6 Linear Mixed Effects Models

6.1 Introduction

In the last chapter, we discussed a general class of ***linear models*** for ***continuous response*** arising from a ***population-averaged*** point of view. Here, ***population mean response*** is represented directly by a ***linear model*** that incorporates ***among-*** and ***within-individual*** covariate information. In keeping with the population-averaged perspective, the ***overall aggregate covariance matrix*** of a response vector is also ***modeled directly***. These models are appropriate when the ***questions of scientific interest*** are questions about features of ***population mean response profiles***.

As we observed, selecting among candidate covariance models to represent the overall covariance structure is an inherent challenge. The ***aggregate pattern*** of variance and correlation may be ***sufficiently complex*** that, for example, standard correlation models like those reviewed in Section 2.5 cannot faithfully represent it.

Moreover, when the number of observations per individual n_i differs across individual and/or the observations are at different time points for different individuals, simple ***exploratory*** approaches like those in Section 2.6 are not possible, and some correlation models may not be feasible. Also, care must be taken in implementation. Of course, as discussed in Section 5.6, the reasons for ***imbalance*** must be carefully considered from the point of view of ***missing data mechanisms***.

In this chapter, we instead take a ***subject-specific perspective*** , which leads to the so-called ***linear mixed effects model*** , the ***most popular*** framework for longitudinal data analysis in practice. Here, ***individual inherent response trajectories*** are represented by a ***linear model*** incorporating covariates, and, as in Chapter 2, ***within-*** and ***among-individual*** sources of correlation are ***explicitly acknowledged and modeled separately***. Following the conceptual point of view in Chapter 2, it is natural to acknowledge individual response profiles in this way, and many scientific questions can be interpreted as pertaining to the "***typical***" features of individual trajectories; e.g., the "typical slope."

As discussed in Section 2.4, because of the use of ***linear models*** , this approach ***implies*** a linear model for ***overall population mean response*** and ***induces*** a model for the ***overall aggregate covariance matrix*** , so that a ***linear population-averaged*** model is a byproduct. Thus, as we noted there, the linear mixed effects model is a relevant framework for addressing questions of ***either*** a subject-specific or population-averaged nature.

Moreover, as we observe shortly, the induced covariance structure ameliorates the problems associated with direct specification of the overall pattern and implementation with ***unbalanced*** data discussed in Section 5.2 when a population-averaged model is adopted directly and offers the analyst ***great flexibility*** for modeling variance and correlation structure.

It follows that the ***same*** methods, namely, ***maximum likelihood*** under the assumption of ***normality*** and ***REML***, can be used to fit a linear mixed effects model, and the same large sample theory results deduced in Section 5.5 hold and are used for the basis approximate inference. Likewise, the same concerns discussed in Section 5.6 regarding ***missing data*** continue to apply.

Unlike the population-averaged approach in Chapter 5, however, because the ***subject-specific*** perspective here represents explicitly ***individual behavior***, it is possible to characterize features of individual behavior and to develop an alternative approach to implementation via maximum likelihood, which we discuss later in this chapter.

6.2 Model specification

BASIC MODEL: For convenience, we restate that the observed data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

We introduce the basic form of the ***linear mixed effects model*** and then present examples that demonstrate how it provides a general framework in which various ***subject-specific*** models can be placed. The model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, m. \tag{6.1}$$

- In (6.1), $\boldsymbol{X}_i$ ($n_i \times p$) and $\boldsymbol{Z}_i$ ($n_i \times q$) are ***design matrices*** for individual i that depend on individual i's ***covariates*** $\boldsymbol{x}_i$ and time; we present examples of how $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ arise from a subject-specific perspective momentarily.
- The vector β ($p \times 1$) in (6.1) is referred to as the ***fixed effects*** parameter.

- $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of ***random effects*** characterizing ***among-individual*** behavior; i.e., where individual i "sits" in the population. The ***standard*** and most basic assumption is that the $\boldsymbol{b}_i$ are ***independent*** of the covariates $\boldsymbol{x}_i$ and satisfy, for $(q \times q)$ ***covariance matrix*** $\boldsymbol{D}$,

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i) = \boldsymbol{D}, \tag{6.2}$$

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}). \tag{6.3}$$

  As we demonstrate, $\boldsymbol{D}$ characterizes variance and correlation due to ***among-individual*** sources. The specifications (6.2) and (6.3) can be relaxed to allow the distribution to differ depending on the values of ***among-individual covariates*** $\boldsymbol{a}_i$, as we discuss shortly, so that

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i), \quad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{a}_i)\}. \tag{6.4}$$

- The ***within-individual deviation*** $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$ represents the ***aggregate effects*** of the ***within-individual realization*** and ***measurement error*** processes operating at the level of the individual. The ***standard*** and most basic assumption is that the $\boldsymbol{e}_i$ are ***independent*** of the random effects $\boldsymbol{b}_i$ and the covariates $\boldsymbol{x}_i$ and satisfy

$$E(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = E(\boldsymbol{e}_i) = \boldsymbol{0}, \quad var(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \text{var}(\boldsymbol{e}_i) = \boldsymbol{R}_i(\gamma). \tag{6.5}$$

  for some $(n_i \times n_i)$ covariance matrix $\boldsymbol{R}_i(\gamma)$ depending on parameters γ. The most common assumption, often adopted ***by default*** without adequate thought, is that

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_{n_i}, \quad \gamma = \sigma^2, \quad \text{for all } i = 1, \ldots, m; \tag{6.6}$$

  we discuss considerations for specification of $\boldsymbol{R}_i(\gamma)$ shortly. Ordinarily, it is further assumed that

$$\boldsymbol{e}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}. \tag{6.7}$$

  (6.5) and (6.7) can be relaxed to allow dependence of $\boldsymbol{e}_i$ on $\boldsymbol{x}_i$ and $\boldsymbol{b}_i$. We consider dependence of $\boldsymbol{e}_i$ on $\boldsymbol{a}_i$ here and defer discussion of more general specifications to Chapter 9.

INTERPRETATION: From the perspective of the ***conceptual model*** (2.9) in Section 2.3,

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \mathcal{B}_i + \boldsymbol{e}_i = \boldsymbol{\mu}_i + \mathcal{B}_i + \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi}, \tag{6.8}$$

inspection of (6.1) shows that we can identify $\boldsymbol{\mu}_i = \boldsymbol{X}_i\beta$ as the $(n_i \times 1)$ overall population mean response vector, $\mathcal{B}_i = \boldsymbol{Z}_i\boldsymbol{b}_i$ as the $(n_i \times 1)$ vector of ***deviations*** from the population mean characterizing where individual i "sits" in the population and thus ***among-individual*** variation, and $\boldsymbol{e}_i$ as the $(n_i \times 1)$ vector of ***within-individual*** deviations due to the realization process and measurement error.

Thus, in the linear mixed effects model (6.1),

$$\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

characterizes the ***individual-specific trajectory*** for individual i. As we demonstrate in examples shortly, this general form offers great latitude for representing individual profiles.

IMPLIED POPULATION-AVERAGED MODEL: It follows from (6.1) and (6.2) – (6.7) that, ***conditional on*** $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$, $\boldsymbol{Y}_i$ is n_i-***variate normal*** with mean vector $\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$ and covariance matrix $\boldsymbol{R}_i(\gamma)$; i.e.,

$$\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i(\gamma)\}.$$

Thus, this conditional distribution characterizes how response observations for individual i vary and covary about the inherent trajectory $\boldsymbol{X}_i\beta+\boldsymbol{Z}_i\boldsymbol{b}_i$ for i due to the ***realization process*** and ***measurement error***.

Letting $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)$ denote the corresponding normal density, and, from (6.3), letting $p(\boldsymbol{b}_i; \boldsymbol{D})$ be the q-variate normal density corresponding to (6.3), the density of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is then given by

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i, \tag{6.9}$$

which is easily shown (try it) to be the density of a n_i-variate normal with mean vector $\boldsymbol{X}_i\beta$ and covariance matrix

$$\boldsymbol{V}_i(\gamma, \boldsymbol{D}, \boldsymbol{x}_i) = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma), \quad \xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T, \tag{6.10}$$

where vech($\boldsymbol{D}$) is the vector of ***distinct*** elements of $\boldsymbol{D}$ (see Appendix A).

Summarizing, the linear mixed effects model framework above implies that

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \quad \boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta, \boldsymbol{V}(\xi, \boldsymbol{x}_i)\}, \quad i = 1, \ldots, m, \tag{6.11}$$

where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is defined in (6.10).

- As in Chapter 5, we will sometimes write $\boldsymbol{V}_i$ and $\boldsymbol{R}_i$ for ***brevity*** , suppressing dependence on parameters for brevity.
- As (6.11) shows, consistent with the discussion above and that in Section 2.4, the subject-specific linear mixed effects model implies a population-averaged model with ***overall population mean*** of the ***same form*** as in (5.4) and ***overall aggregate covariance matrix*** of the particular form (6.10).

- The specific form of the overall covariance matrix (6.10) is ***induced*** by specific choices of $\boldsymbol{R}_i(\gamma)$, reflecting the belief about the nature of the ***within-individual realization and measurement error processes***, and of the covariance matrix $\boldsymbol{D}$ of the random effects, which characterizes ***among-individual variability*** in individual trajectories $\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$.

- This development generalizes in the obvious way when the covariance matrix $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i)$ depends on among-individual covariates $\boldsymbol{a}_i$.

- From the point of view of the conceptual model (6.8), the overall covariance matrix (6.10) is, using the assumptions on independence above,

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \text{var}(\mathcal{B}_i|\boldsymbol{x}_i) + \text{var}(\boldsymbol{e}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma). \tag{6.12}$$

  The correspondence in (6.12) emphasizes that the first term represents the contribution to the induced model for the overall covariance pattern due to ***among-individual*** sources of variance and correlation, and the second term represents the contribution due to ***within-individual*** sources.

  Thus, the induced model allows the data analyst great latitude to think about and incorporate beliefs about these sources ***explicitly***.

MODEL SUMMARY: As in the case of the population-averaged model in Chapter 5, it is convenient to summarize the linear mixed effects model for all $i = 1, \ldots, m$ individuals as follows.

Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \quad (N \times 1), \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} \quad (mq \times 1), \quad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{pmatrix} \quad (N \times 1), \tag{6.13}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \quad (N \times p), \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_m \end{pmatrix} \quad (N \times mq), \tag{6.14}$$

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R}_m \end{pmatrix} \quad (N \times N), \quad \widetilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D} \end{pmatrix} \quad (mq \times mq). \tag{6.15}$$

In (6.13) – (6.15), we suppress dependence of $\boldsymbol{R}_i$ and thus $\boldsymbol{R}$ on γ for brevity.

Using (6.13) – (6.15), we can write the model ***succinctly*** as (verify)

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{6.16}$$

In the literature and most ***software documentation*** , the model is routinely written in the form (6.16).

We now consider several examples that highlight the features of the ***subject-specific linear mixed effects model*** and the considerations involved in model specification.

- As we demonstrated informally in Chapter 2, specification of the model is according to a ***two stage hierarchy*** in which we first represent the form of ***individual inherent trajectories*** in terms of ***individual-specific parameters*** and then "step back" and characterize how these individual-specific parameters ***vary among individuals*** in the population.
- The framework subsumes that of so-called ***random coefficient models***.

