Chapter 5

Random Effects and the Linear Mixed Model

Until now, we have considered primarily the estimation of regression parameters from the linear model, i.e., β in the model

$$E\ (Y_i|X_i) = X_i\beta, \tag{5.1}$$

as well as the variance parameters when var(Y_i) has an arbitrary structure. In this chapter we consider the use of random effects in modeling longitudinal data or clustered outcomes. Many researchers view it as more natural to assume that the mean response depends upon a combination of population parameters β and subject-specific effects. In the setting of the linear model where Y_i is linear in the parameters and the error terms, it is natural to also assume Y_i is linear in the subject-specific effects. As we will show, this still leads to the linear model (5.1), but var(Y_i) now has a special random effects structure.

The use of random effects offers several benefits when modeling longitudinal data. First, it provides a way to model correlation in unbalanced designs. Secondly, random effects can be used to estimate subject-specific effects arising in several applications. Finally, it offers an optimal way to combine within- and between-subject data.

Random effects are useful when strict measurement protocols are not followed and we have measurements made at arbitrary, irregularly spaced intervals. It is not desirable to design a study in this way, but such data sets are not uncommon. It can happen that we start with a strict protocol but because of missingness and missed timing, we end up with measurement times that do not conform to a set of protocols-defined occasions. Use of retrospectively collected records for analysis

also often leads to unbalanced designs. Random effects models handle this in a very natural way. The same is true with clustered data, where units are nested within cluster. Another setting is where the metameter chosen for analysis differs from age or time, as in using current height to predict lung function, i.e., should variance in lung function depend on age, occasion of measurement, height or some combination? Finally, we may choose to rescale the time variable differently for each subject, to reflect time before and after a critical event, such as menarchy or sero-conversion.

This chapter is organized as follows. First we introduce ideas in terms of a two-stage random effects model with both population parameters and subject effects, then we consider general linear mixed model, estimation of β and Σ by ML and REML estimation, and finally estimation of the random effects.

5.1 Two-Stage Random Effects Models

Two-stage random effects models begin by assuming at Stage 1 that each unit has its own design on time, denoted by Z_i, and its own parameter vector β_i. Given Z_i and β_i, we assume that:

Stage 1.

$$\underset{n_i\times 1}{Y_i} = \underset{n_i\times q}{Z_i}\ \underset{q\times 1}{\beta_i} + \underset{q\times 1}{e_i} \tag{5.2}$$

where $e_i \sim N_{n_i}(0, \sigma^2 I)$, and the e_i are independent. The e_{ij}'s are *iid* $N(0, \sigma^2)$, so can be thought of as measurement error. The β_i are often called the "true regression coefficients," since the observed responses for the i^{th} subject are assumed to follow the curve with coefficients β_i, but with added measurement error e_i. This defines Stage 1 of the model.

Here Z_i specifies the growth curve model, such as linear or quadratic (or a spline, etc.) or more generally, the pure "within subject covariates." For example, we may have

$$Y_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ \vdots \\ \vdots \\ e_{in_i} \end{pmatrix}. \tag{5.3}$$

Notice that q, the dimension of β_i, does not vary with i. Apart from the intercept, Z_i contains only within subject covariates. We accommodate

subject-specific covariates in Stage 2. A way to think about the Stage 1 model is that generally we want to be able to obtain estimates of β_i and σ^2 using just the data from the i^{th} subject, i.e.,

$$\widehat{\beta}_i = \left(Z_i^T Z_i\right)^{-1} Z_i^T Y_i,$$

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} Y_i^T \left(I - Z_i \left(Z_i^T Z_i\right)^{-1} Z_i^T\right) Y_i \Big/ \sum_{i=1}^{N} (n_i - q)$$

This may not be possible for each subject (if, e.g., $n_i \leq q$), but it should be feasible in principle. Thus between subject variables, such as sex, cannot be included at Stage 1. Essentially, Stage 1 consists of separate regression models for each subject, with the same set of predictors, but possibly different values for the predictors.

