Chapter 8
Linear Models

8.1 Introduction

In Part III of the book the conditional independence assumptions of Part II are relaxed as we consider models for dependent data. Such data occur in many contexts, with three common situations being when sampling is over time, space, or within families. We do not discuss pure time series applications in which data are collected over a single (usually long) series; this is a vast topic with many specialized texts. Generically, we consider regression modeling situations in which there are a set of units ("clusters") upon which multiple measurements have been collected. For example, when data are available over time for a group of units, we have *longitudinal* (also known as *repeated measures*) data, and each unit forms a cluster. We will often refer to the units as individuals. The methods described in Part II for calculating uncertainty measures (such as standard errors) are not applicable in situations in which the data are dependent.

Throughout Part III we distinguish approaches that specify a full probability model for the data (with likelihood or Bayesian approaches to inference) and those that specify first, and possibly second, moments only (with an estimating function being constructed for inference). As in Part II we believe it will often be advantageous to carry out inference from both standpoints in a complimentary fashion. In some instances the form of the question of interest may be best served by a particular approach, however, and this will be stressed at relevant points.

In this chapter we consider linear regression models. Such models are widely applicable with growth curves, such as the dental data of Sect. 1.3.5, providing a specific example. As another example, in the so-called *split-plot* design, fields are planted with different crops and within each field (unit), different subunits are treated with different fertilizers. We expect crop yields in the same field to be more similar than those in different fields, and yields may be modeled as a linear function of crop and fertilizer effects. With clustered data, we expect measurements on the same unit to exhibit *residual* dependence due to shared unmeasured variables, where the qualifier acknowledges that we have controlled for known regressors.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_8,

The structure of this chapter is as follows. We begin, in Sect. 8.2, with a brief overview of approaches to inference for dependent data, in the context of the dental data of Sect. 1.3.5. Section 8.3 provides a description of the efficiency gains that can be achieved with data collected over time in a longitudinal design. In Figure 8.1(a), linear mixed effects models, in which full probability models are specified for the data, are introduced. In Sects. 8.5 and 8.6, likelihood and Bayesian approaches to inference for these models are described. Section 8.7 discusses the generalized estimating equations (GEE) approach which is based on a marginal mean specification and empirical sandwich estimation of standard errors. We describe how the assumptions required for valid inference may be assessed in Sect. 8.8 and discuss the estimation of longitudinal and cohort effects in Sect. 8.9. Concluding remarks appear in Sect. 8.10 with bibliographic notes in Sect. 8.11.

8.2 Motivating Example: Dental Growth Curves

In Table 1.3 dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure are given for 11 girls and 16 boys, recorded at the ages of 8, 10, 12, and 14 years. In this section we concentrate on the data from the girls only. Figure 8.1(a) plots the dental measurements for each girl versus age. The slopes look quite similar, though there is clearly between-girl variability in the intercepts.

There are various potential aims for the analysis of data such as these:

1. Population inference, in which we describe the average growth as a function of age, for the population from which the sample of children were selected.
2. Assessment of the within- to between-child variability in growth measurements.

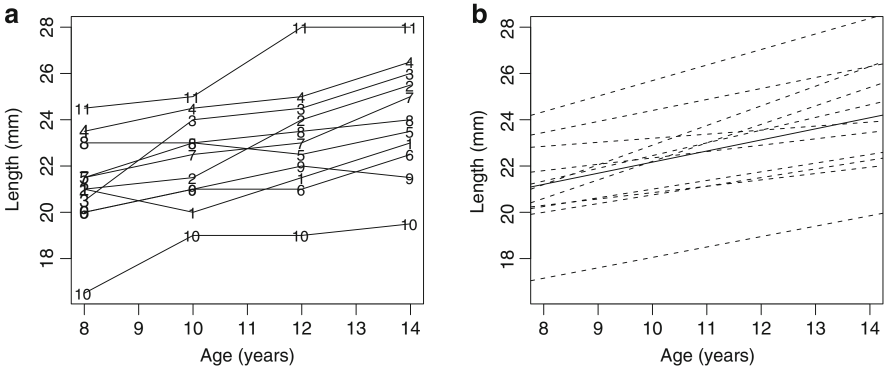


Fig. 8.1 Dental data for girls only: (**a**) individual observed data (with the girl index taken as plotting symbol), (**b**) individual *fitted curves* (*dashed*) and overall *fitted curve* (*solid*)

3. Individual-level inference, either for a child in the sample, or for a new unobserved child (from the same population). The latter could be used to construct a "growth chart" in which the percentile points of children's measurements at different ages are presented.

Part III of the book will provide extensive discussion of *mixed effects* models which contain both *fixed* effects that are shared by all individuals and *random effects* that are unique to particular individuals and are assumed to arise from a distribution. For longitudinal data there are two extreme fixed effects approaches. Proceeding naively, we could assume a single "marginal" curve for *all* of the girls data and carry out a standard analysis assuming independent data. Marginal here refers to *averaging* over girls in the population. At the other extreme we could assume a distinct curve for each girl. Figure 8.1(b) displays the least squares fitted lines corresponding to each of these fixed effects approaches.

Continuing with the marginal approach, let Y_{ij} denote the jth measurement, taken at time t_j on the ith child, $i = 1, \ldots, m = 11, j = 1, \ldots, n_i = 4$. Consider the model

$$\mathrm{E}[Y_{ij}] = \beta_0^{\mathrm{M}} + \beta_1^{\mathrm{M}} t_j \tag{8.1}$$

where β_0^{M} and β_1^{M} represent *marginal* intercept and slope parameters. Then,

$$e_{ij}^{\mathrm{M}} = Y_{ij} - \beta_0^{\mathrm{M}} - \beta_1^{\mathrm{M}} t_j,$$

$i = 1, \ldots, 11; j = 1, \ldots, 4$, denote marginal residuals. In Part II of the book, we emphasized conditional independence, so that observations were independent *given* a set of parameters; due to dependence of observations on the same girl, we would not expect the marginal residuals to be independent.

We fit the marginal model (8.1) to the data from all girls and let

$$\begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \rho_{13} & \rho_{23} & \sigma_3 & \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4 \end{bmatrix} \tag{8.2}$$

represent the standard deviation/correlation matrix of the residuals. Here,

$$\sigma_j = \sqrt{\mathrm{var}(e_{ij}^{\mathrm{M}})},$$

is the standard deviation of the dental length at time t_j and

$$\rho_{jk} = \frac{\mathrm{cov}(e_{ij}^{\mathrm{M}}, e_{ik}^{\mathrm{M}})}{\sqrt{\mathrm{var}(e_{ij}^{\mathrm{M}})\mathrm{var}(e_{ik}^{\mathrm{M}})}},$$

is the correlation between residual measurements taken at times t_j and t_k on the same girl, $j \neq k, j, k = 1, \ldots, 4$. We assume four distinct standard deviations at each of the ages, and distinct correlations between measurements at each of the six combinations of pairs of ages, but assume that these standard deviations and correlations are constant across all girls. We empirically estimate the entries of (8.2) as

$$\begin{bmatrix} 2.12 & & & \\ 0.83 & 1.90 & & \\ 0.86 & 0.90 & 2.36 & \\ 0.84 & 0.88 & 0.95 & 2.44 \end{bmatrix} \tag{8.3}$$

illustrating that, not surprisingly, there is clear correlation between residuals at different ages on the same girl. Fitting a single curve to the totality of the data and using methods for independent data that assume within-girl correlations are zero will clearly give inappropriate standard errors/uncertainty estimates for $\widehat{\beta}_0^{\mathrm{M}}$ and $\widehat{\beta}_1^{\mathrm{M}}$. Fitting such a marginal model is appealing, however, since it allows the *direct* calculation of the average responses at different ages. Fitting a marginal model forms the basis of the GEE approach described in Sect. 8.7.

The alternative fixed effects approach is to assume a fixed curve for each child and analyze each set of measurements separately. However, while providing valid inference for each curve, there is no "borrowing of strength" across children, so that each girl's fit is based solely on her data only and not on the data of other children. We might expect that there is *similarity* between the curves, and therefore, it is reasonable to believe that the totality of data will enhance estimation for each child. In some instances, using the totality of data will be vital. For example, estimating the growth curve for a girl with just a single observation is clearly not possible using the observed data on that girl only. Suppose we are interested in making formal inference for the population of girls from which the $m = 11$ girls are viewed as a random sample; this is not formally possible using the collection of fixed effects estimates from each girl. The basis of the mixed effects model approach described in Sect. 8.4 is to assume a girl-specific set of *random effect* parameters that are assumed to arise from a population. In different contexts, random effects may have a direct interpretation as arising from a population of effects, or may simply be viewed as a convenient modeling tool, in situations in which there is no hypothetical population of effects to appeal to.

Throughout Part III, we will describe mixed effects and GEE approaches to analysis. The mixed effects approach can be seen as having a greater *contextual* basis, since it builds up a model from the level of the unit. In contrast, with a marginal model, as specified in GEE, the emphasis is on population inference based on *minimal* assumptions and on obtaining a reliable standard error via sandwich estimation.

8.3 The Efficiency of Longitudinal Designs

While making inference for dependent data is in general more difficult than for independent data, designs that collect dependent data can be very efficient. For example, in a longitudinal data setting, applying different treatments to the same patient over time can be very beneficial, since each patient acts as his/her own control. To illustrate, we provide a comparison between longitudinal and cross-sectional studies (in which data are collected at a single time point); this section follows the development of Diggle et al. (2002, Sect. 2.3).

We consider a very simple situation in which we wish to compare two treatments, coded as -1 and $+1$, and take four measurements in total. In the cross-sectional study a single measurement is taken on each of four individuals with

$$Y_{i1} = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1}, \tag{8.4}$$

for $i = 1, \dots, m = 4$. The error terms ϵ_{i1} are independent with $\text{E}[\epsilon_{i1}] = 0$ and $\text{var}(\epsilon_{i1}) = \sigma^2$. The design is such that $x_{11} = -1, x_{21} = -1, x_{31} = 1, x_{41} = 1$, so that individuals 1 and 2 (3 and 4) receive treatment -1 ($+1$). With this coding, the treatment effect is

$$\text{E}[Y_1 \mid x = 1] - \text{E}[Y_1 \mid x = -1] = 2\beta_1.$$

The (unbiased) ordinary least squares (OLS) estimators are

$$\widehat{\beta}_0^{\text{C}} = \frac{\sum_{i=1}^{4} Y_{i1}}{4}, \quad \widehat{\beta}_1^{\text{C}} = \frac{Y_{31} + Y_{41} - (Y_{11} + Y_{21})}{4},$$

and, more importantly for our purposes, the variance of the treatment estimator is

$$\text{var}(\widehat{\beta}_1^{\text{C}}) = \frac{\sigma^2}{4}.$$

The subscript here labels the relevant quantities as arising from the cross-sectional design.

For the longitudinal study we assume the model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \delta_{ij},$$

with b_i and δ_{ij} independent and $\text{E}[\delta_{i1}] = 0$, $\text{var}(\delta_{ij}) = \sigma_\delta^2$, $\text{E}[b_i] = 0$, $\text{var}(b_i) = \sigma_0^2$, for $i = 1, 2, j = 1, 2$, so that we record two observations on each of two individuals. The b_i represent random individual-specific parameters and ϵ_{ij} measurement error. Marginally, that is, averaging over individuals, and with $\boldsymbol{Y} = [Y_{11}, Y_{12}, Y_{21}, Y_{22}]^{\text{T}}$, we have $\text{var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{R}$ with

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix} \tag{8.5}$$

where $\sigma^2 = \sigma_0^2 + \sigma_\delta^2$ is the sum of the between- and within-individual variances and $\rho = \sigma_0^2/\sigma^2$ is the correlation between observations on the same individual. Notice that the cross-sectional variance model is a special case of (8.4) with $\epsilon_{i1} = b_i + \delta_{i1}$. We consider two designs. In the first, the treatment is constant over time for each individual: $x_{11} = x_{12} = -1, x_{21} = x_{22} = 1$, while in the second each individual receives both treatments: $x_{11} = x_{22} = 1, x_{12} = x_{21} = -1$. Generalized least squares gives unbiased estimator

$$\widehat{\boldsymbol{\beta}}^{\text{L}} = (\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{Y}, \tag{8.6}$$

with

$$\text{var}(\widehat{\boldsymbol{\beta}}^{\text{L}}) = (\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\sigma^2,$$

and where $\boldsymbol{R}$ is given by (8.5). The variance of the “slope” estimator is

$$\text{var}(\widehat{\beta}_1^{\text{L}}) = \frac{\sigma^2(1-\rho^2)}{4 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}.$$

The *efficiency* of the longitudinal design, as compared to the cross-sectional design, is therefore

$$\frac{\text{var}(\widehat{\beta}_1^{\text{L}})}{\text{var}(\widehat{\beta}_1^{\text{C}})} = \frac{(1-\rho^2)}{1 - \rho(x_{11}x_{12} + x_{21}x_{22})/2}.$$

The efficiency of the longitudinal study with *constant* treatments across individuals is

$$1 + \rho,$$

so that in this case, the cross-sectional study is preferable in the usual situation in which observations on the same individual display positive correlation, that is, $\rho > 0$. When the treatment is constant within individuals, the treatment estimate is based on between-individual comparisons only, and so, it is more beneficial to obtain measurements on additional individuals.

The efficiency of the longitudinal study with treatments *changing* within individuals is

$$1 - \rho,$$

so that the longitudinal study is more efficient when $\rho > 0$, because each individual is acting as his/her own control. That is, we are making within-individual comparisons. If $\rho = 0$, the designs have the same efficiency. In practice, collecting two measurements on different individuals will often be logistically more straightforward than collecting two measurements on the same individual (e.g.,with the possibility of missing data at the second time point), but in pure efficiency terms, the longitudinal design with changing treatment can be very efficient. Clearly, this discussion extends to other longitudinal situations in which covariates are changing over time (and more general situations with covariate variation within clusters).

8.4 Linear Mixed Models

8.4.1 The General Framework

The basic idea behind mixed effects models is to assume that each unit has a regression model characterized by both fixed effects, that are common to all units in the population, and unit-specific perturbations, or random effects. "Mixed" effects refers to the combination of both fixed and random effects. The frequentist interpretation of the random effects is that the units can be viewed as a random sample from a hypothetical super-population of units. A Bayesian interpretation arises through considerations of exchangeability (Sect. 3.9), as we discuss further in Sect. 8.6.2.

Let the multiple responses for the ith unit be $\boldsymbol{Y}_i = [Y_{i1}, \dots, Y_{in_i}]^{\mathrm{T}}$, $i = 1, \dots, m$. We assume that responses on different units are independent but that there is dependence between observations on the same unit. Let $\boldsymbol{\beta}$ represent a $(k+1) \times 1$ vector of fixed effects and $\boldsymbol{b}_i$ a $(q+1) \times 1$ vector of random effects, with $q \leq k$. In this chapter, we assume the mean for Y_{ij} is linear in the fixed and random effects. Let $\boldsymbol{x}_{ij} = [1, x_{ij1}, \dots, x_{ijk}]$ be a $(k+1) \times 1$ vector of covariates measured at occasion j, so that $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i}]$ is the design matrix for the fixed effects for unit i. Similarly, let $\boldsymbol{z}_{ij} = [1, z_{ij1}, \dots, z_{ijq}]^{\mathrm{T}}$ be a $(q+1) \times 1$ vector of variables that are a subset of $\boldsymbol{x}_{ij}$, so that $\boldsymbol{z}_i = [\boldsymbol{z}_{i1}, \dots, \boldsymbol{z}_{in_i}]^{\mathrm{T}}$ is the design matrix for the random effects.

We describe a two-stage linear mixed model (LMM).

Stage One: The response model, *conditional* on random effects $\boldsymbol{b}_i$ is

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{8.7}$$

where $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ zero-mean vector of error terms, $i = 1, \dots, m$.

Stage Two: The random terms in (8.7) satisfy

$$\begin{aligned} \mathrm{E}[\boldsymbol{\epsilon}_i] &= \boldsymbol{0}, \ \ \mathrm{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{E}_i(\boldsymbol{\alpha}), \\ \mathrm{E}[\boldsymbol{b}_i] &= \boldsymbol{0}, \ \ \mathrm{var}(\boldsymbol{b}_i) = \boldsymbol{D}(\boldsymbol{\alpha}), \\ \mathrm{cov}(\boldsymbol{b}_i, \boldsymbol{\epsilon}_{i'}) &= \boldsymbol{0}, \qquad i, i' = 1, \dots, m, \end{aligned}$$

where $\boldsymbol{\alpha}$ is an $r \times 1$ vector containing the collection of variance–covariance parameters. Further, $\mathrm{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{i'}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{b}_i, \boldsymbol{b}_{i'}) = \boldsymbol{0}$, for $i \neq i'$.

The two stages may be collapsed, by averaging over the random effects, to give the marginal model:

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y}_i] &= \boldsymbol{x}_i \boldsymbol{\beta} \\ \mathrm{var}(\boldsymbol{Y}_i) &= \boldsymbol{V}_i(\boldsymbol{\alpha}) \\ &= \boldsymbol{z}_i \boldsymbol{D}(\boldsymbol{\alpha}) \boldsymbol{z}_i^{\mathrm{T}} + \boldsymbol{E}_i(\boldsymbol{\alpha}) \end{aligned} \tag{8.8}$$

for $i = 1, \ldots, m$, so that $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ matrix. The random effects have therefore induced dependence on an individual through the first term in (8.8). However, responses on individuals i and i', $i \neq i'$, are independent:

$$\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = \mathbf{0}$$

where $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'})$ is the $n_i \times n_{i'}$ matrix with element (j, j') corresponding to $\text{cov}(Y_{ij}, Y_{i'j'})$, $j = 1, \ldots, n_i$, $j' = 1, \ldots, n_{i'}$.

8.4.2 Covariance Models for Clustered Data

With respect to model (8.7), a common assumption is that $\boldsymbol{b}_i \sim_{iid} \text{N}_{q+1}(\,\mathbf{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim_{ind} \text{N}_{n_i}(\,\mathbf{0}, \boldsymbol{E}_i)$. A common variance for all individuals and at all measurement occasions, along with uncorrelated errors, gives the simplified form $\boldsymbol{E}_i = \sigma_\epsilon^2 \mathbf{I}_{n_i}$. We will refer to σ_ϵ^2 as the measurement error variance, but as usual, the error terms may include contributions from model misspecification, such as departures from linearity, and data recording errors. The inclusion of random effects *induces* a marginal covariance model for the data. This may be contrasted with the direct specification of a marginal variance model. In this section we begin by deriving the marginal variance structure that arises from two simple random effects models, before describing more general covariance structures. It is important to examine the *marginal* variances and covariances, since these may be directly assessed from the observed data.

We first consider the random intercepts only model $\boldsymbol{z}_i \boldsymbol{b}_i = \mathbf{1}_{n_i} b_i$ with $\text{var}(b_i) = \sigma_0^2$, along with $\boldsymbol{E}_i = \sigma_\epsilon^2 \mathbf{I}_{n_i}$. From (8.8), it is straightforward to show that this stipulation gives the *exchangeable* or *compound symmetry* marginal variance model:

$$\text{var}(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

where $\sigma^2 = \sigma_\epsilon^2 + \sigma_0^2$ and $\rho = \sigma_0^2/\sigma^2$. In this case we have two variance parameters so that $\boldsymbol{\alpha} = [\sigma_\epsilon^2, \sigma_0^2]$. A consequence of between-individual variability in intercepts is therefore constant marginal within-individual correlation. The latter must be nonnegative under this model (since $\sigma_0^2 \geq 0$) which would seem reasonable in most situations.

The exchangeable model is particularly appropriate for clustered data with no time ordering as may arise, for example, in a split-plot design, or for multiple measurements within a family. It may be useful for longitudinal data also, particularly over short time scales. If we think of residual variability as being due to unmeasured variables, then the exchangeable structure is most appropriate when we believe such variables are relatively constant across responses within an individual.

We now consider a model with both random intercepts and random slopes. Such a model is a common choice in longitudinal studies. With respect to (8.7) and for $i = 1, \ldots, m$, the first stage model is

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix} \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}$$

with $\boldsymbol{b}_i = [b_{i0}, b_{i1}]^{\mathrm{T}}$ and $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}$ where

$$\boldsymbol{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}.$$

Therefore, σ_0 is the standard deviation of the intercepts, σ_1 is the standard deviation of the slopes, and σ_{01} is the covariance between the two. This model induces a marginal variance at time t_{ij} which is quadratic in time:

$$\text{var}(Y_{ij}) = \sigma_\epsilon^2 + \sigma_0^2 + 2\sigma_{01} t_{ij} + \sigma_1^2 t_{ij}^2. \tag{8.9}$$

The marginal correlation between observations at times t_{ij} and t_{ik} is

$$\rho_{jk} = \frac{\sigma_0^2 + (t_{ij} + t_{ik})\sigma_{01} + t_{ij} t_{ik} \sigma_1^2}{(\sigma_\epsilon^2 + \sigma_0^2 + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_1^2)^{1/2}(\sigma_\epsilon^2 + \sigma_0^2 + 2t_{ik}\sigma_{01} + t_{ik}^2\sigma_1^2)^{1/2}} \tag{8.10}$$

for $j, k = 1, \ldots, n_i, j \neq k$. Therefore, the assumption of random slopes has induced marginal correlations that vary as a function of the timings of the measurements. After a model is fitted, the variances (8.9) and correlations (8.10) can be evaluated at the estimated variance components and compared to the empirical marginal variance and correlations.

In a longitudinal setting, an obvious extension to model (8.7) is provided by

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}_i + \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i, \tag{8.11}$$

with the error vectors $\boldsymbol{b}_i$, $\boldsymbol{\delta}_i$, and $\boldsymbol{\epsilon}_i$ representing individual-specific random effects, serial dependence, and measurement error. We assume

$$\begin{aligned} \text{E}[\boldsymbol{\epsilon}_i] &= \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \sigma_\epsilon^2 \mathbf{I}_{n_i} \\ \text{E}[\boldsymbol{b}_i] &= \mathbf{0}, \quad \text{var}(\boldsymbol{b}_i) = \boldsymbol{D} \\ \text{E}[\boldsymbol{\delta}_i] &= \mathbf{0}, \quad \text{var}(\boldsymbol{\delta}_i) = \sigma_\delta^2 \boldsymbol{R}_i \\ \text{cov}(\boldsymbol{b}_i, \boldsymbol{\epsilon}_{i'}) &= \mathbf{0}, \qquad i, i' = 1, \ldots, m, \end{aligned}$$

$$\begin{aligned}\text{cov}(\boldsymbol{b}_i, \boldsymbol{\delta}_{i'}) &= \mathbf{0}, \qquad i, i' = 1, \ldots, m,\\ \text{cov}(\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_{i'}) &= \mathbf{0}, \qquad i, i' = 1, \ldots, m,\end{aligned}$$

with $\text{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{i'}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{i'}) = \mathbf{0}$, and $\text{cov}(\boldsymbol{b}_i, \boldsymbol{b}_{i'}) = \mathbf{0}$, for $i \neq i'$. Here, $\boldsymbol{R}_i$ is an $n_i \times n_i$ correlation matrix with elements R_{ijk}, for $j, k = 1, \ldots, n_i$ which correspond to within individual correlations.

In general, it is difficult to identify/estimate all three sources of variability, but this formulation provides a useful conceptual model.