SPECIFICATION OF THE WITHIN-INDIVIDUAL COVARIANCE MATRIX $\boldsymbol{R}_i$: As noted above, the ***within-individual covariance matrix***

$$\boldsymbol{R}_i(\gamma) = \text{var}(\boldsymbol{e}_i|\boldsymbol{b}_i, \boldsymbol{x}_i)$$

represents the ***aggregate effects*** of the ***within-individual realization process*** and the ***measurement error process***. Following the conceptual representation in Chapter 2 as in (2.9), as in (6.8),

$$\boldsymbol{e}_i = \boldsymbol{e}_{Pi} + \boldsymbol{e}_{Mi},$$

where, as we noted in that chapter, we would expect that $\text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, the contribution to $\boldsymbol{R}_i$ due to ***measurement error*** , to be a ***diagonal matrix*** while $\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, the contribution due to the ***realization process*** , may well exhibit ***correlation*** due to the time-ordered nature of the data collection.

Thus, when considering specification of $\boldsymbol{R}_i(\gamma)$, it is fruitful to decompose it as, in obvious notation,

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P) + \boldsymbol{R}_{Mi}(\gamma_M), \qquad \gamma = (\gamma_P^T, \gamma_M^T)^T, \tag{6.17}$$

where $\boldsymbol{R}_{Pi}(\gamma_P)$ is the covariance model for $\text{var}(\boldsymbol{e}_{Pi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$, and $\boldsymbol{R}_{Mi}(\gamma_P)$ is the ***diagonal*** covariance model for $\text{var}(\boldsymbol{e}_{Mi}|\boldsymbol{b}_i, \boldsymbol{x}_i)$.

We now review the considerations involved from the perspective of the representation (6.17).

- First consider the common, often ***default*** specification

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_{n_i}$$

in (6.6). From the perspective of (6.17), this can be viewed as

$$\boldsymbol{R}_i(\gamma) = \sigma_P^2 \boldsymbol{I}_{n_i} + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \sigma^2 = \sigma_P^2 + \sigma_M^2. \tag{6.18}$$

Thus, this specification incorporates the belief that ***serial correlation*** associated with the realization process is ***negligible***, which might be a reasonable assumption if the observation times are ***sufficiently intermittent*** so that such correlation can reasonably be assumed to have "died out." Of course, this assumption should be critically examined.

From (6.18), the default specification also implies the belief noted above and in Chapter 2 that measurement errors are committed ***haphazardly*** and with variance that is ***the same*** regardless of the magnitude of the true realization of the response being measured. We discuss the practical relevance of this latter assumption in later chapters.

Thus, in (6.6),

$$\sigma^2 = \sigma_P^2 + \sigma_M^2$$

and represents variance due to the ***combined effects*** of the realization process and measurement error.

- In general, it is ***commonplace*** to make the assumption that measurement error, if it is thought to exist, occurs ***haphazardly*** with ***constant variance***, and to take

$$\boldsymbol{R}_{Mi}(\gamma_M) = \sigma_M^2 \boldsymbol{I}_{n_i}. \tag{6.19}$$

Thus, it is routine to write (6.17) without comment as

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P) + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \gamma = (\gamma_P^T, \sigma_M^2)^T. \tag{6.20}$$

In applications where the response is ascertained using a ***device*** or ***analytical procedure***, as in the dental study (distance), the hip replacement study (hæmatocrit), or ACTG 193A (CD4 count), it is natural to expect the observed responses to reflect a component of measurement error as in (6.19) and thus to contemplate a model of the form (6.20).

- In some settings, it may be plausible to assume that the response is ascertained ***without measurement error***. For example, in the age-related macular degeneration trial in Section 5.6, we considered the response ***visual acuity*** , which is a count of the number of letters a patient read correctly from a vision chart. Here, it is natural to believe that it is possible to obtain this count ***exactly*** , with no or negligible error.

  In such a situation, the representation of $\boldsymbol{R}_i(\gamma)$ in (6.17) and (6.20) simplifies to

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{R}_{Pi}(\gamma_P), \qquad \gamma = \gamma_P, \tag{6.21}$$

  so that the within-individual covariance matrix model ***reflects entirely*** variation and correlation due to the ***within-individual realization process***.

  Here, plausible models for $\boldsymbol{R}_i(\gamma)$ would naturally be of the form

$$\boldsymbol{R}_i(\gamma) = \boldsymbol{T}_i^{1/2}(\theta)\boldsymbol{\Gamma}_i(\alpha)\boldsymbol{T}_i^{1/2}(\theta), \qquad \gamma = (\theta^T, \alpha^T)^T, \tag{6.22}$$

  where $\boldsymbol{T}_i(\theta)$ is a ***diagonal matrix*** whose diagonal elements reflect the belief about the nature of the ***realization process variance***. For example, assuming that this variance is ***constant over time*** , so that

$$\boldsymbol{T}_i(\theta) = \sigma^2 \boldsymbol{I}_{n_i}, \qquad \theta = \sigma^2,$$

  (6.22) reduces to

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{\Gamma}_i(\alpha), \qquad \gamma = (\sigma^2, \alpha^T)^T, \tag{6.23}$$

  where now σ^2 is the ***assumed constant*** realization variance, and $\boldsymbol{\Gamma}_i(\alpha)$ is a $(n_i \times n_i)$ ***correlation matrix***.

  The specification (6.23) is often assumed by ***default*** , but it is prudent to consider the possibility that, if $n = \max_i(n_i)$ is the largest number of observations across all individuals, which would be the total number of ***intended times*** in a ***prospectively planned*** study, for individual i with n observations,

$$\boldsymbol{T}_i(\theta) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2),$$

  which allows realization variance to exhibit ***heterogeneity*** over time.

- It is commonplace for users who are not well-versed in the underpinnings of the linear mixed model to assume without comment either the default specification (6.6) or possibly (6.23), failing to appreciate the implications of the foregoing discussion and the need to ***distinguish*** the contributions of the realization and measurement error processes to the overall pattern of within-individual variance and correlation.

- Moreover, in much of the literature, these considerations are often not made explicit. When they are, the default specification is usually taken to be

$$\boldsymbol{R}_i(\gamma) = \sigma_P^2 \boldsymbol{\Gamma}_i(\boldsymbol{\alpha}) + \sigma_M^2 \boldsymbol{I}_{n_i}, \qquad \gamma = (\sigma_P^2, \boldsymbol{\alpha}^T, \sigma_M^2)^T. \tag{6.24}$$

EXAMPLE 1, DENTAL STUDY: We considered a subject-specific model for these data in Section 2.4, which we recast now in the context of the linear mixed effects model. Recall that there are no within-individual covariates and one among-individual covariate, gender, $g_i = 0$ if i is a girl and $g_i = 1$ if i is a boy, so that $\boldsymbol{x}_i$ contains g_i and the four time points $(t_1, \ldots, t_4) = (8, 10, 12, 14)$.

From a ***subject-specific*** perspective, the primary question of interest is whether or not the ***typical*** or ***average rate of change*** of dental distance for boys differs from that for girls. In (2.13), we adopted a model for the ***individual trajectory*** for any child that represents it as a ***straight line*** with child-specific intercept and slope, namely

$$Y_{ij} = \beta_{0i} + \beta_{0i} t_{ij} + e_{ij}, \quad i = 1, \ldots, n_i = n = 4, \tag{6.25}$$

so that the question involves the difference in the ***typical*** or ***average slope***.

Define the child-specific "***regression parameter***" for i's straight line trajectory in (6.25) as

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}.$$

We can then summarize (6.25) as

$$\boldsymbol{Y}_i = \boldsymbol{C}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad \boldsymbol{C}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}, \quad i = 1, \ldots, m, \tag{6.26}$$

where, because of the ***balance***, $\boldsymbol{C}_i$ is the ***same*** (4×2) matrix for all i.

As in (2.14), we allow individual-specific intercepts and slopes to vary about ***typical*** or mean values for ***each gender*** according to ***random effects*** with

$$\begin{aligned} \beta_{0i} &= \beta_{0,B} g_i + \beta_{0,G}(1 - g_i) + b_{0i}, \\ \beta_{1i} &= \beta_{1,B} g_i + \beta_{1,G}(1 - g_i) + b_{1i}. \end{aligned} \qquad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}. \tag{6.27}$$

REMARK: In the early longitudinal data literature, a model of the form (6.26) along with a representation for $\boldsymbol{\beta}_i$ as in (6.27) is referred to as a ***random coefficient model***.

We can write (6.27) concisely as (verify)

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \tag{6.28}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0,G} \\ \beta_{1,G} \\ \beta_{0,B} \\ \beta_{1,B} \end{pmatrix}, \quad \boldsymbol{A}_i = \begin{pmatrix} (1-g_i) & 0 & g_i & 0 \\ 0 & (1-g_i) & 0 & g_i \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_2.$$

Substituting (6.28) in (6.26) and rearranging, we have

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{C}_i\boldsymbol{B}_i\boldsymbol{b}_i + \boldsymbol{e}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{6.29}$$

where

$$\boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i, \quad \boldsymbol{Z}_i = \boldsymbol{C}_i\boldsymbol{B}_i.$$

Thus, it is straightforward to deduce that

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & (1-g_i)t_1 & g_i & g_it_1 \\ \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & (1-g_i)t_4 & g_i & g_it_4 \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}. \tag{6.30}$$

Here, $\boldsymbol{X}_i$ is the ***same*** as the design matrix (5.15) in the ***population-averaged model*** in Chapter 5.

To complete the specification , we posit models for ***among-individual covariance matrix*** var($\boldsymbol{b}_i|\boldsymbol{a}_i$) and the ***within-individual covariance matrix*** $\boldsymbol{R}_i(\gamma)$.

- In Section 2.6, empirical exploration of the ***overall aggregate pattern of covariance*** shows evidence that ***overall correlation*** is ***different*** for boys and girls with ***overall variance constant across time*** but possibly ***larger*** for boys than for girls.
- Examination of the ***within-individual residuals*** from ***individual-specific*** fits of model (6.25) to each child ***does not*** show strong evidence of ***within-individual correlation***; we showed this for boys, and the same observation applies to girls.

- Moreover, these residuals suggest for each gender that ***within-child variance*** due to the combined effects of realization and measurement error is ***constant*** over time. Estimates of within-child variance based on pooling the residuals across children of each gender are 2.59 for boys and 0.45 for girls; the ***much larger value for boys*** is likely due in part to the very large fluctuation of distance values within one boy.

- Combining these observations, it may be reasonable to assume that the ***within-child covariance matrix*** is of the general form (6.24) with the correlation matrix $\boldsymbol{\Gamma}_i(\alpha)$ approximately equal to an identity matrix as in (6.18), so that $\boldsymbol{R}_i(\gamma)$ for any child is ***diagonal***.

  However, because the estimates of ***within-child aggregate variance*** are so different, we might consider initially a form of (6.18) that is ***different*** for each gender. That is, relaxing (6.5), so that $\boldsymbol{e}_i$ and $\boldsymbol{a}_i$ are ***not necessarily independent***, a plausible model is, in obvious notation,

$$\begin{aligned} \text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma) &= \sigma^2_{PG}\boldsymbol{I}_4 + \sigma^2_{MG}\boldsymbol{I}_4 \quad \text{if } i \text{ is a girl,} \\ &= \sigma^2_{PB}\boldsymbol{I}_4 + \sigma^2_{MB}\boldsymbol{I}_4 \quad \text{if } i \text{ is a boy,} \end{aligned}$$

  say. This leads to the final specification

$$\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{a}_i) = \{\sigma^2_G I(g_i = 0) + \sigma^2_B I(g_i = 1)\}\boldsymbol{I}_4, \tag{6.31}$$

  where now $\sigma^2_G = \sigma^2_{PG} + \sigma^2_{MG}$ and $\sigma^2_B = \sigma^2_{PB} + \sigma^2_{MG}$ in (6.31) represent ***within-child variance*** due to ***both*** the realization and measurement error processes (rather than overall variance as in Section 5.2).