REMARK. Note that we could allow correlation between $e_{ij}, e_{ij'}$, but this would change the interpretation of the β_i's. Assuming correlation between the error terms implies that e_{ij} is no longer simply "measurement error" but includes model misspecification at the individual level. Alternately, we can have two "error" components, one modeling serial correlation resulting from model misspecification, and one modeling measurement error, as in the Diggle (1988) model discussed in Chapter 1.

The β_i's are random variables; to specify population parameters we model the mean and variance of the random effects at Stage 2. We model variation in the β_i's as a function of subject-specific covariates and residual between subject variation:

Stage 2.

$$\underset{q\times 1}{E\left(\beta_i\right)} = \underset{q\times p}{A_i} \; \underset{p\times 1}{\beta}, \tag{5.4}$$

$$\operatorname{var}(\beta_i) = \underset{q\times q}{D}. \tag{5.5}$$

This completes the Stage 2 model. This model allows some of the variation in the β_i's to be explained by covariates contained in A_i; the remaining variation is measured by D. Note that A_i cannot include within subject covariates, such as time, because the outcome of the model is "β_i" which does not vary over time.

EXAMPLE. Suppose we have two groups, and we model a linear decline in each, with mean slope and intercept depending upon group, i.e., we use model (5.3) to describe individual responses, Y_{ij}. Here t_{ij} is the time (since beginning of study) that the jth measurement was made

for the ith subject. In other settings, t_{ij} might be age of subject. Denote the group by a dummy variable:

$$\begin{aligned} g_i &= 1 \quad \text{if GR 1,} \\ g_i &= 0 \quad \text{if GR 0.} \end{aligned}$$

Consider a model that assumes the mean slope and intercept differs from each group:

$$\begin{aligned} E(\beta_{0i}) &= \beta_1 + \beta_2 g_i, \\ E(\beta_{1i}) &= \beta_3 + \beta_4 g_i. \end{aligned} \tag{5.6}$$

Here β_1 is the the mean intercept for group 0, and $\beta_1 + \beta_2$ is the mean intercept for group 1, so β_2 is the difference in the intercepts in the two groups; β_3 is the mean slope in group 0 and β_4 is the difference on the slopes. Thus β_2 might be considered the main effect of group and β_4 the group $\times$ time interaction effect, where the time trend is assumed to be linear.

In matrix notation, (5.6) implies

$$E(\beta_i) = A_i \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

where

$$A_i = \begin{bmatrix} 1 & g_i & 0 & 0 \\ 0 & 0 & 1 & g_i \end{bmatrix}.$$

Further assume that regardless of treatment group,

$$\text{var}\,(\beta_i) = D = \begin{bmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{bmatrix}$$

where

$$\begin{aligned} d_{00} &= \text{var}(\beta_{0i}), \\ d_{11} &= \text{var}(\beta_{1i}), \\ d_{10} &= \text{cov}(\beta_{0i}, \beta_{1i}). \end{aligned}$$

The variation in β_{0i} after adjusting for person specific covariates is d_{00}, and similarly for d_{11}. Figure 5.1 illustrates a scenario for group 0, where individual curves are plotted using $\beta_{0i} + \beta_{1i}t$. The heavy line represents the curve obtained from plotting $\beta_1 + \beta_3 t$ for group 0.

In this picture there is substantial variability in the intercepts (d_{00}), but apart from two subjects, the slopes are nearly constant ($d_{11} \doteq 0$).

REMARK. The coding of the Stage 1 regression variables is critical for the proper interpretation of the mean and variance effects. Returning to

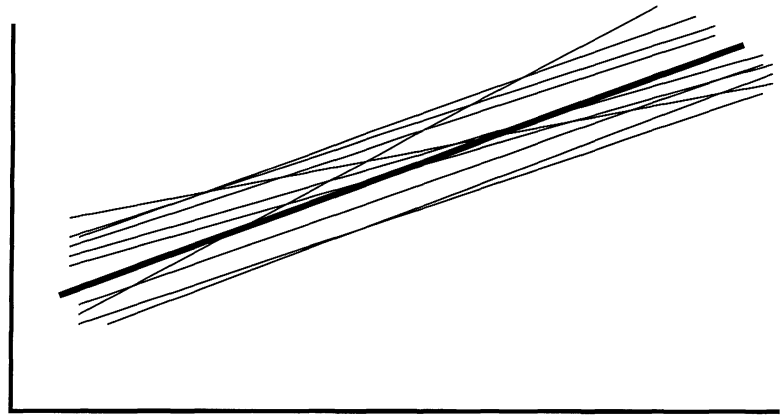

FIGURE 5.1.

