\beta_3 = 0$.

The REML estimates of the fixed effects and the variance components are displayed in Tables 8.6 and 8.7, respectively. Based on the magnitude of the estimate of β_3, relative to its standard error, there is a significant difference between the slopes be-

Table 8.8 Estimated marginal correlations (on the off-diagonals) among repeated measures of percent body fat between 4 years pre- and post-menarche, with estimated variances along the main diagonal.

Time (relative to menarche)								
−4	−3	−2	−1	0	1	2	3	4
61.3	0.82	0.78	0.71	0.61	0.60	0.57	0.52	0.47
0.82	54.9	0.81	0.76	0.70	0.68	0.64	0.60	0.54
0.78	0.81	51.8	0.80	0.76	0.74	0.71	0.66	0.60
0.71	0.76	0.80	52.0	0.81	0.79	0.76	0.71	0.64
0.61	0.70	0.76	0.81	55.4	0.81	0.78	0.73	0.66
0.60	0.68	0.74	0.79	0.81	49.1	0.79	0.76	0.70
0.57	0.64	0.71	0.76	0.78	0.79	44.6	0.77	0.74
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	0.76
0.47	0.54	0.60	0.64	0.66	0.70	0.74	0.76	40.8

fore and after menarche. The estimate of the population mean pre-menarcheal slope is 0.42, which is statistically significant at the 0.05 level. This estimated slope is rather shallow and indicates that the annual rate of body fat accretion is less that 0.5%. Note that the estimated variance of b_{2i} is 1.63, indicating that there is substantial variability from girl to girl in rates of fat accretion and that many girls are losing body fat while others are gaining body fat during the pre-menarcheal period. For example, approximately 95% of girls have changes in percent body fat between −2.09% and 2.92% $\left(\text{i.e., } 0.42 \pm 1.96 \times \sqrt{1.63}\right)$. The estimate of the population mean post-menarcheal slope is 2.46 (with SE = 0.12), which is statistically significant at the 0.05 level. This indicates that the annual rate of body fat accretion is approximately 2.5%, almost six times higher than the corresponding rate in the pre-menarcheal period. The estimated variance of the individual slopes during the post-menarcheal period, $\text{Var}(b_{2i} + b_{3i})$, is 0.88 (or $[1.63 + 2.75 - 2 \times 1.75]$), indicating that there is less variability in the slopes after menarche. For example, approximately 95% of girls have changes in percent body fat between 0.62% and 4.30% $\left(\text{i.e., } 2.46 \pm 1.96 \times \sqrt{0.88}\right)$. In other words, more than 95% of girls are expected to have increases in body fat during the post-menarcheal period, while substantially fewer (approximately 63%) are expected to have increases in body fat during the pre-menarcheal period.

The estimated marginal correlations among annual measurements of percent body fat, based on the estimated covariances among the random effects, are displayed in Table 8.8. These results indicate that there is strong positive correlation (approximately 0.8) between annual measurements of percent body fat. Although the strength of the correlation declines over time, it does not decay to zero even when measure-

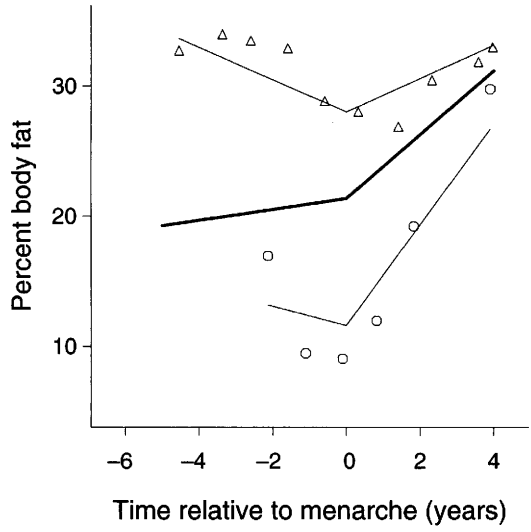

Fig. 8.7 Population average curve (thicker solid line) and empirical BLUPs for two randomly selected girls.

ments are taken eight years apart. In general, the variability of percent body fat is greater in the pre-menarcheal period.

The mixed effects model can be used to obtain estimates of each girl's growth trajectory over time. Figure 8.7 displays the estimated population mean growth curve and the predicted (empirical BLUP) growth curves for two girls, based on the fixed and random effects estimates reported in Tables 8.6 and 8.7. Note that the two girls selected for display in Figure 8.7 differ in the number of measurements that were obtained (with 6 and 10 measurements, respectively). A noticeable feature of the predicted growth curves is that there is more shrinkage toward the population mean curve when fewer data points are available. That is, the predicted growth curve for the girl with only 6 data points is pulled closer to the population mean curve (or further away from her own data points) while the predicted growth curve for the girl with 10 observation follows her data more closely. This feature becomes more apparent when the empirical BLUPs are compared to the ordinary least squares (OLS) estimates based only on the longitudinal observations from each girl (see Figure 8.8). Examination of Figure 8.8 reveals that the empirical BLUP for the girl with 10 observations is largely based on her longitudinal observations. On the other hand, the empirical BLUP for the girl with 6 observations "borrows strength" from the population mean curve. This is a characteristic feature of the empirical BLUPs that was noted in Sections 8.6 and 8.7. When there is less information available for estimating an individual's growth