We now discuss specific choices of $\boldsymbol{R}_i$, beginning with a widely-used time series model, the first-order autoregressive, or AR(1), process. We assume initially that responses are observed at equally spaced times. For $j \geq 2$ and $|\rho| < 1$ suppose

$$\delta_{ij} = \rho \delta_{i,j-1} + u_{ij}, \tag{8.12}$$

with $\boldsymbol{u}_i = [u_{i1}, \ldots, u_{in_i}]^{\text{T}}$, $\text{E}[\boldsymbol{u}_i] = \mathbf{0}$, $\text{var}(\boldsymbol{u}_i) = \sigma_u^2 \mathbf{I}_{n_i}$, and with u_{ij} independent of all other error terms in the model. We first derive the marginal moments corresponding to this model. Repeated application of (8.12) gives, for $k > 0$,

$$\delta_{ij} = u_{ij} + \rho u_{i,j-1} + \rho^2 u_{i,j-2} + \ldots + \rho^{k-1} u_{i,j-k+1} + \rho^k \delta_{i,j-k} \tag{8.13}$$

so that

$$\text{var}(\delta_{ij}) = \sigma_u^2 (1 + \rho^2 + \rho^4 + \ldots + \rho^{2(k-1)}) + \rho^{2k} \text{var}(\delta_{i,j-k}).$$

Taking the limit as $k \to \infty$, and using $\sum_{l=1}^{\infty} x^{l-1} = (1-x)^{-1}$ for $|x| < 1$, gives

$$\text{var}(\delta_{ij}) = \frac{\sigma_u^2}{(1-\rho^2)} = \sigma_\delta^2,$$

which is the marginal variance of all of the δ error terms. Using (8.13),

$$\begin{aligned}\text{cov}(\delta_{ij}, \delta_{i,j-k}) &= \text{E}[\delta_{ij}\delta_{i,j-k}] = \rho^k \text{E}[\delta_{i,j-k}^2] = \rho^k \text{var}(\delta_{i,j-k}^2)\\ &= \rho^k \sigma_\delta^2,\end{aligned}$$

so that the correlations decline as observations become further apart in time. Under this model, the correlation matrix of $\boldsymbol{\delta}_i$ is

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n_i-1} \\ \rho & 1 & \rho & \ldots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \ldots & 1 \end{bmatrix}.$$

The autoregressive model is appealing in longitudinal settings and contains just two parameters, σ_δ^2 and ρ. The model can be extended to unequally spaced times to give covariance

$$\text{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_{ij}-t_{ik}|}. \tag{8.14}$$

A *Toeplitz* model assumes the variance is constant across time and that responses that are an equal distance apart in time have the same correlation.[1] For equally spaced responses in time:

$$\text{var}(\boldsymbol{\delta}_i) = \sigma_\delta^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n_i-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n_i-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i-1} & \rho_{n_i-2} & \rho_{n_i-3} & \cdots & 1 \end{bmatrix}.$$

This model may be useful in situations in which there is a common design across individuals, which allows estimation of the $n_i = n$ parameters ($n-1$ correlations and a variance). The AR(1) model is a special case in which $\rho_k = \rho^k$.

An *unstructured* covariance structure allows for different variances at each occasion $\sigma_{\delta 1}^2, \ldots, \sigma_{\delta n_i}^2$ and distinct correlations for each pair of responses, that is,

$$\text{corr}(\boldsymbol{\delta}_i) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n_i} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \rho_{n_i3} & \cdots & 1 \end{bmatrix}$$

with $\rho_{jk} = \rho_{kj}$, for $j, k = 1, \ldots, n_i$. This model contains $n_i(n_i+1)/2$ parameters per individual, which is a large number if n_i is large. If one has a common design across individuals, it may be plausible to fit this model, but one would still need a large number of individuals m, in order for inference to be reliable. As usual, there is a trade-off between flexibility and parsimony.

8.4.3 Parameter Interpretation for Linear Mixed Models

In this section we discuss how $\boldsymbol{\beta}$ and $\boldsymbol{b}$ may be interpreted in the LMM; this interpretation requires care, as we illustrate in the context of a longitudinal study

[1] In linear algebra, a Toeplitz matrix is a matrix in which each descending diagonal, from left to right, is constant.

with both random intercepts and random slopes. For a generic individual at time t, suppose the model is

$$\mathrm{E}[Y \mid \boldsymbol{b}, t] = (\beta_0 + b_0) + (\beta_1 + b_1)(t - \bar{t})$$

with $\boldsymbol{b} = [b_0, b_1]^{\mathrm{T}}$. The marginal model is

$$\mathrm{E}[Y \mid t] = \beta_0 + \beta_1(t - \bar{t}).$$

so that β_0 is the expected response at $t = \bar{t}$ and the slope parameter β_1 is the expected change in response for a unit increase in time. These expectations are with respect to the distribution of random effects and are averages across the population of individuals.

For a generic individual, $\beta_0 + b_0$ is the expected response at $t = \bar{t}$, and $\beta_1 + b_1$ is the expected change in response for a unit increase in time. In a linear model, β_1 is also the average of the individual slopes, $\beta_1 + b_1$. Consequently, *since the model is linear*, β_1 is both the expected change in the average response in unit time (across individuals) and the average of the individual expected changes in unit time. An alternative interpretation is that β_1 is the change in response for a unit change in t for a "typical" individual, that is, an individual with $b_1 = 0$. In Chap. 9 we will illustrate how the interpretation of parameters in mixed models becomes far more complex when the model is nonlinear in the parameters, and we will see that the consideration of a typical individual is particularly useful in this case.

8.5 Likelihood Inference for Linear Mixed Models

We now turn to inference and first consider likelihood methods for the LMM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i.$$

To implement a likelihood approach, we need to specify a complete probability distribution for the data, and this follows by specifying distributions for $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$, $i = 1, \ldots, m$. A common choice is $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{iid} \mathrm{N}_{n_i}(\,\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i})$ and $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_{q+1}(\,\mathbf{0}, \boldsymbol{D})$ where

$$\boldsymbol{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \cdots & \sigma_{0q} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_q^2 \end{bmatrix},$$

so that the vector of variance–covariance parameters is $\boldsymbol{\alpha} = [\sigma_\epsilon^2, \boldsymbol{D}]$. The marginal mean and variance are

$$\mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{\beta}] = \boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta} \tag{8.15}$$

$$\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}) = \boldsymbol{V}_i(\boldsymbol{\alpha}) = \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}} + \sigma_\epsilon^2\mathbf{I}_{n_i}. \tag{8.16}$$

We have refined notation in this section to explicitly condition on the relevant parameters. In general, inference may be required for the fixed effects regression parameters $\boldsymbol{\beta}$, the variance components $\boldsymbol{\alpha}$, or the random effects, $\boldsymbol{b} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^{\mathrm{T}}$. We consider each of these possibilities in turn.

8.5.1 Inference for Fixed Effects

Likelihood methods have traditionally been applied to nonrandom parameters, and so, we integrate over the random effects in the two-stage model to give

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \int_b p(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \times p(\boldsymbol{b} \mid \boldsymbol{\beta}, \boldsymbol{\alpha})\, d\boldsymbol{b}.$$

Exploiting conditional independencies, we obtain the simplified form

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{m} \int_{b_i} p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma_\epsilon^2) \times p(\boldsymbol{b}_i \mid \boldsymbol{D})\, d\boldsymbol{b}_i$$

and since a convolution of normals is normal, we obtain

$$\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha} \sim \mathrm{N}_{n_i}[\,\boldsymbol{\mu}_i(\boldsymbol{\beta}), \boldsymbol{V}_i(\boldsymbol{\alpha})\,],$$

where the marginal mean $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ and variance $\boldsymbol{V}_i(\boldsymbol{\alpha})$ correspond to (8.15) and (8.16), respectively. The log-likelihood is

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m} \log|\boldsymbol{V}_i(\boldsymbol{\alpha})| - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}). \tag{8.17}$$

The MLEs for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained via maximization of (8.17). The score equations for $\boldsymbol{\beta}$ are

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i - \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i\boldsymbol{\beta} \\ &= \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \qquad (8.18)\end{aligned}$$

and yield the MLE

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i \right), \tag{8.19}$$

which is a generalized least squares estimator. If $\boldsymbol{D} = \mathbf{0}$, then $\boldsymbol{V} = \sigma_\epsilon^2 \mathbf{I}_N$ (where $N = \sum_{i=1}^m n_i$), and $\widehat{\boldsymbol{\beta}}$ corresponds to the ordinary least squares estimator, as we would expect. The variance of $\widehat{\boldsymbol{\beta}}$ may be obtained either directly from (8.19), since the estimator is linear in $\boldsymbol{y}_i$, or from the second derivative of the log-likelihood.

The expected information matrix is block diagonal:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix} \tag{8.20}$$

so there is asymptotic independence between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ and any consistent estimator of $\boldsymbol{\alpha}$ will give an asymptotically efficient estimator for $\boldsymbol{\beta}$ (likelihood-based estimation of $\boldsymbol{\alpha}$ is considered in Sects. 8.5.2 and 8.5.3). Since

$$\boldsymbol{I}_{\beta\beta} = -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right] = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}, \tag{8.21}$$

the observed and expected information matrices coincide. The estimator $\widehat{\boldsymbol{\beta}}$ is linear in the data $\boldsymbol{Y}_i$, and so under normality of the data, $\widehat{\boldsymbol{\beta}}$ is normal also. Under correct specification of the variance model, and with a consistent estimator $\widehat{\boldsymbol{\alpha}}$,

$$\left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i \right)^{1/2} (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \to_d \mathrm{N}_{k+1}(\,\mathbf{0}, \mathbf{I}),$$

as $m \to \infty$. Since $\widehat{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$, it follows immediately that this asymptotic distribution is also appropriate when the data and random effects are not normal. We require the second moments of the data to be correctly specified, however. In Sect. 8.7 we describe how a consistent variance estimator may be obtained when $\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'} \mid \boldsymbol{\alpha}) = \mathbf{0}$, but $\mathrm{var}(\mathbf{Y_i} \mid \boldsymbol{\alpha}) = \mathbf{V}_i(\boldsymbol{\alpha})$ is not necessarily correctly specified.

In terms of the asymptotics it is not sufficient to have m fixed and $n_i \to \infty$ for $i = 1, \ldots, m$. We illustrate for the LMM with $\boldsymbol{z}_i = \boldsymbol{x}_i$, in which case $\boldsymbol{V}_i = \boldsymbol{x}_i \boldsymbol{D} \boldsymbol{x}_i^{\mathrm{T}} + \sigma_\epsilon^2 \mathbf{I}_{n_i}$. Under this setup,

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x}_i \right)^{-1} \\ &= \left(\sum_{i=1}^{m} \left[(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i)^{-1} \sigma_\epsilon^{-2} + \boldsymbol{D} \right]^{-1} \right)^{-1}, \end{aligned}$$

where we have used the matrix identity $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i = \left[(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma_\epsilon^2 + \boldsymbol{D}\right]^{-1}$ (which may be derived from (B.3) of Appendix B). When $n_i \to \infty$,

$$(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i)^{-1} = O(n_i^{-1}) \to \mathbf{0},$$

and if m is fixed,

$$\text{var}(\widehat{\boldsymbol{\beta}}) \to \frac{\boldsymbol{D}}{m},$$

showing that we require $m \to \infty$ for consistency of $\widehat{\boldsymbol{\beta}}$.

Likelihood ratio tests can be used to test hypotheses concerning elements of $\boldsymbol{\beta}$, for fixed $\boldsymbol{\alpha}$ or, in practice, the substitution of an estimate $\boldsymbol{\alpha}$. Various t and F-like approaches have been suggested for correcting for the estimation of $\boldsymbol{\alpha}$, see Verbeeke and Molenberghs (2000, Chap. 6), but if the sample size m is not sufficiently large for reliable estimation of $\boldsymbol{\alpha}$, we recommend resampling methods, or following a Bayesian approach to inference, since this produces inference for $\boldsymbol{\beta}$ that averages over the uncertainty in the estimation of $\boldsymbol{\alpha}$.

For more complex linear models, inference may not be so straightforward. For example, consider the model

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \epsilon_{ij} = \mu_{ij} + \epsilon_{ij}$$

but with nonconstant measurement error variance. A common model is $\text{var}(Y_{ij}) = \sigma_\epsilon^2\mu_{ij}^\gamma$, for known $\gamma > 0$. In this case the MLE for $\boldsymbol{\beta}$ is not available in closed form, and we do not have a diagonal information matrix as in (8.20). An example of the fitting of such a model in a nonlinear setting is given at the end of Sect. 9.20.

Maximum likelihood estimation is also theoretically straightforward for the extended model (8.11) in which we have a richer variance model, but identifiability issues may arise due to the complexity of the error structure.

8.5.2 *Inference for Variance Components via Maximum Likelihood*

The MLE $\widehat{\boldsymbol{\alpha}}$ is obtained from maximization of (8.17), but in general, there is no closed-form solution. However, the expectation-maximization (EM, Dempster et al. 1977) or Newton–Raphson algorithm may be applied to the profile likelihood:

$$l_p(\boldsymbol{\alpha}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\alpha})| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\alpha})^{-1}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}),$$

since recall from Sect. 2.4.2 that the MLE for $\boldsymbol{\alpha}$ is identical to the estimate obtained from the profile likelihood. Under standard likelihood theory,

$$\boldsymbol{I}_{\alpha\alpha}^{1/2}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \to_d \text{N}_r(\mathbf{0}, \mathbf{I}_r),$$

where r is the number of distinct elements of $\boldsymbol{\alpha}$. This distribution provides asymptotic confidence intervals for elements of $\boldsymbol{\alpha}$.

Testing whether random effect variances are zero requires care since the null hypothesis lies on the boundary, and so, the usual regularity conditions are not satisfied. We illustrate by considering the model

$$Y_{ij} = \beta_0 + \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \epsilon_{ij}$$

with $b_i \mid \sigma_0^2 \sim \text{N}(0, \sigma_0^2)$. Suppose we wish to test whether the random effects variance is zero, that is, $H_0 : \sigma_0^2 = 0$ versus $H_1 : \sigma_0^2 > 0$. In this case, the asymptotic null distribution is a 50:50 mixture of χ_0^2 and χ_1^2 distributions, where the former is the distribution that gives probability mass 1 to the value 0. For example, the 95% points of a χ_1^2 and the 50:50 mixture are 3.84 and 2.71, respectively. Consequently, if the usual χ_1^2 distribution is used, the null will be accepted too often, leading to a variance component structure that is too simple.

The intuition behind the form of the null distribution is the following. Estimating σ_0^2 is equivalent to estimating $\rho = \sigma_0^2/\sigma^2$ and setting $\widehat{\rho} = 0$ if the estimated correlation is negative, and under the null, this will happen half the time. If $\widehat{\rho} = 0$, then we recover the null for the distribution of the data, and so, the likelihood ratio will be 1. This gives the mass at the value 0, and combining with the usual χ_1^2 distribution gives the 50:50 mixture.

If H_0 and H_1 correspond to models with k and $k+1$ random effects, respectively, each with general covariance structures, then the asymptotic distribution is a 50:50 mixture of χ_k^2 and χ_{k+1}^2 distributions. Hence, for example, if we wish to test random intercepts only versus correlated random intercepts and random slopes (with $\boldsymbol{D}$ having elements $\sigma_0^2, \sigma_{01}, \sigma_1^2$), then the distribution of the likelihood ratio statistic is a 50:50 mixture of χ_1^2 and χ_2^2 distributions. Similar asymptotic results are available for more complex models/hypotheses; see, for example, Verbeeke and Molenberghs (2000).

8.5.3 Inference for Variance Components via Restricted Maximum Likelihood

While MLE for variance components yields consistent estimates under correct model specification, the estimation of $\boldsymbol{\beta}$ is not acknowledged, in the sense that inference proceeds as if $\boldsymbol{\beta}$ were known. We have already encountered this in Sect. 2.4.2 for the simple linear model where it was shown that the MLE of σ^2 is RSS$/n$, while the unbiased version is RSS$/(n - k - 1)$, where RSS is the residual sum of squares and k is the number of covariates. An alternative, and often preferable, method that acknowledges estimation of $\boldsymbol{\beta}$ is provided by restricted (or residual) maximum likelihood (REML). We provide a Bayesian justification for REML in Sect. 8.6 and here provide another derivation based on marginal likelihood.

Recall the definition of marginal likelihood from Sect. 2.4.2. Let $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, be minimal sufficient statistics and suppose

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \propto p(\boldsymbol{s}_1, \boldsymbol{s}_2 \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) = p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda}) p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{\lambda}, \boldsymbol{\phi}) \tag{8.22}$$

where $\boldsymbol{\lambda}$ represents the parameters of interest and $\boldsymbol{\phi}$ the remaining (nuisance) parameters. Inference for $\boldsymbol{\lambda}$ may be based on the marginal likelihood $L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda})$. We discuss how marginal likelihoods may be derived for general LMMs.

To derive a marginal likelihood, we need to find a function of the data, $\boldsymbol{U} = f(\boldsymbol{Y})$, whose distribution does not depend upon $\boldsymbol{\beta}$. We briefly digress to discuss an *error contrast*, $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}$, which is defined by the property that $\mathrm{E}[\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}] = 0$ for all values of $\boldsymbol{\beta}$, with $\boldsymbol{C}$ an N-dimensional vector. For the LMM

$$\mathrm{E}[\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}] = 0 \;\text{ for all } \boldsymbol{\beta} \text{ if and only if } \;\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x} = 0.$$

When $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0}$,

$$\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{z}\boldsymbol{b} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\epsilon},$$

which does not depend on $\boldsymbol{\beta}$, suggesting that the marginal likelihood could be based on error contrasts. If $\boldsymbol{x}$ is of full rank, that is, is of rank $k+1$, there are exactly $N-k-1$ linearly independent error contrasts (since $k+1$ fixed effects have been estimated, which induces dependencies in the error contrasts). Let $\boldsymbol{B} = [\boldsymbol{C}_1, \ldots, \boldsymbol{C}_{N-k-1}]$ denote an error contrast matrix. Given two error contrast matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, it can be shown that there exists a full rank, $(N-k-1) \times (N-k-1)$ matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{B}_1^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{B}_2^{\mathrm{T}}$. Therefore, likelihoods based on $\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{Y}$ or on $\boldsymbol{B}_2^{\mathrm{T}}\boldsymbol{Y}$ will be proportional, and estimators based on either will be identical. Let $\boldsymbol{H} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}$, and choose $\boldsymbol{B}$ such that $\mathbf{I} - \boldsymbol{H} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ and $\mathbf{I} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$. It is easily shown that $\boldsymbol{B}$ is an error contrast matrix since

$$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{B}^{\mathrm{T}}(\mathbf{I} - \boldsymbol{H})\boldsymbol{x} = \boldsymbol{0}.$$

The function of the data we consider is therefore $\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$ which may be written as

$$\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}(\mathbf{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{r},$$

where $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$, and $\widehat{\boldsymbol{\beta}}_{\mathrm{O}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ is the OLS estimator, showing that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$ is a linear combination of residuals (hence the name "residual" maximum likelihood). Since $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0}$, we can confirm that

$$\boldsymbol{U} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{z}\boldsymbol{b} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\epsilon},$$

with $\mathrm{E}[\boldsymbol{U}] = \boldsymbol{0}$. Further, the distribution of $\boldsymbol{U}$ does not depend upon $\boldsymbol{\beta}$, as required for a marginal likelihood.

We now derive the distribution of $\boldsymbol{U}$ by considering the transformation from $\boldsymbol{Y} \to [\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}}] = [\boldsymbol{B}^{\text{T}}\boldsymbol{Y}, \boldsymbol{G}^{\text{T}}\boldsymbol{Y}]$, where

$$\widehat{\boldsymbol{\beta}}_{\text{G}} = \boldsymbol{G}^{\text{T}}\boldsymbol{Y} = (\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}$$

is the generalized least squares (GLS) estimator. We derive the Jacobian of the transformation, using (B.1) and (B.2) in Appendix B:

$$\begin{aligned} |\boldsymbol{J}| &= \left| \frac{\partial(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}})}{\partial \boldsymbol{Y}} \right| = |\boldsymbol{B} \;\; \boldsymbol{G}| = \left| \begin{bmatrix} \boldsymbol{B}^{\text{T}} \\ \boldsymbol{G}^{\text{T}} \end{bmatrix} [\boldsymbol{B} \;\; \boldsymbol{G}] \right|^{1/2} \\ &= |\; \boldsymbol{B}^{\text{T}}\boldsymbol{B} \;|^{1/2} |\; \boldsymbol{G}^{\text{T}}\boldsymbol{G} - \boldsymbol{G}^{\text{T}}\boldsymbol{B}(\boldsymbol{B}^{\text{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\text{T}}\boldsymbol{G} \;|^{1/2} \\ &= 1 \times |\; \boldsymbol{G}^{\text{T}}\boldsymbol{G} - \boldsymbol{G}^{\text{T}}(\mathbf{I} - \boldsymbol{H})\boldsymbol{G} \;|^{1/2} \\ &= \boldsymbol{G}^{\text{T}}\boldsymbol{H}\boldsymbol{G} = |\; \boldsymbol{x}^{\text{T}}\boldsymbol{x} \;|^{-1/2} \neq 0 \end{aligned}$$

which implies that $[\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}}]$ is of full rank (and equal to N). The vector $[\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}}]$ is a linear combination of normals and so is normal, and

$$\begin{aligned} \text{cov}(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}}) &= \text{E}[\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{\text{G}} - \boldsymbol{\beta})^{\text{T}}] \\ &= \text{E}[\boldsymbol{B}^{\text{T}}\boldsymbol{Y}\boldsymbol{Y}^{\text{T}}\boldsymbol{G}] - \text{E}[\boldsymbol{B}^{\text{T}}\boldsymbol{Y} - \boldsymbol{\beta}^{\text{T}}] \\ &= \boldsymbol{B}^{\text{T}}\left[\text{var}(\boldsymbol{Y}) + \text{E}(\boldsymbol{Y})\text{E}(\boldsymbol{Y}^{\text{T}})\right]\boldsymbol{G} + \boldsymbol{B}^{\text{T}}\boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{\beta}^{\text{T}} \\ &= \boldsymbol{B}^{\text{T}}\boldsymbol{V}\boldsymbol{G}^{\text{T}} + \boldsymbol{B}^{\text{T}}\boldsymbol{x}\boldsymbol{\beta}(\boldsymbol{x}\boldsymbol{\beta})^{\text{T}} \\ &= \; \mathbf{0}, \end{aligned}$$

where we have repeatedly used $\boldsymbol{B}^{\text{T}}\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{V} = \text{var}(\boldsymbol{Y})$. So $\boldsymbol{U}$ and $\widehat{\boldsymbol{\beta}}_{\text{G}}$ are uncorrelated and, since they are normal, independent also. Consequently,

$$\begin{aligned} p(\boldsymbol{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\text{G}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{J} \mid \\ &= p(\boldsymbol{U} \mid \widehat{\boldsymbol{\beta}}_{\text{G}}, \boldsymbol{\beta}) p(\widehat{\boldsymbol{\beta}}_{\text{G}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \, |\boldsymbol{J}| \\ &= p(\boldsymbol{U} \mid \boldsymbol{\alpha}) p(\widehat{\boldsymbol{\beta}}_{\text{G}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \, |\boldsymbol{J}|. \end{aligned} \tag{8.23}$$

By comparison with (8.22), we have $\boldsymbol{s}_1 = \boldsymbol{U}$, $\boldsymbol{s}_2 = \widehat{\boldsymbol{\beta}}_{\text{G}}$, $\boldsymbol{\lambda} = \boldsymbol{\alpha}$, and $\boldsymbol{\phi} = \boldsymbol{\beta}$, and $p(\boldsymbol{U} \mid \boldsymbol{\alpha})$ is a marginal likelihood. Rearrangement of (8.23) gives

$$p(\boldsymbol{U} \mid \boldsymbol{\alpha}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\widehat{\boldsymbol{\beta}}_{\text{G}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})} |\boldsymbol{J}|^{-1}.$$

Since

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = (2\pi)^{-N/2} |\boldsymbol{V}|^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\text{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) \right],$$

and

$$p(\widehat{\boldsymbol{\beta}}_{\text{G}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = (2\pi)^{-(k+1)/2} |\boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x}|^{1/2} \exp\left[-\frac{1}{2} (\widehat{\boldsymbol{\beta}}_{\text{G}} - \boldsymbol{\beta})^{\text{T}} \boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x} (\widehat{\boldsymbol{\beta}}_{\text{G}} - \boldsymbol{\beta}) \right]$$

we obtain the marginal likelihood

$$p(\boldsymbol{U} \mid \boldsymbol{\alpha}) = c \frac{|\boldsymbol{x}^{\text{T}} \boldsymbol{x}|^{1/2} |\boldsymbol{V}|^{-1/2}}{|\boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x}|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}})^{\text{T}} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}}) \right]$$

with $c = (2\pi)^{-(N-k-1)/2}$, which (as already mentioned) does not depend upon $\boldsymbol{B}$. Hence, we can choose any linearly independent combination of the residuals.