our example, if t_{ij} is time since baseline (so $t = 0$ denotes baseline), then β_{0i} represents the expected response at baseline. If t_{ij} is age (say a_{ij}), then the β_{0i}'s will rarely have a useful interpretation because they represent expected response at age zero. In this setting one could contemplate coding t_{ij} as $(a_{ij} - \overline{a}_{i\cdot})$ or $(a_{ij} - \overline{a}_{\cdot\cdot})$ or $(a_{ij} - A)$, where A is some fixed age in the age range of participants. This last choice will sometimes be most attractive since it makes the β_{0i}'s interpretable as the individual's expected response at the common age A, and d_{00} is the variation around the response at age A. Notice that a response at A *need not be* observed for any subject as the model is used to predict the response at A for each subject. If $t_{ij} = (a_{ij} - \overline{a}_{i\cdot})$ then β_{0i} estimates the subject's response at their mean age over the period of follow-up. This may vary considerably from subject to subject, and create a large value for d_{00} which is not meaningful. This is rarely a good choice for centering.

In some circumstances, it may be useful to use $t_{ij} = (a_{ij} - a_i^*)$, where a_i^* is the age that the ith individual experiences a benchmark event. Namova *et al.* (2001) use random effects models to characterize change in body fat in girls before and after menarchy. The study was designed to begin annual follow-up prior to menarchy and continue for four years after menarchy. For analysis, time is coded as time since menarchy, and could be positive or negative. The intercepts, β_{0i}'s, provide an estimate of the individuals body fat at menarchy, even though no actual measures of body fat were made at this occasion, except by chance.

Returning to the two-stage model, we use the identities

$$E(Y_i|X_i) = E(E(Y_i|Z_i, \beta_i))$$

and

$$\text{var}(Y_i|X_i) = E\ \text{var}(Y_i|Z_i, \beta_i) + \text{var}E(Y_i|Z_i, \beta_i)$$

where X_i contains the covariates in Z_i and A_i, to show that

$$E(Y_i|X_i) = E(Z_i\beta_i) = Z_iA_i\beta$$

and

$$\text{var}(Y_i|X_i) = Z_iDZ_i^T + \sigma^2 I. \tag{5.7}$$

Hence we have the same general linear model for correlated data, except:

1. We have a special structure for X_i given by $X_i = Z_iA_i$, which implies also that all covariates can all be classified as either "within" or "between" subject covariates. As discussed in the next section, this can be a serious limitation.

2. We have a special random-effects structure for $\text{var}(Y_i) = \Sigma_i$ as $Z_iDZ_i^T + \sigma^2 I$. Notice that this allows variances and covariances to depend on individual times of measurements. Notice also that the number of variance-covariance parameters does not depend upon the n_i.

Returning to the example, where $Z_{ij} = (1, t_{ij})$, (5.7) implies that

$$\begin{aligned}\text{var}\,(Y_{ij}) &= (1, t_{ij})\, D \begin{pmatrix} 1 \\ t_{ij} \end{pmatrix} + \sigma^2 \\ &= d_{00} + 2t_{ij}d_{10} + t_{ij}^2 d_{11} + \sigma^2\end{aligned}$$

depends upon t_{ij}, it is easily shown that $\text{cov}\left(Y_{ij}, Y_{ij'}\right)$ depends upon $t_{ij}, t_{ij'}$. Hence $\Sigma_i(\theta)$ depends upon the pattern of observation times, as well as the variance and covariance parameters. Notice that only four parameters $(\sigma^2, d_{00}, d_{10}, d_{11}$, or in general, $1 + q(q+1)/2)$ are needed to model $\text{var}(Y_i)$.