  If the much larger estimated within-child variance for boys is mainly an ***artifact*** of the unusual pattern for one boy, an alternative model is the ***default*** (6.6), $\boldsymbol{R}_i(\gamma) = \sigma^2\boldsymbol{I}_4$. Here, one might want to examine sensitivity of fitted models to the data from the "unusual" boy by, for example, deleting him from the analysis.

- Because there is not strong evidence of within-child correlation, it is natural to attribute the overall pattern of correlation mainly to ***among-child sources***. We can examine the ***induced representation*** of the component of overall covariance structure due to among-child sources as follows. For illustration, take for each i

$$\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix}.$$

It is then straightforward to show that (try it), with $\boldsymbol{Z}_i$ as in (6.30), $\boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T$ has ***diagonal elements***

$$D_{11} + D_{22} t_j^2 + 2 D_{12} t_j, \quad j = 1, \ldots, 4, \tag{6.32}$$

and (j, j') off-diagonal element

$$D_{11} + D_{22} t_j t_{j'} + D_{12}(t_j + t_{j'}) \quad j, j' = 1, \ldots, 4. \tag{6.33}$$

(6.32) shows that this component of the induced overall covariance structure allows for among-individual variance that possibly changes with time, and (6.33) imposes a rather complicated pattern of ***among-individual covariance and correlation*** that is clearly ***nonstationary***. Thus, this component of the model is sufficiently flexible to capture complex covariance patterns.

Because the evidence is suggestive of an overall pattern that may be ***different by gender***, one possibility is to take $\text{var}(\boldsymbol{b}_i | \boldsymbol{x}_i)$ to depend on $\boldsymbol{a}_i$ (gender) as in (6.4) and

$$\text{var}(\boldsymbol{b}_i | \boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i) = \boldsymbol{D}_G I(g_i = 0) + \boldsymbol{D}_B I(g_i = 1). \tag{6.34}$$

However, it is hard to judge to what extent the empirical evidence reflects a ***real difference***. The simpler common model $\text{var}(\boldsymbol{b}_i | \boldsymbol{a}_i) = \boldsymbol{D}$ may well be sufficient.

HIP REPLACEMENT STUDY: Recall from Section 5.2 that, for $m = 30$ subjects undergoing hip replacement (13 male, 15 female), hæmatocrit was measured at week 0, prior to surgery, and then ideally at weeks 1 2, and 3 thereafter, where some subjects are ***missing*** the week 2 and possibly baseline measure. Also available is patient age, so that $\boldsymbol{a}_i = (g_i, a_i)^T$, where gender $g_i = 0$ for females and $g_i = 1$ for males; and a_i is the age of the patient (years).

We can interpret the primary question of interest from a SS perspective to determine if there are differences between genders in ***individual-specific features*** of the pattern of change of hæmatocrit following hip replacement. As we demonstrate, we can also investigate associations between these features and age.

Taking this point of view, from Figure 5.2, a natural model for the individual subject trajectories is

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}, \tag{6.35}$$

which allows each subject to have his/her own specific ***quadratic*** profile. The model (6.35) can be written succinctly as

$$\boldsymbol{Y}_i = \boldsymbol{C}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{2i} \end{pmatrix}, \quad \boldsymbol{C}_i = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{pmatrix}. \tag{6.36}$$

As for the dental study, we can allow individual-specific intercepts, linear terms, and quadratic terms to vary about ***typical*** or mean values for ***each gender***, and we can further allow ***typical*** or mean hæmatocrit at ***baseline*** to depend on ***age*** through the model specification

$$\begin{aligned} \beta_{0i} &= \{\beta_{0,M}(1-g_i) + \beta_{0,F}g_i\} + \{\beta_{3,M}(1-g_i) + \beta_{3,F}g_i\}a_i + b_{0i} \\ \beta_{1i} &= \beta_{1,M}(1-g_i) + \beta_{1,F}g_i + b_{1i} \\ \beta_{2i} &= \beta_{2,M}(1-g_i) + \beta_{2,F}g_i + b_{2i}, \end{aligned} \tag{6.37}$$

where $\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i})^T$ is a vector of ***random effects***. The models for the individual-specific linear (β_{1i}) and quadratic (β_{2i}) terms could be modified to also depend on age. Letting

$$\boldsymbol{\beta}_i = (\beta_{0i}\,,\ \beta_{1i}\,,\ \beta_{2i})^T,$$

the model (6.37) can be represented as

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i,$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0,M} \\ \beta_{0,F} \\ \beta_{1,M} \\ \beta_{1,F} \\ \beta_{2,M} \\ \beta_{2,F} \\ \beta_{3,M} \\ \beta_{3,F} \end{pmatrix}, \quad \boldsymbol{A}_i = \begin{pmatrix} (1-g_i) & g_i & 0 & 0 & 0 & 0 & (1-g_i)a_i & g_i a_i \\ 0 & 0 & (1-g_i) & g_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & (1-g_i) & g_i & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_3. \tag{6.38}$$

Upon substitution into (6.36), we have (verify) that $\boldsymbol{Z}_i = \boldsymbol{C}_i$ and $\boldsymbol{X}_i$ is the $(n_i \times 8)$ matrix

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & g_i & (1-g_i)t_{i1} & g_i t_{i1} & (1-g_i)t_{i1}^2 & g_i t_{i1}^2 & (1-g_i)a_i & g_i a_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & g_i & (1-g_i)t_{in_i} & g_i t_{in_i} & (1-g_i)t_{in_i}^2 & g_i t_{in_i}^2 & (1-g_i)a_i & g_i a_i \end{pmatrix}.$$

Specification of the ***within-individual*** covariance matrix $\boldsymbol{R}_i(\gamma)$ and the covariance matrix $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i)$ proceeds according to the same considerations as above.

Assuming for illustration that we take $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}$ for all i, $\boldsymbol{D}$ is a (3×3) matrix, and the component $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$ of $\boldsymbol{V}_i$ corresponding to ***among-individual sources*** is a $(n_i \times n_i)$ matrix whose elements have a rather ***complicated*** form (try it), depending on the six distinct elements of $\boldsymbol{D}$ as well as functions of time.

In many applications that, although ***in principle*** we expect that ***all of*** individual-specific intercepts, linear terms, and quadratic terms ***vary*** in the population, practically speaking, the ***induced overall covariance model*** $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma)$ depends on a rather large vector ξ of covariance parameters. Accordingly, the induced overall covariance structure is ***highly parameterized*** and is capable of representing complex true patterns of overall variance and correlation.

It may well be that, even though quadratic terms β_{2i} ***do vary*** in the population, ***relative*** to the extent of variation in intercepts and linear terms β_{0i} and β_{1i}, this variation is ***practically negligible***. Accordingly, it is not uncommon under ***quadratic*** and ***higher-order polynomial*** individual-specific models to ***simplify*** the model for β_i by ***eliminating*** random effects associated with quadratic and higher terms. This entails redefining $\boldsymbol{Z}_i$ and $\boldsymbol{D}$ accordingly.

The resulting $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$ may still be ***sufficiently rich*** to approximate the true component of among-individual covariance, and the induced overall structure may still be sufficiently parametrized to capture the true overall pattern. In addition, from a ***computational perspective***, the resulting model is likely to be less burdensome and problematic to fit; see below.

We demonstrate by ***eliminating*** the random effect b_{2i} in the specification for β_{2i} in (6.37), replacing it by

$$\beta_{2i} = \beta_{2,M}(1 - g_i) + \beta_{2,F}g_i. \tag{6.39}$$

Strictly speaking, (6.39) implies that the quadratic term in (6.35) is the ***same*** for all males and for all females. While this is likely an oversimplification, as an approximation it enjoys the advantages noted above. Under (6.39),

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \text{so that} \quad \boldsymbol{B}_i\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ 0 \end{pmatrix}. \tag{6.40}$$

Of course, from a ***subject-specific*** perspective, this is strictly an approximation of ***convenience***, as we most certainly do not really believe that individuals of each gender have ***individual-specific trajectories*** characterized by ***exactly the same*** quadratic component.

RELATIVE MAGNITUDES OF AMONG-INDIVIDUAL VARIATION: This foregoing demonstration with the hip replacement study scenario exemplifies an important ***general consideration*** when specifying linear mixed effects models. Although, ***conceptually***, from a SS point of view, all individual-specific parameters are expected to exhibit variation in the population, it is their ***relative magnitudes of variation*** that are practically important.

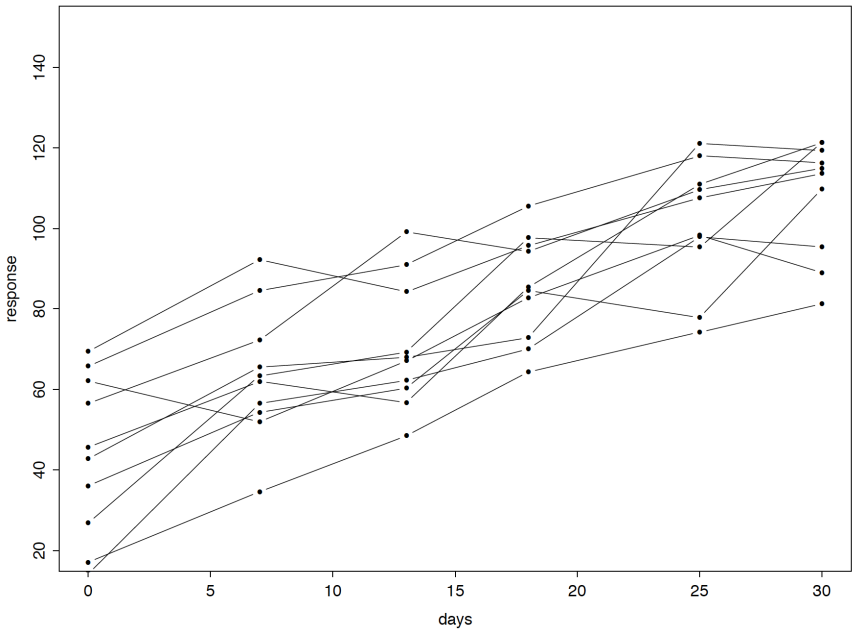

Figure 6.1: *Longitudinal data where variation in slope may be negligible.*

Consider the situation in Figure 6.1, depicting trajectories for 10 individuals for which a ***straight line inherent trend*** is a reasonable characterization. The ***individual-specific intercepts*** clearly vary substantially, but the assumed underlying lines appear to have ***very similar slopes***. Although scientifically it is reasonable to expect that individual rates of change ***should vary*** , e.g., as would be expected with patterns of growth across individual subjects or plots, ***relative*** to the variation in intercepts, the variation in slopes may well be ***orders of magnitude*** smaller.

For simplicity, assume there are no covariates. Letting β_{0i} and β_{1i} be the intercept and slope for individual i, if we assume

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1 + b_{1i},$$

$\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$, if we take var($\boldsymbol{b}_i$) = $\boldsymbol{D}$, D_{11} represents the variance of intercepts and D_{22} that of slopes. If D_{11} is ***nonnegligible*** relative to the mean intercept β_0, then intercepts vary perceptibly, but if D_{22} is ***virtually negligible*** relative to the size of the mean slope β_1, then variation in slopes is almost undetectable.

In such a situation, optimization algorithms involved in the implementation of inference by ML or REML, as discussed in the next section, can fail, as D_{22} and in fact the covariance D_{12} are not ***practically identifiable*** under these circumstances.