The restricted log-likelihood upon which inference for $\boldsymbol{\alpha}$ may be based is

$$l_R(\boldsymbol{\alpha}) = -\frac{1}{2} \log |\boldsymbol{x}^{\text{T}} \boldsymbol{V}(\boldsymbol{\alpha})^{-1} \boldsymbol{x}| - \frac{1}{2} \log |\boldsymbol{V}(\boldsymbol{\alpha})| - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}})^{\text{T}} \boldsymbol{V}(\boldsymbol{\alpha})^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}}).$$

Comparison with the profile log-likelihood for $\boldsymbol{\alpha}$,

$$l_P(\boldsymbol{\alpha}) = -\frac{1}{2} \log |\boldsymbol{V}(\boldsymbol{\alpha})| - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}})^{\text{T}} \boldsymbol{V}(\boldsymbol{\alpha})^{-1} (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}}),$$

shows that we have an additional term, $-\frac{1}{2} \log |\boldsymbol{x}^{\text{T}} \boldsymbol{V}(\boldsymbol{\alpha})^{-1} \boldsymbol{x}|$, that may be viewed as accounting for the degrees of freedom lost in estimation of $\boldsymbol{\beta}$. Computationally, finding REML estimators is as straightforward as their ML counterparts, as the objective functions differ simply by a single term. Both ML and REML estimates may be obtained using EM or Newton–Raphson algorithms; see Pinheiro and Bates (2000) for details.

In general, REML estimators have finite sample bias, but they are less biased than ML estimators, particularly for small samples. So far, as estimation of the variance components are concerned, the asymptotic distribution of the REML estimator is normal, with variance given by the inverse of the Fisher's information matrix, where the latter is based on $l_R(\boldsymbol{\alpha})$.

REML is effectively based on a likelihood with data constructed from the distribution of the residuals $\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\text{G}}$. Therefore, when two regression models are to be compared, the data under the two models are different; hence, REML likelihood ratio tests for elements of $\boldsymbol{\beta}$ cannot be performed. Consequently, when a likelihood ratio test is required to formally compare two nested regression models, maximum likelihood must be used to fit the models. Likelihood ratio tests for variance components are valid under restricted maximum likelihood, however, since the covariates, and hence residuals, are constant in both models.

Example: One-Way ANOVA

The simplest example of a LMM is the balanced one-way random effects ANOVA model:

$$Y_{ij} = \beta_0 + b_i + \epsilon_{ij},$$

with b_i and ϵ_{ij} independent and distributed as $b_i \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$ and $\epsilon_{ij} \mid \sigma_\epsilon^2 \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2)$, with n observations on each unit and $i = 1, \dots, m$ to give $N = nm$ observations in total. In this example, $\boldsymbol{\beta} = \beta_0$ and $\boldsymbol{\alpha} = [\sigma_\epsilon^2, \sigma_0^2]$. This model was considered briefly in Sect. 5.8.4.

The model can be written in the form of (8.7) as

$$\boldsymbol{y}_i = \mathbf{1}_n \beta_0 + \mathbf{1}_n b_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{y}_i = [y_{i1}, \dots, y_{in}]^{\mathrm{T}}$ and $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \dots, \epsilon_{in}]^{\mathrm{T}}$. Marginally, this specification implies that the data are normal with $\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{\beta}] = \mathbf{1}_N \beta_0$ and $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{\alpha}) = \mathrm{diag}(\boldsymbol{V}_1, \dots, \boldsymbol{V}_m)$ where

$$\boldsymbol{V}_i = \mathbf{1}_n \, \mathbf{1}_n^{\mathrm{T}} \sigma_0^2 + \mathbf{I}_n \sigma_\epsilon^2,$$

for $i = 1, \dots, m$. In the case of $n = 3$ observations per unit, this yields the $N \times N$ marginal variance

$$\boldsymbol{V} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \rho & \rho & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho & \rho & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \rho & 1 & \rho & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \rho & \rho & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & \rho & \rho \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \rho & \rho & 1 \end{bmatrix},$$

where $\sigma^2 = \sigma_\epsilon^2 + \sigma_0^2$ is the marginal variance of each observation, and

$$\rho = \frac{\sigma_0^2}{\sigma^2} = \frac{\sigma_0^2}{\sigma_\epsilon^2 + \sigma_0^2}$$

is the marginal correlation between two observations on the same unit. The correlation, ρ, is induced by the shared random effect and is referred to as the *intra-class correlation coefficient*.

For some data/mixed effects model combinations, there are more combined fixed and random effects than data points, which is at first sight disconcerting, but the random effects have a special status since they are tied together through a common distribution. In the above ANOVA model, we have $m + 3$ unknown quantities if we include the random effects, but these random effects may be integrated from the model so that the distribution of the data may be written in terms of the three parameters, $[\beta_0, \sigma_0^2, \sigma_\epsilon^2]$ only, without reference to the random effects, that is,

$$\boldsymbol{Y} \mid \beta_0, \sigma_0^2, \sigma_\epsilon^2 \sim \mathrm{N}_N \left[\, \mathbf{1}\beta_0, \boldsymbol{V}(\sigma_0^2, \sigma_\epsilon^2) \, \right].$$

A fixed effects model with a separate parameter for each group has $m+1$ parameters, which shows that the mixed effects model can offer a parsimonious description.

The MLE for β_0 is given by the grand mean, i.e., $\widehat{\beta}_0 = \overline{Y}_{\cdot\cdot}$. With balanced data the ML and REML estimators for the variance components are available in closed form (see Exercise 8.2). We define the between- and within-group mean squares as

$$\text{MSA} = \frac{n\sum_{i=1}^{m}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2}{m-1}, \quad \text{MSE} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{y}_{ij} - \overline{y}_{i\cdot})^2}{m(n-1)}.$$

The MLEs of the variance components are

$$\widehat{\sigma}_\epsilon^2 = \text{MSE},$$
$$\widehat{\sigma}_0^2 = \max\left(0, \frac{(1-1/m)\text{MSA} - \text{MSE}}{n}\right).$$

The REML for $\widehat{\sigma}_\epsilon^2$ is the same as the MLE, but the REML estimate for σ_0^2 is

$$\widehat{\sigma}_0^2 = \max\left(0, \frac{\text{MSA} - \text{MSE}}{n}\right),$$

which is slightly larger than the ML estimate, having accounted for the estimation of β_0. Notice that the ML and REML estimators for σ_0^2 may be zero.

Example: Dental Growth Curves

We consider the full data and fit a model with distinct fixed effects (intercepts and slopes) for boys and girls and with random intercepts and slopes but with a common random effects distribution for boys and girls. Specifically, at stage one,

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_j + \epsilon_{ij}$$

for boys, $i = 1, \ldots, 16$, and

$$Y_{ij} = (\beta_0 + \beta_2 + b_{i0}) + (\beta_1 + \beta_4 + b_{i1})t_j + \epsilon_{ij}$$

for girls, $i = 17, \ldots, 27$. At stage two,

$$\boldsymbol{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \mid \boldsymbol{D} \sim_{iid} \text{N}_2\left(\, \boldsymbol{0}, \boldsymbol{D}\,\right), \quad \boldsymbol{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01}^2 & \sigma_1^2 \end{bmatrix}$$

for $i = 1, \ldots, 27$. We take $[t_1, t_2, t_3, t_4] = [-2, -1, 1, 2]$ so that we have centered by the average age of 11 years. In the generic notation introduced in Sect. 8.4, the above model translates to

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i + \epsilon_{ij}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^{\mathrm{T}}$, and the design matrices for the fixed and random effects are

$$\boldsymbol{x}_{ij} = \begin{cases} [1, t_j, 0, 0\,] & \text{for } i = 1, \ldots, 16 \\ [1, t_j, 1, t_j] & \text{for } i = 17, \ldots, 27, \end{cases}$$

and $\boldsymbol{z}_{ij} = [1, t_j]$, where $j = 1, 2, 3, 4$. Therefore, β_0 is the average tooth length at 11 years for boys, β_1 is the slope for boys (specifically the average change in tooth length between two populations of boys whose ages differ by 1 year), β_2 is the difference between the average tooth lengths of girls and boys at 11 years, and β_3 is the average difference in slopes between girls and boys. The intercept random effects b_{i0} may be viewed as the accumulation of all unmeasured variables that contribute to the tooth length for child i differing from the relevant (boy or girl) population average length (measured at 11 years). The slope random effects b_{i1} are the child by time interaction terms and summarize all of the unmeasured variables for child i that lead to the rate of change in growth for this child differing from the relevant (boy or girl) population average.

Fitting this model via REML yields

$$\widehat{\boldsymbol{\beta}} = [25, 0.78, -2.3, -0.31]^{\mathrm{T}}$$

with standard errors

$$[0.49, 0.086, 0.76, 0.14].$$

The asymptotic 95% confidence interval for the average difference in tooth lengths at 11 years is $[-3.8, -0.83]$, from which we conclude that the average tooth lengths at 11 years is greater for boys than for girls. The 95% interval for the slope difference is $[-0.57, -0.04]$ suggesting that the average rate of growth is greater for boys also.

There are a number of options to test whether gender-specific slopes are required, that is, to decide on whether $\beta_4 = 0$. A Wald test using the REML estimates gives a p-value of 0.026 (so that one endpoint of a 97.4% confidence interval is zero), which conventionally would suggest a difference in slopes. To perform a likelihood ratio test, we need to carry out a fit using ML, since REML is not valid, as explained in Sect. 8.5.2. Fitting the models with and without distinct slopes gives a change in twice the log-likelihood of 5.03, with an associated p-value of 0.036, which is consistent with the Wald test. Hence, there is reason to believe that the slopes for boys and girls are unequal, with the increase in the average growth over 1 year being estimated as 0.3 mm greater for boys than for girls.

The estimated variance–covariance matrices of the random effects, $\widehat{\boldsymbol{D}}$, under REML and ML are

$$\begin{bmatrix} 1.84^2 & 0.21 \times 1.84 \times 0.18 \\ 0.21 \times 1.84 \times 0.18 & 0.18^2 \end{bmatrix},$$

and

$$\begin{bmatrix} 1.75^2 & 0.23 \times 1.75 \times 0.15 \\ 0.23 \times 1.75 \times 0.15 & 0.15^2 \end{bmatrix}$$

so that, as expected, the REML estimates are slightly larger. Although $\widehat{\boldsymbol{\beta}}$ depends on $\widehat{\boldsymbol{D}}$, the point estimates of $\boldsymbol{\beta}$ are identical under ML and REML here, because of the balanced design. The standard errors for elements of $\boldsymbol{\beta}$ are slightly larger under REML, due to the differences in $\widehat{\boldsymbol{V}}$.

Under REML, the estimated standard deviations of the distributions of the intercepts and slopes are $\widehat{\sigma}_0 = 1.84$ and $\widehat{\sigma}_1 = 0.18$, respectively. Whether these are "small" or "not small" relates to the scale of the variables with which they are associated. Interpretation of elements of $\boldsymbol{D}$ depends, in general, on how we parameterize the time variable. For example, if we changed the time scale via a location shift, we would change the definition of the intercept. As parameterized above, the off-diagonal term D_{01} describes the covariance between the child-specific responses at 11 years and the child-specific slopes (the REML estimates of the correlation between these quantities is 0.23).

Suppose we reparameterize stage one of the model as

$$\mathrm{E}[Y_{ij} \mid \boldsymbol{b}_i^\star] = (\beta_0^\star + b_{i0}^\star) + (\beta_1 + b_{i1})t_j^\star$$

with $[t_1^\star, t_2^\star, t_3^\star, t_4^\star] = [8, 10, 12, 14]$ and $\boldsymbol{b}_i^\star = [b_{i0}^\star, b_{i1}]^{\mathrm{T}}$. Then $\beta_0^\star = \beta_0 - \beta_1\bar{t}$, $b_{i0}^\star = b_{i0} - b_{i1}\bar{t}$, and

$$\begin{aligned} D_{00}^\star &= D_{00} - 2\bar{t}D_{01} + \bar{t}^2 D_{11} \\ D_{01}^\star &= D_{01} - \bar{t}D_{11} \\ D_{11}^\star &= D_{11}. \end{aligned}$$

Consequently, only the interpretation of the variance of the slopes remains unchanged, when compared with the previous parameterization.

We return to the original parameterization and examine further the fitting of this model. Since we have assumed a common measurement error variance σ_ϵ^2, and common random effects variances $\boldsymbol{D}$ for boys and girls, the implied marginal standard deviations and correlations are the same for boys and girls and may be estimated from (8.9) and (8.10). Under REML, $\widehat{\sigma}_\epsilon = 1.31$ and the standard deviations (on the diagonal) and correlations (on the off-diagonal) are

$$\begin{bmatrix} 2.23 & & & \\ 0.65 & 2.23 & & \\ 0.64 & 0.65 & 2.30 & \\ 0.62 & 0.65 & 0.68 & 2.35 \end{bmatrix}. \tag{8.24}$$

We see that the standard deviations increases slightly over time, and the correlations decrease only slightly for observations further apart in time, suggesting that the random slopes are not contributing greatly to the fit. Fitting a random-intercepts-only model to these data produced a marginal variance estimate of 2.28^2 and common within-child correlations of 0.63.

The empirical standard deviations and correlations for boys and girls are given, respectively, by

$$\begin{bmatrix} 2.45 & & & \\ 0.44 & 2.14 & & \\ 0.56 & 0.39 & 2.65 & \\ 0.32 & 0.63 & 0.59 & 2.09 \end{bmatrix}, \quad \begin{bmatrix} 2.12 & & & \\ 0.83 & 1.90 & & \\ 0.86 & 0.90 & 2.36 & \\ 0.84 & 0.88 & 0.95 & 2.44 \end{bmatrix}$$

which suggests that our model needs refinement, since clearly the correlations for girls are greater than for boys.

8.5.4 Inference for Random Effects

In some situations, interest will focus on inference for the random effects. For example, for the dental data, we may be interested in the growth curve of a particular child. Estimates of random effects are also important for model checking.

Various approaches to inference for random effects have been proposed. The simplest, which we describe first, is to take an empirical Bayes approach. From a Bayesian standpoint, there is no distinction inferentially between fixed and random effects (the distinction is in the priors that are assigned). Consequently, inference is simply based on the posterior distribution $p(\boldsymbol{b}_i \mid \boldsymbol{y})$. Consider the LMM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

and assume $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent with $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \text{N}_{q+1}(\,\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{ind} \text{N}_{n_i}(\,\boldsymbol{0}, \sigma_\epsilon^2\mathbf{I})$, so that $\boldsymbol{\alpha} = [\sigma_\epsilon^2, \boldsymbol{D}]$. We begin by considering the simple, albeit unrealistic, situation, in which $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are known. Letting $\boldsymbol{y}_i^\star = \boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}$, we have

$$\begin{aligned} p(\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) &\propto p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) \times \pi(\boldsymbol{b}_i \mid \boldsymbol{\alpha}) \\ &\propto \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y}_i^\star - \boldsymbol{z}_i\boldsymbol{b}_i)^{\text{T}}(\boldsymbol{y}_i^\star - \boldsymbol{z}_i\boldsymbol{b}_i) - \frac{1}{2}\boldsymbol{b}_i^{\text{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i\right] \end{aligned}$$

which we recognize as a multiple linear regression with a zero-centered normal prior on the parameters $\boldsymbol{b}_i$ (this model is closely linked to that used in ridge regression, see Sect. 10.5.1). Using a standard derivation, (5.7),

$$\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha} \sim \text{N}_{q+1}\left[\,\text{E}(\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}), \text{var}(\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha})\,\right]$$

with mean and variance

$$\begin{aligned} \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}] &= \left(\frac{\boldsymbol{z}_i^{\text{T}}\boldsymbol{z}_i}{\sigma_\epsilon^2} + \boldsymbol{D}^{-1}\right)^{-1} \frac{\boldsymbol{z}_i^{\text{T}}}{\sigma_\epsilon^2}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \\ &= \boldsymbol{D}\boldsymbol{z}_i^{\text{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \end{aligned} \tag{8.25}$$

$$\begin{aligned}\text{var}(\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \left(\frac{\boldsymbol{z}_i^{\text{T}} \boldsymbol{z}_i}{\sigma_\epsilon^2} + \boldsymbol{D}^{-1} \right)^{-1} \\ &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} \boldsymbol{z}_i \boldsymbol{D}, \end{aligned} \tag{8.26}$$

see Exercise 8.4. As we will see in this section, the estimate (8.25) may be derived under a number of different formulations.

A fully Bayesian approach would consider

$$p(\boldsymbol{b} \mid \boldsymbol{y}) = \int \int p(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{y}) \, d\boldsymbol{\beta} d\boldsymbol{\alpha},$$

which emphasizes that the uncertainty in $\boldsymbol{\beta}, \boldsymbol{\alpha}$ is not acknowledged in the derivation of (8.25) and (8.26).

We now demonstrate how we may account for estimation of $\boldsymbol{\beta}$ with a flat prior on $\boldsymbol{\beta}$ and assuming $\boldsymbol{\alpha}$ known. The posterior mean and variance of $\boldsymbol{\beta}$ are

$$\begin{aligned}\text{E}[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha}] &= \widehat{\boldsymbol{\beta}}_{\text{G}} \\ \text{var}(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha}) &= (\boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x})^{-1}\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{\text{G}}$ is the GLS estimator (these forms are derived for more general priors later, see (8.35) and (8.36)). Consequently,

$$\begin{aligned}\text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\alpha}] &= \text{E}_{\beta \mid y, \alpha} \left[\text{E}(\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\alpha}) \right] \\ &= \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}_{\text{G}}) \end{aligned} \tag{8.27}$$

$$\begin{aligned}\text{var}(\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\alpha}) &= \text{E}_{\beta \mid y, \alpha}[\text{var}(\boldsymbol{b}_i \mid \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\alpha})] + \text{var}_{\beta \mid y, \alpha}(\text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\alpha}]) \\ &= \text{E}_{\beta \mid y, \alpha}[\boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} \boldsymbol{z}_i \boldsymbol{D}] + \text{var}_{\beta \mid y, \alpha}(\boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta})) \\ &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} \boldsymbol{z}_i \boldsymbol{D} + \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i (\boldsymbol{x}^{\text{T}} \boldsymbol{V}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}_i^{\text{T}} \boldsymbol{V}_i^{-1} \boldsymbol{z}_i \boldsymbol{D}. \end{aligned} \tag{8.28}$$

Therefore, we can easily account for the estimation of $\boldsymbol{\beta}$, but no such simple development is available to account for estimation of $\boldsymbol{\alpha}$.

From a frequentist perspective, inference for random effects is often viewed as *prediction* rather than estimation, since $[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$ are random variables and not unknown constants. Many different criteria may be used to find a predictor $\widehat{\boldsymbol{b}} = f(\boldsymbol{Y})$ of $\boldsymbol{b}$, for a generic unit.

We begin by defining the optimum predictor as that which minimizes the mean squared error (MSE). Let $\boldsymbol{b}^\star$ represent a general predictor and consider the MSE:

$$\text{MSE}(\boldsymbol{b}^\star) = \text{E}_{\boldsymbol{y}, \boldsymbol{b}}[(\boldsymbol{b}^\star - \boldsymbol{b})^{\text{T}} \boldsymbol{A} (\boldsymbol{b}^\star - \boldsymbol{b})],$$

where we emphasize that the expectation is with respect to both $\boldsymbol{y}$ and $\boldsymbol{b}$, and $\boldsymbol{A}$ is any positive definite symmetric matrix. We show that the MSE is minimized by $\widehat{\boldsymbol{b}} = \mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$. For the moment, we suppress the dependence on any additional parameters. We can express the MSE in terms of $\boldsymbol{b}^\star$ and $\widehat{\boldsymbol{b}}$:

$$\begin{aligned}\mathrm{MSE}(\boldsymbol{b}^\star) &= \mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\boldsymbol{b}^\star - \boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{b}^\star - \boldsymbol{b})]\\ &= \mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}} + \widehat{\boldsymbol{b}} - \boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}} + \widehat{\boldsymbol{b}} - \boldsymbol{b})]\\ &= \mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})] + 2 \times \mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})\boldsymbol{A}(\widehat{\boldsymbol{b}} - \boldsymbol{b})]\\ &\quad + \mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\widehat{\boldsymbol{b}} - \boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{b}} - \boldsymbol{b})]. \end{aligned} \tag{8.29}$$

The third term does not involve $\boldsymbol{b}^\star$, and we may write the second expectation as

$$\begin{aligned}\mathrm{E}_{\boldsymbol{Y},\boldsymbol{b}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})\boldsymbol{A}(\widehat{\boldsymbol{b}} - \boldsymbol{b})] &= \mathrm{E}_{\boldsymbol{Y}}\{\mathrm{E}_{\boldsymbol{b}|\boldsymbol{y}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})\boldsymbol{A}(\widehat{\boldsymbol{b}} - \boldsymbol{b}) \mid \boldsymbol{y}]\}\\ &= \mathrm{E}_{\boldsymbol{Y}}[(\boldsymbol{b}^\star - \widehat{\boldsymbol{b}})\boldsymbol{A}(\widehat{\boldsymbol{b}} - \widehat{\boldsymbol{b}})] = \boldsymbol{0}\end{aligned}$$

and so, minimizing MSE corresponds to minimizing the first term in (8.29). This quantity must be nonnegative, and so, the solution is to take $\boldsymbol{b}^\star = \widehat{\boldsymbol{b}}$. The latter is the solution irrespective of $\boldsymbol{A}$. So the best prediction is that which estimates the random variable $\boldsymbol{b}$ by its conditional mean. We now examine properties of $\widehat{\boldsymbol{b}}$.

The usual frequentist optimality criteria for a fixed effect $\boldsymbol{\theta}$ concentrate upon unbiasedness and upon the variance of the estimator, $\mathrm{var}(\widehat{\boldsymbol{\theta}})$, see Sect. 2.2. When inference is required for a random effect $\boldsymbol{b}$, these criteria need adjustment. Specifically, an unbiased predictor $\widehat{\boldsymbol{b}}$ is such that

$$\mathrm{E}[\widehat{\boldsymbol{b}} - \boldsymbol{b}] = \boldsymbol{0},$$

to give

$$\mathrm{E}[\widehat{\boldsymbol{b}}] = \mathrm{E}[\boldsymbol{b}]$$

so that the expectation of the predictor is equal to the expectation of the random variable that it is predicting. For $\widehat{\boldsymbol{b}} = \mathrm{E}[\boldsymbol{b} \mid \boldsymbol{y}]$,

$$\mathrm{E}_{\boldsymbol{Y}}[\widehat{\boldsymbol{b}}] = \mathrm{E}_{\boldsymbol{Y}}[\mathrm{E}_{\boldsymbol{b}|\boldsymbol{y}}(\boldsymbol{b} \mid \boldsymbol{y})] = \mathrm{E}_{\boldsymbol{b}}[\boldsymbol{b}]$$

where the first step follows on substitution of $\widehat{\boldsymbol{b}}$ and the second from iterated expectation; therefore, we have an unbiased predictor. We emphasize that we do not have an unbiased estimator in the usual sense, and in general, $\widehat{\boldsymbol{b}}$ will display *shrinkage* toward zero, as we illustrate in later examples.

The variance of a random variable is defined with respect to a fixed number, the mean. In the context of prediction of a random variable, a more relevant summary of the variability is

$$\mathrm{var}(\widehat{\boldsymbol{b}} - \boldsymbol{b}) = \mathrm{var}(\widehat{\boldsymbol{b}}) + \mathrm{var}(\boldsymbol{b}) - 2 \times \mathrm{cov}(\widehat{\boldsymbol{b}}, \boldsymbol{b}).$$

If this quantity is small, then the predictor and the random variable are moving in a stochastically similar way. We have

$$\begin{aligned}\text{cov}_{\hat{\boldsymbol{b}},\boldsymbol{b}}(\widehat{\boldsymbol{b}},\boldsymbol{b}) &= \text{E}_{\boldsymbol{Y}}[\text{cov}(\widehat{\boldsymbol{b}},\boldsymbol{b} \mid \boldsymbol{y})] + \text{cov}_{\boldsymbol{Y}}(\text{E}[\widehat{\boldsymbol{b}} \mid \boldsymbol{y}],\text{E}[\boldsymbol{b} \mid \boldsymbol{y}]) \\ &= \text{E}_{\boldsymbol{Y}}[\text{cov}(\widehat{\boldsymbol{b}},\boldsymbol{b} \mid \boldsymbol{y})] + \text{cov}_{\boldsymbol{Y}}(\widehat{\boldsymbol{b}},\widehat{\boldsymbol{b}}) \\ &= \text{var}(\widehat{\boldsymbol{b}}), \end{aligned} \tag{8.30}$$

since the first term in (8.30) is the covariance between the constant $\text{E}[\widehat{\boldsymbol{b}} \mid \boldsymbol{y}]$ (since $\boldsymbol{y}$ is conditioned upon) and $\widehat{\boldsymbol{b}}$, and so is zero. To obtain the form of the second term in (8.30), we have used $\text{E}[\widehat{\boldsymbol{b}} \mid \boldsymbol{y}] = \text{E}[\text{E}[\boldsymbol{b} \mid \boldsymbol{y}] \mid \boldsymbol{y}] = \widehat{\boldsymbol{b}}$. Hence,

$$\text{var}(\widehat{\boldsymbol{b}} - \boldsymbol{b}) = \text{var}(\boldsymbol{b}) - \text{var}(\widehat{\boldsymbol{b}}) = \boldsymbol{D} - \text{var}(\widehat{\boldsymbol{b}}).$$

In order to determine the form of $\widehat{\boldsymbol{b}} = \text{E}[\boldsymbol{b} \mid \boldsymbol{y}]$ and evaluate $\text{var}(\widehat{\boldsymbol{b}} - \boldsymbol{b})$, we need to provide more information on the model that is to be used, so that the form of $p(\boldsymbol{b} \mid \boldsymbol{y})$ can be determined.