REMARK. The case where $n_i = n$, and $Z_i = Z$, is very special. Here each subject has exactly the same design on time. In this setting it is well known that in the absence of subject-specific covariates, the population mean curve is the same as the average of the individual curves. In addition, OLS estimates of each subject have identical precision (assuming σ^2 is the same for each subject).

If in addition, A_i can be formulated as $A_i = a_i^T \otimes I$, where a_i is a $k \times 1$ vector of subject-specific covariates and I is a $q \times q$ identity matrix, then the regression of each component of β_i on the covariates has the same

design matrix; i.e., for some partition of β we can write the following:

$$\begin{aligned} E(\beta_{0i}) &= a_i^T \beta^0 \\ E(\beta_{1i}) &= a_i^T \beta^1 \\ &\vdots \\ E(\beta_{qi}) &= a_i^T \beta^q, \end{aligned}$$

where the β^0, β^1, ..., β^q are distinct vectors, each of length $k \times 1$, so that β is $gk \times 1$. Essentially $A_i = a_i^T \otimes I$ means that the covariate models are the same for each random effect, i.e., if sex affects the intercept, it affects the slope, etc., as well.

In this special setting, one can sometimes reduce the estimation problem to a series of univariate regressions without any loss of efficiency. First, each β_i is estimated by OLS for each subject, then each element of $\widehat{\beta}_i$ is analyzed using univariate regression methods. This approach is fully efficient if the assumption that $\Sigma = Z\,D\,Z^T + \sigma^2\,I$ holds, and it may be reasonably efficient under weaker assumptions. Of course, one cannot use univariate methods to test for global covariate effects on time, and there will be further loss of efficiency with unbalanced designs of missing data, or settings where the design matrix departs from $a_i^T \otimes I$. But it does suggest that this simple two-step strategy is worth considering if the design is nearly balanced, and the covariates affect all of the model coefficients.

5.2 A Linear Mixed Model (LMM)

In this section we reformulate the two-stage random effects model as a more general mixed model that allows greater flexibility in handling all types of covariates. We rewrite β_i as

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

where $b_i \sim N_q(0, D)$. Hence b_i gives the coefficients for an individual's residual curve after the covariate effects have been accounted for. For the example in 5.1, b_{0i} is an individual's deviation from the mean intercept for their treatment group, and b_{1i} is their residual from the mean slope.

We can now combine Stages 1 and 2 of the random effects model to write

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

Thus we have partitioned Y_i into three components:

$$Y_i = \text{mean} + \begin{matrix}\text{"between subject} \\ \text{residual"}\end{matrix} + \begin{matrix}\text{"within subject} \\ \text{residual"}\end{matrix}.$$

The between subject residual is b_i with zero mean and variance D, and e_i is the within subject residual with zero mean and variance $\sigma^2 I_{n_i}$. It follows that

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

and

$$\text{var}\,(Y_i) = Z_i D Z_i^T + \sigma^2 I.$$

Notice that this two-stage derivation requires that $X_i = Z_i A_i$ have a special structure which is inconvenient, since A_i must have only non-time-varying covariates and Z_i has only time-varying covariates (except the intercept). In order to allow for a sufficiently complex structure for the mean response $(Z_i A_i)\beta$, it may be necessary to include many variables in Z_i, requiring an equally complex Σ. Note that the same Z_i appears in both $E(Y_i)$ and $\text{var}(Y_i)$. Simply modifying the model to allow X_i to include whatever covariates we want without changing Z_i gives more flexibility. For example, in modeling mean lung function in children as linear in age and height, using the two-stage model would require D to be at least 3×3 since both age and height vary over time within a subject.

Once we get away from the concept of two-stage, not only can we let X_i be arbitrary, we can also choose q arbitrarily, i.e., we could assume the intercepts vary randomly but the slopes do not, in which case Z_i is just an $n_i \times 1$ vector of ones. This yields compound symmetry:

$$\begin{aligned}\text{var}(y_{ij}) &= d_{00} + \sigma^2, \\ \text{cov}(y_{ij},\ y_{ij'}) &= d_{00}.\end{aligned}$$

Thus we define the linear mixed model (LMM) as follows:

$$\boxed{\underset{n_i \times 1}{Y_i} = \underset{n_i \times p}{X_i}\ \underset{p \times 1}{\beta} + \underset{n_i \times q}{Z_i}\ \underset{q \times 1}{b_i} + \underset{n_i \times 1}{e_i}.} \tag{5.8}$$

Both b_i and e_i are independent, zero mean error terms. The only constraint on X_i and Z_i is that Z_i be a subset of the columns of X_i. This is because we think of Z_i b_i as zero mean residuals, so they should "deviate" from some corresponding nonzero mean. For example, it would be counterintuitive to let the individual slopes deviate in the population and assume that the population mean slope is zero.