It is commonplace under these conditions to invoke an ***approximation*** analogous to that in (6.39) to achieve numerical stability, namely,

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1. \tag{6.41}$$

This does not mean that we "***believe***" slopes do not vary ***at all*** in the population; rather, this is an ***approximation*** recognizing that their magnitude of variation is inconsequential ***relative*** to that of other phenomena, which allows implementation of the model to be feasible. The inclusion of the design matrix $\boldsymbol{B}_i$ in the general model specification accommodates this possibility.

In a model like (6.41), it is popular to distinguish between individual-specific features being "***fixed***" or "***random***;" in (6.41), β_{0i} would be said to be "random" while β_{1i} would be referred to as "fixed."

In Section 6.6, we discuss this and related issues further.

HIV CLINICAL TRIAL: It should be clear that the ***hierarchical framework*** of the model offers great latitude for thinking about and representing individual-specific and population-level phenomena. As a final brief example, consider ACTG Study 193A, introduced in Section 5.2. Here, subjects were ***randomized*** to four treatment regimens, with age and gender recorded at baseline, so that $\boldsymbol{a}_i = (g_i, a_i, \delta_{i1}, \ldots, \delta_{i4})^T$, where $g_i = 0$ (1) for a female (male) subject; a_i is age; and $\delta_{i\ell} = 1$ if subject i was randomized to treatment regimen ℓ and 0 otherwise, $\ell = 1, \ldots, 4$.

From Figure 5.3, a reasonable approximation is to assume that each subject has his/her own inherent underlying ***straight line*** log(CD4+1) trajectory,

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

where now β_{0i} represents individual i's inherent mean log(CD4+1) immediately prior to initiation of therapy. Although we thus would not expect the β_{0i} to be associated with randomized treatment, they may be associated with individual characteristics such as gender and age.

If interest focuses on comparing the ***patterns of change*** of log(CD4 +1) among the four regiimens, from a SS point of view, this can be cast as comparing the ***typical*** or mean slopes under the four regimens. A model that incorporates baseline associations with covariates and allows typical slopes to differ across treatments is

$$\begin{aligned} \beta_{0i} &= \beta_{00} + \beta_{01} a_i + \beta_{02} g_i + b_{0i}, \\ \beta_{1i} &= \beta_{10} + \beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3} + b_{1i}. \end{aligned}$$

This model could be further modified to allow way in which slopes differ across treatments to be different for each gender or to depend on age.

HIERARCHICAL MODEL SUMMARY: The ***linear mixed effects*** model is often presented formally as a ***two-stage hierarchy*** as follows. In its usual general form , for each $i = 1, \ldots, m$,

Stage 1 - Individual model.

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\beta_i + \boldsymbol{e}_i \quad (n_i \times 1), \qquad \boldsymbol{e}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}, \tag{6.42}$$

where $\boldsymbol{C}_i$ is a $(n_i \times k)$ ***design matrix*** ordinarily depending on the ***times*** $t_{i1}, \ldots, t_{in_i}$, and β_i is a $(k \times 1)$ vector of ***individual-specific regression parameters***. The regression parameter β_i can be viewed as determining individual i's ***inherent trajectory***.

The default is that $\boldsymbol{e}_i$ is independent of $\boldsymbol{x}_i$ and β_i and thus $\boldsymbol{b}_i$, , although, as we have observed, (6.42) is often generalized to allow dependence on $\boldsymbol{a}_i$, so that $\boldsymbol{R}_i(\gamma)$ depends on $\boldsymbol{a}_i$. In more general versions of this hierarchy discussed in Chapter 9, dependence on β_i and thus $\boldsymbol{b}_i$ is also allowed.

In addition, $\boldsymbol{R}_i(\gamma)$ can be decomposed into components due to the ***within-individual realization*** and ***measurement error processes***, as in (6.17.

Stage 2- Population model.

$$\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i \quad (k \times 1), \qquad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \quad (q \times 1), \tag{6.43}$$

where β $(p \times 1)$ is a vector of ***fixed effects***; $\boldsymbol{A}_i$ $(k \times p)$ and $\boldsymbol{B}_i$ $(k \times q)$ are ***design matrices***; and $k = q$ in many cases, although models with $k > q$ are sometimes specified when some components of β_i are thought to vary ***negligibly*** among individuals. Typically, $\boldsymbol{A}_i$ incorporates ***among-individual covariates*** , while $\boldsymbol{B}_i$ is comprised of 0s and 1s and serves to indicate which elements of β_i are treated as "***random***" and which are treated as "***fixed***."

The default is that $\boldsymbol{b}_i$ and $\boldsymbol{x}_i$ are independent, but, as we have seen, this can be relaxed to allow dependence on $\boldsymbol{a}_i$, so that $\boldsymbol{D}(\boldsymbol{a}_i)$ depends on $\boldsymbol{a}_i$.

Substituting the ***population model*** (6.43) in the ***individual model*** (6.42) yields the ***linear mixed effects model***

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \boldsymbol{b}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{e}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{R}_i(\gamma)\}, \tag{6.44}$$

where $\boldsymbol{X}_i = \boldsymbol{C}_i\boldsymbol{A}_i$ $(n_i \times p)$ is the ***fixed effects design matrix*** , $\boldsymbol{Z}_i = \boldsymbol{C}_i\boldsymbol{B}_i$ $(n_i \times q)$ is the ***random effects design matrix*** , and the usual assumptions on the conditional distributions of $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ can be relaxed if need be.

6.3 Inference and considerations for missing data

IMPLIED POPULATION-AVERAGED MODEL: As shown in (6.10) and (6.11), given a particular specification of the ***two-stage hierarchy*** in (6.42) and (6.43) leading to a ***linear mixed effects model*** as in (6.44), we are led to a ***population-averaged*** model of the form

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \quad \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i(\gamma), \quad \xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T, \tag{6.45}$$

where $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are ***independent***. The model (6.45) is of course of the same form as the models considered in Chapter 5, where the covariance matrix $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is of the ***particular form*** in (6.45). The model (6.45) can be expressed succinctly as in (6.16) as

$$E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{6.46}$$

The specifications in (6.45) and (6.46) can of course be ***generalized*** to allow a more general ***among-individual covariance matrix*** of the form $\text{var}(\boldsymbol{b}_i|\boldsymbol{z}_i) = \boldsymbol{D}(\boldsymbol{a}_i)$.

ESTIMATION OF β AND ξ: From (6.45) and (6.46), it should be clear that, under the ***normality*** assumptions at each stage of the hierarchy (6.42) and (6.43), it follows that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is assumed to be n_i-variate normal, i.e.,

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta, \boldsymbol{V}(\xi, \boldsymbol{x}_i)\}.$$

It follows that estimators for β and ξ can be obtained by appealing to the developments in Sections 5.3 and 5.4. That is, β and ξ can be estimated by solving the estimating equations corresponding to ***maximum likelihood*** or ***REML***.

LARGE SAMPLE INFERENCE: Moreover, the ***large sample*** results in Section 5.5 go through unchanged. Thus, the ***approximate*** sampling distributions for the estimator $\widehat{\beta}$ obtained using either ML or REML can be used as described in that section. Namely, the ***model-based*** result in (5.68),

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \qquad \widehat{\boldsymbol{\Sigma}}_M = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right)^{-1} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}, \tag{6.47}$$

can be used as the basis for inference on β.

Likewise, the ***robust*** or ***empirical*** result in (5.81) and (5.82),

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_R), \tag{6.48}$$

$$\widehat{\boldsymbol{\Sigma}}_R = \tag{6.49}$$
$$\left\{\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(Y_i - \boldsymbol{X}_i\widehat{\beta})(Y_i - \boldsymbol{X}_i\widehat{\beta})^T \boldsymbol{V}^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i \left\{\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}.$$

can also be used. Both (6.47) and (6.48) and (6.49) can be used for inference on ***linear*** functions $\boldsymbol{L}\beta$ as in (5.85). This inference can be from a SS or PA perspective in accordance with the scientific questions. The analyst should be careful to be clear about this.

As with the models in Chapter 5, the true distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ need not be ***normal*** for these approximations to be valid (except when there are missing data; see below).

INFORMATION CRITERIA: The ***information criteria*** (5.90) – (5.92) discussed in Section 5.5 can also be used to compare models that are not nested and in particular to compare different specifications of the ***overall covariance structure*** that are ***induced*** by combinations of choices of models for, say var($\boldsymbol{e}_i|\boldsymbol{x}_i$) and var($\boldsymbol{b}_i|\boldsymbol{a}_i$).

- For example, for the dental study, one could compare the specifications of a ***common among-individual*** covariance matrix, var($\boldsymbol{b}_i|\boldsymbol{a}_i$) = var($\boldsymbol{b}_i$) = $\boldsymbol{D}$ to taking

  $$\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}(\boldsymbol{a}_i) = \boldsymbol{D}_G I(g_i = 0) + \boldsymbol{D}_B I(g_i = 1)$$

  as in (6.34).

- Likewise, one could compare taking var($\boldsymbol{e}_i|\boldsymbol{x}_i$) = $\sigma^2\boldsymbol{I}_4$ for all children versus allowing a separate ***within-child variance*** for each gender,

  $$\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma, \boldsymbol{a}_i) = \{\sigma_G^2 I(g_i = 0) + \sigma_B^2 I(g_i = 1)\}\boldsymbol{I}_4$$

  as in (6.31).

MISSING DATA: The ***same*** implications of ***missing data*** discussed in Section 5.6 apply to the linear mixed effects model. In particular, ***under the assumptions of a MAR mechanism*** and ***normality*** , the estimators for β and ξ are ***consistent*** , and the large sample approximation to the sampling distribution of $\widehat{\beta}$ as in (6.47) can be used, but with, ideally, $\widehat{\boldsymbol{\Sigma}}_M$ replaced by the appropriate element of the inverse of the ***observed information matrix***. The approximation in (6.48) and (6.49) should ***not*** be used, as discussed in Section 5.6.

BALANCED DATA: There is an interesting curiosity in the case of ***balanced*** data, so that $\boldsymbol{Y}_i$ is $(n \times 1)$ for all $i = 1, \ldots, m$, with components observed at the ***same*** n time points. In this case, $\boldsymbol{Z}_i = \boldsymbol{Z}^*$, say, is the same for all i (verify). If the linear mixed effects model specification is such that

$$\boldsymbol{R}_i(\gamma) = \sigma^2 \boldsymbol{I}_n,$$

the induced overall covariance matrix for each i is

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{V}^* = \boldsymbol{Z}^* \boldsymbol{D} \boldsymbol{Z}^{*T} + \sigma^2 \boldsymbol{I}_n, \tag{6.50}$$

say. Then, under certain conditions, letting

$$\widehat{\boldsymbol{V}}^* = \boldsymbol{Z}^* \widehat{\boldsymbol{D}} \boldsymbol{Z}^{*T} + \widehat{\sigma}^2 \boldsymbol{I}_n,$$

where $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}^2$ are the estimators for $\boldsymbol{D}$ and σ^2 obtained by ML or REML, the estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}^{*-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}^{*-1} \boldsymbol{Y}_i$$

and the ***ordinary least squares*** estimator

$$\widehat{\beta}_{OLS} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{Y}_i$$

are ***numerically identical***.

- This follows because it can be shown by cleverly applying ***matrix inversion results*** given in Appendix A that, with overall covariance structure $\boldsymbol{V}$ as in (6.50), the expressions

  $$\left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{*-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{*-1} \boldsymbol{Y}_i \quad \text{and} \quad \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{Y}_i$$

  are ***equivalent***.