For the LMM,

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{Y}_i \end{bmatrix} \sim \text{N}_{q+1+n_i}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_i\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{D} & \boldsymbol{D}\boldsymbol{z}_i^{\text{T}} \\ \boldsymbol{z}_i\boldsymbol{D} & \boldsymbol{V}_i \end{bmatrix} \right)$$

since

$$\text{cov}(\boldsymbol{b}_i, \boldsymbol{Y}_i) = \text{cov}(\boldsymbol{b}_i, \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i) = \text{cov}(\boldsymbol{b}_i, \boldsymbol{z}_i\boldsymbol{b}_i) = \boldsymbol{D}\boldsymbol{z}_i^{\text{T}},$$

(Appendix B), and similarly, $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{b}_i) = \boldsymbol{z}_i\boldsymbol{D}$. The conditional distribution of a multivariate normal distribution is normal also (Appendix D) with mean

$$\widehat{\boldsymbol{b}}_i = \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}_i] = \boldsymbol{D}\boldsymbol{z}_i^{\text{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \tag{8.31}$$

which coincides with the Bayesian derivation earlier, (8.25). From a frequentist perspective, (8.25) is known as the best linear unbiased predictor (BLUP), where unbiased refers to it satisfying $\text{E}[\widehat{\boldsymbol{b}}_i] = \text{E}[\boldsymbol{b}_i]$.

The form (8.31) is not of practical use since it depends on the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; instead, we use

$$\widehat{\boldsymbol{b}}_i = \text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}] = \widehat{\boldsymbol{D}}\boldsymbol{z}_i^{\text{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{\text{G}}) \tag{8.32}$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{D}(\widehat{\boldsymbol{\alpha}})$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\alpha}})$. The implications of the substitution of $\widehat{\boldsymbol{\beta}}_{\text{G}}$ are not great, since it is an unbiased estimator and appears in (8.31) in a linear fashion, but the use of $\widehat{\boldsymbol{\alpha}}$ is more problematic. In particular the predictor $\widehat{\boldsymbol{b}}_i$ is no longer linear in the data, so that exact properties can no longer be derived.

The uncertainty in the prediction, accounting for the estimation of $\boldsymbol{\beta}$, is

$$\begin{aligned}\text{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i) &= \boldsymbol{D} - \text{var}(\widehat{\boldsymbol{b}}_i)\\ &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D} + \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}\end{aligned}$$

after tedious algebra (Exercise 8.5), so that (8.27) is recovered. We again emphasize that this estimate of variability of prediction does not acknowledge the uncertainty in $\widehat{\boldsymbol{\alpha}}$. Given correct specification of the marginal variance model, $\text{var}(\boldsymbol{Y} \mid \boldsymbol{\alpha}) = \boldsymbol{V}(\boldsymbol{\alpha})$, and a consistent estimator of $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{b}}_i$ is asymptotically normal with a known distribution, which can be used to form interval estimates. As an alternative to the use of (8.25), we can implement a fully Bayesian approach (Sect. 8.6), though no closed-form solution emerges.

As a final derivation, rather than assume normality, we could consider estimators that are *linear* in $\boldsymbol{y}$. Exercise 8.6 shows that this again leads to

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}).$$

The best linear predictor is therefore identical to the best predictor under normality. For general distributions, $\text{E}[\boldsymbol{b}_i \mid \boldsymbol{y}_i]$ will not necessarily be linear in $\boldsymbol{y}_i$.

Since we now have a method for predicting $\boldsymbol{b}_i$, we can examine fitted values:

$$\begin{aligned}\widehat{\boldsymbol{Y}}_i &= \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i\\ &= \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\left[\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\right]\\ &= (\mathbf{I}_{n_i} - \boldsymbol{W}_i)\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{W}_i\boldsymbol{y}_i,\end{aligned}$$

with $\boldsymbol{W}_i = \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}$, so that we have a weighted combination of the population profile and the unit's data. If $\boldsymbol{D} = \boldsymbol{0}$, we obtain $\widehat{\boldsymbol{Y}}_i = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$, and if $\boldsymbol{D}$ is "small," the fitted values are close to the population curve, which is reasonable if there is little between-unit variability. If elements of $\boldsymbol{D}$ are large, the fitted values are closer to the observed data.

Example: One-Way ANOVA

For the simple balanced ANOVA model previously considered, the calculation of $\text{E}[b_i \mid \boldsymbol{y}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}]$ results in

$$\widehat{b}_i = \frac{n\widehat{\sigma}_0^2}{\widehat{\sigma}_\epsilon^2 + n\widehat{\sigma}_0^2}(\overline{y}_i - \widehat{\beta}_0)$$

to give a predictor that is a weighted combination of the "residual" $\overline{y}_i - \widehat{\beta}_0$ and zero. For finite n, the predictor is biased towards zero. As $n \rightarrow \infty$, $\widehat{b}_i \rightarrow \overline{y}_i - \widehat{\beta}_0$, so that $\widehat{\beta}_0 + \widehat{b}_i \rightarrow \overline{y}_i$, illustrating that the shrinkage disappears as the number of observations on a unit n increases, as we would hope.

8.6 Bayesian Inference for Linear Mixed Models

8.6.1 A Three-Stage Hierarchical Model

We consider the LMM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent and distributed as $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \text{N}_{q+1}(\,\mathbf{0}, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{ind} \text{N}_{n_i}(\,\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$, $i = 1, \ldots, m$.

The second stage assumption for $\boldsymbol{b}_i$ can be motivated using the concept of exchangeability that we encountered in Sect. 3.9. If we believe a priori that $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ are exchangeable (and are considered within a hypothetical infinite sequence of such random variables), then it can be shown using representation theorems (Sect. 3.9) that the prior has the form

$$p(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m) = \int \prod_{i=1}^{m} p(\boldsymbol{b}_i \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi})\, d\boldsymbol{\phi},$$

so that the collection $[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$ are conditionally independent, given *hyperparameters* $\boldsymbol{\phi}$, with the hyperparameters having a distribution known as a *hyperprior*.

Hence, we have a two-stage (hierarchical) prior:

$$\begin{aligned} \boldsymbol{b}_i \mid \boldsymbol{\phi} &\sim_{iid} p(\cdot \mid \boldsymbol{\phi}), \quad i = 1, \ldots, m \\ \boldsymbol{\phi} &\sim_{iid} \pi(\cdot). \end{aligned}$$

Parametric choices for $p(\cdot \mid \boldsymbol{\phi})$ and $\pi(\cdot)$ are based on the application, though computational convenience may also be a consideration (as we discuss in Sect. 8.6.3). We initially consider the multivariate normal prior $\text{N}_{q+1}(\,\mathbf{0}, \boldsymbol{D})$ so that $\boldsymbol{\phi} = \boldsymbol{D}$. The practical importance of this representation is that under exchangeability the beliefs about each of the unit-specific parameters must be identical. For example, for the dental data, if we do not believe that the individual-specific deviations from the average intercepts and slopes for boys and girls are exchangeable, then we should consider separate prior specifications for each gender. In general, if collections of units cluster due to an observed covariate that we believe will influence $\boldsymbol{b}_i$, then our prior should reflect this. This framework contrasts with the sampling theory approach in which the random effects are assumed to be a random sample from a hypothetical infinite population.

The three-stage model is

Stage One: Likelihood:

$$p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma_\epsilon^2), \quad i = 1, \ldots, m.$$

Stage Two: Random effects prior:

$$p(\boldsymbol{b}_i \mid \boldsymbol{D}), \quad i = 1, \ldots, m.$$

Stage Three: Hyperprior:

$$p(\boldsymbol{\beta}, \boldsymbol{D}, \sigma_\epsilon^2).$$

8.6.2 Hyperpriors

It is common to assume independent priors:

$$\pi(\boldsymbol{\beta}, \boldsymbol{D}, \sigma_\epsilon^2) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{D})\pi(\sigma_\epsilon^2).$$

A multivariate normal distribution or $\boldsymbol{\beta}$ and an inverse gamma distribution for σ_ϵ^2 are often reasonable choices, since they are flexible enough to reflect a range of prior information. The data are typically informative on $\boldsymbol{\beta}$ and σ_ϵ^2 also. These choices also lead to conditional distributions that have convenient forms for Gibbs sampling (Sect. 3.8.4). The prior specification for $\boldsymbol{D}$ is less straightforward.

If $\boldsymbol{D}$ is a diagonal matrix with elements σ_k^2, $k = 0, 1, \ldots, q$, then an obvious choice is

$$\pi(\sigma_0^2, \ldots, \sigma_q^2) = \prod_{k=0}^{q} \text{IGa}(a_k, b_k),$$

where IGa(a_k, b_k) denotes the inverse gamma distribution with prespecified parameters a_k, b_k, $k = 0, \ldots, q$. These choices also lead to conjugate conditional distributions for Gibbs sampling. Other choices are certainly possible, however, for example, those contained in Gelman (2006). A prior for non-diagonal $\boldsymbol{D}$ is more troublesome; there are $(q+2)(q+1)/2$ elements, with the restriction that the matrix of elements is positive definite. The inverse Wishart distribution is the conjugate choice and is the only distribution for which any great practical experience has been gathered.

We digress to describe how the Wishart distribution can be motivated. Suppose $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_r \sim_{iid} \text{N}_p(\,\boldsymbol{0}, \boldsymbol{S})$, with $\boldsymbol{S}$ a non-singular variance–covariance matrix, and let

$$\boldsymbol{W} = \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\text{T}}. \tag{8.33}$$

Then $\boldsymbol{W}$ follows a Wishart distribution, denoted $\text{Wish}_p(r, \boldsymbol{S})$, with probability density function

$$p(\boldsymbol{w}) = c^{-1} \mid \boldsymbol{w} \mid^{(r-p-1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{w}\boldsymbol{S}^{-1})\right]$$

where

$$c = 2^{rp/2}\mathit{\Gamma}_p(r/2) \mid \boldsymbol{S} \mid^{r/2}, \tag{8.34}$$

with

$$\mathit{\Gamma}_p(r/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \mathit{\Gamma}[(r+1-j)/2]$$

the generalized gamma function. We require $r > p - 1$ for a proper density. The mean is

$$\text{E}[\boldsymbol{W}] = r\boldsymbol{S}.$$

Taking $p = 1$ yields

$$p(w) = \frac{(2S)^{-r/2}}{\mathit{\Gamma}(r/2)} w^{r/2-1} \exp(-w/2S),$$

for $w > 0$, revealing that the Wishart distribution is a multivariate version of the gamma distribution, parameterized as $\text{Ga}[r/2, 1/(2S)]$. Further, taking $S = 1$ gives a χ_r^2 random variable, which is clear from (8.33).

If $\boldsymbol{W} \sim \text{Wish}_p(r, \boldsymbol{S})$, the distribution of $\boldsymbol{D} = \boldsymbol{W}^{-1}$ is known as the inverse Wishart distribution, denoted $\text{InvWish}_p(r, \boldsymbol{S})$, with density

$$p(\boldsymbol{d}) = c^{-1} \mid \boldsymbol{d} \mid^{-(r+p+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{d}^{-1}\boldsymbol{S})\right],$$

where c is again given by (8.34). We denote this random variable by $\boldsymbol{D}$ in anticipation of subsequently specifying an inverse Wishart distribution as prior for the variance–covariance matrix of the random effects $\boldsymbol{D}$. The mean is

$$\text{E}[\boldsymbol{D}] = \frac{\boldsymbol{S}^{-1}}{r - p - 1}$$

and is defined for $r > p + 1$. If $p = 1$, we recover the inverse gamma distribution $\text{IGa}(r/2, 1/2S)$ with

$$\text{E}[D] = \frac{1}{S(r-2)}$$

$$\text{var}(D) = \frac{1}{S^2(r-2)(r-4)},$$

so that small r gives a more dispersed distribution (which is true for general p). One way of thinking about prior specification is to imagine that the prior data for the precision consists of observing r multivariate normal random variables with empirical variance–covariance matrices $\boldsymbol{R} = \boldsymbol{S}^{-1}$. See Appendix D for further properties of the Wishart and inverse Wishart distributions.

Returning from our digression, within the LMM, we specify $\boldsymbol{W} = \boldsymbol{D}^{-1} \sim \text{W}_{q+1}(r, \boldsymbol{R}^{-1})$ where we have taken $\boldsymbol{S} = \boldsymbol{R}^{-1}$ to aid in prior specification. We require choices for r and $\boldsymbol{R}$. Since

$$\text{E}[\boldsymbol{D}] = \frac{\boldsymbol{R}}{r - q - 2},$$

$\boldsymbol{R}$ may be scaled to be a prior estimate of $\boldsymbol{D}$, with r acting as a strength of belief in the prior, with large r placing more mass close to the mean.

One method of specification that attempts to minimize the influence of the prior is to take $r = q + 3$ the smallest integer that gives a proper prior to give $\text{E}[\boldsymbol{D}] = \boldsymbol{R}$, as the prior guess for $\boldsymbol{D}$. We now describe another way of specifying a Wishart prior, based on Wakefield (2009b). Marginalization over $\boldsymbol{D}$ gives $\boldsymbol{b}_i$ as multivariate Student's t with location $\boldsymbol{0}$, scale matrix $\boldsymbol{R}/(r - p + 1)$, and degrees of freedom $d = r - q + 2$. The margins of a multivariate Student's t are t also, which allows r and $\boldsymbol{R}$ to be chosen via specification of an interval for the jth element of $\boldsymbol{b}_i$, b_{ij}. Specifically, b_{ij} follows a univariate Student's t distribution with location 0, scale $R_{jj}/(r - q + 2)$, and degrees of freedom $d = r - q$. For a required range of $[-V, V]$ with probability 0.95, we use the relationship $\pm t^d_{0.025}\sqrt{D}_{jj} = \pm V$, where t^d_p is the $100 \times p$th quantile of a Student's t random variable with d degrees of freedom. Picking the smallest integer that results in a proper prior gives $r = q + 1$ so that $d = 1$ and $R_{jj} = V^2 d/2(t^d_{1-(1-p)/2})^2$.

As an example of this procedure, consider a single random effect ($q = 0$). We specify a Ga$[r/2, 1/(2S)]$ prior for σ_0^{-2}, so that marginally, b_i is a Student's t distribution with location 0, scale r/S, and degrees of freedom r. The above prescription gives $r = 1$ and $S = (t^d_{1-(1-p)/2})^2/V^2$. In the more conventional Ga(a, b) parameterization, we obtain $a = 0.5$ and $b = V^2/[2(t^d_{1-(1-p)/2})^2]$. For example, for the dental data, if we believe that a 95% range for the intercepts, about the population intercept, is $\pm V = \pm 0.2$, we obtain the choice Ga$(0.5, 0.000124)$ for σ_0^{-2}. This translates into a prior for σ_0 (which is more interpretable) with 5%, 50%, and 95% points of [0.008, 0.023, and 0.25]. An important point to emphasize is that within the LMM, a proper prior is required for $\boldsymbol{D}$ to ensure propriety of the posterior distribution.

A weakness with the Wishart distribution is that it is deficient in second moment parameters, since there is only a single degrees of freedom parameter r. So, for example, it is not possible to have differing levels of certainty in the tightness of the prior distribution for different elements of $\boldsymbol{D}$. This contrasts with the situation in which $\boldsymbol{D}$ is diagonal, and we specify independent inverse gamma priors, which gives separate precision parameters for each variance.

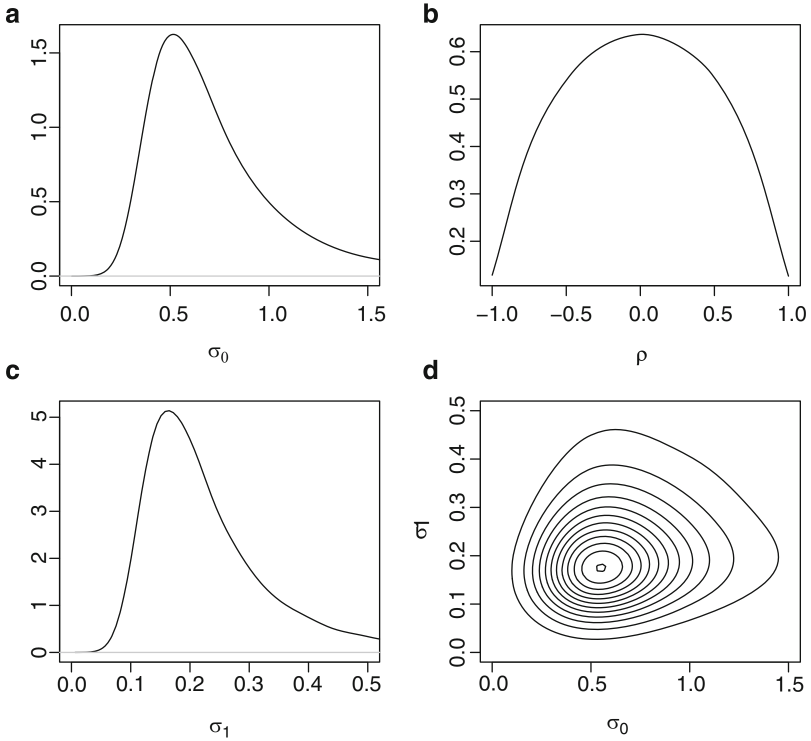


Fig. 8.2 Prior summaries for the prior $\boldsymbol{D}^{-1} \sim \mathrm{W}_2(r, \boldsymbol{R}^{-1})$ with $r = 4$ and $\boldsymbol{R}$ containing elements $[1.0, 0, 0, 0.1]$. Univariate marginal densities for (**a**) σ_0, (**b**) ρ, (**c**) σ_1, and the bivariate density for (**d**) (σ_0, σ_1)

Figure 8.2 displays summaries for an example with a 2 × 2 variance–covariance matrix (so that $q = 1$). We assume $\boldsymbol{D}^{-1} \sim W_2(r, \boldsymbol{R}^{-1})$ with $r = 4$ and $\mathrm{E}[\boldsymbol{D}] = \frac{\boldsymbol{R}}{4-1-2} = \boldsymbol{R}$ with $\boldsymbol{R} = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}$. We summarize samples from the Wishart via marginal distributions for σ_0, σ_1, and ρ since these are more interpretable. These plots were obtained by simulating samples for $\boldsymbol{D}^{-1}$ from the Wishart prior and then converting these samples to the required functions of interest. Finally, we smooth the sample histograms and scatter plots to produce Fig. 8.2. As we would expect, the prior on the correlation is symmetric about 0. Examination of intervals for σ_0, σ_1 can inform on whether we believe the prior is suitable for any given application. Going one step further, we could then simulate random effects from the zero mean normal with variance $\boldsymbol{D}$, the latter being a draw from the prior; we might also continue to simulate data, though this would require draws from the other priors too.

8.6.3 Implementation

For simplicity, we suppose that $\boldsymbol{x}_i = \boldsymbol{z}_i$. It is convenient in what follows to reparameterize in terms of the set $[\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W}]$ where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$, $\tau = \sigma_\epsilon^{-2}$, and $\boldsymbol{W} = \boldsymbol{D}^{-1}$. The joint posterior is

$$p(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W} \mid \boldsymbol{y}) \propto \prod_{i=1}^{m} \left[p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W}) \right] \pi(\boldsymbol{\beta}) \pi(\tau) \pi(\boldsymbol{W}),$$

with priors:

$$\boldsymbol{\beta} \sim \mathrm{N}_{q+1}(\boldsymbol{\beta}_0, \boldsymbol{V}_0), \quad \tau \sim \mathrm{Ga}(a_0, b_0), \quad \boldsymbol{W} \sim \mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1}).$$

Marginal distributions, and summaries of these distributions, are not available in closed form. Various approaches to obtaining quantities of interest are available. The INLA procedure described in Sect. 3.7.4 is ideally suited to the LMM. As an alternative, we describe an MCMC strategy using Gibbs sampling (Sect. 3.8.4). The required conditional distributions are

- $p(\boldsymbol{\beta} \mid \tau, \boldsymbol{W}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \boldsymbol{y})$.
- $p(\tau \mid \boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \boldsymbol{y})$.
- $p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \tau, \boldsymbol{W}, \boldsymbol{y})$, $i = 1, \dots, m$.
- $p(\boldsymbol{W} \mid \boldsymbol{\beta}, \tau, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \boldsymbol{y})$.

where we block update $\boldsymbol{\beta}$, $\boldsymbol{W}$, and $\boldsymbol{\beta}_i$ to reduce dependence in the Markov chain.

The conditional distributions for $\boldsymbol{\beta}$, τ, and $\boldsymbol{\beta}_i$ are straightforward to derive (Exercise 8.10) and are given, respectively, by

$$\begin{aligned}
\boldsymbol{\beta} \mid \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m, \boldsymbol{W} &\propto \prod_{i=1}^{m} p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W}) \pi(\boldsymbol{\beta}) \\
&\sim \mathrm{N}_{q+1} \Bigg[\left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1} \right)^{-1} \left(\boldsymbol{W} \sum_{i=1}^{m} \boldsymbol{\beta}_i + \boldsymbol{V}_0^{-1} \boldsymbol{\beta}_0 \right), \\
&\qquad \left(m\boldsymbol{W} + \boldsymbol{V}_0^{-1} \right)^{-1} \Bigg]
\end{aligned}$$

$$\begin{aligned}
\tau \mid \boldsymbol{\beta}_i, \boldsymbol{y} &\propto \prod_{i=1}^{m} p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) \pi(\tau) \\
&\sim \mathrm{Ga} \left[a_0 + \frac{\sum_{i=1}^{m} n_i}{2}, b_0 + \frac{1}{2} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i)^{\mathrm{T}} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_i) \right]
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\beta}_i \mid \tau, \boldsymbol{W}, \boldsymbol{y} &\propto \prod_{i=1}^{m} p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W}) \\
&\sim \mathrm{N}_{q+1} \left[(\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{y}_i + \boldsymbol{W}\boldsymbol{\beta}), (\tau \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{W})^{-1} \right].
\end{aligned}$$

Conditional independencies have been exploited, and in each case, the notation explicitly conditions on only those parameters on which the conditional distribution depends. For example, to derive the conditional distribution for $\boldsymbol{\beta}$, we only require $[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$ and $\boldsymbol{W}$. The conditional for $\boldsymbol{\beta}_i$ is, once we reparameterize, identical to the empirical Bayes estimates derived for the random effects in Sect. 8.5.4 (Exercise 8.11). This comparison illustrates how the uncertainty in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = [\tau, \boldsymbol{W}]$ is accounted for across iterates of the Gibbs sampler.

Deriving the conditional distribution for $\boldsymbol{W}$ is a little more involved. First, note that

$$(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta}_i - \boldsymbol{\beta}) = \mathrm{tr}[(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta}_i - \boldsymbol{\beta})] = \mathrm{tr}[\boldsymbol{W}(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}].$$

Then

$$\begin{aligned}
\boldsymbol{W} \mid \boldsymbol{\beta}_i, \boldsymbol{\beta} &\propto \prod_{i=1}^{m} p(\boldsymbol{\beta}_i \mid \boldsymbol{W}) \times \pi(\boldsymbol{W}) \\
&\propto \mid \boldsymbol{W} \mid^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2}\left[\sum_{i=1}^{m}(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta}_i - \boldsymbol{\beta}) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{R})\right]\right\} \\
&= \mid \boldsymbol{W} \mid^{(m+r-q-1-1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{W}\left[\sum_{i=1}^{m}(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{R}\right]\right)\right\}
\end{aligned}$$

to give the conditional distribution

$$\boldsymbol{W} \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\beta} \sim \mathrm{W}_{q+1}\left[r+m, \left(\boldsymbol{R} + \sum_{i=1}^{m}(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}\right)^{-1}\right].$$

This illustrates how r and $\boldsymbol{R}$ are comparable to m and the between-unit sum of squares, respectively, which aids in prior specification. Since

$$\mathrm{E}[\boldsymbol{D} \mid \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \boldsymbol{\beta}] = \frac{\boldsymbol{R} + \sum_{i=1}^{m}(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})^{\mathrm{T}}}{r+m-q-2}$$

the form of the conditional distribution suggests that it is better to err on the side of picking $\boldsymbol{R}$ too small, since a large $\boldsymbol{R}$ will always dominate the sum of squares. If m is small, the prior is always influential.