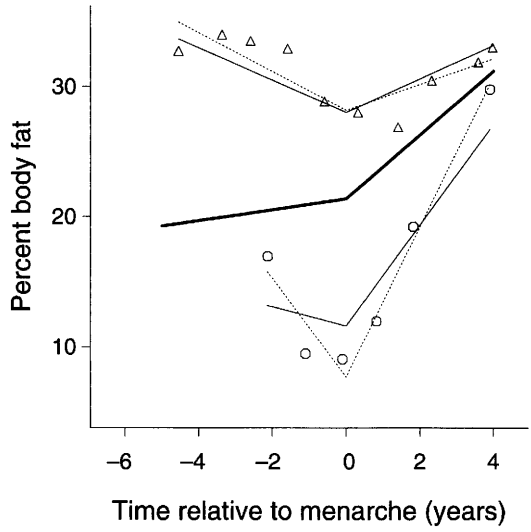

Fig. 8.8 Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

curve, there is a greater "borrowing of strength" from the data obtained on all girls in the study.

Finally, we can use these data to illustrate a hybrid random effects and covariance pattern model by fitting the following model to the percent body fat:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + U_i(t_{ij}) + \epsilon_{ij},$$

where the $U_i(t_{ij})$ are assumed to have a normal distribution, with zero mean, variance σ_u^2, and correlation

$$\text{Corr}\{U_i(t_{ij}), U_i(t_{ik})\} = \rho(|t_{ij} - t_{ik}|).$$

The $U_i(t_{ij})$ induce serial correlation among the responses, such that the correlation becomes weaker as the time separation increases. Two popular choices for $\rho(|t_{ij} - t_{ik}|)$ are the exponential correlation function,

$$\rho(|t_{ij} - t_{ik}|) = e^{-\alpha|t_{ij} - t_{ik}|},$$

and the Gaussian correlation function,

$$\rho(|t_{ij} - t_{ik}|) = e^{-\alpha|t_{ij} - t_{ik}|^2},$$

for some $\alpha > 0$. Finally, the ϵ_{ij} are the usual sampling or measurement errors and these are assumed to be independent, with mean zero and variance σ^2.

Table 8.9 Comparison of the maximized (REML) log-likelihoods and AIC for the mixed effects model and the hybrid models with exponential and Gaussian serial correlation.

Model	−2 (REML) Log-Likelihood	AIC
Mixed Effects	6062.4	6076.4
Hybrid: Exponential Serial Correlation	5999.9	6007.9
Hybrid: Gaussian Serial Correlation	5991.2	5999.2

Table 8.10 Estimated regression coefficients (fixed effects) and standard errors for the hybrid model with Gaussian serial correlation.

Variable	Estimate	SE	Z
Intercept	21.2918	0.5400	39.43
Time	0.2168	0.1439	1.51
$(\text{Time})_+$	2.1655	0.2331	9.29

We considered the goodness of fit of the hybrid model when the serial correlation function is exponential and Gaussian. Table 8.9 displays the maximized (REML) log-likelihood and AIC for the hybrid models with exponential and Gaussian serial correlation; the maximized (REML) log-likelihood and AIC for the mixed effects model considered previously are also displayed. These results indicate that the hybrid model with Gaussian serial correlation fits the data best, since it has the largest maximized log-likelihood (when compared to the hybrid model with exponential serial correlation) and the smallest AIC (when compared to the mixed effects model).

The REML estimates of the fixed effects from the hybrid model with Gaussian serial correlation are displayed in Table 8.10. The estimates of β are similar to those reported in Table 8.6. In particular, the estimate of β_3 is very similar, and when compared to its standard error, there is a significant difference between the slopes before and after menarche. On the other hand, the estimate of the population mean pre-menarcheal slope is 0.22, and is no longer statistically significant at the 0.05 level. Overall, the substantive conclusions are very similar in the two sets of analyses: there is at most a very weak pre-menarcheal slope, indicating that the annual rate of body rate accretion is very modest (0.2–0.4%), while the annual rate of fat accretion during the post-menarcheal period is discernibly greater (approximately

2.4–2.5%) than the corresponding rate in the pre-menarcheal period. Of note, an attempt to fit an extended mixed effects model (with randomly varying intercepts and pre- and post-menarcheal slopes) by incorporation of a Gaussian serial correlation component failed to converge. This lack of convergence was taken as an indication that the observed data simply do not support the need for both randomly varying slopes and serially correlated residuals. As was mentioned in Section 8.2, there can be identifiability problems with the hybrid model unless the random effects structure is kept very simple (e.g., random intercepts only). That is, there may be insufficient information in the data at hand to support separate estimation of randomly varying slopes, serially correlated residuals, and measurement errors.