- This continues to hold even if σ^2 and $\boldsymbol{D}$ in (6.50) take on different values corresponding to different levels of an ***among-individual covariate***, as for the dental study, where these are taken to ***differ by gender***.

- Demonstration of this equivalence is left as an exercise for the ***diligent student***.

- Note that, although the ML/REML estimator and the OLS estimator are numerically equivalent, this does ***not*** mean that one can ***disregard*** the need to characterize covariance structure and just take all N observations to be ***mutually independent***.

  Correct characterization of the ***sampling distribution*** of the estimator requires that the overall covariance be ***acknowledged and modeled***, and the large sample approximate sampling distribution depends on this assumed structure.

POPULATION-AVERAGED VERSUS SUBJECT-SPECIFIC PERSPECTIVE: As we have observed, the linear mixed effects model can be viewed in different ways.

- We motivated the model from a ***subject-specific perspective***, which dovetails naturally with the ***conceptual framework*** for longitudinal data we introduced in Section 2.3. This perspective underlies the view of the model as a ***two-stage hierarchy***, as presented in Section 6.2, which involves an ***individual-level model*** expressed in terms of ***individual-specific regression parameters*** and a ***population-level model*** that characterizes how these parameters ***vary*** in the population of individuals due to (i) ***systematic associations*** with among-individual covariates and (ii) "***unexplained***" or "***natural***" sources, such as biological differences or unobserved covariates.

  This view is natural when the questions of scientific interest involve ***subject-specific phenomena***.

- This formulation also ***implies*** a ***population-averaged model***, where the form of the ***overall covariance structure*** incorporating components due to ***among-*** and ***within-individual sources*** is ***induced***. Thus, an alternative perspective on the linear mixed effects model is as a population-averaged model for which specification of a form for the overall covariance structure is facilitated "automatically" rather than chosen explicitly by the data analyst. This relieves the analyst from the often ***challenging task*** of specifying a suitable overall structure. Moreover, the induced form for the overall covariance structure dictated by the linear mixed model is ***sufficiently rich***, involving a number of parameters, that it is likely able to represent well very ***complicated, nonstationary*** patterns of overall variance and correlation, as exemplified by (6.32) and (6.33).

  Thus, it is common to adopt a linear mixed effects model even when the questions of scientific interest involve ***population-averaged phenomena***.

- As we have already emphasized, the ***fixed effects*** β and questions posed in terms of them can be interpreted from either perspective.

- However, the perspective under which the model is adopted has ***implications for inference***, in particular in regard to the interpretation and fitting of the ***overall covariance structure***. Clearly, from either perspective, we desire a model that captures the ***salient features*** of covariance so that inferences on β will be reliable. At the same time, the model should not involve more parameters to be estimated than necessary, which in finite samples can ***degrade precision of estimation*** of β (despite the optimistic first-order asymptotic theory).

- As noted above, from a ***population-averaged*** perspective, the ***induced*** form of the overall covariance structure is a convenient and flexible way of represented what might possibly be a complex structure. From this point of view, $\xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ is simply a vector of parameters that characterizes the structure, and thus there are ***no restrictions*** on possible values of ξ. In particular, $\boldsymbol{D}$ need not be restricted to be a legitimate covariance matrix, with non-negative diagonal elements. Likewise, γ need not be restricted to take on values that render $\boldsymbol{R}_i(\gamma)$ a legitimate covariance matrix. What matters is that the parameterization in terms of ξ can represent a legitimate ***overall covariance structure***.

- From a subject-specific perspective, however, the separate components $\boldsymbol{D}$ and $\boldsymbol{R}_i(\gamma)$ ***are interpreted*** as covariance matrices corresponding to ***among-*** and ***within-individual*** sources of variation and correlation. Thus, from this point of view, there ***are restrictions*** on the parameter space of $\xi = \{\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$ that ensure that these are legitimate covariance matrices, that is, positive (semi-) definite matrices. Thus, for example, the diagonal elements of $\boldsymbol{D}$ are restricted to be nonnegative.

- Accordingly, which perspective is relevant ***will dictate*** how assessment of and inference on the assumed covariance structure takes place. We discuss this in more detail in Section 6.6.

6.4 Best linear unbiased prediction and empirical Bayes

RANDOM EFFECTS: Ordinarily, the primary objective of an analysis is to address questions of scientific interest expressed in terms of the ***fixed effects*** β, which may have ***either*** a population-averaged or subject-specific interpretation.

When a ***subject-specific*** perspective is adopted, the ***two-stage hierarchical interpretation*** of the linear mixed effects model reflects the belief that each individual has specific regression parameters β_i characterizing his/her ***inherent trajectory***. The β_i are then represented in the population model as depending on individual-specific ***random effects*** $\boldsymbol{b}_i$ that reflect how *i*'s regression parameters deviate from the "***typical***" values and likewise how *i*'s inherent trajectory deviates from the overall population mean profile. The $\boldsymbol{b}_i$ are ***random vectors*** assumed to arise from a probability distribution(s) that characterizes the extent of variation in these features in the population.

In the ***standard version*** of the linear mixed effects model we discuss in this chapter, the distribution of the $\boldsymbol{b}_i$ is taken to be ***q-variate normal***, with mean zero with covariance matrix $\boldsymbol{D}$ (which of course can be relaxed to allow ***separate*** distributions for each level of an among-individual covariate). For the discussion here, we take

$$\boldsymbol{b}_i|\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad i = 1, \ldots, m; \tag{6.51}$$

the developments below of course can be generalized.

Thus, from a subject-specific point of view, it is often of interest to "***estimate***" $\boldsymbol{b}_i$ for each individual. These estimates can be used for ***diagnostic purposes***, e.g., to identify individuals or groups of individuals whose profiles over time may be ***outlying*** relative to the bulk of the population. They can also be used to characterize ***individual-specific trajectories***.

Because the $\boldsymbol{b}_i$ are random vectors, each corresponding to a ***randomly chosen individual*** from the population, characterizing $\boldsymbol{b}_i$ is akin to ***predicting*** the value taken on by a random vector corresponding to a randomly chosen individual. Thus, inference on $\boldsymbol{b}_i$ is often regarded as a ***prediction problem***. Because $\boldsymbol{Y}_i$ contains information about $\boldsymbol{b}_i$, it is natural to view this prediction problem as characterizing $\boldsymbol{b}_i$ ***given*** that we have observed $\boldsymbol{Y}_i = \boldsymbol{y}_i$. The usual approach is to use as a predictor the value that is "***most likely***" given that we have observed $\boldsymbol{Y}_i = \boldsymbol{y}_i$.

BAYESIAN PERSPECTIVE: It is thus natural to consider this problem based on a ***Bayesian*** formulation and to "***estimate***" $\boldsymbol{b}_i$ by the value that ***maximizes*** the ***posterior distribution*** of $\boldsymbol{b}_i$ given $\boldsymbol{Y}_i$ evaluated at $\boldsymbol{y}_i$; that is, finding the ***posterior mode***.

- In the Bayesian view of the linear mixed effect model, the $\boldsymbol{b}_i$ are regarded as ***parameters***, and the probability distribution (6.51) is referred to as the ***prior distribution*** for them.
- For the discussion here, we do not consider the parameters β and ξ from the classical Bayesian perspective as random quantities with suitable prior distributions, but treat them as ***fixed and known***; more on this momentarily.

Taking this point of view, let as in (6.9)

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma) \tag{6.52}$$

be the density of the assumed conditional normal distribution

$$\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i \sim \mathcal{N}\{\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i(\gamma)\}.$$

Let

$$p(\boldsymbol{b}_i; \boldsymbol{D})$$

be the density corresponding to (6.51). Then, identifying this as the "***prior***" and (6.52) as the "***likelihood***," by Bayes' theorem, the ***posterior density*** of $\boldsymbol{b}_i$ conditional on observing $\boldsymbol{Y}_i = \boldsymbol{y}_i$ is given by

$$p(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})}{p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D})}, \tag{6.53}$$

where, from (6.9),

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta, \gamma, \boldsymbol{D}) = \int p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{b}_i; \beta, \gamma)\, p(\boldsymbol{b}_i; \boldsymbol{D})\, d\boldsymbol{b}_i.$$

It is straightforward to verify (do it) that, under the normal specifications (6.51) and (6.52), the ***posterior distribution*** with density (6.53) is ***also normal*** with ***mean***

$$\boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\boldsymbol{y}_i - \boldsymbol{X}_i\beta\}. \tag{6.54}$$

Because the ***mean*** of a normal distribution is also the ***mode*** of the density, the expression (6.54) also satisfies the requirement that it ***maximizes*** the posterior density.

EMPIRICAL BAYES: From (6.54), it is natural to substitute estimators $\widehat{\beta}$ and $\widehat{\xi}$ for β and ξ, which yields the so-called ***empirical Bayes*** "***estimator***" for $\boldsymbol{b}_i$ given by

$$\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{D}}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}\}, \tag{6.55}$$

where we have written (6.55) as depending on the response vector $\boldsymbol{Y}_i$ with the understanding that the actual observed value of $\boldsymbol{Y}_i$ is substituted in forming the "***estimate***."

If ξ were ***known***, so that (6.55) becomes

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}\}, \tag{6.56}$$

it is straightforward (try it) to show that (conditional on $\tilde{\boldsymbol{x}}$) $\widehat{\boldsymbol{b}}_i$ in (6.56) has mean zero and covariance matrix

$$\text{var}(\widehat{\boldsymbol{b}}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \right\} \boldsymbol{Z}_i \boldsymbol{D}, \tag{6.57}$$

where we have used the streamlined notation for $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$.

Because what we really are doing is ***prediction*** of the "***moving target***" $\boldsymbol{b}_i$, which is a random rather than fixed quantity, (6.57) is known to ***understate*** the variability in $\widehat{\boldsymbol{b}}_i$.

Accordingly, it is recommended to instead use

$$\text{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}^{-1} \right\} \boldsymbol{Z}_i \boldsymbol{D} \tag{6.58}$$

(verify). Of course, in practice, ξ is replaced by its estimator (ML or REML), in which case (6.57) and (6.58) both understate the variability in $\widehat{\boldsymbol{b}}_i$ as a ***predictor*** of $\boldsymbol{b}_i$.

Laird and Ware (1982) and Davidian and Giltinan (1995, Section 3.3) offer more discussion.

REMARK: It is possible to arrive at (6.55) directly by an argument similar to that above using Bayes theorem, where ξ is treated as known but β is viewed instead as a ***random vector independent*** of $\boldsymbol{b}_i$ with ***prior density*** $p(\beta|\beta^*, \boldsymbol{H})$ depending on ***hyperparameters*** β^* and $\boldsymbol{H}$ corresponding to the $\mathcal{N}(\beta^*, \boldsymbol{H})$ distribution.

Under these conditions, the posterior densities of $\boldsymbol{b}_i$ and β can be derived. If one assumes ***vague*** prior information on β by setting $\boldsymbol{H}^{-1} = \boldsymbol{0}$, it can be shown that mean of the posterior density for β is $\widehat{\beta}$ and that for $\boldsymbol{b}_i$ is (6.55). The details are presented in Davidian and Giltinan (1995, Section 3.3).

