4

Linear Mixed Models for Gaussian Longitudinal Data

4.1 Introduction

Although this book focuses on models for repeated categorical data, it is helpful to first consider some key topics in the analysis of continuous longitudinal data, where most parametric models are based on underlying normality assumptions. Two general extensions of the univariate linear regression models to repeated measures can be distinguished. First, a multivariate model can be formulated, in which each component is modeled using a univariate linear regression model, and with the association structure directly modeled through a marginal covariance matrix. Second, a random-effects approach can be followed. In the next sections, these two model families will be discussed in turn. We will compare both approaches, and we will summarize how estimation and inference proceeds.

Ideas will be illustrated in the simple context of a response Y measured repeatedly on a homogeneous set of subjects i, $i = 1, \ldots, N$, and where it is believed that Y evolves linearly over time. This can immediately be generalized to more complex settings with non-linear trends and/or to models in which covariates are included to model the believe that trends may depend on subject-specific covariates.

4.2 Marginal Multivariate Model

Let Y_{ij} be the jth measurement available for the ith subject or cluster, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Further, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is the n_i-dimensional vector with all observations available for subject i. Assuming an average linear trend for Y as a function of time, a multivariate regression model can be obtained by assuming that the elements Y_{ij} in $\boldsymbol{Y}_i$ satisfy $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$, with the assumption that the error components ε_{ij} are normally distributed with mean zero. In vector notation, we get $\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ for an appropriate design matrix X_i, with $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ and with $\boldsymbol{\varepsilon}_i' = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})$. The model is completed by specifying an appropriate covariance matrix V_i for $\boldsymbol{\varepsilon}_i$, leading to the multivariate model

$$\boldsymbol{Y}_i \quad \sim \quad N(X_i\boldsymbol{\beta}, V_i). \tag{4.1}$$

Let I_{n_i} denote the identity matrix of dimension n_i, then we have that $V_i = \sigma^2 I_{n_i}$ corresponds to the univariate linear regression model, assuming all repeated measurements Y_{ij} to be independent, i.e., ignoring the fact that repeated measures within subjects may be (highly) correlated. In the case of balanced data, i.e., when a fixed number n of measurements is taken for all subjects, and when measurements are taken at fixed time-points $t_1, \ldots, t_n$, a useful covariance model is $V_i = V$, where V is a general (unstructured) $n \times n$ positive definite covariance matrix. This yields the classical mulivariate regression model (Seber 1984, Chapter 8).

Depending on the context and the actual data at hand, other choices may be appropriate. For example, a first-order autoregressive model assumes that the covariance between two measurements Y_{ij} and Y_{ik} from the same subject i is of the form $\sigma^2\rho^{|t_{ij}-t_{ik}|}$ for unknown parameters σ^2 and ρ. Another example is compound symmetry, which assumes the covariance to be of the form $\sigma^2 + \gamma\delta_{jk}$ for unknown parameters σ^2 and $\gamma > -\sigma^2$, and where δ_{jk} equals one for $j = k$ and zero otherwise. These are examples of homogeneous covariance structures since they assume the variance of all Y_{ij} to be equal. Heterogeneous versions can be formulated as well (Verbeke and Molenberghs 2000).

4.3 The Linear Mixed Model

The random-effects approach toward extending the univariate linear regression model to longitudinal settings is based on the assumption that, for every subject, the response can be modeled by a linear regression model, but with subject-specific regression coefficients. Continuing our simple example, suppose that the individual trajectories of the response Y are of the type as shown in Figure 4.1. Obviously, a linear regression model with intercept and linear time effect seems plausible to describe the data of

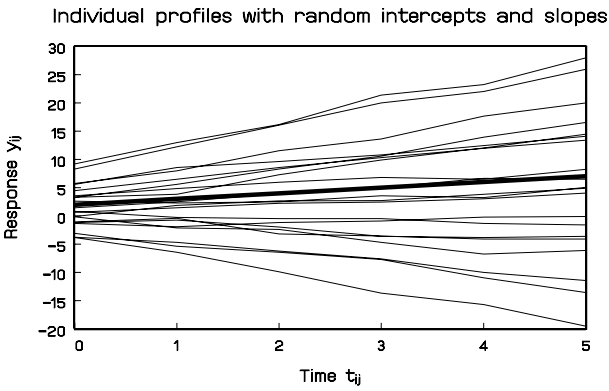

FIGURE 4.1. *Hypothetical example of continuous longitudinal data that can be well described by a linear mixed model with random intercepts and random slopes. The thin lines represent the observed subject-specific evolutions. The bold line represents the population-averaged evolution.*

each person separately. However, different persons tend to have different intercepts and different slopes. One can therefore assume that the outcome Y_{ij}, measured at time t_{ij} satisfies $Y_{ij} = \widetilde{\beta}_{i0} + \widetilde{\beta}_{i1} t_{ij} + \varepsilon_{ij}$. As before, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})'$ is assumed to be normally distributed with mean vector zero, and some covariance matrix which we now denote by Σ_i.