Notice that (5.8) is like the LMCD for correlated data in that

$$E\left(Y_{n_i \times 1}\right) = X_i \beta$$

but now

$$\text{var}\left(Y_{n_i \times 1}\right) = Z_i D Z_i^T + \sigma^2 I$$

has the characteristic random effects structure. We also have random effects (b_i's) which can be estimated. We can generalize this model even further to allow $\text{var}(e_i) = R_i$. Notice that R_i cannot be completely unstructured because we will have overparameterized. Generally, both the dimension and the elements of R_i may depend on unique times of measurement.

Sometimes, as in the Diggle (1988) model, R_i is expanded further to include both a serial measurement component, say S_i and a measurement, or sampling error component, say M_i, with $\text{var}(M_i) = \sigma^2 I_{n_i \times n_i}$ and var $S_i = \tau^2 \Omega_i$; Ω_i may have one of the structures considered in Section 1.4. Hence $R_i = \sigma^2 I + \tau^2 \Omega_i$. As noted above, for arbitrary Z_i, D and Ω_i, there are identifiability problems which are not readily quantified for the general case, hence the different components of

$$\text{var}\left(Y_i\right) = Z_i D Z_i^T + R_i$$

should be kept relatively simple. McCulloch and Searle (2001) discuss several models for $\text{var}(Y_i)$ in this setting.

5.3 ML Estimation for the LMM

For the purpose of estimating β and θ, where θ is the vector of parameters in D and R_i, the LMM can be viewed as a special case of the LMCD where Σ_i has a variance component structure. Notice that specification of the LMM does not require normally distributed error terms. When Σ_i is known, the optimal estimator (and also the ML assuming normality) of β is $\widehat{\beta}(\Sigma_i^{-1})$. Method-of-moment, ML and REML estimates for θ have been proposed (Jennrich and Schluchter, 1986; Laird and Ware, 1984; Vonesh and Carter, 1982). We will discuss only ML and REML estimation; software is readily available for ML and REML. As in the general case, ML and REML esimates are consistent even in the absence of normality. In some settings we may also want to estimate the individual random effects (b_i's). A special theory is needed for these; it will be discussed in 5.5. For the remainder of this section we assume that $b_i \sim N_q(0, D)$, $e_i \sim N_{n_i}(0, R_i)$ and given X_i, Z_i the Y_i's are independent, $i = 1, \ldots, N$ so that we may derive ML and REML estimates of θ.

First consider estimating $\widehat{\theta}$ via ML (assuming $R_i = \sigma^2 I$). By definition $(\widehat{\beta}_{ML}, \widehat{\theta}_{ML})$ are obtained by maximizing $\mathcal{L}(\beta, \theta)$ over (β, θ) where

$$\mathcal{L}(\beta, \theta) = \prod_{i=1}^{N} \frac{1}{|\Sigma_i|^{1/2}} e^{-1/2(Y_i - X_i\beta)^T \Sigma_i^{-1}(Y_i - X_i\beta)}$$

$$\Sigma_i = Z_i D Z_i^T + \sigma^2 I,$$

and

$$\theta^T = (\sigma^2,\ d_{00},\ d_{11},\ d_{10},\ \text{etc.}).$$

As before, $\widehat{\beta}(\Sigma_i^{-1})$ can be obtained in closed form for given θ. While expressions for the likelihood equations for θ can be found by directly differentiating the likelihood and substituting in $\widehat{\beta}(\Sigma_i^{-1})$ for β, it is instructive, and easier, to use the EM approach to derive them, even though most computing packages do not routinely use the EM for the computations.