BEST LINEAR UNBIASED PREDICTION (BLUP): Putting the Bayesian interpretation aside, we consider another perspective on (6.55). A standard principle in statistics is that a "***best***" predictor is one that ***minimizes mean squared error***. Namely, here, $\boldsymbol{c}(\boldsymbol{Y}_i)$ is the best predictor if it minimizes

$$E[\{\boldsymbol{c}(\boldsymbol{Y}_i) - \boldsymbol{b}_i\}^T \boldsymbol{A} \{\boldsymbol{c}(\boldsymbol{Y}_i) - \boldsymbol{b}_i\}], \tag{6.59}$$

where this expectation is with respect to the joint distribution of $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$, and $\boldsymbol{A}$ is any positive definite symmetric matrix. It is a fundamental result that the ***best predictor*** in the sense of minimizing (6.59) is

$$E(\boldsymbol{b}_i|\boldsymbol{Y}_i), \tag{6.60}$$

which does not depend on $\boldsymbol{A}$. The argument is straightforward and proceeds by ***adding and subtracting*** (6.60) to each of the terms in braces in (6.59) and rearranging to show that $\boldsymbol{c}(\boldsymbol{Y}_i) = E(\boldsymbol{b}_i|\boldsymbol{Y}_i)$ (the diligent student will be sure to try this).

Thus, ***under the usual normality assumptions*** for the linear mixed model, the developments above show that (6.55) with β replacing $\widehat{\beta}$ and ξ ***known*** is "***best***" in this sense. Because (6.55) is also ***linear*** in $\boldsymbol{Y}_i$, it is the best ***linear function*** of $\boldsymbol{Y}_i$ to use as a predictor under normality.

In general, the best predictor (6.60) ***need not*** be linear. However, if attention is restricted to predictors $\boldsymbol{c}(\boldsymbol{Y}_i)$ that are ***linear*** functions of $\boldsymbol{Y}_i$, it is can be shown that, ***without any normality assumptions*** (and ξ known), (6.55) is the ***best linear unbiased predictor*** for $\boldsymbol{b}_i$ in the sense that it minimizes the mean squared error, is a ***linear*** function of $\boldsymbol{Y}_i$, and is such that $E(\widehat{\boldsymbol{b}}_i) = E(\boldsymbol{b}_i) = \boldsymbol{0}$.

We do not provide the argument here; Searle, Casella, and McCulloch (2006, Chapter 7) and Robinson (1991) offer detailed derivations.

In practice, ξ is replaced by the ML or REML estimator $\widehat{\xi}$, in which case some authors have referred to the resulting predictor as an ***estimated best linear unbiased predictor*** or ***EBLUP***.

In the linear mixed effects model literature, the term ***BLUP*** , ***empirical Bayes estimator*** , and ***EBLUP*** are often used ***interchangeably***.

HENDERSON'S MIXED MODEL EQUATIONS: Yet another approach to deducing a ***predictor*** for $\boldsymbol{b}_i$ is due to Henderson (1984). It is customary to present this using the "stacked" notation in (6.13) – (6.15). Here, we treat ξ and thus γ, $\boldsymbol{R}$, and $\boldsymbol{D}$ as ***known***.

For known ξ, Henderson proposes to "***estimate***" the $\boldsymbol{b}_i$, $i = 1, \ldots, m$, which are stacked in the vector $\boldsymbol{b}$, jointly with β, by minimizing in β and $\boldsymbol{b}$ the ***objective function***

$$\log|\widetilde{\boldsymbol{D}}| + \boldsymbol{b}^T\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{b} + \log|\boldsymbol{R}| + (\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}), \tag{6.61}$$

which under normality is twice the negative log of the ***posterior density*** of $\boldsymbol{b}$ for fixed β and twice the negative loglikelihood for β holding $\boldsymbol{b}$ fixed.

Differentiating (6.61) with respect to β and $\boldsymbol{b}$ using the matrix differentiation rules in Appendix A and setting equal to zero yields

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) &= \boldsymbol{0} \\ \widetilde{\boldsymbol{D}}\boldsymbol{b} - \boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) &= \boldsymbol{0}, \end{aligned}$$

which can be rearranged to yield (verify) the so-called ***mixed model equations***

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \widetilde{\boldsymbol{D}}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \end{pmatrix}. \tag{6.62}$$

It can be shown by demonstrating that

$$\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \widetilde{\boldsymbol{D}}^{-1})\boldsymbol{Z}^T\boldsymbol{R}^{-1} = (\boldsymbol{R} + \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T)^{-1} = \boldsymbol{V}^{-1},$$

which can be derived using matrix inversion results in Appendix A, that the solutions to (6.62) are

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}, \qquad \widehat{\boldsymbol{b}} = \widetilde{\boldsymbol{D}}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}),$$

from whence the expression (6.55) for $\widehat{\boldsymbol{b}}_i$ follows.

SHRINKAGE: We demonstrate that empirical Bayes estimators (BLUPs) have the well-known property of "***shrinking***" predictions toward the mean in the sense we now describe. Consider (6.55) with ξ ***known***, that is

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}). \tag{6.63}$$

First consider the simplest special case of the linear mixed model, where $\boldsymbol{X}_i = \boldsymbol{1}_{n_i}$ for all i and $p = 1$ and $\boldsymbol{Z}_i = \boldsymbol{1}_{n_i}$ for all i and $q = 1$, so that the Y_{ij} have common scalar mean β for $j = 1, \ldots, n_i$, and the random effect is a ***scalar***; that is,

$$\boldsymbol{Y}_i = \boldsymbol{1}_{n_i}\beta + \boldsymbol{1}_{n_i}b_i + \boldsymbol{e}_i, \tag{6.64}$$

where $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \text{var}(\boldsymbol{e}_i) = \sigma^2\boldsymbol{I}_{n_i}$ and $\text{var}(b_i|\boldsymbol{x}_i) = \text{var}(b_i) = D$, a scalar. Then $\boldsymbol{V}_i = D\boldsymbol{J}_{n_i} + \sigma^2\boldsymbol{I}_{n_i}$, which of course has ***compound symmetric*** correlation structure. It can be shown that

$$\boldsymbol{V}_i^{-1} = \sigma^{-2}\left(\boldsymbol{I}_{n_i} - \frac{D}{\sigma^2 + n_iD}\boldsymbol{J}_{n_i}\right)$$

(verify). Then, defining $\overline{Y}_i = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$ to be the simple average of the elements of $\boldsymbol{Y}_i$ and noting that $\widehat{\beta}$ is a weighted average of the $\overline{Y}_i$ (verify), it is straightforward to show that the BLUP (6.63) is

$$\widehat{b}_i = \frac{n_iD}{\sigma^2 + n_iD}(\overline{Y}_i - \widehat{\beta}). \tag{6.65}$$

Several insights follow from (6.65):

- First, note that we can write (6.65) as

  $$\widehat{b}_i = w_i(\overline{Y}_i - \widehat{\beta}) + (1 - w_i)0, \quad w_i = \frac{n_iD}{\sigma^2 + n_iD} < 1,$$

  so that $\widehat{b}_i$ can be interpreted as a ***weighted average*** of the estimated overall deviation $(\overline{Y}_i - \widehat{\beta})$, which is our ***best guess*** for where i "sits" in the population relative to the overall mean β based solely on the data, and 0, the mean of b_i.

  The "weight" $w_i < 0$ thus moves $\widehat{b}_i$ away from being solely based on the data and toward the mean of b_i (0).

The more data we have on i, reflected by larger n_i, the closer w_i is to 1, and the ***more weight*** is put on $(\overline{Y}_i - \widehat{\beta})$ as being a reflection of where i "sits." Likewise, if ***among-individual variation*** is large ***relative*** to ***within-individual variation***, so that D/σ^2 is large, again, $\widehat{b}_i$ puts ***more weight*** on the data from i in predicting where i sits. If, on the other hand, n_i is small and/or among-individual variation is small relative to within-individual variation, the information in the data about where i "sits" is ***not of high quality***, so $\widehat{b}_i$ puts more weight toward 0.

- In (6.64), i's individual-specific mean at any time point is $\beta + b_i$. If we were to ***predict*** this individual-specific mean from (6.64), we would naturally use $\widehat{\beta} + \widehat{b}_i$, which, from (6.65), can be written as (verify)

$$\widehat{\beta} + \widehat{b}_i = w_i \overline{Y}_i + (1 - w_i)\widehat{\beta} = \frac{n_i D}{\sigma^2 + n_i D}\overline{Y}_i + \frac{\sigma^2}{\sigma^2 + n_i D}\widehat{\beta}. \tag{6.66}$$

  In (6.66), if w_i is close to 1, then the prediction is based mainly on the data from i, $\overline{Y}_i$. This will be the case if n_i is large and/or D is large relative to σ^2, in which case the quality of information from i is high and/or there is ***little to be learned about a specific individual*** from the population. If w_i is close to 0, then the prediction is based mainly on the estimated overall population mean $\widehat{\beta}$. This will be the case if n_i is small and/or if among-individual variation, as reflected by D, is small relative to within-individual variation, reflected by σ^2, in which case the poor quality of information on i and the fact that individuals in the population do not vary much suggest that there is ***little to be learned*** about i from the data.

- The foregoing phenomena are usually referred to a ***shrinkage*** in the sense that, in predicting where an individual "sits" in the population and thus his/her individual-specific trajectory, the information from the data is "***shrunk***" toward the overall population mean.

These observations of course extend to the ***general form*** of the linear mixed model. In particular, the obvious predictor of the individual-specific trajectory $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ is

$$\begin{aligned}
\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i &= \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1})\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i \\
&= \boldsymbol{R}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \boldsymbol{R}_i\boldsymbol{V}_i^{-1})\boldsymbol{Y}_i.
\end{aligned} \tag{6.67}$$

Analogous to (6.66), (6.67) can be interpreted as a ***weighted average*** of the estimated overall population mean profile $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ and the data $\boldsymbol{Y}_i$ on i. If $\boldsymbol{R}_i$, which reflects ***within-individual variation***, is ***large*** relative to among-individual variation, (6.67) puts more weight on the ***population mean profile***; the opposite will be true if ***among-individual variation*** is relatively large.

Similarly, viewing the model as a ***two-stage hierarchy*** , with ***stage 2 population model*** $\beta_i = \boldsymbol{A}_i\beta + \boldsymbol{B}_i\boldsymbol{b}_i$ as in (6.43), by similar reasoning, if we form "estimates" of the individual-specific parameters β_i as

$$\widehat{\beta}_i = \boldsymbol{A}_i\widehat{\beta} + \boldsymbol{B}_i\widehat{\boldsymbol{b}}_i,$$

we would expect analogous "shrinkage" in the sense that the $\widehat{\beta}_i$ will tend to be "shrunk" toward $\boldsymbol{A}_i\widehat{\beta}$.

CAVEATS ON DIAGNOSTICS USING EMPIRICAL BAYES ESTIMATES: It is tempting, and indeed popular, in practice to use the $\widehat{\boldsymbol{b}}_i$ for ***diagnostic purposes***.

- It is common to construct ***histograms and scatterplots*** of the $\widehat{\boldsymbol{b}}_i$ to identify individuals who may be regarded as ***unusual*** relative to the rest of the individuals from the relevant populations from which they arise. For example, such individuals may have individual-specific trajectories that ***evolve differently*** from those for the bulk of the other individuals in the population.
- It is also common to use the $\widehat{\boldsymbol{b}}_i$ to evaluate the relevance of the ***normality assumption*** on the random effects $\boldsymbol{b}_i$ by plotting histograms and scatterplots as well as ***normal quantile plots*** of the components of the $\widehat{\boldsymbol{b}}_i$.

There are several caveats one must bear in mind when inspecting such graphical diagnostics.