Because subjects are randomly sampled from a population of subjects, it is natural to assume that the subject-specific regression coefficients $\widetilde{\boldsymbol{\beta}}_i = (\widetilde{\beta}_{i0}, \widetilde{\beta}_{i1})'$ are randomly sampled from a population of regression coefficients. It is customary to assume the $\widetilde{\boldsymbol{\beta}}_i$ to be (multivariate) normal, but extensions can be formulated (Verbeke and Lesaffre 1996, Magder and Zeger 1996). Assuming $\widetilde{\boldsymbol{\beta}}_i$ to be bivariate normal with mean $(\beta_0, \beta_1)'$ and 2×2 covariance matrix D we can reformulate the model as

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{pi} + \varepsilon_{ij}, \tag{4.2}$$

with $\widetilde{\beta}_{i0} = \beta_0 + b_{i0}$ and $\widetilde{\beta}_{i1} = \beta_1 + b_{i1}$, and the new random effects $\boldsymbol{b}_i = (b_{i0}, b_{i1})'$ are now normal with mean zero and covariance D. The population-averaged profile is linear, with intercept β_0 and slope β_1, and is represented by the bold line in Figure 4.1.

The above model is a special case of the general linear mixed model which assumes that the vector $\boldsymbol{Y}_i$ of repeated measurements for the ith subject satisfies

$$\boldsymbol{Y}_i | \boldsymbol{b}_i \sim N(X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i, \Sigma_i) \tag{4.3}$$

$$\boldsymbol{b}_i \sim N(\boldsymbol{0}, D), \tag{4.4}$$

for $n_i \times p$ and $n_i \times q$ known design matrices X_i and Z_i, for a p-dimensional vector $\boldsymbol{\beta}$ of unknown regression coefficients, for a q-dimensional vector $\boldsymbol{b}_i$

of subject-specific regression coefficients assumed to be sampled from the q-dimensional normal distribution with mean zero and covariance D, and with Σ_i a covariance matrix parameterized through a set of unknown parameters. The components in $\boldsymbol{\beta}$ are called 'fixed effects,' the components in $\boldsymbol{b}_i$ are called 'random effects.' The fact that the model contains fixed as well as random effects motivates the term 'mixed models.'

Unless the model is fitted in a Bayesian framework (Gelman *et al* 1995), estimation and inference are based on the marginal distribution for the response vector $\boldsymbol{Y}_i$. Let $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ and $f(\boldsymbol{b}_i)$ be the density functions corresponding to (4.3) and (4.4), respectively, the marginal density function of $\boldsymbol{Y}_i$ is

$$f_i(\boldsymbol{y}_i) \quad = \quad \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)\ f(\boldsymbol{b}_i)\ d\boldsymbol{b}_i,$$

which can easily be shown to be the density function of an n_i-dimensional normal distribution with mean vector $X_i\boldsymbol{\beta}$ and with covariance matrix $V_i = Z_i D Z_i' + \Sigma_i$. Note that the linear mixed model implies a marginal model of the form (4.1), but with a very specific parametric form for the marginal covariance matrix V_i, easily allowing highly unbalanced data. In this respect, the linear mixed model can be interpreted as a procedure to obtain flexible multivariate marginal models. As was already shown in our earlier example, the fixed effects describe the population-averaged evolution.

Because the mixed model is defined through the distributions $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ and $f(\boldsymbol{b}_i)$, this will be called the hierarchical formulation of the linear mixed model. The corresponding marginal normal distribution with mean $X_i\boldsymbol{\beta}$ and covariance $V_i = Z_i D Z_i' + \Sigma_i$ is called the marginal formulation of the model. Note that, although the marginal model naturally follows from the hierarchical one, both these models are not equivalent. Indeed, different random-effects models can produce the same marginal model. To see this, consider the case where every subject is measured twice ($n_i = 2$). First, assume that the random-effects structure is confined to a random intercept ($\boldsymbol{b}_i$ is scalar), and the residual error structure $\Sigma_i = \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ (Model I). The resulting marginal covariance matrix is

$$V = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (d) \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} d+\sigma_1^2 & d \\ d & d+\sigma_2^2 \end{pmatrix}. \quad (4.5)$$