Randomized Study of Dual or Triple Combinations of HIV-1 Reverse Transcriptase Inhibitors

The final illustration uses data from a randomized, double-blind, study of AIDS patients with advanced immune suppression (CD4 counts of $\leq$ 50 cells/mm^3) (Henry et al., 1998). Patients in AIDS Clinical Trial Group (ACTG) Study 193A were randomized to dual or triple combinations of HIV-1 reverse transcriptase inhibitors. Specifically, patients were randomized to one of four daily regimens containing 600 mg of zidovudine: zidovudine alternating monthly with 400 mg didanosine, zidovudine plus 2.25 mg of zalcitabine, zidovudine plus 400 mg of didanosine, or zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine (triple therapy). For the analyses presented here, we focus on the comparison of the first three treatment regimens (dual therapy) with the fourth (triple therapy).

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to mistimed measurements and missing data that resulted from skipped visits and dropout (see data for four randomly selected subjects presented in Table 8.11). The number of measurements of CD4 counts during the first 40 weeks of follow-up varied from 1 to 9, with a median of 4. The goal of our analyses is to compare the dual and triple therapy groups in terms of short-term changes in CD4 counts from baseline to week 40 (approximately 10 months of follow-up). The analyses are based on log transformed CD4 counts, log(CD4 counts +1), available on 1309 patients.

In Figure 8.9 the trend in the mean response in the dual and triple therapy groups is assessed using *lowess* smoothed curves. The curves reveal a modest decline in the mean response during the first 16 weeks for the dual therapy group, followed by a steeper decline from week 16 to week 40. In contrast, for the triple therapy group, the mean response increases during the first 16 weeks and declines thereafter. The rate of decline from week 16 to week 40 appears to be similar for the two groups. A note of caution: because there is a substantial amount of missing data, the plot of the mean response over time can be potentially misleading unless the data are missing completely at random (MCAR). When data are missing, and especially when there is attrition of subjects whose responses are different from those who remain in the study, a plot of the mean response over time can be deceptive; the observed changes

Table 8.11 Data on log CD4 counts for four randomly selected subjects from ACTG study 193A.

Subject ID	Group	Time	log(CD4 + 1)
56	0	0.0	1.7047
56	0	8.1	1.7918
56	0	16.1	0.6932
56	0	25.4	1.0986
56	0	33.4	0.6932
56	0	39.1	0.6932
544	1	0.0	3.3844
544	1	7.6	3.2189
544	1	15.9	2.1972
544	1	31.9	1.6094
736	0	0.0	3.7495
736	0	8.9	3.4965
736	0	18.9	3.1780
736	0	30.9	2.7726
986	1	0.0	4.4659
986	1	17.4	3.3322
986	1	30.9	3.5553
986	1	39.6	3.3673

Note: Group = 1 if randomized to triple therapy, Group = 0 if randomized to dual therapy.

in the mean response may reflect the pattern of missingness or the attrition, and not within-individual change. (See Chapters 17 and 18 for a more detailed discussion of this issue.)

Next we consider a model for the mean response that allows the rates of change before and after week 16 to differ within and between groups. Specifically, we assume that each patient has a piecewise linear spline with a knot at week 16. That is, each patient's response trajectory can be described with an intercept and two slopes—one slope for the changes in response before week 16, another slope for the changes in response after week 16. The average slopes for changes in response before and after week 16 are allowed to vary by group. Because this is a randomized study, the mean response at baseline is assumed to be the same in the two groups.

Letting t_{ij} denote the time since baseline for the j^{th} measurement on the i^{th} subject (with $t_{ij} = 0$ at baseline), we consider the following linear mixed effects model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \text{Group}_i \times t_{ij} + \beta_5 \text{Group}_i \times (t_{ij} - 16)_+ + b_{1i} + b_{2i}\, t_{ij} + b_{3i}\, (t_{ij} - 16)_+,$$