If we collapse over $\boldsymbol{\beta}_i$, $i = 1, \ldots, m$, we obtain the two-stage model with

Stage One: Marginal likelihood:

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \tau, \boldsymbol{W} \sim \mathrm{N}_N(\boldsymbol{x}\boldsymbol{\beta}, \boldsymbol{V}),$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{W}, \tau)$.

Stage Two: Priors:

$$\pi(\boldsymbol{\beta})\pi(\boldsymbol{W})\pi(\tau).$$

An MCMC algorithm iterates between

- $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{W}, \tau)$
- $p(\tau \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{W})$
- $p(\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta})$

This approach is appealing since it is over a reduced parameter space, but the form of $p(\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta}, \tau)$ is extremely awkward. The conditional for $\boldsymbol{\beta}$ offers some intuition on the Bayesian approach, however. Specifically, writing $\boldsymbol{\alpha} = [\tau, \boldsymbol{W}]$, we obtain the conditional distribution:

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha} \sim \mathrm{N}_{q+1}\left[\mathrm{E}(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha}), \mathrm{var}(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha})\right]$$

where the mean and variance can be written in the weighted forms

$$\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha}] = \boldsymbol{w} \times \widehat{\boldsymbol{\beta}}_{\mathrm{G}} + (\mathbf{I} - \boldsymbol{w}) \times \boldsymbol{\beta}_0 \tag{8.35}$$

$$\mathrm{var}(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha}) = \boldsymbol{w} \times \mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{G}}). \tag{8.36}$$

Here, $\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y}$ is the GLS estimator with variance $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{G}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}$, and the $(q+1) \times (q+1)$ weight matrix is

$$\boldsymbol{w} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x} + \boldsymbol{V}_0^{-1})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}.$$

As the prior becomes more diffuse, that is, as $\boldsymbol{V}_0^{-1} \to \mathbf{0}$, the weight $\boldsymbol{w} \to \mathbf{I}$, the conditional posterior mean approaches the GLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$, and the conditional posterior variance approaches $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{G}})$. In contrast, as $\boldsymbol{V}^{-1} \to \mathbf{0}$, so that the prior becomes more concentrated about $\boldsymbol{\beta}_0$, $\boldsymbol{w} \to \mathbf{0}$ and the conditional posterior moments approach the prior distribution. Since

$$\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{y}] = \mathrm{E}_{\boldsymbol{\alpha}|\boldsymbol{y}}\left[\mathrm{E}(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\alpha})\right],$$

the posterior mean is the conditional posterior mean averaged over $\boldsymbol{\alpha} \mid \boldsymbol{y}$. As is typical, the Bayesian estimate integrates over $\boldsymbol{\alpha}$, while the GLS estimator conditions on $\widehat{\boldsymbol{\alpha}}$ for evaluation of $\boldsymbol{V}$. We would expect likelihood and Bayesian point and interval estimates to be similar for large samples because the posterior $\boldsymbol{\alpha} \mid \boldsymbol{y}$ will become increasingly concentrated about $\widehat{\boldsymbol{\alpha}}$.

8.6.4 Extensions

Computationally, under a Bayesian approach via MCMC, it is relatively straightforward to extend the basic LMM. The conditional distributions may not be of conjugate form, but Metropolis–Hastings steps can be substituted (Sect. 3.8.2). For example, great flexibility in the distributional assumptions and error models is available, though prior specification will usually require greater care. To automatically protect against outlying measurements/individuals, Student's t errors may be specified at stage one/stage two of the hierarchy, though when regression is the focus of the analysis, the greatest effort should be concentrated upon specifying appropriate mean–variance relationships at the two stages.

With the advent of MCMC, there is a temptation to fit complex models that attempt to reflect every possible nuance of the data. However, the statistical properties of complex models (such as consistency of estimation under incorrect model specification) are difficult to determine, as are the implied marginal distributions for the data (which can aid in model assessment). Overfitting is also always a hazard. Consequently, caution should be exercised in model refinement. One of the arts of statistical analysis is deciding on when model refinement is warranted.

Example: Dental Growth Curves

We analyze the data from the $m = 11$ girls only and adopt the following three-stage hierarchical model:

Stage One: As likelihood, we assume

$$y_{ij} = \beta_{i0} + \beta_{i1} t_j + \epsilon_{ij},$$

with $\epsilon_{ij} \mid \tau \sim_{iid} \mathrm{N}(0, \tau^{-1}), j = 1, \ldots, 4, i = 1, \ldots, 11.$
Stage Two: Let

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix},$$

with random effects prior

$$\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} \sim \mathrm{N}_2(\boldsymbol{\beta}, \boldsymbol{D}), \quad i = 1, \ldots, m.$$

Stage Three: As hyperprior, we assume

$$\pi(\tau, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) = \pi(\tau) \times \pi(\boldsymbol{\beta}) \times \pi(\boldsymbol{D}^{-1})$$

with improper priors on τ and $\boldsymbol{\beta}$:

$$\pi(\tau) \propto \tau^{-1}, \quad \pi(\boldsymbol{\beta}) \propto 1$$

and

$$\boldsymbol{D}^{-1} \sim W_2(r, \boldsymbol{R}^{-1}).$$

In the LMM, there is typically abundant information in the data with respect to τ and $\boldsymbol{\beta}$. By placing a flat prior on $\boldsymbol{\beta}$ (which are often the parameters of interest), we are also basing inference on the data alone (in nonlinear models, more care is required since a proper prior is often required to ensure propriety of the posterior).

With just 11 girls, we would expect inference for $\boldsymbol{D}$ to be sensitive to the prior, and so, we consider three choices of r and $\boldsymbol{R}$. Each prior has the same mean of

$$\mathrm{E}[\boldsymbol{D}] = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix} = \frac{\boldsymbol{R}}{r - q - 2} \tag{8.37}$$

with $q = 1$ here. The above specification corresponds to an a priori belief that the spread of the expected response at 11 years across girls is

$$\pm 1.96 \mathrm{E}[\sigma_0] \approx \pm 1.96 \sqrt{R_{11}} = \pm 1.96$$

and the variability in slopes across girls is expected to be

$$\pm 1.96 \mathrm{E}[\sigma_1] \approx \pm 1.96 \sqrt{R_{22}} = \pm 0.62.$$

The exact intervals can be evaluated in an obvious fashion using simulation. The off-diagonal of $\boldsymbol{R}$ is 0 as we assume there is no reason to believe the correlation between intercepts and slopes will be positive or negative.

The degrees of freedom r is on the same scale as m and may be viewed as a prior sample size. We pick $r = 4, 7, 28$, and to obtain the same prior mean, (8.37), $\boldsymbol{R}$ is specified as

$$\begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad \begin{bmatrix} 4.0 & 0 \\ 0 & 0.4 \end{bmatrix}, \quad \begin{bmatrix} 25 & 0 \\ 0 & 2.5 \end{bmatrix},$$

for each of $r = 4, 7, 28$, respectively. To obtain a proper posterior, we require $r > 1$. We pick $r = 4$ as our smallest choice since the mean exists for this value. Samples from this prior are displayed in Fig. 8.2.

We present the results in terms of elements of $\boldsymbol{D}$, for direct comparison with the prior. If we were reporting substantive conclusions, we would choose σ_0, σ_1, ρ, or interval estimates for $\boldsymbol{\beta}_{i^\star} = [\beta_{i^\star 0}, \beta_{i^\star 1}]$, the parameters of a new girl who is exchangeable with those in the study. Table 8.1 gives posterior medians and 95% interval estimates for the fixed effects and variance components. We see sensitivity to the prior with respect to inference for $\boldsymbol{D}$. As r increases, the posterior medians draw closer to the prior means of 1.0 and 0.1. For β_0 and β_1, the medians are robust to the prior specification, while the width of the intervals for β_0 and β_1 change in proportion to the behavior of σ_0^2 and σ_1^2, respectively. The interval estimates for β_0 narrow, while those for β_1 widen, though the changes are modest. With only 11 subjects, we would expect sensitivity to the prior on $\boldsymbol{D}$. For $r = 7$, the "total degrees of freedom" is 18 with a prior contribution of 7 and a data contribution of 11.

Table 8.1 Posterior medians and 95% intervals for fixed effects and variance components, under three priors for the dental growth data for girls

	Prior	$r = 4$		$r = 7$		$r = 28$	
β_0	–	22.6	[21.4,23.8]	22.6	[21.5,23.7]	22.6	[21.8,23.5]
β_1	–	0.48	[0.33,0.63]	0.48	[0.31,0.65]	0.48	[0.28,0.67]
σ_0^2	1.0	3.48	[1.66,8.75]	2.97	[1.51,6.63]	1.78	[1.14,2.97]
σ_{01}	0.0	0.13	[$-$0.10,0.54]	0.10	[$-$0.14,0.46]	0.04	[$-$0.10,0.20]
σ_1^2	0.1	0.03	[0.01,0.10]	0.05	[0.02,0.12]	0.08	[0.05,0.14]

The population intercept is β_0 and the population slope is β_1. The variances of the random intercepts and random slopes are σ_0^2 and σ_1^2, respectively, and the covariance between the two is σ_{01}

8.7 Generalized Estimating Equations

8.7.1 Motivation

We now describe the GEE approach to modeling/inference. GEE attempts to make minimal assumptions about the data-generating process and is constructed to answer population-level, rather than individual-level, questions. There are some links with the quasi-likelihood approach described in Sect. 2.5 in that, rather than specify a full probability model for the data, only the first two moments are specified. GEE is motivated by dependent data situations, however, and exploits replication across units to empirically estimate standard errors through sandwich estimation. GEE uses a “working” second moment assumption; “working” refers to the choice of a variance model that may not necessarily correspond to exactly the form we believe to be true but rather to be a choice that is statistically convenient (we elaborate on this point subsequently). Any discrepancies from the truth are corrected using sandwich estimation to give a procedure that gives a consistent estimator of both the regression parameters and the standard errors (so long as we have independence between individuals).

We assume the marginal mean model

$$\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta},$$

and consider the $n_i \times n_i$ *working* variance–covariance matrix:

$$\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i \tag{8.38}$$

with $\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = \boldsymbol{0}$ for $i \neq i'$, so that observations on different individuals are assumed uncorrelated. To motivate GEE, we begin by assuming that $\boldsymbol{W}_i$ is known and does not depend on unknown parameters. In this case the GLS estimator minimizes

$$\sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

and is given by the solution to the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}) = \mathbf{0},$$

which is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{Y}_i.$$

We have $\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, and if the information about $\boldsymbol{\beta}$ grows with increasing m, then $\widehat{\boldsymbol{\beta}}$ is consistent. The vital observation is that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator for *any* fixed $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{W}_1, \dots, \boldsymbol{W}_m)$. The weighting of observations by the latter dictates the efficiency of the estimator but not its consistency. The variance, $\mathrm{var}(\widehat{\boldsymbol{\beta}})$, is

$$\left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1}. \tag{8.39}$$

If the assumed variance–covariance matrix is substituted, that is, $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i$, then we obtain the model-based variance

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1}.$$

A Gauss–Markov theorem shows that, in this case, the estimator is efficient amongst linear estimators *if* the variance model (8.38) is correct (Exercise 8.6). The novelty of GEE is that rather than depend on a correctly specified variance model, sandwich estimation, via (8.39), is used to repair any deficiency in the working variance model.

8.7.2 The GEE Algorithm

We now suppose that $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ are unknown parameters in the variance–covariance model. A common approach is to assume

$$\boldsymbol{W}_i(\boldsymbol{\alpha}) = \alpha_1 \boldsymbol{R}_i(\boldsymbol{\alpha}_2),$$

where $\alpha_1 = \mathrm{var}(Y_{ij})$ is the variance of the response, for all i and j, and $\boldsymbol{R}_i(\boldsymbol{\alpha}_2)$ is a working correlation matrix that depends on parameters $\boldsymbol{\alpha}_2$. There are a number of choices for $\boldsymbol{R}_i$, including independence, exchangeable and AR(1) models (as described in Sect. 8.4.2). For known $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\beta}}$ is the root of the estimating equation

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \mathbf{0}. \tag{8.40}$$

When $\boldsymbol{\alpha}$ is unknown, we require an estimator $\widehat{\boldsymbol{\alpha}}$ that converges to "something" so that, informally speaking, we have a stable weighting matrix $\boldsymbol{W}(\widehat{\boldsymbol{\alpha}})$ in the estimating equation.

The sandwich variance estimator is

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \mathrm{var}(\boldsymbol{Y}_i) \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \tag{8.41}$$

where $\widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i(\widehat{\boldsymbol{\alpha}})$ and $\mathrm{var}(\boldsymbol{Y}_i)$ is estimated by the variance–covariance matrix of the residuals:

$$(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^{\mathrm{T}}. \tag{8.42}$$

This produces a consistent estimate of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$, so long as we have independence between units, that is, $\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = 0$ for $i \neq i'$. It is the replication across units that produces consistency, and so, the approach cannot succeed if we have no replication. Exercise 8.12 shows that we cannot estimate $\mathrm{var}(\boldsymbol{Y})$ using the analog of (8.42) when there is dependence between units.

For inference, we may use the asymptotic distribution

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}_{k+1}(\,\mathbf{0}, \mathbf{I}),$$

where we emphasize that the asymptotics are in the number of units, m. The variance estimator is sometimes referred to as *robust*, but *empirical* is a more appropriate description since the form can be highly unstable for small m.

In the most general case of working variance model specification, we may allow the working variance model to depend on $\boldsymbol{\beta}$ also, so that we have $\boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to allow mean–variance relationships. For example, in a longitudinal setting, the variance may depend on the square of the marginal mean μ_{ij} with an autoregressive covariance model:

$$\begin{aligned} \mathrm{var}(Y_{ij}) &= \alpha_1 \mu_{ij}^2 \\ \mathrm{cov}(Y_{ij}, Y_{ik}) &= \alpha_1 \alpha_2^{|t_{ij} - t_{ik}|} \mu_{ij} \mu_{ik} \\ \mathrm{cov}(Y_{ij}, Y_{i'k}) &= 0, \quad i \neq i' \end{aligned}$$

with $j = 1, \dots, n_i$, $k, k' = 1, \dots, n_{i'}$ and where t_{ij} is the time associated with response Y_{ij}. In this model, α_1 is the component of the variance that does not depend on the mean (and is assumed constant across time and across individuals), α_2 is the correlation between responses on the same individual which are one unit of time apart and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$. In general the roots of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha},\boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \boldsymbol{0} \tag{8.43}$$

are not available in closed form when $\boldsymbol{\beta}$ appears in $\boldsymbol{W}$.

We can write the $(k+1)\times 1$ estimating function in a variety of forms, for example:

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})$$

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \boldsymbol{x}_{ij} W_i^{jk}(Y_{ik} - \boldsymbol{x}_{ik}\boldsymbol{\beta})$$

where W_i^{ij} denotes entry (i,j) of $\boldsymbol{W}_i^{-1}$. We will often use the middle form, since this emphasizes that the basic unit of replication (upon which the asymptotic properties depend) is indexed by i.

The GEE approach is constructed to carry out marginal inference, and so we cannot perform individual-level inference. For a linear model, marginalizing a LMM produces a marginal model identical to that used in a GEE approach. As a consequence, parameter interpretation, as discussed in Sect. 8.4.3 in the marginal setting, is identical in the LMM and in GEE. When nonlinear models are considered in Chap. 9 there is no equivalence and the differences between the conditional and marginal approaches to inference becomes more pronounced. For the linear model, sandwich estimation may be applied to the MLE of $\boldsymbol{\beta}$.

So far, as the choice of "working" correlation structure is concerned, we encounter the classic efficiency/robustness trade-off. If we choose a simple structure, there are few elements in $\boldsymbol{\alpha}$ to estimate, but there is a potential loss of efficiency. A more complex model may provide greater efficiency if the variance model is closer to the true data-generating mechanism but more instability in estimation of $\boldsymbol{\alpha}$. Clearly, this choice should be based on the sample size, with relatively sparse data encouraging the use of a simple model.

We summarize the GEE approach to modeling/estimation when the working variance model depends on $\boldsymbol{\alpha}$ and not on $\boldsymbol{\beta}$. The steps of the approach are:

1. Specification of a mean model, $\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$.
2. Specification of a working variance model, $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$.
3. From (1) and (2), an estimating function is constructed, and sandwich estimation is applied to the variance of the resultant estimator.

In general, iteration is needed to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate, set $t = 0$, and iterate between:

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(t)}) = \boldsymbol{0}$, with $\boldsymbol{G}$ given by (8.40), to give $\widehat{\boldsymbol{\beta}}^{(t+1)}$.
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(t+1)}$ based on $\widehat{\boldsymbol{\beta}}^{(t+1)}$.

Set $t \to t+1$, and return to 1.

Example: Linear Regression

We illustrate the use of a working variance assumption in an independent data situation. Suppose

$$\mathrm{E}[Y_i] = \boldsymbol{x}_i\boldsymbol{\beta},$$

for $i = 1, \ldots, n$. Under the working *independence* variance model, var$(\boldsymbol{Y}) = \alpha\mathbf{I}$, the OLS estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$$

is recovered. The sandwich form of variance estimate is

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y})\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}. \tag{8.44}$$

Assuming the working variance is "true" gives the model-based estimate

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\alpha,$$

and α may be estimated by

$$\widehat{\alpha} = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2,$$

which is formerly equivalent to quasi-likelihood. If we replace var$(\boldsymbol{Y})$ in (8.44) by a diagonal matrix with diagonal elements $(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2$, then we obtain a variance estimator that protects (asymptotically) against errors with nonconstant variance. We cannot protect against correlated outcomes, however, since there is no replication.

8.7.3 Estimation of Variance Parameters

To formalize the estimation of $\boldsymbol{\alpha}$, we may introduce a second estimating equation. In the context of data with $\mu_{ij} = \mathrm{E}[Y_{ij}]$ and var$(Y_{ij}) \propto v(\mu_{ij})$, we define residuals $R_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}$. Recall that $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of parameters, and suppose $\boldsymbol{\alpha}$ is an $r \times 1$ vector of variance parameters. We then consider the pair of estimating equations:

$$\boldsymbol{G}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \tag{8.45}$$

$$\boldsymbol{G}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{E}_i^{\mathrm{T}}\boldsymbol{H}_i^{-1}\left[\boldsymbol{T}_i - \boldsymbol{\Sigma}_i(\boldsymbol{\alpha})\right] \tag{8.46}$$

where the "data" in the second estimating equation are

$$\boldsymbol{T}_i^{\mathrm{T}} = [R_{i1}R_{i2}, \ldots, R_{in_i-1}R_{in_i}, R_{i1}^2, \ldots, R_{in_i}^2],$$

an $[n_i + n_i(n_i - 1)/2]$-dimensional vector with

$$\boldsymbol{\Sigma}_i(\boldsymbol{\alpha}) = \mathrm{E}[\boldsymbol{T}_i]$$

a model for the variances of, and correlations between, the residuals. In (8.46), $\boldsymbol{E}_i = \partial \Sigma_i / \partial \alpha$ is the $[n_i + n_i(n_i - 1)/2] \times r$ vector of derivatives, and $\boldsymbol{H}_i = \mathrm{cov}(\boldsymbol{T}_i)$ is the $[n_i + n_i(n_i - 1)/2] \times [n_i + n_i(n_i - 1)/2]$ working covariance model for the squared and cross residual terms. If $\boldsymbol{G}_2$ is correctly specified, then there will be efficiency gains. A further advantage of this approach is that it is straightforward to incorporate a regression model for the variance–covariance parameters, that is, $\boldsymbol{\alpha} = g(\boldsymbol{x})$, for some link function $g(\cdot)$. For general $\boldsymbol{H}$, we will require the estimation of fourth order statistics, that is, var($\boldsymbol{T}$), which is a highly unstable endeavor unless we have an abundance of data. For this reason, working independence, $\boldsymbol{H}_i = \mathbf{I}$, is often used.

If $\mathrm{E}[\boldsymbol{T}] \neq \boldsymbol{\Sigma}$, then we will not achieve consistent estimation of the true variance model but, crucially, consistency of $\boldsymbol{\beta}$ through $\boldsymbol{G}_1$ is guaranteed, so long as $\widehat{\boldsymbol{\alpha}}$ converges to "something." We reiterate that a consistent estimate of var($\widehat{\boldsymbol{\beta}}$) is guaranteed through the use of sandwich estimation, so long as units are independent.

As an illustration of the approach, assume for simplicity $n_i = n = 3$ so that

$$\boldsymbol{T}_i^{\mathrm{T}} = [R_{i1}R_{i2}, R_{i1}R_{i3}, R_{i2}R_{i3}, R_{i1}^2, R_{i2}^2, R_{i3}^2].$$

With an exchangeable variance model:

$$\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})^{\mathrm{T}} = \mathrm{E}[\boldsymbol{T}_i^{\mathrm{T}}] = [\alpha_1\alpha_2, \alpha_1\alpha_2, \alpha_1\alpha_2, \alpha_1, \alpha_1, \alpha_1]$$

so that α_1 is the marginal variance, and α_2 is the correlation between observations on the same unit. With $\boldsymbol{H}_i = \mathbf{I}$, that is, a working independence model for the variance parameters, the estimating function for $\boldsymbol{\alpha}$ is

$$\boldsymbol{G}_2(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \begin{bmatrix} \alpha_2 & \alpha_2 & \alpha_2 & 1 & 1 & 1 \\ \alpha_1 & \alpha_1 & \alpha_1 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} R_{i1}R_{i2} \\ R_{i1}R_{i3} \\ R_{i2}R_{i3} \\ R_{i1}^2 \\ R_{i2}^2 \\ R_{i3}^2 \end{bmatrix} - \begin{bmatrix} \alpha_1\alpha_2 \\ \alpha_1\alpha_2 \\ \alpha_1\alpha_2 \\ \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{bmatrix} \right).$$

We therefore need to simultaneously solve the two equations:

$$\sum_{i=1}^{m} \widehat{\alpha}_2 \left[\sum_{j<k} R_{ij} R_{ik} - \widehat{\alpha}_1 \widehat{\alpha}_2 \right] + \sum_{j=1}^{3} (R_{ij}^2 - \widehat{\alpha}_1) = 0$$

$$\sum_{i=1}^{m} \widehat{\alpha}_1 \left[\sum_{j<k} R_{ij} R_{ik} - \widehat{\alpha}_1 \widehat{\alpha}_2 \right] = 0.$$

Dividing the second of these by $\widehat{\alpha}_1$ shows that

$$\widehat{\alpha}_1 \widehat{\alpha}_2 = \frac{1}{3m} \sum_{i=1}^{m} \sum_{j<k} R_{ij} R_{ik}$$

and substituting this into the first equation gives

$$\widehat{\alpha}_1 = \frac{1}{3m} \sum_{i=1}^{m} \sum_{j<k} R_{ij}^2,$$

to yield a pair of method of moments estimators.

Example: Dental Growth Curve

We use a GEE approach with the marginal model:

$$\mathrm{E}[Y_{ij}] = \boldsymbol{x}_{ij} \boldsymbol{\beta},$$

and interactions so that

$$\boldsymbol{x}_{ij} = \begin{cases} [1, t_j, 0, 0\,] & \text{for } i = 1, \ldots, 16 \\ [1, t_j, 1, t_j] & \text{for } i = 17, \ldots, 27, \end{cases}$$

where $j = 1, 2, 3, 4$ and $[t_1, t_2, t_3, t_4] = [-2, -1, 1, 2]$. Table 8.2 summarizes analyses with independence and exchangeable working correlation models, including standard errors under the assumption that the working model is correct (the "model" standard errors) and under sandwich estimation.