Here it is most convenient to let the complete data be $(Y_i, b_i, e_i), i = 1, \ldots, N$, since this gives an easy maximization for θ at the M-step. The observed data are just the Y_i's. It is then convenient to first obtain the joint distribution of (Y_i, b_i, e_i), as

$$\begin{pmatrix} Y_i \\ b_i \\ e_i \end{pmatrix} = N\left[\begin{pmatrix} X_i\beta \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Z_i D Z_i^T + \sigma^2 I & Z_i D & \sigma^2 I \\ D Z_i^T & D & 0 \\ \sigma^2 I & 0 & \sigma^2 I \end{pmatrix}\right].$$

Notice that $\text{cov}(Y_i, b_i, e_i)^T$ is singular, $(\text{cov}(Y_i, Z_i b_i + e_i) = I)$ and given (b_i, e_i), Y_i contributes nothing to the estimation of θ. However, the marginals, (Y_i, b_i) and (Y_i, e_i) are not singular and these are needed in our calculations.

As in ML estimation with the LMCD, given Σ, the ML estimate of β is WLS(Σ). Hence we derive likelihood equations for the θ parameters only. Thus the relevant piece of the complete data likelihood is

$$\prod_{i=1}^{N} \frac{e^{-b_i^T D^{-1} b_i/2}}{|D|^{1/2}} \quad \frac{e^{-e_i^T e_i/2\sigma^2}}{\sigma^{n_i}},$$

so that $\sum_{i=1}^{N} b_i b_i^T$ and $\sum_{i=1}^{N} e_i^T e_i$ are the complete data "sufficient statistics." It follows that the M-step for θ is

$$\widehat{D} = \sum_{i=1}^{N} b_i b_i^T \Big/ N$$

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} e_i^T e_i \Big/ \Sigma n_i.$$

At the E-step, given $\widehat{\theta}$ and $\widehat{\beta} = \beta(\Sigma(\widehat{\theta})^{-1})$ calculate

$$S_1 = \sum_{i=1}^{N} E(b_i b_i^T | Y_i, \widehat{\theta}, \widehat{\beta})$$

and

$$S_2 = \sum_{i=1}^{N} E(e_i^T e_i | Y_i, \widehat{\theta}, \widehat{\beta}).$$

Given the joint moments of (Y_i, b_i, e_i), it is straightforward to show that the conditional moments of e_i and b_i given Y_i are:

(a) $E(b_i|Y_i) = 0 + DZ_i^T \Sigma_i^{-1}(Y_i - X_i\beta)$

(b) $\text{var}(b_i|Y_i) = D - DZ_i^T(\Sigma_i)^{-1}Z_iD$

(c) $E(b_i b_i^T|Y_i) = E(b_i|Y_i)E(b_i|Y_i)^{-1} + \text{var}(b_i|Y_i)$

$$= DZ_i^T\Sigma_i^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^{-1}\Sigma_i^{-1}Z_iD + D - DZ_i^T\Sigma_i^{-1}Z_iD$$

(d) $E(e_i|Y_i) = 0 + \sigma^2\Sigma_i^{-1}(Y_i - X_i\beta)$

(e) $\text{var}(e_i|Y_i) = \sigma^2 I - \sigma^4\Sigma_i^{-1}$

(f) $E(e_i e_i^T|Y_i) = \sigma^4\Sigma_i^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T\Sigma_i^{-1}$

$$+ \sigma^2\{I - \sigma^2\Sigma_i^{-1}\}$$

(g) $E(e_i^T e_i|Y_i) = \text{tr } E\{e_i e_i^T|Y_i\}$.