Second, consider the random effects to consist of a random intercept and a random slope ($\boldsymbol{b}_i = (b_{0i}, b_{1i})'$), mutually uncorrelated, with residual error structure $\Sigma_i = \Sigma = \sigma^2 I_2$ (Model II). The resulting covariance matrix now equals

$$V \quad = \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

$$= \begin{pmatrix} d_1 + \sigma^2 & d_1 \\ d_1 & d_1 + d_2 + \sigma^2 \end{pmatrix}. \tag{4.6}$$

Obviously, the parametric models (4.5) and (4.6) for the marginal covariance are equivalent: $d_1 = d$, $d_2 = \sigma_2^2 - \sigma_1^2$, and $\sigma^2 = \sigma_1^2$. Thus, (at least) two different hierarchical models can produce the same marginal model, illustrating that a good fit of the marginal model should not be seen as equally strong evidence for any of the mixed models. Arguably, a satisfactory treatment of the random-effects model is only possible within a Bayesian context.

In addition, it is important to see that there are even marginal models that are not implied by a mixed model. The simplest example is found by considering the marginal model with compound symmetric covariance structure (Section 4.2). If the within-subject correlation is positive ($\gamma \geq 0$), this model could have been implied by a mixed model with random intercepts b_i that are normally distributed with mean 0 and variance γ, and with uncorrelated error components with common variance σ^2. However, if the within-cluster correlation is negative ($\gamma < 0$), the resulting marginal model cannot be implied by an appropriate random-effects model. This would be the case, for example, in a context of competition such as when littermates compete for the same food resources.

4.4 Estimation and Inference for the Marginal Model

As indicated earlier, the fitting of a linear mixed model is usually based on the marginal model that, for subject i, is multivariate normal with mean $X_i\boldsymbol{\beta}$ and covariance $V_i(\boldsymbol{\alpha}) = Z_i D Z_i' + \Sigma_i$, hereby explicitly denoting that V_i depends on an unknown vector $\boldsymbol{\alpha}$ of parameters in the covariance matrices D and Σ_i. The parameters in $\boldsymbol{\alpha}$ are usually called 'variance components.' The classical approach to estimation and inference is based on maximum likelihood (ML). Assuming independence across subjects, the likelihood takes the form

$$L_{\text{ML}}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} \left|V_i(\boldsymbol{\alpha})\right|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2} \left(\boldsymbol{Y_i} - X_i\boldsymbol{\beta}\right)' V_i^{-1}(\boldsymbol{\alpha}) \left(\boldsymbol{Y_i} - X_i\boldsymbol{\beta}\right) \right] \right\}. \tag{4.7}$$

Estimation of $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\alpha}')$ requires joint maximization of (4.7) with respect to all elements in $\boldsymbol{\theta}$. In general, no analytic solutions are available, calling for numerical optimization routines.

Conditionally on $\boldsymbol{\alpha}$, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is given by (Laird and Ware 1982):

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i \boldsymbol{Y_i}, \tag{4.8}$$

where W_i equals V_i^{-1}. In practice, $\boldsymbol{\alpha}$ is not known and can be replaced by its MLE $\widehat{\boldsymbol{\alpha}}$. However, one often also uses the so-called restricted maximum likelihood (REML) estimator for $\boldsymbol{\alpha}$ (Harville 1974), which allows to estimate $\boldsymbol{\alpha}$ without having to estimate the fixed effects in $\boldsymbol{\beta}$ first. It is known from simpler models, such as linear regression models, that, while classical ML estimators are biased downwards, this is not the case for the REML estimators (Verbeke and Molenberghs 2000, Section 5.3).

When it comes to inference, in practice, the fixed effects in $\boldsymbol{\beta}$ are often of primary interest, as they describe the average evolution in the population. Conditionally on $\boldsymbol{\alpha}$, the maximum likelihood (ML) estimate for $\boldsymbol{\beta}$ is given by (4.8), which is normally distributed with mean

$$E\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})\right] = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i E\left[\boldsymbol{Y_i}\right] = \boldsymbol{\beta}, \tag{4.9}$$

and covariance

$$\begin{aligned} \mathrm{Var}\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})\right] &= \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \\ &\quad \times \left(\sum_{i=1}^{N} X_i' W_i \mathrm{Var}\left[\boldsymbol{Y_i}\right] W_i X_i \right) \\ &\quad \times \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \qquad (4.10) \\ &= \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1}, \qquad (4.11) \end{aligned}$$

provided that the mean and covariance were correctly specified in our model, i.e., provided that $E(\boldsymbol{Y}_i) = X_i\boldsymbol{\beta}$ and $\mathrm{Var}(\boldsymbol{Y}_i) = V_i = Z_i D Z_i' + \Sigma_i$. Approximate Wald-type tests for components in $\boldsymbol{\beta}$ can now easily be obtained.