The point estimates and model-based standard errors under working independence always correspond to those from an OLS fit. The point estimates under the two working models are also identical here due to the balanced design. This agreement will not hold in general. The marginal variance is estimated as 2.26, and the correlation parameter under the exchangeable model as 0.61. These are in very close agreement with the equivalent values of 2.28 and 0.63 obtained from the random intercepts LMM. As we would expect for these data, the model-based and sandwich

Table 8.2 Summaries for the dental growth data of fixed effects from GEE analyses, under independence and exchangeable working correlation matrices; β_0 and β_1 are the population intercept and population slope for boys and $\beta_0 + \beta_2$ and $\beta_1 + \beta_3$ are the population intercept and population slope for girls

	Independence			Exchangeable		
		Standard error			Standard error	
	Estimate	Model	Sandwich	Estimate	Model	Sandwich
β_0	25.0	0.28	0.44	25.0	0.47	0.44
β_1	0.78	0.13	0.098	0.78	0.079	0.098
β_2	−2.32	0.44	0.75	−2.32	0.74	0.75
β_3	−0.31	0.20	0.12	−0.31	0.12	0.12

standard errors are quite similar under the exchangeable working model, because we have seen that the empirical estimates of the second moments are close to those of an exchangeable correlation structure. In contrast, the working independence standard errors change quite considerably. The sandwich standard errors are larger for the time static intercepts and smaller for the parameters associated with time (the two slopes).

Likelihood inference for a LMM with random intercepts and slopes produced identical point estimates to those in Table 8.2 and standard errors of [0.49, 0.086, 0.76, 0.14], which are in reasonable agreement with the sandwich standard errors reported in the table.

Example: Dental Data, Reduced Dataset

In the dental example the balanced design and relative abundance of data leads to summaries that might suggest that the alternative methods we have described are always in complete agreement. To correct this illusion, we now report summaries from an artificially created dental growth curve data set in which it is assumed that children randomly drop out of the study at some point after the first measurement. This yielded the data in Fig. 8.3 with 39 measurements on boys (previously there were 64) and 25 on girls (previously there were 44).

We analyze these data using GEE and LMMs, the latter via likelihood and Bayesian approaches to inference. For GEE, we implement independence and exchangeable working correlation structures. Table 8.3 gives point estimates along with uncertainty measures. For GEE, we report sandwich standard errors, for the likelihood LMM model-based standard errors and for the Bayes LMM posterior (model-based) standard deviations. The posterior distributions for the regression parameters were close to normal, with interval estimates based on a normal approximation virtually identical to those based directly on samples from the posterior. For the Bayesian analysis, we used a flat prior on $\boldsymbol{\beta}$, and the Wishart prior for $\boldsymbol{D}^{-1}$ had prior mean (8.37), with $r = 4$.

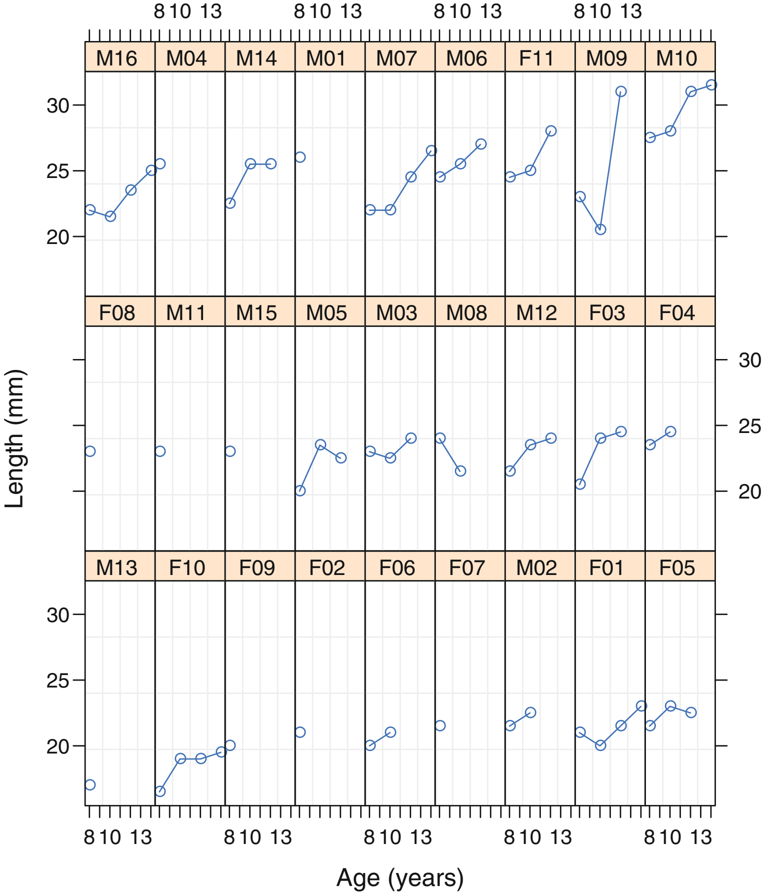


Fig. 8.3 Distance versus age for reduced dental data

Table 8.3 Summaries for the reduced dental growth data of fixed effects from GEE under independent and exchangeable working correlation matrices and likelihood and Bayesian LMMs; β_0 and β_1 are the population intercept and population slope for boys and $\beta_0 + \beta_2$ and $\beta_1 + \beta_3$ are the population intercept and population slope for girls

	GEE independence		GEE exchangeable		LMM likelihood		LMM Bayesian	
	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.d.
β_0	24.9	0.75	24.8	0.63	24.7	0.65	24.8	0.63
β_1	0.77	0.20	0.71	0.11	0.70	0.14	0.70	0.16
β_2	−2.70	1.23	−2.01	0.97	−1.92	1.04	−1.98	1.02
β_3	−0.53	0.27	−0.21	0.15	−0.17	0.23	−0.19	0.26

For these data, none of the analyses are completely satisfactory since the small number of observations does not give confidence in the sandwich standard errors, nor are the data sufficiently abundant to allow any reliable evaluation of assumptions for the LMM analyses. The exchangeable standard errors appear too small for the slope parameters, though the point estimates are in reasonable agreement with their LMM counterparts. The GEE independence standard errors are more in line with the LMM analyses, though the point estimates are quite different for β_2 and β_3. As expected under these priors, the likelihood and Bayes analyses are in reasonable agreement.

8.8 Assessment of Assumptions

8.8.1 Review of Assumptions

Each of the approaches to modeling that we have described depend, to a varying degree, upon assumptions. To ensure that inference is accurate, we need to check that these assumptions are at least approximately valid. We begin by reviewing the assumptions, starting with GEE (since it depends on the fewest assumptions).

For GEE, we have the marginal mean model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{e}_i,$$

and working covariance var$(\boldsymbol{e}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$, $i = 1, \ldots, m$. The first consideration is whether the marginal model $\text{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate. In particular, one must check whether the model requires refinement by, for example, the addition of quadratic terms or interactions. We may also examine whether additional variables, such as confounders, are required in the model. These considerations are common to all approaches. If the mean model is inadequate, but all other assumptions are satisfied, then we will still have a consistent estimator of the assumed form, but the relevance of inference is open to question. For example, suppose the true relationship is quadratic, but we incorrectly assume a linear model. The linear association will still be consistently estimated but may be very misleading. Deciding on a course of action if the mean model is inadequate depends on the nature of the analysis. If we are in exploratory mode, then fitting different models is not problematic. But if we are in confirmatory mode, then we would want to minimize changes to the model, though knowing of inadequacies is important for interpretation.

The use of a sandwich estimate for the standard errors is reliable in the sense of giving consistent estimates regardless of whether the working covariance model mimics the truth, but a working model that is far from the truth will lead to a loss of efficiency (so that the standard errors are bigger than they need to be), which suggests one should examine whether the assumed working model is far from that suggested by the data. In addition, if the number of units m is not large, then the estimate of the sandwich standard errors could be very unstable, and asymptotic inference may be inappropriate. As usual, there is no easy recipe for deciding whether m is "sufficiently large", since this depends on the design across individuals in the sample. The decision may be based on simulation, though experience with similar datasets is beneficial.

For the LMM, the usual model is

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_{q+1}(\,\boldsymbol{0}, \boldsymbol{D}\,)$, $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{ind} \mathrm{N}_{n_i}(\,\boldsymbol{0}, \sigma_\epsilon^2\mathbf{I}\,)$, and $\boldsymbol{b}_i, \boldsymbol{\epsilon}_i$ independent, $i = 1, \ldots, m$. This leads to the marginal model $\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha} \sim \mathrm{N}_{n_i}(\,\boldsymbol{x}_i\boldsymbol{\beta}, \boldsymbol{V}_i\,)$ and estimator

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y} \tag{8.47}$$

with

$$\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\right)^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}_{k+1}(\,\boldsymbol{0}, \mathbf{I}\,).$$

Therefore, if m is large, we do not require the data or the random effects to be normally distributed since the estimator is linear in the data, and so we can appeal to a central limit theorem. For an accurate standard error, we require the model-based form of the variance to be close to the truth, however. It is particularly important that there are no unmodeled mean–variance relationships. Another key requirement is that the random effects arise from a common distribution. Often, unit-specific covariates will be available, and these may define subpopulations that have different distributions (e.g., differing variance–covariance matrices $\boldsymbol{D}$) in covariate-defined subpopulations. If m is small we require, in addition, the data to be "close to normal" for valid inference. Sandwich estimation can be easily applied to obtain an empirical standard error, keeping in mind the caveats expressed above with regard to the need for sufficiently large m.

For prediction of the random effects, we have seen that the BLUP estimator is optimal under a number of different criteria. Normality of the random effects or the errors is not required, though an appropriate variance model is again important.

A Bayesian analysis of the LMM adds hyperpriors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ to the two-stage likelihood model. Each of the modeling assumptions required for likelihood-based inference are needed for a Bayesian analysis. However, asymptotic inference is not needed if, for example, MCMC is used. Accurate inference requires checking of the first and second stage assumptions because inference relies on the model being correct (or in practice, close to correct). Also, thought is required when priors are specified because inference may well be sensitive to the choices made. In particular, care is called for in the specification for $\boldsymbol{D}$. We emphasize that normality of the data and the random effects is not needed for a valid analysis if the sample size is large. For example, for inference with respect to $\boldsymbol{\beta}$, the posterior for $\boldsymbol{\beta}$ will be accurate so long as the asymptotic distribution of the estimator, (8.47), is faithful. Essentially, the asymptotic distribution replaces the likelihood contribution to the posterior.

8.8.2 Approaches to Assessment

For those individuals with sufficient data, individual-specific models may be fitted to allow examination of the appropriateness of initially hypothesized models in terms of the linear component and assumptions about the errors, such as constant variance, serial correlation, and normality if m is small. Following the fitting of marginal or mixed models, the assumptions may then be assessed further, with examination of residuals a useful exercise.

Residuals may be defined with respect to different levels. With respect to the usual LMM, a vector of unstandardized *population-level* (marginal) residuals is

$$\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}$$

and these are most useful for analyses based on the marginal (GEE) approach. A vector of unstandardized *unit-level* (stage one) residuals is

$$\boldsymbol{\epsilon}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b}_i.$$

The vector of random effects $\boldsymbol{b}_i$ is also a form of (stage two) residual. Estimated versions of these residuals are

$$\widehat{\boldsymbol{e}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} \tag{8.48}$$

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i \tag{8.49}$$

and $\widehat{\boldsymbol{b}}_i$, $i = 1, \ldots, m$.

We first discuss the population residuals (8.48). Recall, from consideration of the ordinary linear model (Sect. 5.11), that estimated residuals have dependencies induced by replacement of parameters by their estimates. The situation is far worse for dependent data because we would expect the population residuals to be dependent, even if the true parameter values were known. If $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is the true error structure, then

$$\text{var}(\boldsymbol{e}_i) = \boldsymbol{V}_i \;\text{ and }\; \text{var}(\widehat{\boldsymbol{e}}_i) \approx \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}}),$$

showing the dependence of the residuals under the model. This means that, when working with $\boldsymbol{e}_i$, it is difficult to check whether the covariance model is correctly specified. Plotting $\widehat{e}_{ij}$ versus the lth covariate x_{ijl}, $l = 1, \ldots, k$ may also be misleading due to the dependence within the residuals. Therefore, standardization is essential to remove the dependence.

Let $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i^{\mathsf{T}}$ be the Cholesky decomposition of $\widehat{\boldsymbol{V}}_i = \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})$. We can use this decomposition to form

$$\widehat{\boldsymbol{e}}_i^{\star} = \boldsymbol{L}_i^{-1}\widehat{\boldsymbol{e}}_i = \boldsymbol{L}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})$$

so that var($\widehat{\boldsymbol{e}}_i^{\star}$) $\approx \mathbf{I}_{n_i}$. We may then work with the model

$$Y_i^{\star} = \boldsymbol{x}_i^{\star}\boldsymbol{\beta} + \boldsymbol{e}_i^{\star}$$

where $\boldsymbol{Y}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{Y}_i$, $\boldsymbol{x}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{x}_i$, and $\boldsymbol{e}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{e}_i$. Plots of $\widehat{e}_{ij}^\star$ against $x_{ijl}^\star$, $l = 1, \dots, k$ should not show systematic patterns if the assumed linear form is correct.

QQ plots of $\widehat{e}_{ij}^\star$ versus the expected residuals from a normal distribution can be used to assess normality (unless m is small, normal errors are not required for accurate inference, but the closer to normality are the data, the smaller the m required for the asymptotics to have practically "kicked in"). If e_{ij} are normal, then standardized residuals will be normally distributed also, since $e_{ij}^\star$ is a linear combination of elements of $\boldsymbol{e}_i$.

The correctness of the mean–variance relationship can be assessed by plotting $\widehat{e}_{ij}^{\star 2}$ (or $|\widehat{e}_{ij}^\star|$) against fitted values $\widehat{\mu}_{ij}^\star = \boldsymbol{x}_{ij}^\star \widehat{\boldsymbol{\beta}}$. Any systematic (non-horizontal) trends suggest problems. Local smoothers (as described in Chap. 11) can be added to plots to aid interpretation and plotting symbols such as unit or observation number can also be useful to identify collections of observations for which the model is not adequate.

For the LMM with $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{iid} \mathrm{N}_{n_i}(\,\boldsymbol{0}, \sigma_\epsilon^2 \mathbf{I}\,)$, the stage one residuals (8.49) may be formed. Standardized versions are $\widehat{\epsilon}_{ij}^\star = \widehat{\epsilon}_{ij}/\widehat{\sigma}_\epsilon$. As usual, these residuals may be plotted against covariates. One may construct normal QQ plots, though a correct mean–variance relationship is more influential than lack of normality (so long as the sample size is not small). The constant variance assumption may be examined via a plot of $\widehat{\epsilon}_{ij}^{\star 2}$ (or $|\widehat{\epsilon}_{ij}^\star|$) versus $\widehat{\mu}_{ij} = \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}\widehat{\boldsymbol{b}}_i$.

Recall the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i, \tag{8.50}$$

introduced in Sect. 8.4.2, with $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_{q+1}(\,\boldsymbol{0}, \boldsymbol{D}\,)$ and $\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \sim_{iid} \mathrm{N}_{n_i}(\,\boldsymbol{0}, \sigma_\epsilon^2\mathbf{I}\,)$ representing random effects and measurement error and δ_{ij} being zero-mean normal error terms with serial dependence in time. A simple and commonly used form for serial dependence is the AR(1) model (also described in Sect. 8.4.2) which gives covariances

$$\mathrm{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_{ij}-t_{ik}|} = \sigma_\delta^2 R_{ijk}.$$

Conditional on $\boldsymbol{b}_i$, this leads to the variance–covariance for responses on unit i:

$$\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i) = \boldsymbol{V}_i = \sigma_\delta^2 \boldsymbol{R}_i + \sigma_\epsilon^2 \mathbf{I}_{n_i}. \tag{8.51}$$

If model (8.50) is fitted, then residuals of the form (8.49) may be formed, but these should be standardized in the same way as just described for population residuals (i.e., using the decomposition $\widehat{\boldsymbol{V}_i} = \boldsymbol{L}_i\boldsymbol{L}_i^{\mathrm{T}}$) since they will have marginal variance (8.51).

In a temporal setting, one may want to detect whether serial correlation is present in the residuals. Two tools for such detection are the *autocorrelation function* and the *semi-variogram*. We describe the autocorrelation function and the semi-variogram generically with respect to the model

$$Y_t = \mu_t + \epsilon_t,$$

for $t = 1, \ldots, n$. We assume the error terms ϵ_t are *second-order stationary*, which means that $\text{E}[\epsilon_t] = \mu$ is constant, and $\text{cov}(\epsilon_t, \epsilon_{t+d}) = C(d)$, where $d \geq 0$, that is, the covariance only depends on the temporal spacing between the variables. This implies that the variance of ϵ_t is constant, and equal to $C(0)$, for all t. The autocorrelation function (ACF) is defined, for time points $d \geq 0$ apart, as

$$\rho(d) = \frac{\text{cov}(\epsilon_t, \epsilon_{t+d})}{\sqrt{\text{var}(\epsilon_t)\text{var}(\epsilon_{t+d})}} = \frac{C(d)}{C(0)},$$

for all t. Now, suppose we have estimates of the errors $\widehat{\epsilon}_t$ for responses equally spaced over time, which we label as $t = 1, \ldots, n$. The *empirical* ACF is defined as

$$\widehat{\rho}(d) = \frac{\widehat{C}(d)}{\widehat{C}(0)} = \frac{\sum_{t=1}^{n-d} \widehat{\epsilon}_t \, \widehat{\epsilon}_{t+d}/(n-d)}{\sum_{t=1}^{n} \widehat{\epsilon}_t^2/n},$$

for $d = 0, 1, \ldots, n-1$. A *correlogram* plots $\widehat{\rho}(d)$ versus d for $d = 0, 1, 2, \ldots, n-1$. If the residuals are a white noise process (i.e., uncorrelated), then asymptotically

$$\sqrt{n}\,\widehat{\rho}(d) \;\to_d\; \text{N}(0, 1),$$

for $d = 1, 2, \ldots$, to give, for example, 95% confidence bands of $\pm 1.96/\sqrt{n}$.

We now turn to a description of the semi-variogram, a tool which was introduced by Matheron (1971) in the context of spatial analysis (more specifically, geostatistics) and is described in the context of longitudinal data by Diggle et al. (2002, Chap. 3.4). Define the semi-variogram of the residuals ϵ_t, as

$$\gamma(d) = \frac{1}{2}\text{var}(\epsilon_t - \epsilon_{t+d}) = \frac{1}{2}\text{E}\left[(\epsilon_t - \epsilon_{t+d})^2\right]$$

for $d \geq 0$. The reason for the $1/2$ term will soon become apparent. The semi-variogram exists under weaker conditions than the ACF, specifically under *intrinsic stationarity*, which means that ϵ_t has constant mean and $\text{var}(\epsilon_t - \epsilon_{t+d})$ only depends on d (so that the covariance need not be defined). For zero-mean error terms and under second-order stationarity,

$$\begin{aligned}
\gamma(d) &= \frac{1}{2}\text{var}(\epsilon_t) + \frac{1}{2}\text{var}(\epsilon_{t+d}) - \text{cov}(\epsilon_t, \epsilon_{t+d}) \\
&= C(0) - C(d) \\
&= C(0)[1 - \rho(d)].
\end{aligned}$$

Suppose we now have estimated errors $\widehat{\epsilon}_l$, along with associated times t_l, $l = 1, \ldots, n$. The sample semi-variogram uses the empirical halved squared differences between pairs of residuals

$$v_{ll'} = \frac{1}{2}(\widehat{\epsilon}_l - \widehat{\epsilon}_{l'})^2,$$

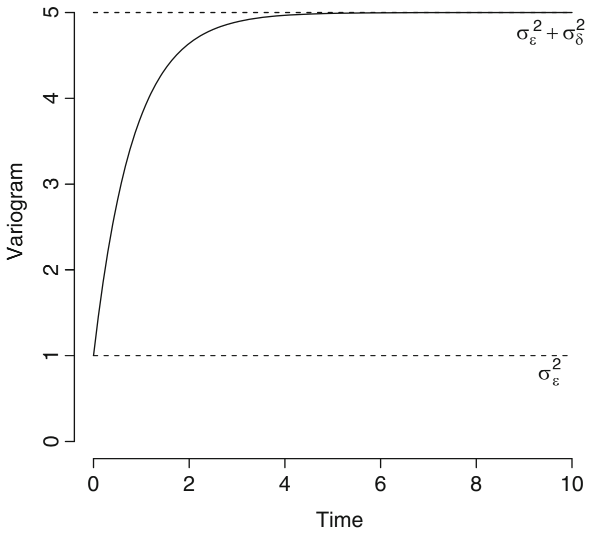


Fig. 8.4 Theoretical (semi-)variogram corresponding to (8.53) with $\sigma_\epsilon^2 = 1$, $\sigma_\delta^2 = 4$ and $\rho = 0.3$

along with the spacings $d_{ll'} = |t_l - t_{l'}|$ for $l = 1, \ldots, n$ and $l < l' = 1, \ldots, n$. With irregular sampling times, the variogram can be estimated from the pairs $(d_{ll'}, v_{ll'})$, with the resultant plot being smoothed.[2] An example of such a plot is given in Fig. 8.9. Under normality of the data, the marginal distribution of each $v_{ll'}$ is $C(0)\chi_1^2$, and this large variability can make the variogram difficult to interpret. In addition, because each residual contributes to $n-1$ terms in the empirical cloud of points, the points are not independent, and a single outlying point can influence the plot at different time lags.

Suppose now we are in a longitudinal setting, in which response y_{ij} is observed at time t_{ij}, and we fit the LMM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{8.52}$$

with the usual forms for $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$. After fitting, we form the stage one residuals (8.49), that is, $\widehat{\epsilon}_{ij} = y_{ij} - \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_{ij}\widehat{\boldsymbol{b}}_i$. We might believe the serial dependence takes the same form across individuals. For equally spaced times, we can examine the empirical ACF of the residuals where, for simplicity, we assume that there are n responses on each of the m individuals,

$$\widehat{\rho}(d) = \frac{\sum_{i=1}^m \sum_{j=1}^{n-d} \widehat{\epsilon}_{ij}\, \widehat{\epsilon}_{i,j+d}/(n-d)}{\sum_{i=1}^m \sum_{j=1}^n \widehat{\epsilon}_{ij}^2/n},$$

for $d = 0, 1, \ldots, n-1$.

[2] For unequally spaced times, the longitudinal data literature often recommends the construction of the empirical semi-variogram (Diggle et al. 2002, Sect. 3.4; Fitzmaurice et al. 2004, Sect. 9.4), though one could construct and smooth the empirical covariance function in a similar fashion.

Now suppose that we again fit model (8.52), and we have n_i responses for individual i with sampling times t_{ij}. We then define the semi-variogram for the ith individual as

$$\gamma_i(d_{ijk}) = \frac{1}{2}\text{E}\left[(\epsilon_{ij} - \epsilon_{ik})^2\right]$$

where $d_{ijk} = |t_{ij} - t_{ik}|$. We now form

$$v_{ijk} = \frac{1}{2}(\widehat{\epsilon}_{ij} - \widehat{\epsilon}_{ik})^2$$

and the semi-variogram can then be estimated by plotting the pairs (d_{ijk}, v_{ijk}) for $i = 1, \dots, m$ and $j < k = 1, \dots, n_i$ and smoothing. If no serial dependence is present, the smoother should be roughly horizontal.

Consider the interpretation of the variogram when model (8.50) is the "truth," but suppose we fit a LMM without the autocorrelated terms. We consider stage one residuals, which under (8.50) will take the form

$$\epsilon'_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} - \boldsymbol{z}_{ij}\boldsymbol{b}_i = \delta_{ij} + \epsilon_{ij}.$$

For differences in residuals on the same individual,

$$\begin{aligned}\epsilon'_{ij} - \epsilon'_{ik} &= \delta_{ij} + \epsilon_{ij} - \delta_{ik} - \epsilon_{ik}\\ &= (\delta_{ij} - \delta_{ik}) + (\epsilon_{ij} - \epsilon_{ik}),\end{aligned}$$

and so the semi-variogram takes the form

$$\begin{aligned}\gamma_i(d_{ijk}) &= \frac{1}{2}\text{E}\left[(\epsilon'_{ij} - \epsilon'_{ik})^2\right]\\ &= \frac{1}{2}\text{E}\left[(\delta_{ij} - \delta_{ik})^2 + (\epsilon_{ij} - \epsilon_{ik})^2\right]\\ &= \sigma_\delta^2[1 - \rho(d_{ijk})] + \sigma_\epsilon^2. \qquad (8.53)\end{aligned}$$

As $d_{ijk} \to 0$, $\gamma_i(d_{ijk}) \to \sigma_\epsilon^2$. The rate at which asymptote $\sigma_\delta^2 + \sigma_\epsilon^2$ is reached as $d_{ijk} \to \infty$ is determined by ρ. This variogram is illustrated in Fig. 8.4.