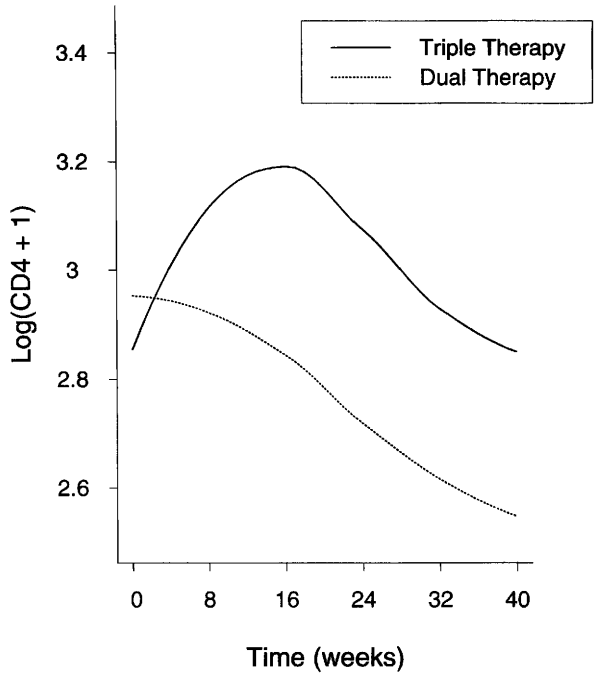

Fig. 8.9 Lowess smoothed curves of log(CD4 + 1) against time (in weeks), for subjects in the dual and triple therapy groups in ACTG study 193A.

where $\text{Group}_i = 1$ if the i^{th} subject is randomized to triple therapy, and $\texttt{Group}_\texttt{i} = 0$ otherwise; $(t_{ij} - 16)_+ = t_{ij} - 16$ if $t_{ij} > 16$ and $(t_{ij} - 16)_+ = 0$ if $t_{ij} \leq 16$. In this model, $(\beta_1 + b_{1i})$ is the intercept for the i^{th} subject and has interpretation as the true log CD4 count at baseline (when $t_{ij} = 0$). Similarly $(\beta_2 + b_{2i})$ is the i^{th} subject's slope, or rate of change in log CD4 counts from baseline to week 16, if randomized to dual therapy; $(\beta_2 + \beta_4 + b_{2i})$ is the i^{th} subject's slope if randomized to triple therapy. Finally, the i^{th} subject's slope from week 16 to week 40 is given by $\{(\beta_2 + \beta_3) + (b_{2i} + b_{3i})\}$ if randomized to dual therapy and $\{(\beta_2 + \beta_3 + \beta_4 + \beta_5) + (b_{2i} + b_{3i})\}$ if randomized to triple therapy. The null hypothesis of no treatment group differences in the changes in log CD4 counts can be expressed as H_0: $\beta_4 = \beta_5 = 0$.

The REML estimates of the fixed effects are displayed in Table 8.12. A test of H_0: $\beta_4 = \beta_5 = 0$ yields a Wald test, $W^2 = 59.21$, with 2 degrees of freedom ($p < 0.0001$); the corresponding likelihood ratio test yields $G^2 = 57.99$, with 2 degree of freedom ($p < 0.0001$). Based on the magnitude of the estimate of β_4, relative to its standard error, there is a significant group difference in the rates of change from baseline to week 16. In the dual therapy group, there is a significant decrease in the mean of the log CD4 counts from baseline to week 16. The estimated change during the first 16 weeks is -0.12, or 16×-0.0073. On the untransformed scale, this

Table 8.12 Estimated regression coefficients (fixed effects) and standard errors for the log CD4 counts.

Variable	Estimate	SE	Z
Intercept	2.9415	0.0256	114.81
t_{ij}	−0.0073	0.0020	−3.70
$(t_{ij} - 16)_+$	−0.0120	0.0032	−3.79
Group × t_{ij}	0.0269	0.0039	6.98
Group × $(t_{ij} - 16)_+$	−0.0277	0.0062	−4.47

corresponds to an approximate 10% decrease in CD4 counts (since $e^{-0.12} = 0.89$). In contrast, in the triple therapy group, there is a significant increase in the mean response. The estimated change during the first 16 weeks in the triple therapy group is 0.31, or $16 \times (-0.0073 + 0.0269)$; the estimated slope for the triple therapy group, 0.0196, has a standard error of 0.0033. On the untransformed scale, this corresponds to an approximate 35% increase in CD4 counts (since $e^{0.31} = 1.36$).

The lowess curves in Figure 8.9 suggest that the rate of decline from week 16 to week 40 is similar for the two groups. The null hypothesis of no treatment group differences in the rates of change in log CD4 counts from week 16 to week 40 can be expressed as H_0: $\beta_4 + \beta_5 = 0$ (or H_0: $\beta_4 = -\beta_5$). The estimates of β_4 and β_5 in Table 8.12 appear to support the null hypothesis since they are of similar magnitude but opposite signs. A test of the null hypothesis, H_0: $\beta_4 + \beta_5 = 0$, yields a Wald test, $W^2 = 0.07$, with 1 degree of freedom ($p > 0.75$); the corresponding likelihood ratio test yields $G^2 = 0.07$, with 1 degree of freedom ($p > 0.75$).