Note however, that these Wald tests are based on standard errors obtained from replacing $\boldsymbol{\alpha}$ in (4.11) by its ML or REML estimate and therefore underestimate the true variability in $\widehat{\boldsymbol{\beta}}$ because they do not take into account the variability introduced by estimating $\boldsymbol{\alpha}$. Therefore, the classical normal or chi-squared reference distributions are often replaced by t or F-distributions, with the same numerator degrees of freedom as the original

chi-squared distribution. The denominator degrees of freedom need to be estimated from the data. This is often based on so-called Satterthwaite-type approximations (Satterthwaite 1941), and is only fully developed for the case of linear mixed models. We refer to Verbeke and Molenberghs (2000, Section 6.2) for more information on this aspect. In most longitudinal applications, different persons contribute independent information, which results in numbers of denominator degrees of freedom which are typically large enough, whatever estimation method is used, to lead to very similar p-values. Only for very small samples in terms of independent replicates, or when mixed models would be used with crossed random effects (random effects for persons as well as for items) different estimation methods for degrees of freedom may lead to severe differences in the resulting p-values.

Note also that the standard errors based on (4.11) are valid, only if the mean and covariance were correctly specified, while the only condition for $\widehat{\boldsymbol{\beta}}$ to be unbiased is that the mean is correctly specified. Because in practice, it is often difficult to assess correct specification of the covariance structure, one often prefers standard errors to be based on (4.10), rather than (4.11), but with $\mathrm{Var}\,(\boldsymbol{Y_i})$ estimated by $(\boldsymbol{y}_i - X_i\widehat{\boldsymbol{\beta}})(\boldsymbol{y}_i - X_i\widehat{\boldsymbol{\beta}})'$ rather than $\widehat{V}_i$. The so-called robust or empirical standard errors are consistent, as long as the mean is correctly specified. This procedure is a special case of the theory on generalized estimating equations (GEE), introduced by Liang and Zeger (1986) which will be applied in Chapter 8 in the context of discrete outcomes.

When interest is also in inference for some of the variance components in $\boldsymbol{\alpha}$, classical asymptotic Wald, likelihood ratio, and score tests can be used. However, due to restrictions on the parameter spaces, some hypotheses of interest may be on the boundary of the parameter space, implying that classical testing procedures are no longer valid. In some special but important cases, analytic results are available on how to correctly test such hypotheses. We herefore refer to Stram and Lee (1994, 1995) for results on the likelihood ratio test, and to Verbeke and Molenberghs (2003) for results on the score test. A detailed discussion on inference for the marginal linear mixed model can be found in Verbeke and Molenberghs (2000, Chapter 6).

4.5 Inference for the Random Effects

Although in practice, one is usually primarily interested in estimating the parameters in the marginal model, it is often useful to calculate estimates for the random effects $\boldsymbol{b}_i$ as well, as they reflect how much the subject-specific profiles deviate from the overall average profile. Such estimates can then be interpreted as residuals which may be helpful for detecting special

profiles (i.e., outlying individuals) or groups of individuals evolving differently over time. Also, estimates for the random effects are needed whenever interest is in prediction of subject-specific evolutions. Obviously, it is then no longer sufficient to assume that the data can be described well by the marginal model $N(X_i\boldsymbol{\beta}, V_i)$. Instead, one has to explicitly assume that the hierarchical model specification (4.3) and (4.4) is appropriate. Because random effects represent a natural heterogeneity between the subjects, this assumption will often be justified for data where the between-subjects variability is large in comparison to the within-subject variability.