- The $\widehat{\boldsymbol{b}}_i$ have ***different distributions*** for each i unless the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are ***the same*** for all individuals. Thus, for unbalanced data, graphics based on the raw $\widehat{\boldsymbol{b}}_i$ may be ***uninterpretable***. One approach to addressing this is to ***standardize*** the $\widehat{\boldsymbol{b}}_i$ using (6.58).
- An even more ominous concern that persists even if the $\widehat{\boldsymbol{b}}_i$ all have the same distribution is ***shrinkage***. Histograms and other graphics of the $\widehat{\boldsymbol{b}}_i$ will reflect ***less variability*** than is ***actually present*** in the distribution of the true $\boldsymbol{b}_i$. In particular, the $\boldsymbol{b}_i$ have true covariance matrix $\boldsymbol{D}$, but as in (6.58),

  $$\text{var}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i|\tilde{\boldsymbol{x}}) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T \left\{ \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{X}_i^T \boldsymbol{V}^{-1} \right\} \boldsymbol{Z}_i\boldsymbol{D}.$$

  Thus, such graphical displays will ***not necessarily*** reflect the true random effects distribution. In particular, the $\widehat{\boldsymbol{b}}_i$ will tend to be "pulled in" toward the center, so that the usefulness of such plots for, for example, detecting ***departures from normality*** is suspect.

As we have demonstrated, the $\widehat{\boldsymbol{b}}_i$ can be viewed as minimizing the mean square error (6.59), which involves the ***squared-error loss function***, which also follows from normality. Louis (1984) and Shen and Louis (1991) discuss developing alternatives to the usual empirical Bayes estimators that are based on other loss functions.

The bottom line is that, while it is not entirely useless to inspect diagnostics based on $\widehat{\boldsymbol{b}}_i$, these ***potential drawbacks*** need to be kept in mind.

6.5 Implementation via the EM algorithm

With today's computational power, obtaining the ML and REML estimates of the model parameters is ***straightforward*** using standard optimization techniques such as Newton-Raphson and variants to maximize the ML and REML objective functions. However, an alternative computational approach that was popular before the advent of modern computing was to use the ***Expectation-Maximization (EM) algorithm***, as demonstrated by Laird and Ware (1982).

The EM algorithm is a ***computational technique*** to maximize an objective function and can be motivated generically from a missing data perspective in a MAR context, starting from the ***observed data likelihood*** as in (5.107); the details are presented, for example, in Section 3.4 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data." If the optimization problem can be cast cleverly as a "missing data" or "latent unobserved variable" problem, then the EM algorithm mechanics can be applied to derive an iterative scheme that, under reasonable conditions, should ***converge*** to the values of the model parameters maximizing the objective function and is ***guaranteed*** to increase toward the maximum at each iteration.

We do not attempt to derive the implementation of the EM algorithm for maximizing the ML and REML objective functions for a linear mixed effects models here from these first principles. Rather, we simply ***sketch heuristically*** the rationale for and form of the algorithm in the case of maximizing the ML objective function.

For definiteness, consider the form of the linear mixed model given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i}), \quad , i = 1, \ldots, m,$$

so with $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ independent of $\boldsymbol{x}_i$ and $\boldsymbol{R}_i(\gamma) = \sigma^2\boldsymbol{I}_{n_i}$, the usual default specification.

In this situation, the algorithm follows by analogy to a missing data problem from viewing the ***full data*** as $(\boldsymbol{Y}_i, \boldsymbol{x}_i, \boldsymbol{b}_i)$, $i = 1, \ldots, m$, and the ***observed data*** as $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, so that the $\boldsymbol{b}_i$, $i = 1, \ldots, m$, are "***missing***" for all i. As we have all along, we condition on the $\boldsymbol{x}_i$.

The joint density of $(\boldsymbol{Y}_i, \boldsymbol{b}_i)$ conditional on $\boldsymbol{x}_i$, $i = 1, \ldots, m$, under the above conditions is easily seen to be ***proportional to*** (check)

$$\prod_{i=1}^{m} \sigma^{-1} \exp\{-(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i)/(2\sigma^2)\}|\boldsymbol{D}|^{-1/2} \exp(-\boldsymbol{b}_i^T\boldsymbol{D}^{-1}\boldsymbol{b}_i/2). \qquad (6.68)$$

If β were known, the unknown parameters in ξ are σ^2 and $\boldsymbol{D}$, and it is straightforward to observe from (6.68) that ***sufficient statistics*** for σ^2 and $\boldsymbol{D}$ are then

$$T_1 = \sum_{i=1}^{m} \boldsymbol{e}_i^T\boldsymbol{e}_i, \qquad \boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{b}_i, \qquad \boldsymbol{T}_2 = \sum_{i=1}^{m} \boldsymbol{b}_i\boldsymbol{b}_i^T. \qquad (6.69)$$

Note that the quantities in (6.69) for β known would be calculable if we had the "full data" available; that is, if we could observe $\boldsymbol{b}_i$ and $\boldsymbol{Y}_i$ and thus $\boldsymbol{e}_i$ for $i = 1, \ldots, m$. In this case, the estimators for $\boldsymbol{D}$ and σ^2 would be

$$\widehat{\sigma}^2 = T_1/N, \qquad \widehat{\boldsymbol{D}} = \boldsymbol{T}_2/m. \qquad (6.70)$$

As can be seen in Section 3.4 of the above-mentioned notes, under these conditions, the EM algorithm is based on repeated evaluation of the ***conditional expectations*** of the "full data" sufficient statistics in (6.69) given the "observed data" $\boldsymbol{Y}_i$, $i = 1, \ldots, m$ (also conditional on $\boldsymbol{x}_i$). Thus, we must derive these conditional expectations.

One way to do this is to write down the (degenerate) joint distribution of $(\boldsymbol{Y}_i^T, \boldsymbol{b}_i^T, \boldsymbol{e}_i^T)^T$, conditional on $\boldsymbol{x}_i$, and then deduce the required quantities by appealing to standard formulæ for the ***conditional moments*** of components of a multivariate normal. This joint distribution is

$$\left(\begin{array}{c} \boldsymbol{Y}_i \\ \boldsymbol{b}_i \\ \boldsymbol{e}_i \end{array}\middle|\, \boldsymbol{x}_i \right) \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{X}_i\beta \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i} & \boldsymbol{Z}_i\boldsymbol{D} & \sigma^2\boldsymbol{I}_{n_i} \\ \boldsymbol{D}\boldsymbol{Z}_i^T & \boldsymbol{D} & \boldsymbol{0} \\ \sigma^2\boldsymbol{I} & \boldsymbol{0} & \sigma^2\boldsymbol{I}_{n_i} \end{pmatrix} \right\}. \qquad (6.71)$$

The marginal joint distributions of $(\boldsymbol{Y}_i, \boldsymbol{b}_i)$ and $(\boldsymbol{Y}_i, \boldsymbol{e}_i)$ given $\boldsymbol{x}_i$ are embedded in (6.71). We have already seen that

$$E(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{x}_i) = \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta),$$

and it follows from standard calculations for conditional moments (verify) that

$$\text{var}(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{x}_i) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\boldsymbol{D}.$$

Of course

$$E(\boldsymbol{b}_i\boldsymbol{b}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i) = E(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)E(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)^T + \text{var}(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i). \tag{6.72}$$

Similarly, it can be verified that

$$E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) = \sigma^2\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)$$

$$\text{var}(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) = \sigma^2(\boldsymbol{I}_{n_i} - \sigma^2\boldsymbol{V}_i^{-1}),$$

and, from standard results for quadratic forms,

$$\begin{aligned} E(\boldsymbol{e}_i^T\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i) &= \text{tr}\{E(\boldsymbol{e}_i\boldsymbol{e}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i)\} \\ E(\boldsymbol{e}_i\boldsymbol{e}_i^T|\boldsymbol{Y}_i,\boldsymbol{x}_i) &= E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)E(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i)^T + \text{var}(\boldsymbol{e}_i|\boldsymbol{Y}_i,\boldsymbol{x}_i). \end{aligned} \tag{6.73}$$

Based on (6.72) and (6.73) and some algebra, the algorithm proceeds as follows. Given starting values $\sigma^{2(0)}$ and $\boldsymbol{D}^{(0)}$, at the ℓth iteration, with $\sigma^{2(\ell)}$ and $\boldsymbol{D}^{(\ell)}$ the current iterates and $\boldsymbol{V}_i^{(\ell)} = \sigma^{2(\ell)}\boldsymbol{I}_{n_i} + \boldsymbol{Z}_i\boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T$, carry out the following two steps:

1. Calculate
$$\beta^{(\ell)} = \left(\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{X}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{Y}_i.$$

2. Define
$$\boldsymbol{r}_i^{(\ell)} = \boldsymbol{Y}_i - \boldsymbol{X}_i\beta^{(\ell)}, \qquad \boldsymbol{b}_i^{(\ell)} = \boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{r}_i^{(\ell)}, \qquad i = 1,\ldots,m.$$

   Then update $\sigma^{2(\ell)}$ and $\boldsymbol{D}^{(\ell)}$ as

$$\sigma^{2(\ell+1)} = N^{-1}\sum_{i=1}^m\{(\boldsymbol{r}_i^{(\ell)} - \boldsymbol{Z}_i\boldsymbol{b}_i^{(\ell)})^T(\boldsymbol{r}_i^{(\ell)} - \boldsymbol{Z}_i\boldsymbol{b}_i^{(\ell)}) + \sigma^{2(\ell)}\text{tr}(\boldsymbol{I}_{n_i} - \sigma^{2(\ell)}\boldsymbol{V}_i^{(\ell)-1})\},$$

$$\boldsymbol{D}^{(\ell+1)} = m^{-1}\sum_{i=1}^m(\boldsymbol{b}_i^{(\ell)}\boldsymbol{b}_i^{(\ell)\,T} + \boldsymbol{D}^{(\ell)} - \boldsymbol{D}^{(\ell)}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{(\ell)-1}\boldsymbol{Z}_i\boldsymbol{D}^{(\ell)}).$$

Iterate between steps 1 and 2 until convergence. See Laird, Lange, and Stram (1987) for details of implementation; these authors also present an algorithm for maximizing the REML objective function.

As is well known, this algorithm can be ***very slow*** to reach convergence; however, a purported advantage relative to direct maximization is that the value of the objective function is guaranteed to increase at every iteration. Frankly, the implementations of direct optimization in SAS and R have been optimized to the point that it is unusual to encounter computational difficulties; however, in this event, the EM algorithm is an alternative approach.

6.6 Testing variance components

As we discussed at the end of Section 6.3, it is possible to take ***either*** a ***population-averaged*** or a ***subject-specific*** perspective on the linear mixed effects model, which we reiterate briefly.

- Under a ***subject-specific perspective***, we explicitly adopt the ***hierarchical*** interpretation of the model, where individuals are acknowledged to have their own individual-specific trajectories governed by individual-specific parameters β_i. Questions of scientific interest have to do with the properties of the ***distributions*** of β_i. Thus, the fixed effects β represent features relevant to the ***mean*** or "***typical***" value of β_i (possibly for different among-individual covariate values). The covariance matrix $\boldsymbol{D}$ (and generalizations thereof) represents the acknowledged variation of these features in the populations of interest. Accordingly, the diagonal elements of $\boldsymbol{D}$ are interpreted as explicitly reflecting the variances of these features, while the off-diagonal elements reflect how these features co-vary in populations of interest. From this point of view, $\boldsymbol{D}$ is a ***legitimate covariance matrix*** in the sense that, at the very least, it is ***nonnegative definite*** (positive semidefinite).

  Likewise, the matrices $\boldsymbol{R}_i$ are acknowledged to also be ***legitimate covariance matrices*** reflecting within-individual variance and correlation. Thus, for example, σ^2 in the simplest specification $\sigma^2 \boldsymbol{I}_{n_i}$ is the total within-individual variance (assumed constant over time) dictating how responses on an individual vary about his/her individual-specific trajectory due to the realization process and measurement error, and it is natural that we believe that $\sigma^2 \geq 0$.