The estimated variances of the random effects in Table 8.13 indicate that there is substantial variability from patient to patient in baseline CD4 counts and the rates of change in CD4 counts. For example, although many patients randomized to triple therapy show increases in CD4 counts during the first 16 weeks, some patients have declining CD4 counts. Specifically, approximately 95% of patients randomized to triple therapy are expected to have changes in log CD4 counts from baseline to week 16 between −0.64 and 1.27 $\left(\text{or } 16 \times \left[0.0196 \pm 1.96 \times \sqrt{0.000923}\right]\right)$. That is, approximately 26% of patients are expected to have decreases in CD4 counts during the first 16 weeks of triple therapy. There is also a substantial component of variability due to measurement error.

In a clinical trial it is often of interest to predict the direction and magnitude of the treatment effect for patients with specific covariate values. In the physician–patient context, for example, these predictions can be used to identify those patients who do not respond well to their assigned therapy. When there is interest in subject-specific predictions, we must consider the relative magnitudes of the between-subject and within-subject variability. When the within-subject or measurement error variability

Table 8.13 Estimated covariance ($\times 1000$) of the random effects and standard errors for the log CD4 counts.

Parameter	Estimate	SE
$\text{Var}(b_{1i}) = g_{11}$	585.742	34.754
$\text{Var}(b_{2i}) = g_{22}$	0.923	0.160
$\text{Var}(b_{3i}) = g_{33}$	1.240	0.395
$\text{Cov}(b_{1i}, b_{2i}) = g_{12}$	7.254	1.805
$\text{Cov}(b_{1i}, b_{3i}) = g_{13}$	−12.348	2.730
$\text{Cov}(b_{2i}, b_{3i}) = g_{23}$	−0.919	0.236
$\text{Var}(\epsilon_i) = \sigma^2$	306.163	10.074

is relatively large, the observed response profile for a subject is unreliable and a better prediction can be obtained by "borrowing strength" from the data on all of the subjects. Next we consider the prediction of patients' response trajectories from the following linear mixed effects model that also includes gender and baseline age:

$$\begin{aligned} E(Y_{ij}|b_i) &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \text{Group}_i \times t_{ij} - \beta_4 \text{Group}_i \times (t_{ij} - 16)_+ \\ &\quad + \beta_5 \text{Age}_i + \beta_6 \text{Gender}_i + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+ \\ &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 \text{Group}_i \times \{t_{ij} - (t_{ij} - 16)_+\} \\ &\quad + \beta_5 \text{Age}_i + \beta_6 \text{Gender}_i + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+, \end{aligned}$$

where Age_i is the baseline age (in years) of the patient, $\text{Gender}_i = 1$ if the i^{th} patient is male, and $\text{Gender}_i = 0$ otherwise. In this model the mean rate of change from baseline to week 16 can differ in the two groups (with slopes of β_2 and $\beta_2 + \beta_4$ respectively), but the mean rate of change from week 16 to week 40 is assumed to be the same (with slope of $\beta_2 + \beta_3$).

The REML estimates of the fixed effects are displayed in Table 8.14 and the substantive conclusions about the treatment group comparisons are similar to those obtained from Table 8.12. Controlling for gender and age at baseline, there is a 10% decrease in CD4 counts (since $e^{16 \times -0.0072} = e^{-0.12} = 0.89$) in the dual therapy group. In contrast, in the triple therapy group, there is a 35% increase in CD4 counts (since $e^{16 \times (-0.0072 + 0.0263)} = e^{0.31} = 1.36$). In both treatment groups, there is a significant decline in the mean response from week 16 to week 40, corresponding to an approximate 40% decrease in CD4 counts (since $e^{24 \times (-0.0072 - 0.0124)} = e^{-0.47} = 0.63$).

Table 8.14 Estimated regression coefficients (fixed effects) and standard errors for the revised model for the log CD4 counts.

Variable	Estimate	SE	Z
Intercept	2.6457	0.1280	20.67
t_{ij}	−0.0072	0.0019	−3.71
$(t_{ij} - 16)_+$	−0.0124	0.0029	−4.33
Group × $\{t_{ij} - (t_{ij} - 16)_+\}$	0.0263	0.0034	7.68
Age	0.0100	0.0030	3.31
Gender	−0.0927	0.0754	−1.23

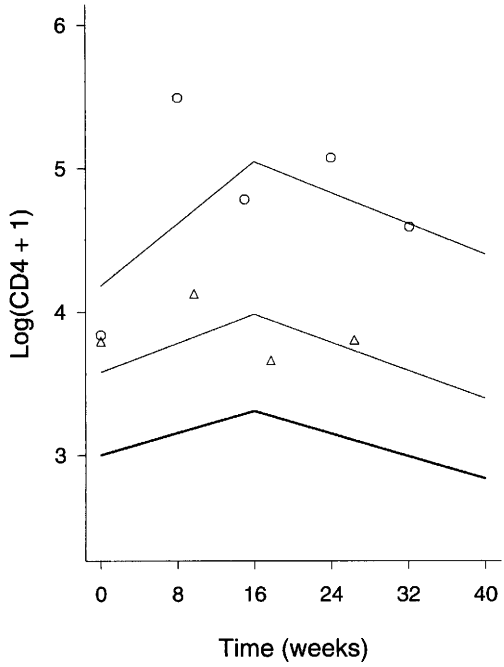

Fig. 8.10 Population average curve (thicker solid line) and empirical BLUPs for two male patients, aged 45, with similar baseline CD4 counts, and treated with triple therapy.