Because the subject-specific parameters $\boldsymbol{b}_i$ are assumed random, it is most natural to estimate them using Bayesian techniques (Box and Tiao 1992, Gelman *et al* 1995). Conditional on $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ follows a multivariate normal distribution with mean vector $X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$ and with covariance matrix Σ_i. In combination with the distributional assumptions for $\boldsymbol{b}_i$, one can easily derive (Smith 1973, Lindley and Smith 1972) that, conditionally on $\boldsymbol{Y}_i = \boldsymbol{y}_i$, $\boldsymbol{b}_i$ follows a multivariate normal posterior distribution with mean $\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta}) = DZ_i'V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y}_i - X_i\boldsymbol{\beta})$, which is used in practice as an estimator for $\boldsymbol{b}_i$. Its covariance estimator is equal to

$$\begin{aligned} &\mathrm{var}(\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta})) \\ &\quad = \quad DZ_i' \left\{ V_i^{-1} - V_i^{-1}X_i \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i \right)^{-1} X_i'V_i^{-1} \right\} Z_iD. \end{aligned} \tag{4.12}$$

Note that (4.12) underestimates the variability in $\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta}) - \boldsymbol{b}_i$ since it ignores the variation of $\boldsymbol{b}_i$. Therefore, inference for $\boldsymbol{b}_i$ is usually based on

$$\mathrm{var}[\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta}) - \boldsymbol{b}_i] \quad = \quad D - \mathrm{var}[\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta})] \tag{4.13}$$

as an estimator for the variation in $\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta}) - \boldsymbol{b}_i$ (Laird and Ware 1982).

So far, all calculations were performed conditionally upon the vector $\boldsymbol{\theta}$ of parameters in the marginal model. In practice, the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in $\widehat{\boldsymbol{b}_i}(\boldsymbol{\theta})$, (4.12), and (4.13) are replaced by their maximum or restricted maximum likelihood estimates. The resulting estimates for the random effects are called "Empirical Bayes" (EB) estimates, which we will denote by $\widehat{\boldsymbol{b}_i}$. Note that (4.12) and (4.13) then underestimate the true variability in the obtained estimate $\widehat{\boldsymbol{b}_i}$ because they do not take into account the variability introduced by replacing the unknown parameter $\boldsymbol{\theta}$ by its estimate. Similarly as for the fixed effects, inference is therefore often based on approximate t-tests or F-tests, rather than on traditional Wald tests.

It immediately follows from (4.13) that for any linear combination $\boldsymbol{\lambda}\boldsymbol{b}_i$ of the random effects, $\mathrm{var}(\boldsymbol{\lambda}'\widehat{\boldsymbol{b}_i}) \leq \mathrm{var}(\boldsymbol{\lambda}'\boldsymbol{b}_i)$, indicating that the EB estimates show less variability than actually present in the random-effects population. This phenomenon is usually referred to as shrinkage (Carlin and Louis 1996, Strenio, Weisberg, and Bryk 1983). The shrinkage is also seen in the

prediction $\widehat{\boldsymbol{y}}_i \equiv X_i\widehat{\boldsymbol{\beta}}+Z_i\widehat{\boldsymbol{b}}_i$ of the ith profile, which can be rewritten as $\widehat{\boldsymbol{y}}_i = \Sigma_i V_i^{-1} X_i\widehat{\boldsymbol{\beta}}+[I_{n_i}-\Sigma_i V_i^{-1}]\boldsymbol{y}_i$. Note that $\widehat{\boldsymbol{y}}_i$ can be interpreted as a weighted average of the population-averaged profile $X_i\widehat{\boldsymbol{\beta}}$ and the observed data $\boldsymbol{y}_i$, with weights $\Sigma_i V_i^{-1}$ and $I_{n_i} - \Sigma_i V_i^{-1}$, respectively. The "numerator" of $\Sigma_i V_i^{-1}$ is the residual covariance matrix Σ_i and the "denominator" is the overall covariance matrix V_i. Hence, severe shrinkage is to be expected when the residual variability is large in comparison to the between-subject variability (modeled by the random effects), whereas little shrinkage will occur if the opposite is true.

In practice, one often uses histograms and scatter plots of components of $\widehat{\boldsymbol{b}}_i$ for diagnostic purposes, such as the detection of outliers, which are subjects who seem to evolve differently from the other subjects in the data set. Examples and more details on the use of EB estimates can be found in Verbeke and Molenberghs (2000, Chapter 7) or in DeGruttola, Lange, and Dafni (1991) and Waternaux, Laird, and Ware (1989). It should be emphasized that the EB estimates cannot be used to check the underlying normality assumption about the random effects. Verbeke and Lesaffre (1996) have shown that, in some cases with severe deviations from normality, the normality assumption forces the EB estimates to look like a normal distribution. They propose to use more general random-effects distributions, such as mixtures of normals. In Chapter 23, we will use related ideas in the context of models for non-continuous responses.