We now briefly consider the use of population residuals, starting with the random intercepts model:

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \delta_{ij} + \epsilon_{ij},$$

with $b_i \mid \sigma_0^2 \sim_{iid} \text{N}(0, \sigma_0^2)$ and the AR(1) model for δ_{ij}. The population residuals under this model are

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} = b_i + \delta_{ij} + \epsilon_{ij},$$

$i = 1, \dots, m; j = 1, \dots n_i$. For differences in residuals on the same individual,

$$\begin{aligned} e_{ij} - e_{ik} &= b_i + \delta_{ij} + \epsilon_{ij} - b_i - \delta_{ik} - \epsilon_{ik} \\ &= (\delta_{ij} - \delta_{ik}) + (\epsilon_{ij} - \epsilon_{ik}), \end{aligned}$$

and so we obtain the same semi-variogram, (8.53), as before. Since b_i is constant for individual i, its variance does not appear.

In general, the variogram is limited in its use for population residuals for the LMM, as we now illustrate. Consider the LMM with random intercepts and independent random slopes:

$$b_{i0} \mid D_0 \sim \text{N}(0, D_0), \quad b_{i1} \mid D_1 \sim \text{N}(0, D_1).$$

This leads to marginal variance

$$\text{var}(Y_{ij}) = \sigma_\epsilon^2 + D_0 + D_1 t_{ij}^2,$$

which is not constant over time. Therefore, a semi-variogram of population residuals should not be constructed, because we do not have second-order stationarity.

Predictions of the random effects $\widehat{\boldsymbol{b}}_i$ may be used to assess assumptions associated with the random effects distribution, though since these have undergone shrinkage, they may be deceptive. One may instead carry out individual fitting and then use the resultant estimates to assess the normality assumption. The latter may be assessed via QQ plots, but the interpretation of plots requires care since estimates and not observed quantities are being plotted; see Lange and Ryan (1989). We may also assess whether the variance of the random effects is independent of covariates $\boldsymbol{x}_i$. If the spread of the random effects distribution depends on the levels of covariates, and this is missed, then inaccurate inference can result (Heagerty and Kurland 2001). For the LMM, it is better to examine stage one and stage two residuals separately, rather than population residuals, since the latter are a mixture of the two, and so, if something appears amiss, it is difficult to determine the stage at which the inadequacy is occurring. As usual, as discussed in Sect. 4.9, the implications of changing the model should be carefully considered, and one should avoid the temptation to model every nuance of the data.

Example: FEV1 Over Time

The dental data that have formed our running illustration are balanced, and there are few individuals and time points, and so, these data are not ideal for illustrating model checking. Hence, we introduce data from an epidemiological study described by van der Lende et al. (1981). We analyze a sample of 133 men and women, initially aged 15–44, from the rural area of Vlagtwedde in the Netherlands. Study participants were followed over time to obtain information on the prevalence of, and risk factors for, chronic obstructive lung diseases. These data were previously

Table 8.4 Mean FEV1 (and sample size) by smoking status and time

Time	Former smoker	Current smoker
0	3.52 (23)	3.23 (85)
3	3.58 (27)	3.12 (95)
6	3.26 (28)	3.09 (89)
9	3.17 (30)	2.87 (85)
12	3.14 (29)	2.80 (81)
15	2.87 (24)	2.68 (73)
19	2.91 (28)	2.50 (74)

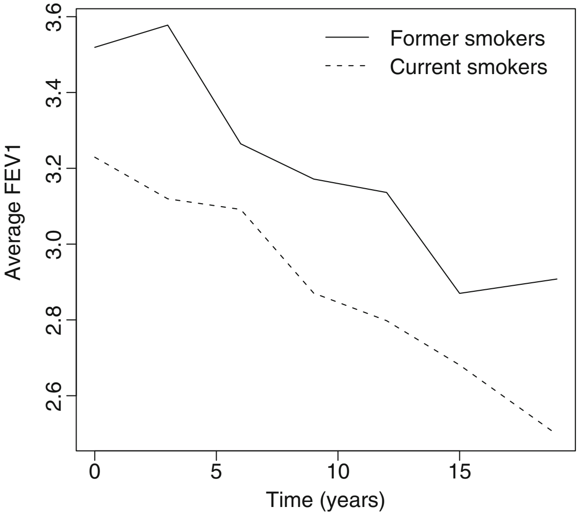


Fig. 8.5 Mean FEV1 profiles versus time for two smoking groups

analyzed by Fitzmaurice et al. (2004). Follow-up surveys provided information on respiratory symptoms and smoking status. Pulmonary function was measured by spirometry, and a measure of forced expiratory volume (FEV1) was obtained every 3 years for the first 15 years of the study and also at year 19. Each study participant was either a current or a former smoker, with current smoking defined as smoking at least one cigarette per day. In this dataset, FEV1 was not recorded for every subject at each of the planned measurement occasions so that the number of measurements of FEV1 on each subject varied between 1 and 7. Table 8.4 shows the numbers of observations available at each time point. There are 32 former smokers and 101 current smokers in total, and we see that the numbers with missing observations at each time point are not drastically different.

Figure 8.5 plots the mean FEV1 profiles versus time for former smokers (solid line) and current smokers (dashed line). It is clear that there is a difference in the overall level, with former smokers having higher responses. Whether the rate of decline in FEV1 is different in the two groups is not so obvious. Figure 8.6 plots the individual trajectories versus time for former smokers (solid lines) and current smokers (dashed lines). There is clearly large between-individual variability in levels so that observations on the same individual will be correlated.

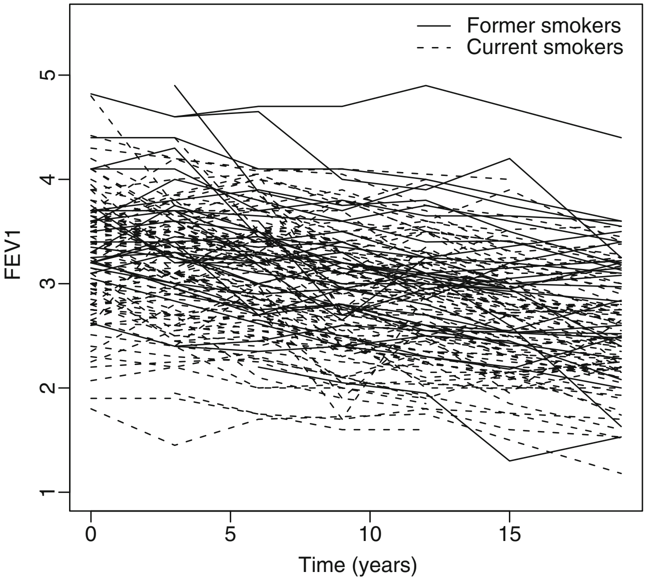


Fig. 8.6 FEV1 versus time for 133 individuals, former and current smokers are indicated by *solid* and *dashed lines* respectively

Let Y_{ij} represent the FEV1 on individual i at time (from baseline) t_{ij} (in years), with $S_i = 0/1$ indicating former/current smoker. We treat this example as illustrative only and therefore fit various models to examine the effects on inference and to demonstrate model assessment and comparison. We initially fit the following three models using REML:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij} \tag{8.54}$$

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + b_i + \epsilon_{ij} \tag{8.55}$$

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 S_i \times t_{ij} + b_i + \epsilon_{ij} \tag{8.56}$$

with $b_i \mid \sigma_0^2 \sim_{iid} \text{N}(0, \sigma_0^2)$ and $\epsilon_{ij} \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ and with ϵ_{ij} and b_i independent, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. We emphasize that the random effect distribution is assumed common to both former and current smokers. Estimates and standard errors for β_1, β_2, and β_3 are given in Table 8.5. We include an ordinary least squares (OLS) fit of the model $\text{E}[Y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 S_i \times t_{ij}$. This model is clearly inappropriate since it assumes independent observations but, when compared to the equivalent LMM, (8.56), illustrates that the standard errors of the estimates corresponding to time-varying covariates (time β_1 and the interaction β_3) are reduced under the LMM. This behavior occurs because within-individual comparisons are more efficient in a longitudinal study (as discussed in Sect. 8.3).

Table 8.5 Results of various LMM analyses and an ordinary least squares (OLS) fit to the FEV1 data

Model	β_1 (Time)	s.e.	β_2 (Smoke)	s.e.	β_3 (Inter)	s.e.
LMM TIME	−0.037	0.0013	–	–	–	–
LMM TIME + SMOKE	−0.037	0.0013	−0.31	0.11	–	–
LMM TIME × SMOKE	−0.034	0.0026	−0.27	0.11	−0.0046	0.0030
OLS TIME × SMOKE	−0.038	0.0067	−0.31	0.085	−0.00041	0.0077

Table 8.6 Results of LMM (likelihood and Bayesian) and GEE analyses for the FEV1 data

Model	β_1 (Time)	s.e.	β_2 (Smoke)	s.e.	σ_0
Likelihood LMM	−0.037	0.0013	−0.31	0.11	0.53
Bayes LMM	−0.037	0.0013	−0.31	0.12	0.53
GEE	−0.037	0.0015	−0.31	0.11	–
Likelihood LMM AR(1)	−0.037	0.0013	−0.31	0.11	0.53

σ_0 is the standard deviation of the random intercepts

To compare the three LMMs in Table 8.5, we must use MLE for likelihood ratio tests, since the data are not constant under the different models under REML (due to different $\widehat{\boldsymbol{\beta}}_{\text{G}}$, Sect. 8.5.3). For

$$H_0 : \text{ Model (8.54) versus } H_1 : \text{ Model (8.55)}$$

we have a likelihood ratio statistic of 8.22 on 1 degree of freedom and a p-value of 0.0042. Hence, there is strong evidence to reject the null, and we conclude that there are differences in intercepts for former and current smokers (as we suspected from Fig. 8.5). For

$$H_0 : \text{ Model (8.55) versus } H_1 : \text{ Model (8.56)}$$

we have a likelihood ratio statistic of 2.29 on 1 degree of freedom and a p-value of 0.13. Hence, under conventional levels of significance, there is no reason to reject the null, and we conclude that the interaction is not needed, so that the decline in FEV1 with time is the same for both former and current smokers.

We now report a Bayesian analysis of model (8.55) with improper flat priors on $\beta_0, \beta_1, \beta_2$, the improper prior $\sigma_\epsilon^2 \propto \frac{1}{\sigma_\epsilon^2}$ and $\sigma_0^{-2} \sim \text{Ga}(0.5, 0.02)$. The latter prior gives 95% of its mass for σ_0, the standard deviation of the between-individual intercepts, between 0.09 and 6.5. Table 8.6 gives the results, which are very similar to those of the likelihood-based approach, which is reassuring.

We now fit the marginal model version of (8.55) using GEE. We use an exchangeable correlation structure, since clearly we have dependence between measurements on the same individual at different times, but the exact form of the correlation is not clear. The results are given in Table 8.6 and again show good agreement for the regression coefficients. In the exchangeable correlation structure, there are two components to $\boldsymbol{\alpha}$, a marginal variance, α_1, and a common marginal correlation, α_2. The model may be compared to the random intercepts

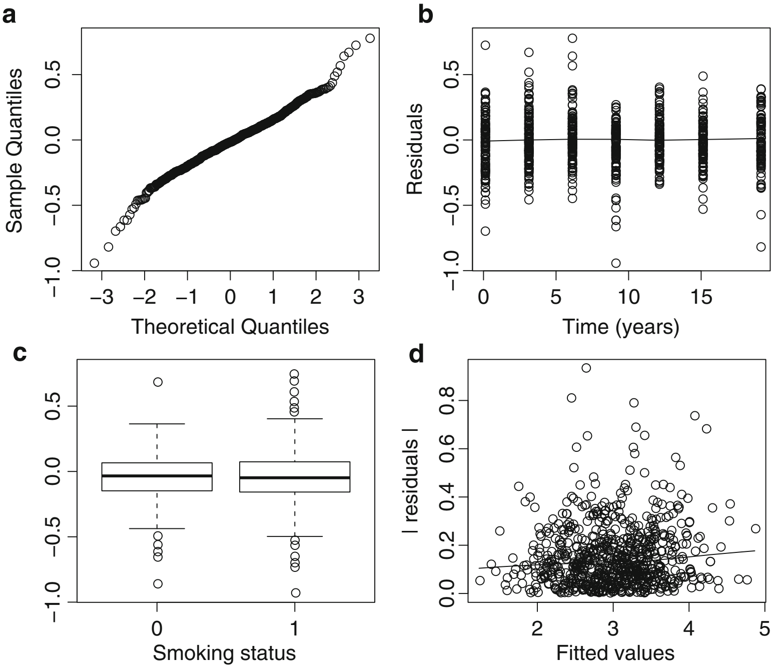


Fig. 8.7 Stage one residual plots for the FEV1 data: (**a**) normal QQ plot, (**b**) residuals versus time, (**c**) residuals as a function of smoking status (0 = former smoker, 1 = current smoker), (**d**) absolute value of residuals versus fitted values

model in which we have marginal variance $\alpha_1 = \sigma_0^2 + \sigma_\epsilon^2$ and marginal correlation $\alpha_2 = \sigma_0^2/(\sigma_0^2 + \sigma_\epsilon^2)$. From the GEE analysis, $\widehat{\alpha}_1 = 0.31$ and $\widehat{\alpha}_2 = 0.82$ to give $\sqrt{\widehat{\alpha}_1 \times \widehat{\alpha}_2} = 0.50$, which is comparable to the estimates of $\widehat{\sigma}_0 = 0.53$ in Table 8.6.

We now examine the assumptions of the various approaches. We focus on the linear model that includes time and smoking (but no interaction). Figure 8.7 summarizes the stage one residuals:

$$\widehat{\epsilon}_{ij} = y_{ij} - \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i.$$

Panel (a) shows that the distribution of the errors is symmetric but heavier tailed than normal. With such a large sample, there is nothing troubling in this plot, and, there are no outlying points. Panels (b) and (c) plot the residuals against time and smoking status. We see no nonlinear behavior in the time plot and no great divergence from constant variance in either plot. A very important assumption in mixed effects modeling is that a common random effects distribution across covariates is appropriate. To examine this assumption, separate analyses were carried out for former and current smokers. The estimates of the variance components for former smokers were $\widehat{\sigma}_\epsilon = 0.22$ and $\widehat{\sigma}_0 = 0.58$ and for current smokers, $\widehat{\sigma}_\epsilon = 0.21$ and $\widehat{\sigma}_0 = 0.51$. The differences between estimates in the two groups are small, and we

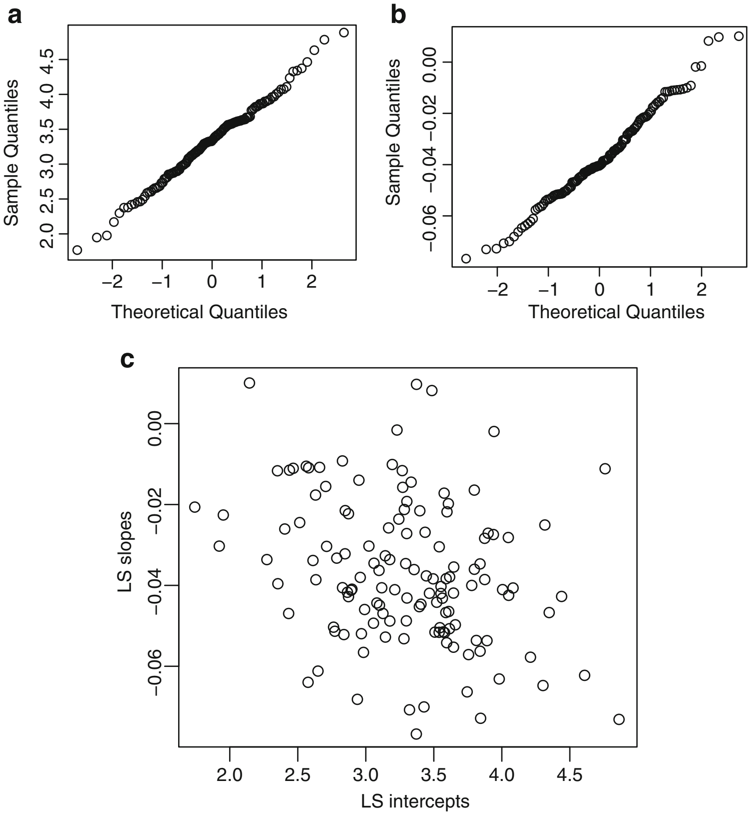


Fig. 8.8 Normal QQ plots of OLS estimates for the FEV1 data for (**a**) intercepts, (**b**) slopes, and (**c**) scatterplot of pairs of least square estimates

conclude that a common random effects distribution is reasonable. Panel (d) plots the absolute value of the residuals versus the fitted values $\boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} + \widehat{b}_i$, along with a smoother. If the variance function is correctly specified, then we should see no systematic pattern. Here, there is nothing to be too concerned about since there is only a slight increase in variability as the mean increases. These residual plots are based on residuals from the likelihood analysis (the Bayesian versions are similar).

For the 132 individuals with more than a single response, individual OLS fits were performed. Figure 8.8 shows normal QQ plots of the intercept and slope parameter estimates in panels (a) and (b) and a bivariate scatter plot of the pairs of estimates in panel (c). The estimates look remarkably normal, at least in (a) and (b), and there are no outlying individuals.

Finally, we examine the residuals for serial correlation. Figure 8.9 gives the semi-variogram of the stage one residuals along with a smoother and indicates some evidence of dependence. In panel (a), the pattern is not apparent, but in panel (b), the semi-variance axis is reduced for clarity, which allows the trend to be more

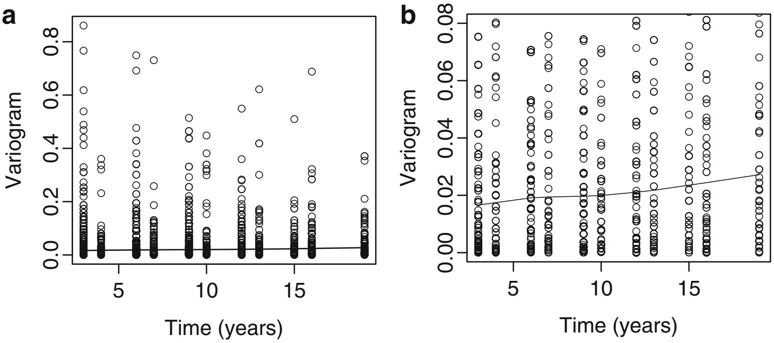


Fig. 8.9 For the FEV1 data: (**a**) the (semi)-variogram of stage one residuals, (**b**) on a truncated semi-variogram scale

clearly seen. Consequently, we fit an AR(1) model to the residuals (Sect. 8.4.2), using restricted maximum likelihood, and obtain the parameter estimates in the last row of Table 8.6. This model is a significant improvement over the non-serial correlation model (as measured by a likelihood ratio test, $p = 0.0002$). However, there is virtually no change in the estimates/standard errors, since the AR correlation parameter is just 0.20, with an asymptotic 95% confidence interval of [0.087, 0.30].

We may also examine whether random slopes are required. Fitting this model via restricted likelihood gave a standard deviation of $\widehat{\sigma}_1 = 0.0099$. The likelihood ratio statistic test for correlated random intercepts and slopes, versus random intercepts only, is 10.9 which is significant at around the 0.0025 level (where the distribution under the null is a mixture of χ_1^2 and χ_2^2 distributions, see Sect. 8.5.2).

In terms of the fixed effects, there is little sensitivity to the assumed random effects structure. Inference under the random intercepts and slopes models is similar to the random intercepts only model, since the between-individual variability in slopes is small (though statistically significant). The population change in FEV1 is a drop of 0.0371 per year, with a standard error of 0.0013–0.0015 depending on the model. The posterior median for the intraindividual correlation, $\sigma_0^2/(\sigma_\epsilon^2 + \sigma_0^2)$, is 0.84 with 95% interval [0.82, 0.89] suggesting that the majority of the variability in FEV1 is between individual.

8.9 Cohort and Longitudinal Effects

We now describe another benefit of longitudinal studies, the ability to estimate both longitudinal and cohort effects. We frame the discussion around the modeling of Y = FEV1 as a function of age. We might envisage that FEV1 changes as age increases within an individual and that individuals may have different baseline levels of FEV$_1$ due to "cohort" effects. A birth cohort is a group of individuals who were

Fig. 8.10 Three population (cohort) trajectories over time

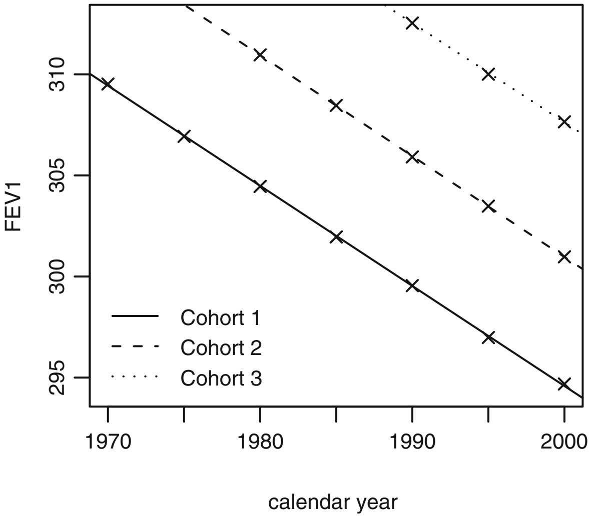
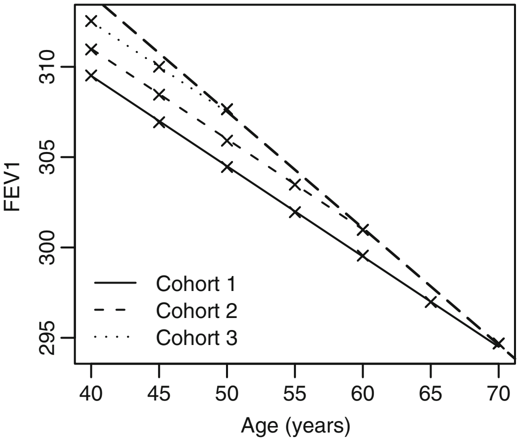


Fig. 8.11 Relationship between cross-sectional and longitudinal effects in a hypothetical example with three populations. The *dashed line* (which is the top line) represents the cross-sectional slope



born in the same year. Cohort effects may include the effects of environmental pollutants and differences in lifestyle choices or medical treatment received. In a cross-sectional study, a group of individuals are measured at a single time point. A great advantage of longitudinal studies, as compared to cross-sectional studies, is that both cohort and aging (longitudinal) effects may be estimated.

As an illustration, Fig. 8.10 shows the trajectories of three hypothetical individuals as a function of calendar time. The starting positions are different due to cohort effects. Figure 8.11 shows the same individuals but with trajectories plotted versus age. The cross-sectional association, which would result from observing the final measurement only, is highlighted and displays a steeper decline than seen in the longitudinal slope.

To examine in more detail the issues, consider the model

$$\mathrm{E}[Y_{ij} \mid x_{ij}, x_{i1}] = \beta_0 + \beta_{\mathrm{C}} x_{i1} + \beta_{\mathrm{L}}(x_{ij} - x_{i1})$$

where Y_{ij} is the jth FEV1 measurement on individual i and x_{ij} is the age of the individual at occasion j, with x_{i1} being the age on a certain day (so that all the individuals are comparable). At the first occasion,

$$\mathrm{E}[Y_{i1} \mid x_{i1}] = \beta_0 + \beta_{\mathrm{C}} x_{i1},$$

so that β_{C} is the average change in response between two populations who differ by one unit in their baseline ages. Said another way, we are examining the differences in FEV1 between two birth cohorts a year apart, so that β_{C} is the *cohort effect*.