The inclusion of the random effects in the model allows each patient's response trajectory to be described with an intercept and two slopes, one slope for the changes in response before week 16, another slope for the changes in response after week 16. Based on the REML estimates of the fixed effects and variance components, the predicted (or BLUP) trajectory for each patient can be obtained. Figure 8.10 displays the estimated population mean curve and the predicted curves for two male patients, aged 45, and with similar baseline CD4 counts, who were randomized to triple therapy.

In general, the empirical BLUPs, or the predictions of summary measures (e.g., predicted area under the curve for a patient), can be used to identify those patients who have or have not responded well to their assigned therapy. In the physician–patient context, these predictions may be far more relevant than knowledge of the population mean curve. The appealing feature of the linear mixed effects model analysis is that it allows inferences about both the population trends and individual-specific trajectories.

8.9 COMPUTING: FITTING LINEAR MIXED EFFECTS MODELS USING PROC MIXED IN SAS

To fit linear mixed effects models we need to make use of the RANDOM statement in PROC MIXED. The RANDOM statement is used to define all effects that are considered to be random. Specifically, the RANDOM statement is used to define the covariates in the design matrix, Z_i, for the random effects, b_i. Ordinarily these will be a subset of the covariates included on the MODEL statement. (Recall that the covariates in the design matrix, X_i, for the fixed effects appear in the MODEL statement.) While the MODEL statement is used to define the design matrix for the fixed effects and the RANDOM statement is used to define the design matrix for the random effects, note that an intercept is included by default in the former but not the latter. That is, unlike the MODEL statement, PROC MIXED does not include an intercept in the RANDOM statement by default. However, you can specify INTERCEPT (or INT) as a random effect on the RANDOM statement. The RANDOM statement is also used to specify the structure of the covariance matrix for the random effects, G. The structure of G is specified using the TYPE=option. The random effects can be assumed to be correlated (TYPE=UN) or uncorrelated (TYPE=VC); ordinarily, covariance pattern models are not used to account for the covariance among the random effects. For reasons discussed in Section 8.2, we recommend using an unstructured covariance matrix (TYPE=UN) for G. To ensure that the unstructured covariance matrix for the random effects is constrained to be positive-definite, the TYPE=FAO(q) option can be used (where q is the number of random effects). The latter option can be useful when the TYPE=UN option yields an estimated G matrix that is not positive-definite.

For example, to fit a model with randomly varying intercepts and slopes to data from two or more groups measured repeatedly over time, we can use the illustrative SAS commands given in Table 8.15. Note that the SUBJECT option on the RANDOM statement is used in the same manner as on the REPEATED statement and denotes a

Table 8.15 Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
    CLASS  id  group;
    MODEL  y=group  time  group*time / S  CHISQ;
    RANDOM  INTERCEPT  time / TYPE=UN  SUBJECT=id  G  V;
```

variable that distinguishes clusters of correlated responses. By including a variable in the SUBJECT option (e.g., a subject identifier), pairs of observations with distinct values of that variable are regarded as independent. Pairs of observations with the same values of that variable share common values of the random effects.

Various options can be included on the RANDOM statement. The option G requests that the estimates of the variances and covariances of the random effects be displayed. The option GCORR requests that the estimates of the correlations among the random effects be displayed. The option V requests that the estimates of the marginal covariance matrix, averaged over the distribution of the random effects, be displayed for the first subject. That is, the option V produces estimates of $\Sigma_i = \text{Cov}(Y_i) = Z_i G Z_i' + \sigma^2 I_{n_i}$. Finally, when there is interest in predicting the random effects, the SOLUTION (or S) option can be used to request that the estimated BLUPs for the random effects, $\widehat{b_i}$, be displayed (in addition to standard errors for predictions based on the expression for $\text{Var}(\widehat{b_i} - b_i)$ given in Section 8.7). Alternatively, the predicted values of the response, $\widehat{Y_i} = X_i\widehat{\beta} + Z_i\widehat{b_i}$, can be requested by using the OUTPRED (or OUTP) option on the MODEL statement. This option specifies a SAS data-set than contains the predicted values of Y_i, denoted by the variable name `Pred`, and some related quantities. For example, to obtain the estimated BLUPs for the random effects and predicted values of the response, $\widehat{Y_i}$, we can use the illustrative SAS commands given in Table 8.16. The OUTPRED option specifies an output SAS data-set containing the predicted values, $\widehat{Y_i}$, whereas the SOLUTION option on the RANDOM statement requests that the estimated BLUPs be produced as part of the standard output from PROC MIXED. Inclusion of the Output Delivery System (ODS) statement creates a SAS data-set containing the estimated BLUPs. Predicted values of the outcome, at occasions other than those actually observed, can also be obtained by including "pseudo-observations" in the data set that have missing values for the outcome variable and the desired values of the covariates.

The alert reader would have noticed that the residual error variance, σ^2, has not been included on the RANDOM statement. Instead, it is included in an implicit REPEATED statement. Recall that the repeated statement is used to specify assumptions about the nature of the covariance among the errors. When the REPEATED statement is not included in PROC MIXED, it is assumed, by default, that the covariance

Table 8.16 Illustrative commands for obtaining the estimated BLUPs and the predicted responses from a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
ODS OUTPUT SOLUTIONR=bluptable;

PROC MIXED;
   CLASS  id  group;
   MODEL  y=group  time  group*time / S  CHISQ  OUTPRED=yhat;
   RANDOM INTERCEPT  time / TYPE=UN SUBJECT=id SOLUTION G V;

PROC PRINT DATA=yhat;
   VAR  id  group  time  y  Pred;

PROC PRINT DATA=bluptable;
```