- Under a ***population-averaged perspective***, questions of scientific interest have to do with features of ***overall mean response profiles***. Here, we view the hierarchical formulation as not necessarily representing phenomena of interest but rather as a convenient mechanism to ***induce*** a rich and flexible ***overall covariance structure*** that can handle ***unbalanced data*** where responses are ascertained at possibly different time points for different individuals and that accommodates possibly ***nonstationary*** patterns of overall correlation. Thus, as we noted at the end of Section 6.3, the matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i$ are simply building blocks of an ***overall*** legitimate covariance structure, and thus ***need not*** be legitimate covariance matrices themselves.

These considerations emphasize that it is ***imperative*** that the analyst acknowledge the modeling perspective taken when it comes to making ***inferences*** about covariance structure or, more precisely, inferences on the covariance parameters $\xi = (\gamma^T, \text{vech}(\boldsymbol{D})^T\}^T$, as we now describe.

EXAMPLE: For definiteness, consider the situation of the hip replacement data in Section 6.2. Suppose that we assume as in (6.35) that

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}, \quad \boldsymbol{\beta}_i^T = (\beta_{0i}, \beta_{1i}, \beta_{2i})^T,$$

and then take $\boldsymbol{\beta}_i$ to be as in (6.37), so that we can write as in (6.38)

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{B}_i \boldsymbol{b}_i,$$

where

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix}, \quad \boldsymbol{B}_i = \boldsymbol{I}_3. \tag{6.74}$$

If we take the $\boldsymbol{b}_i$ to be independent of $\boldsymbol{a}_i$ with $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}$, then $\boldsymbol{D}$ is a (3×3) matrix, which involves ***six*** distinct parameters. If we further assume that $\text{var}(\boldsymbol{e}_i) = \sigma^2 \boldsymbol{I}_{n_i}$, which involves an additional parameter, then this of course induces an overall covariance structure of the form

$$\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}^T + \sigma^2 \boldsymbol{I}_{n_i},$$

involving seven parameters, so that ξ is (7×1).

- From a ***PA*** perspective, this model is a way to ***induce*** a quadratic PA population mean model and an overall covariance structure depending on ξ. Because under the PA perspective $\boldsymbol{D}$ is ***not required*** to be nonnegative definite and σ^2 is ***not required*** to be ≥ 0, there are no restrictions on ξ.
- From a ***SS*** perspective, this model embodies the belief that each individual in the population has his/her own individual-specific quadratic trajectory and that individual-specific intercepts, linear components and quadratic components ***vary*** and ***co-vary*** in the population according to the ***covariance*** matrix $\boldsymbol{D}$; in addition, individual-specific responses vary about individual-specific trajectories with ***variance*** σ^2. Here, $\boldsymbol{D}$ is ***required*** to be nonnegative definite and σ^2 is ***required*** to be ≥ 0 for this perspective to be reasonable.

Now consider ***eliminating*** b_{2i} from the model as in (6.40) and taking instead

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{6.75}$$

so that $\text{var}(\boldsymbol{b}_i) = \boldsymbol{D}_2$ is now a (2×2) matrix with three distinct parameters and ξ is then (4×1).

- From a ***PA*** perspective, the specification (6.75) is a way to ***induce*** a ***more parsimonious*** overall covariance structure with ***fewer parameters***.

- From a ***SS*** perspective, (6.75) embodies the assumption that, while individual-specific intercepts and linear components ***vary nonegligibly*** in the population of individuals, individual-specific ***quadratic*** components either do not vary at all or, relative to the variation in intercepts and linear components, exhibit ***negligible variation*** among individuals.

  Thus, as we discussed in Section 6.2, it is popular to view this as asking whether the individual-specific quadratic components are "***fixed***" or "***random***."

Thus, from either perspective, it is of interest to evaluate whether or not (6.75) is adequate to represent the true state of affairs or if (6.74) is required.

- From a ***PA*** perspective, this corresponds to asking whether or not a ***simpler representation*** of the ***overall covariance structure*** based on fewer parameters is adequate or if the richer induced structure involving more parameters is required.

- From a ***SS*** perspective, this corresponds to what we believe about the ***relative magnitude of variation*** in individual-specific quadratic components.

To address this ***formally*** from either perspective, we might want to carry out a ***hypothesis test*** of whether or not (6.75) is sufficient to represent the situation relative to (6.74).

It is straightforward to show that (6.75) can be ***equivalently represented*** by taking the (3×3) matrix $\boldsymbol{D}$ corresponding to $\text{var}(\boldsymbol{b}_i)$ in (6.74) to be of the form

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.76}$$

say. (The diligent student will want to verify this.)

Thus, we can address this issue by testing the null hypothesis that in fact

$$H_0 : \boldsymbol{D} = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{12} & D_{22} & D_{23} \\ D_{13} & D_{23} & D_{33} \end{pmatrix} = \begin{pmatrix} D_{11} & D_{12} & 0 \\ D_{12} & D_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_2 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.77}$$

against an appropriate alternative.

As the model (6.75) is ***nested*** within the model (6.74), it is natural to consider using a ***likelihood ratio test*** for this purpose, constructed from the loglikelihoods fitting the "***full***" model under specification (6.74) and the "***reduced***" model under specification (6.75).

VALIDITY OF TEST PROCEDURES: The key issue is whether or not this likelihood ratio test is a ***valid test*** of H_0. One of the ***regularity conditions*** required for usual ***large sample theory approximations*** to hold is that the true value of a parameter is not on the ***boundary of its parameter space*** but rather lies in its ***interior***. In the context of ***hypothesis testing*** , the value of the parameter ***under the null hypothesis*** cannot be on the boundary of the parameter space but must be in the ***interior of the parameter space*** for usually asymptotic arguments leading to tests to be valid.

In particular, the ***normal approximation*** to the sampling distribution of an estimator, which is used to form Wald and F-type tests, and the ***chi-square approximation*** to the sampling distribution of the likelihood ratio test statistic ***rely critically*** on this condition.

- If we regard (3×3) matrix $\boldsymbol{D}$ in (6.77) as a symmetric matrix, whose parameters simply serve to characterize an overall covariance structure, as we do from a ***PA*** perspective, then there is ***no restriction*** on the values taken on by D_{33} (or any of the parameters, for that matter). Under this perspective, the value of D_{33} in (6.77) under H_0 (0) is in the ***interior*** of the parameter space.
- If we regard the (3×3) matrix $\boldsymbol{D}$ in (6.77) as a ***legitimate covariance matrix*** , as we do from a ***SS*** perspective, then D_{33} is a ***variance*** and, for $\boldsymbol{D}$ to nonnegative definite, it must be that $D_{33} \geq 0$. Under this perspective, the value of D_{33} under H_0 is thus on the ***boundary*** of the parameter space.

We are thus led to the following.

POPULATION-AVERAGED PERSPECTIVE: In the example, comparing the usual ***likelihood ratio test statistic*** described above to the appropriate ***chi-square critical value*** will yield a ***valid test*** of H_0 in (6.77), whose interpretation is as above.

In general, if a PA perspective is taken, the matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i(\gamma)$ are ***not required*** to be nonnegative definite, so that there are no restrictions on ξ. Thus, the values of ξ under a null hypothesis representing simpler structure will not lie on the boundary of the parameter space, and testing whether or not there is evidence a more complex ***induced overall covariance structure*** is preferred over a simpler one can be conducted in the usual way.

SUBJECT-SPECIFIC PERSPECTIVE: In the example, as noted above, H_0 places at least one parameter on the boundary of the parameter space. Thus, carrying out the likelihood ratio test in the usual way ***will not*** lead to a valid test.

To achieve a valid test, one must appeal to specialized theoretical results for ***nonstandard testing situations*** in a classic paper by Self and Liang (1987). Stram and Lee (1994) used this theory to demonstrate that, when $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$, the large sample distribution of the likelihood ratio test statistic is, under reasonable conditions, a ***mixture of chi-squared distributions***.

For $\boldsymbol{D}$ $(q+1 \times q+1)$, for testing a general null hypothesis of the form

$$H_0 : \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_q & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix},$$

where $\boldsymbol{D}_q$ is a $(q \times q)$ postitive definite matrix versus the alternative that $\boldsymbol{D}$ is a general $(q+1 \times q+1)$ nonnegative definite matrix, the large sample distribution of the likelihood ratio test statistic under H_0 is a mixture of a χ^2_{q+1} distribution and a χ^2_q distribution with equal weights of 0.5.

- Our example is the special case of $q = 2$.
- The effect is to ***reduce*** the p-value that results relative to that that would be obtained if one (incorrectly) used the likelihood ratio testing procedure in the usual way. Thus, ignoring the "***boundary problem***" will lead in general to rejection of H_0 of less often and to possibly adopting models that are too parsimonious.

RESULT: We do not discuss this further here; details can be found in Stram and Lee (1994) and Section 6.3 of Verbeke and Molenberghs (2000); see also Verbeke and Molenberghs (2003).

The takeaway message is that ***faithfully acknowledging*** the perspective to be taken (PA vs. SS) on the scientific questions is critical to achieving reliable inferences. The data analyst must probe his or her scientific collaborators to ensure that the appropriate perspective is taken.

STANDARD ERRORS FOR COVARIANCE PARAMETERS: Testing as discussed above is usually carried out to refine the model with the goal of improving inferences on the overall population mean structure from a PA perspective or to assist interpretation from a SS perspective. From a PA perspective, reducing the number of covariance parameters and thereby achieving a ***more parsimonious*** representation of overall covariance structure will hopefully lead to inferences on β and the population mean that are ***more efficient in finite samples***. Here, the covariance parameters are ordinarily ***not*** of scientific interest in their own right.

From a SS perspective, this testing provides insight on the ***relative magnitudes of variation*** of features of ***individual inherent trajectories*** (e.g., individual-specific intercepts and slopes) in the population of individuals. In this case, the diagonal elements of the matrix $\boldsymbol{D}$ represent the magnitudes of variation of these features, and the off-diagonal elements represent how these co-vary in the population of individuals. Thus, ***scientific questions*** may involve characterizing these magnitudes of variation and thus may be stated formally in terms of the diagonal elements of $\boldsymbol{D}$.

When a model specification is adopted such that the diagonal elements of $\boldsymbol{D}$ are assumed to be non-zero, ***estimates*** of these elements characterize the variation in the features to which they correspond on the individual trajectory model (that is, the β_i). Thus, it is of interest to report these estimates accompanied by ***appropriate standard errors***.

In principle , calculation of standard errors for these elements and more generally for all components of the covariance parameter ξ can be based on a large sample approximation to the sampling distribution of the estimator $\widehat{\xi}$. Such an approximation can be derived by an estimating equation argument similar to that for $\widehat{\beta}$ if one is willing to assume that the n_i are ***fixed*** (so no missing data as discussed in Section 5.6). A key issue is that the covariance matrix of the asymptotic distribution of the estimator $\widehat{\xi}$ depends on the ***third*** and ***fourth moments*** of the true distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$. If one is willing to assume ***normality*** of the response, this covariance matrix can be derived from the information matrix in (5.109) and depends on the ***fourth moment*** of a normal distribution. If the true distribution of the response is ***not normal*** , then the approximate sampling distribution for ξ so obtained and thus approximate standard errors derived from it can be ***very unreliable***.

We do not present details here. This discussion underscores the general issue that inference on second moment properties is more problematic than inference on first moment properties.