Since

$$\mathrm{E}[Y_{ij} \mid x_{ij}, x_{i1}] - \mathrm{E}[Y_{i1} \mid x_{i1}] = \beta_{\mathrm{L}}(x_{ij} - x_{i1})$$

it is evident that β_{L} is the *longitudinal effect*, that is, the change in the average FEV1 between two populations who are in the same birth cohort and whose ages differ by 1 year. The usual cross-sectional model is

$$\begin{aligned} \mathrm{E}[Y_{ij} \mid x_{ij}] &= \beta_0 + \beta_1 x_{ij} \\ &= \beta_0 + \beta_1 x_{i1} + \beta_1(x_{ij} - x_{i1}) \end{aligned} \tag{8.57}$$

so that the model implicitly assumes equal longitudinal and cohort effects, that is, $\beta_1 = \beta_{\mathrm{C}} = \beta_{\mathrm{L}}$.

In the cross-sectional study with model (8.57),

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})(Y_{ij} - \overline{Y})}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2}$$

with $\overline{x} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij}$ and $\overline{Y} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}$ with $N = \sum_{i=1}^{m} n_i$. The expected value of this estimator is

$$\mathrm{E}[\widehat{\beta}_1] = \beta_{\mathrm{L}} + \frac{\sum_{i=1}^{m} n_i (x_{i1} - \overline{x}_1)(\overline{x}_i - \overline{x})}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2} (\beta_{\mathrm{C}} - \beta_{\mathrm{L}}) \tag{8.58}$$

(Exercise 8.15) so that the estimate is a combination of cohort and longitudinal effects. The cross-sectional regression model will give an unbiased estimate of the longitudinal association if $\beta_{\mathrm{L}} = \beta_{\mathrm{C}}$ or if $\{x_{i1}\}$ and $\{\overline{x}_i\}$ are orthogonal. To conclude, longitudinal studies can be powerfully employed to separate cohort and longitudinal effects.

8.10 Concluding Remarks

In this chapter we have described two approaches to fitting linear models to dependent data: LMMs and GEE. GEE has the fewest assumptions and is designed for population-level inference. Asymptotics are required for inference, and so, GEE is less appealing when the number of individuals m is small. A sufficiently large sample size is required for both normality of the estimator and reliability of the sandwich variance estimator. The use of the sandwich variance estimator makes GEE the most dependable method in large sample situations. However, there can be losses in efficiency if we choose a working correlation matrix that is far from reality. With GEE, it is not possible to make inference for individuals or incorporate prior information.

LMMs are more flexible than GEE in terms of the questions that can be addressed with the data, but this flexibility comes at the price of a greater number of assumptions. For likelihood inference, as with GEE, we require the number of units m to be sufficiently large for asymptotic inference. Prior information cannot be incorporated in a likelihood analysis; for that, we need a Bayesian approach. For a small number of individuals, a Bayesian approach fully captures the uncertainty, but inference is completely model-based, and with a small number of individuals, it is unlikely that we will be able to check the modeling assumptions.

8.11 Bibliographic Notes

For descriptions of linear mixed effects models, see Hand and Crowder (1996, Chap. 5), Diggle et al. (2002, Sects. 4.4 and 4.5), and Verbeeke and Molenberghs (2000). Covariance models are described in Verbeeke and Molenberghs (2000, Chap. 10), Pinheiro and Bates (2000, Chap. 5), and Diggle et al. (2002, Chap. 5). Demidenko (2004) provides theory for mixed models, including the linear case. Robinson (1991) provides an interesting discussion of BLUP estimates. Two early influential references on the LMM from Bayesian and likelihood perspectives, respectively, are Lindley and Smith (1972) and Laird and Ware (1982).

The name GEE was coined by Liang and Zeger (1986) and Zeger and Liang (1986). See also Gourieroux et al. (1984) who considered sandwich estimation for regression parameters with a consistent estimator of additional parameters. Prentice (1988) introduced a second estimating equation for estimation of $\boldsymbol{\alpha}$. Crowder (1995) points out that the existence of the $\boldsymbol{\alpha}$ parameters in the working covariance matrix is not guaranteed, in which case the asymptotics break down. Fitzmaurice et al. (2004) is an excellent practical text on longitudinal modeling.

8.12 Exercises

8.1 *A Gauss–Markov Theorem for Dependent Data:* Suppose $\text{E}[\boldsymbol{Y}] = \boldsymbol{x\beta}$ and $\text{var}(\boldsymbol{Y}) = \boldsymbol{V}$, with $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\text{T}}, \dots, \boldsymbol{Y}_m^{\text{T}}]^{\text{T}}$ and where $\boldsymbol{Y}_i = [Y_{i1}, \dots, Y_{in_i}]^{\text{T}}$ and $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_m]^{\text{T}}$ is $N \times (k+1)$ with $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i}]$, $\boldsymbol{x}_{ij} = [1, x_{ij1}, \dots, x_{ijk}]^{\text{T}}$, $N = \sum_i n_i$ and $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of regression coefficients.

Consider linear estimators of the form

$$\widetilde{\boldsymbol{\beta}}_{\text{W}} = (\boldsymbol{x}^{\text{T}}\boldsymbol{W}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{W}^{-1}\boldsymbol{Y},$$

where $\boldsymbol{W}$ is symmetric and positive definite. Show that:

(a) $\text{E}[\widetilde{\boldsymbol{\beta}}_{\text{W}}] = \boldsymbol{\beta}$.
(b) $\text{var}(\widetilde{\boldsymbol{\beta}}_{\text{V}}) \leq \text{var}(\widetilde{\boldsymbol{\beta}}_{\text{W}})$.

[Hint: In (b), show that $\text{var}(\widetilde{\boldsymbol{\beta}}_{\text{W}}) - \text{var}(\widetilde{\boldsymbol{\beta}}_{\text{V}})$ is positive semi-definite.]

8.2 Consider the data in Table 5.4 (from Davies 1967) that were presented in Sect. 5.8.1. These data consist of the yield in grams from six randomly chosen batches of raw material, with five replicates each. The aim of this experiment was to find out to what extent batch-to-batch variation was responsible for variation in the final product yield.

One possibility for a model for these data is the one-way analysis of variance with

$$y_{ij} = \mu + b_i + \epsilon_{ij},$$

with $j = 1, \dots, n$, replicates on $i = 1, \dots, m$, batches, $b_i \mid \sigma_0^2 \sim_{iid} \text{N}(0, \sigma_0^2)$, $\epsilon_{ij} \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$, with b_i and ϵ_{ij} independent.

In what follows the following identity is useful. Let $\mathbf{I}_n$ denote the $n \times n$ identity matrix and $\boldsymbol{J}_n$ the $n \times n$ matrix of 1's. Then

$$(a\mathbf{I}_n + b\boldsymbol{J}_n)^{-1} = \frac{1}{a}\left(\mathbf{I}_n - \frac{b}{a+nb}\boldsymbol{J}_n\right), \quad a \neq 0, \quad a \neq -nb,$$

and

$$|a\mathbf{I}_n + b\boldsymbol{J}_n| = a^{n-1}(a+nb).$$

(a) Derive the log-likelihood for $\mu, \sigma_0^2, \sigma_\epsilon^2$.
(b) Differentiate the log-likelihood, and show that the MLEs are

$$\begin{aligned}\hat{\mu} &= \bar{y}_{\cdot\cdot},\\ \widehat{\sigma}_\epsilon^2 &= \text{MSE},\\ \widehat{\sigma}_0^2 &= \frac{(1-1/m)\text{MSA} - \text{MSE}}{n},\end{aligned}$$

where MSA= $n\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2/(m-1)$ and MSE= $\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2/[m(n-1)]$.
[Hint: Life is easier if the model is parameterized in terms of $\lambda = \sigma_\epsilon^2 + n\sigma_0^2$.]

(c) Obtain the form of var($\widehat{\mu}$), and give an estimator of this quantity.
(d) Find the REML estimators of σ_0^2 and σ_ϵ^2.
(e) In the one-way random effects model with balanced data, it can be shown that

$$\frac{\text{MSA}/(n\sigma_0^2 + \sigma_\epsilon^2)}{\text{MSE}/\sigma_\epsilon^2} \sim F_{m-1,m(n-1)},$$

the F distribution on $m-1$ and $m(n-1)$ degrees of freedom. Use this result to explain why F = MSA/MSE may be compared with an $F_{m-1,m(n-1)}$ distribution to provide a test of $H_0 : \sigma_0^2 = 0$.
(f) Using the last part, show that the probability that the REML estimator $\widehat{\sigma}_0^2$ is negative is the probability that an $F_{m(n-1),(m-1)}$ random variable is bigger than $1 + n\sigma_0^2/\sigma_\epsilon^2$.
(g) Numerically obtain an MLE, with associated standard error, for μ. Additionally, find ML and REML estimates of σ_0^2 and σ_ϵ^2.
(h) Confirm these estimates using a statistical package.

8.3 Consider the so-called Neymann–Scott problem (previously considered in Exercises 2.6 and 3.3) in which

$$Y_{ij} \mid \mu_i, \sigma^2 \sim_{ind} \text{N}(\mu_i, \sigma^2),$$

for $i = 1, \ldots, n, j = 1, 2$.

(a) Obtain the MLE for σ^2, and show that it is inconsistent. Why are there problems here?
(b) Consider a REML approach. Assign an improper uniform prior to $\mu_1, \ldots, \mu_n$, and integrate out these parameters. Obtain the REML of σ^2, and show that it is an unbiased estimator.

8.4 Derive (8.25) and (8.26).
[Hint: The identities

$$\left(\frac{\boldsymbol{z}_i^{\text{T}}\boldsymbol{z}_i}{\sigma_\epsilon^2} + \boldsymbol{D}^{-1}\right)^{-1}\frac{\boldsymbol{z}_i^{\text{T}}}{\sigma_\epsilon^2} = \boldsymbol{D}\boldsymbol{z}_i^{\text{T}}\boldsymbol{V}_i^{-1}$$

$$\left(\boldsymbol{D}^{-1} + \frac{\boldsymbol{z}_i^{\text{T}}\boldsymbol{z}_i}{\sigma_\epsilon^2}\right)^{-1} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{z}_i\boldsymbol{D}$$

are useful. These follow from

$$(\boldsymbol{E}+\boldsymbol{F})^{-1}\boldsymbol{E} = \mathbf{I} - (\boldsymbol{E}+\boldsymbol{F})^{-1}\boldsymbol{F}$$
$$(\boldsymbol{G}+\boldsymbol{E}\boldsymbol{F}\boldsymbol{E}^{\mathrm{T}})^{-1} = \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1}\boldsymbol{E}(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{E}+\boldsymbol{F}^{-1})^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G}^{-1},$$

respectively.]

8.5 Show that

$$\begin{aligned}\operatorname{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i) &= \operatorname{var}(\boldsymbol{b}_i) - \operatorname{var}(\widehat{\boldsymbol{b}}_i) = \boldsymbol{D} - \operatorname{var}(\widehat{\boldsymbol{b}}_i)\\ &= \boldsymbol{D} - \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D} + \boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{z}_i\boldsymbol{D}.\end{aligned}$$

8.6 Consider the class of linear predictors $\boldsymbol{b}^*(\boldsymbol{y}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}$, where $\boldsymbol{a}$ and $\boldsymbol{B}$ are constants of dimensions $(q+1)\times 1$ and $(q+1)\times n$. Let $\boldsymbol{W} = \boldsymbol{b} - \boldsymbol{B}\boldsymbol{y}$, and show that

$$\mathrm{E}[(\boldsymbol{b}^* - \boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{b}^* - \boldsymbol{b})] = [\boldsymbol{a} - \mathrm{E}(\boldsymbol{W})]^{\mathrm{T}}\boldsymbol{A}\,[\boldsymbol{a} - \mathrm{E}(\boldsymbol{W})] + \operatorname{tr}[\boldsymbol{A}\operatorname{var}(\boldsymbol{W})]\,.$$

Deduce that this expression is minimized by taking $\boldsymbol{a} = -\boldsymbol{B}\boldsymbol{x}\boldsymbol{\beta}$ and $\boldsymbol{B} = \boldsymbol{D}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{V}^{-1}$. Hence, show that

$$\boldsymbol{D}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})$$

is the best linear predictor of $\boldsymbol{b}$, whatever the distributions of $\boldsymbol{b}$ and $\boldsymbol{y}$.

8.7 Prove that if the prior distribution for $\boldsymbol{\theta}^{\mathrm{T}} = [\theta_1, \ldots, \theta_m]$ can be written as

$$p(\boldsymbol{\theta}) = \int \prod_{i=1}^{m} p(\theta_i \mid \phi)p(\phi)\,d\phi,$$

then the covariances $\operatorname{cov}(\theta_i, \theta_j)$ are all nonnegative.
[Hint: You may assume that $\mathrm{E}[\theta_i \mid \phi] = \mathrm{E}[\theta_j \mid \phi]$ for $i \neq j$.]

8.8 We return to the yield data of Exercise 8.2.

(a) Numerically evaluate the formula

$$\widehat{\boldsymbol{b}}_i = \mathrm{E}[\boldsymbol{b}_i \mid \boldsymbol{y}_i] = \widehat{\boldsymbol{D}}\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})$$

in your favorite package, and obtain predictions for the yield data.
(b) Obtain measures of the variability of the prediction via $\operatorname{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i)$.
(c) Confirm your predictions using LMM software.

8.9 A Bayesian analysis of the yield data of Exercise 8.2 will now be performed. In terms of the parameters $\beta_0, \sigma_\epsilon^2$, and $\lambda = \sigma_\epsilon^2 + n\sigma_0^2$, the likelihood is

$$\begin{aligned}p(\boldsymbol{y} \mid \beta_0, \sigma_\epsilon^2, \lambda) &= (2\pi)^{-nm/2}(\sigma_\epsilon^2)^{-m(n-1)/2}\lambda^{-m/2}\\ &\quad\times \exp\left\{-\frac{1}{2}\left[\frac{nm(y_{++} - \beta_0)^2}{\lambda} + \frac{\mathrm{SS_B}}{\lambda} + \frac{\mathrm{SS_W}}{\sigma_\epsilon^2}\right]\right\}\end{aligned}$$

where

$$\text{SS}_{\text{B}} = n\sum_{i=1}^{m}(y_{i+} - y_{++})^2, \quad \text{SS}_{\text{W}} = \sum_{i=1}^{m}\sum_{i=1}^{n}(y_{ij} - y_{i+})^2.$$

Assume the improper prior

$$\pi(\beta_0, \sigma_\epsilon^2, \lambda) \propto \frac{1}{\sigma_\epsilon^2 \lambda}.$$

(a) Integrate β_0 from the joint posterior $p(\beta_0, \sigma_\epsilon^2, \lambda \mid \boldsymbol{y})$ to obtain $p(\sigma_\epsilon^2, \lambda | \, \boldsymbol{y})$. Show that this distribution has the form of a product of independent inverse gamma distributions with an additional term that is due to the constraint $\lambda > \sigma_\epsilon^2 > 0$.

(b) Obtain the distribution of $p(\beta_0 \mid \sigma_\epsilon^2, \lambda, \boldsymbol{y})$.

(c) Give details of a composition algorithm (as described in Sect. 3.8.4) for simulating from the posterior $p(\beta_0, \sigma_\epsilon^2, \lambda \mid \boldsymbol{y})$.

(d) Implement the algorithm for the yield data.

 (i) Give histograms and 5%, 50%, 95% quantile summaries of the univariate posterior distributions for $\beta_0, \sigma_\epsilon^2, \lambda, \sigma_0^2$, and $\rho{=}\sigma_0^2/(\sigma_0^2{+}\sigma_\epsilon^2)$.

 (ii) Obtain a bivariate scatterplot representation of the posterior distribution $p(\sigma_\epsilon^2, \sigma_0^2 \mid \boldsymbol{y})$.

 (iii) Using samples from the distribution for ρ, answer the original question concerning the extent of batch-to-batch variability that is contributing to the total variability.

(e) Obtain the distribution of $p(b_i \mid \beta_0, \sigma_\epsilon^2, \sigma_0^2, \boldsymbol{y})$. Hence, describe an algorithm for simulating from the posterior $p(b_i \mid \boldsymbol{y})$. Implement this algorithm for the yield data, and give 5%, 50%, 95% quantile summaries for $p(b_i \mid \boldsymbol{y})$, $i = 1, \ldots, m$.

(f) Now, consider an alternative computational approach assuming independent priors with an improper flat prior on μ, the improper prior $\pi(\sigma_e^2) \propto \sigma_e^{-2}$, and a Ga$(0.05, 0.01)$ prior for σ_0^{-2}. Implement a Gibbs sampling algorithm for sampling from the conditionals:

- $\mu \mid \sigma_e^2, \sigma_0^2, \boldsymbol{b}, \boldsymbol{y}$
- $\sigma_e^2 \mid \mu, \sigma_0^2, \boldsymbol{b}, \boldsymbol{y}$
- $\sigma_0^2 \mid \mu, \sigma_e^2, \boldsymbol{b}, \boldsymbol{y}$
- $\boldsymbol{b} \mid \mu, \sigma_e^2, \sigma_0^2, \boldsymbol{y}$,

where $\boldsymbol{b} = [b_1, \ldots, b_m]^{\text{T}}$.

Report posterior medians and 90% credible intervals for $\mu, \sigma_0^2, \sigma_e^2, \boldsymbol{b}$, and ρ, and compare your answers with those using the alternative priors derived in the earlier part of the question.

8.10 Derive the conditional distributions, given in Sect. 8.6.3, that are required for Gibbs sampling in the LMM.

8.11 Show the equivalence of the BLUP predictor $\widehat{\boldsymbol{b}}_i$ and the Gibbs conditional distribution $\boldsymbol{b}_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}$.

8.12 Consider the tooth growth data that were analyzed in this chapter. These data are available in the `R` package `nlme` as `Orthodont`. Let Y_{ij} denote the growth (in mm) at occasion t_j (in years) for boy i, $i = 1, \ldots, m$, $j = 1, \ldots, 4$, with $t_1 = 8$, $t_2 = 10$, $t_3 = 12$, $t_4 = 14$.

(a) Code up a GEE algorithm with working independence in your favorite package, and report $\widehat{\boldsymbol{\beta}}$ and $\text{var}(\widehat{\boldsymbol{\beta}})$.

(b) Using an available option in a statistical package such as `R` confirm the results of the previous part.

(c) Show that $\text{var}(\boldsymbol{Y}) = (\boldsymbol{Y} - \boldsymbol{x\beta})^{\text{T}}(\boldsymbol{Y} - \boldsymbol{x\beta}) = \boldsymbol{0}$ if we attempt to use sandwich estimation in the situation in which $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) \neq \boldsymbol{0}$.

8.13 In this question the effect of using different correlation structures, designs, and sample sizes in the GEE approach will be examined. Let Y_{ij} represent the observed growth on individual i at time x_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, n_i$. Let $N = \sum_{i=1}^{m} n_i$.

Assume the marginal model is

$$\text{E}[Y_{ij}] = \beta_0 + \beta_1 x_{ij},$$

so that $\text{E}[\boldsymbol{Y}] = \boldsymbol{x\beta}$ where $\boldsymbol{Y}$ is of dimension $N \times 1$, $\boldsymbol{x}$ is $N \times 2$, and $\boldsymbol{\beta}$ is 2×1. Consider the estimating function

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}),$$

with working covariances $\boldsymbol{W}_i$ of dimension $n_i \times n_i$, $i = 1, .., m$.

Assume throughout that $\beta_0 = 18$, $\beta_1 = 0.5$, $\alpha_1 = 1$, and α_2 is set to either 0.5 or 0.9. Simulate data from the multivariate normal distribution $\boldsymbol{Y}_i \sim \text{N}_{n_i}(\boldsymbol{x}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$, with the form of $\boldsymbol{V}_i$ taken as either the exchangeable or the AR(1) matrices $\boldsymbol{W}_i$ that are given below, for $i = 1, \ldots, m$. Examine the efficiency of these working models as a function of:

- The number of individuals, with $m = 8, 20, 60$
- Two designs:
  - *Design I:* Balanced with $n_i = n = 4$, $i = 1, \ldots, m$ and $x_1 = 8$, $x_2 = 10$, $x_3 = 12$, $x_4 = 14$ for all individuals
  - *Design II:* Unbalanced with $n_i = n = 3$, $i = 1, \ldots, m$ and

$$x_{i1} = 8, \quad x_{i2} = 10, \quad x_{i3} = 12 \qquad \text{for } i = 1, \ldots, m/4$$

$$x_{i1} = 8, \quad x_{i2} = 10, \quad x_{i3} = 14, \qquad \text{for } i = m/4 + 1, \ldots, m/2$$

$$x_{i1} = 8, \quad x_{i3} = 12, \quad x_{i3} = 14, \qquad \text{for } i = m/2 + 1, \ldots, 3m/4$$

$$x_{i2} = 10, \quad x_{i3} = 12, \quad x_{i4} = 14, \qquad \text{for } i = 3m/4 + 1, \ldots, m$$

- The working covariance structure $\alpha_1 \boldsymbol{W}_i$ with:
  - *Independence:* $\boldsymbol{W}_i = \mathbf{I}_{n_i}$ where $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix.
  - *Exchangeable:* $\boldsymbol{W}_i$ has diagonal elements 1 and off-diagonal elements α_2.
  - *First-order autocorrelation:* $\boldsymbol{W}_i$ has diagonal elements 1 and off-diagonal elements $W_{ijk} = \alpha_2^{|x_{ij} - x_{ik}|}$, $j, k = 1, \ldots, n_i$, $j \neq k$, $i = 1, \ldots, m$.

In total, there are $3 \times 2 \times 4 = 24$ sets of simulations, and for each you should:

(a) Report the 95% confidence interval coverage for β_1.

(b) Report the standard errors and efficiencies. For each working covariance model, there are two standard error calculations; the "true" standard errors are obtained across simulations while $\overline{\text{var}(\widehat{\beta}_1)}$ describes the average (across simulations) of the reported squared standard error, where the latter is calculated using the sandwich formula. To evaluate the efficiencies, the (sandwich) variance of the estimators under each of the working models should be calculated.

8.14 Crowder and Hand (1990) describe data on the body weight of rats measured over 64 days. These data are available in the `R` package `nlme` and are named `BodyWeight`. Body weight is measured (in grams) on day 1, and every 7 days subsequently until day 64, with an extra measurement on day 44. There are 3 groups of rats, each on a different diet; 8 rats are on a control diet, and two sets of 4 rats are each on a different treatment.

(a) Fit LMMs to these data using ML/REML, with the primary aim being to determine whether there are differences in intercepts and slopes for each of the diets. Repeat this procedure using GEE.

(b) Carefully describe the models that you fit, in particular the choice of random effects structure in the LMM, and summarize your findings in simple terms.

(c) Now, analyze the first group of rats using a Bayesian analysis. Specifically, suppose Y_{ij} is the body weight of rat i at time t_j, and consider the three-stage model:
Stage One:

$$Y_{ij} = \beta_0 + b_i + \beta_1 t_j + \epsilon_{ij}$$

with $\epsilon_{ij} \mid \tau \sim_{iid} \mathrm{N}(0, \tau^{-1})$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

Stage Two: $b_i \mid \tau_0 \sim_{iid} \text{N}(0, \tau_0^{-1})$, with b_i independent of the ϵ_{ij}, $i{=}1, \dots, m, j = 1, \dots, n$.
Stage Three: Independent hyperpriors with:

$$\pi(\boldsymbol{\beta}) \propto 1,$$

$$\pi(\tau) \propto \tau^{-1},$$

$$\pi(\tau_0) \sim \text{Ga}(0.1, 0.5)$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\text{T}}$.

(d) Find the form of the conditional distributions that are required for constructing a Gibbs sampling algorithm to explore the posterior distribution $p(\boldsymbol{\beta}, \tau, b_1, \dots, b_m, \tau_0 \mid \boldsymbol{y})$:

- $p(\boldsymbol{\beta} \mid \tau, b_1, \dots, b_m, \tau_0, \boldsymbol{y})$.
- $p(\tau \mid \boldsymbol{\beta}, b_1, \dots, b_m, \tau_0, \boldsymbol{y})$.
- $p(\tau_0 \mid \boldsymbol{\beta}, \tau, b_1, \dots, b_m, \boldsymbol{y})$.
- $p(b_i \mid \boldsymbol{\beta}, \tau, b_j, j \neq i, \tau_0, \boldsymbol{y})$, $i = 1, \dots, m$.

(e) Implement this algorithm for the data on the 8 rats in the control group. Provide trace plots of selected parameters to provide evidence of convergence of the Markov chain. Report two sets of summaries, consisting of the 5%, 50%, 95% quantiles, from two chains started from different values.

(f) Check your answers using available software, such as `INLA` or `WinBUGS`.

8.15 Prove (8.58).