Table 8.17 Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, within-subject errors with an exponential covariance, and independent measurement errors using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS  id  group;
   MODEL  y=group  time  group*time / S  CHISQ;
   REPEATED / TYPE=SP(EXP)(time)  LOCAL  SUBJECT=id;
   RANDOM INTERCEPT  time / TYPE=UN  SUBJECT=id  G  V;
```

among the errors, $R_i = \sigma^2 I_{n_i}$. To fit hybrid models that include both random effects and correlated errors, it is necessary to include both the RANDOM statement and the REPEATED statement. For example, to fit a hybrid model with (1) randomly varying intercepts and slopes, (2) within-subject errors with an exponential covariance structure, and (3) independent measurement or sampling errors, we can use the illustrative SAS commands given in Table 8.17. On the REPEATED statement we use the option TYPE=SP(EXP)(time) to specify an exponential covariance structure for the within-subject errors that depends on time. This command exploits the spatial covariance structures option built into PROC MIXED. Finally, the option LOCAL requests that a diagonal matrix, $\sigma^2 I_{n_i}$, be added to the exponential covariance structure for R_i.

8.10 FURTHER READING

Useful reviews of the linear mixed effects models, targeted at applied researchers, can be found in the articles by Feldman (1988), Gibbons et al. (1988), Naumova et al. (2001), and Chapters 3 and 4 of Singer and Willett (2003). A comprehensive, but more mathematically challenging discussion of linear mixed effects models can be found in Chapter 3 of Verbeke and Molenberghs (2000) and in the review article by Cnaan *et al.* (1997).

An excellent, non-technical, discussion of the notion of shrinkage can be found in Efron and Morris (1977); also, see the discussion of prediction of random effects in Naumova et al. (2001).

Finally, a tutorial description of fitting linear mixed effects models using PROC MIXED in SAS can by found in Singer (1998); also see Chapters 6 and 7 of Littell et al. (1996) and Chapter 8 of Verbeke and Molenberghs (2000).

Bibliographic Notes

Harville (1977) introduced a general class of linear mixed effects models suitable for the analysis of repeated measures and growth curves; also see Hartley and Rao (1967). The idea of allowing certain regression coefficients to vary randomly across individuals was also a recurring theme in the early contributions to growth curve analysis by Wishart (1938), Box (1950), Rao (1958), Potthoff and Roy (1964), and Grizzle and Allen (1969); these early contributions to growth curve modeling laid the foundation for the linear mixed effects model. Laird and Ware (1982), Jennrich and Schluchter (1986), Laird et al. (1987), Lindstrom and Bates (1988), Diggle (1988), Chi and Reinsel (1989), and others, drew upon this family to propose a general class of models for longitudinal data. Ware (1985) provides a general overview of the application of linear mixed effects models to repeated measures and longitudinal data; also see Chapter 3 of Davidian and Giltinan (1995) for a concise review of linear mixed effects models for repeated measures data.

The notion of shrinkage was first introduced in a seminal paper by Stein (1955). Best linear unbiased prediction (BLUP) is discussed in Henderson (1963); see Robinson (1991) for an interesting review of the prediction of random effects.

Problems

8.1 In a study of exercise therapies, 37 patients were assigned to one of two weightlifting programs (Freund *at al.*, 1986). In the first program (treatment 1), the number of repetitions was increased as subjects became stronger. In the second program (treatment 2), the number of repetitions was fixed but the amount of

weight was increased as subjects became stronger. Measures of strength were taken at baseline (day 0), and on days 2, 4, 6, 8, 10, and 12.

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

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

ID Treatment Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7

Note: The categorical variable Treatment is coded 1 = Program 1 (increase number of repetitions), 2 = Program 2 (increase amount of weight).

8.1.1 On a single graph, construct a time plot that displays the mean strength versus time (in days) for the two treatment groups. Describe the general characteristics of the time trends for the two exercise programs.

8.1.2 Read the data from the external file and put the data in a "univariate" or "long" format, with 7 "records" per patient.

8.1.3 Fit a model with randomly varying intercepts and slopes, and allow the mean values of the intercept and slope to depend on treatment group (i.e., include main effect of treatment, a linear time trend, and a treatment by linear time trend interaction as fixed effects).

(a) What is the estimated variance of the random intercepts?

(b) What is the estimated variance of the random slopes?

(c) What is the estimated correlation between the random intercepts and slopes?

(d) Give an interpretation to the magnitude of the estimated variance of the random intercepts. For example, "approximately 95% of subjects have baseline measures of strength between `a` and `b`" (calculate the limits of the interval between `a` and `b`).

(e) Give an interpretation to the magnitude of the estimated variance of the random slopes.

8.1.4 Is a model with only randomly varying intercepts defensible? Explain?

8.1.5 What are the mean intercept and slope in the two exercise programs.

8.1.6 Based on the previous analysis, interpret the effect of treatment on changes in strength. Does your analysis suggest a difference between the two groups?

8.1.7 What is the estimate of $\text{Var}(Y_{i1}|b_i)$? What is the estimate of $\text{Var}(Y_{i1})$? Explain the difference.

8.1.8 Obtain the predicted (empirical BLUP) intercept and slope for each subject.

8.1.9 Using any standard linear regression procedure, obtain the ordinary least squares (OLS) estimates of the intercept and slope from the regression of strength on time (in days) for subject 24 (ID = 24). That is, restrict the analysis to data on subject 24 only and estimate that subject's intercept and slope.

8.1.10 For subject 24 (ID = 24), compare the predicted intercepts and slopes obtained in Problems 8.1.8 and 8.1.9. How and why might these differ?

8.2 AIDS Clinical Trial Group (ACTG) study 193A was a randomized, double-blind, study of AIDS patients with advanced immune suppression (CD4 counts of ≤ 50 cells/mm^3) (Henry et al., 1998). Patients were randomized to one of four daily regimens containing 600 mg of zidovudine:

(1) zidovudine alternating monthly with 400 mg didanosine;

(2) zidovudine plus 2.25 mg of zalcitabine;

(3) zidovudine plus 400 mg of didanosine;

(4) zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine.

In the analyses reported in Section 8.8, the first three treatment groups were combined and compared to the fourth.

Measurements of CD4 counts were scheduled to be collected at baseline and at 8-week intervals during follow-up. However, the CD4 count data are unbalanced due to mistimed measurements and missing data that resulted from skipped visits and dropout. The number of measurements of CD4 counts during the first 48 weeks of follow-up varied from 1 to 9, with a median of 4. CD4 count refers to the number of T-lymphocyte cells in the body; these cells are directly affected by the HIV virus. A normal CD4 count is approximately 800 to 1000; a CD4 count below 200 is one of the diagnostic criteria for AIDS established by the Centers for Disease Control and Prevention (CDC).

The raw data are stored in an external file: `cd4.dat`

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

ID Group Age Gender Week Log(CD4 + 1)

Note: The categorical variable Group is coded 1 = zidovudine alternating monthly with 400 mg didanosine, 2 = zidovudine plus 2.25 mg of zalcitabine, 3 = zidovudine plus 400 mg of didanosine, and 4 = zidovudine plus 400 mg of didanosine plus 400 mg of nevirapine. The variable Week represents time since baseline (in weeks).

8.2.1 On a single graph, construct a smoothed time plot that displays the mean log CD4 counts versus time (in weeks) for the four treatment groups. Describe the general characteristics of the time trends for the four groups.

8.2.2 Fit a model where each patient's response trajectory is represented by a randomly varying piecewise linear spline with a knot at week 16. That is, fit a model with random intercepts and two randomly varying slopes, one slope for the changes in log CD4 counts before week 16, another slope for the changes in response after week 16. Allow the average slopes for changes in response before and after week 16 to vary by group, but assume the mean response at baseline is the same in the four groups.

8.2.3 Is a model with only randomly varying intercepts defensible? Explain?

8.2.4 Construct a 6-degrees-of-freedom test of the null hypothesis of no treatment group differences in the changes in log CD4 counts.

8.2.5 Based on the previous analysis, interpret the effects of treatment on changes in log CD4 counts from baseline to week 16.

8.2.6 Based on the previous analysis, interpret the effects of treatment on changes in log CD4 counts from week 16 to week 40.

8.2.7 Using the estimates of the fixed effects from the previous analysis, construct a time plot that displays the *estimated* mean log CD4 counts versus time (in weeks) for the four treatment groups. Does the plot suggest that one treatment regimen is superior to the others in terms of short-term (40 weeks) changes in CD4 counts?