It follows that the likelihood equations for $\widehat{\theta}_{ML}$ and $\widehat{\beta}_{ML}$ are the solution to the following:

$$\widehat{D} = \sum_{i=1}^{N} \left[\widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}\left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)^T\widehat{\Sigma}_i^{-1}Z_i\widehat{D} + \widehat{D} - \widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}Z_i\widehat{D}\right] \Big/ N, \tag{5.9}$$

$$\widehat{\sigma}^2 = \text{tr} \sum_{i=1}^{N} \left[\widehat{\sigma}^4\widehat{\Sigma}_i^{-1}\left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)^T\widehat{\Sigma}_i^{-1} + \sigma^2\left(I - \widehat{\sigma}^2\widehat{\Sigma}_i^{-1}\right)\right] \Big/ n_+ \tag{5.10}$$

and

$$\widehat{\beta} = \left(\Sigma X_i^T \widehat{\Sigma}_i^{-1} X_i\right)^{-1} \Sigma X_i^T \widehat{\Sigma}_i^{-1} Y_i \tag{5.11}$$

where

$$\widehat{\Sigma}_i = Z_i \widehat{D} Z_i^T + \widehat{\sigma}^2 I.$$

An iterative algorithm for $\theta^{(p)} \longrightarrow \theta^{(p+1)}$ can be obtained by evaluating the right-hand side of equations (5.9)–(5.10) at $(\beta^{(p)}, \theta^{(p)})$, and setting $\widehat{\sigma}^2$ and $\widehat{D}$ in the left-hand side to $\theta^{(p+1)}$ and using (5.11) to define $\beta^{(p+1)}$ given $\theta^{(p+1)}$. Using this EM algorithm to compute ML's is available as an option in some software packages but convergence can be very slow, especially when the between subject variance (D) is large relative to the within subject variance (σ^2). (This corresponds to a large fraction of missing information.) Alternatively, we can simply use "EM" approach to get expressions for first and second derivatives and then use Newton-Raphson for the computations. However, obtaining second derivatives for θ is more complex but needed for standard errors of $\widehat{\theta}$. See, for example, McCulloch and Searle (2001).

5.4 REML Estimation in the LMM

Recall that the "Bayesian" definition of REML is to maximize the marginal posterior likelihood giving β a flat prior:

$$\underbrace{\mathcal{L}(\theta; Y_1, \ldots, Y_N)}_{\text{REML likelihood}} = \int_{R^p} \underbrace{f(Y_1, \ldots, Y_N | \beta, \theta)}_{\text{ML likelihood}} \, d\beta.$$

This is just like an "incomplete data likelihood" where the complete data is $\mathcal{Z} = (Y_1, \ldots, Y_N, \beta)$ and the incomplete data is $\mathcal{Y} = (Y_1, \ldots, Y_N)$. From section 3.2, we have that

$$\frac{\partial \ln \mathcal{L}(\mathcal{Y}; \Phi)}{\partial \Phi} = E\left\{ \frac{\partial \ln \mathcal{L}(\mathcal{Z}; \Phi)}{\partial \Phi} | \mathcal{Y}, \Phi \right\}$$

where $\mathcal{L}(\mathcal{Y}; \Phi)$ and $\mathcal{L}(\mathcal{Z}; \Phi)$ are the incomplete and complete data likelihoods. Thus for this case, we can write

$$\frac{\partial \ln \mathcal{L}(\theta; Y_1, \ldots, Y_N)}{\partial \theta} = E\left\{ \frac{\partial \ln \mathcal{L}(\theta; Y_1, \ldots, Y_N, \beta)}{\partial \theta} | Y_1, \ldots, Y_N, \theta \right\},$$

where expectation is over the posterior distribution of β given the Y_i's and θ.

But $\partial \ln \mathcal{L}(\theta; Y_1, \ldots, Y_N, \beta)/\partial\theta$ gives the ordinary ML likelihood equations. Setting these likelihood equations to zero yields equations (5.9)–(5.10), for θ derived in the previous section, except that $\widehat{\beta}$ in (5.9–5.10) is just β, since the derivation is with respect to θ given β.

So to get REML equations, we take expectations of (5.9) and (5.10) (with all estimates replaced by their estimands) with respect to the posterior of β, given $Y_1, \ldots, Y_N$, which is distributed as

$$N\left(\widehat{\beta},\ \left(\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i\right)^{-1}\right).$$

Because of joint normality, the conditional means of b_i and e_i are linear in β and the conditional variance does not depend upon β. Hence it is straightforward to see that taking expectations and setting $\theta = \widehat{\theta}_{\text{REML}}$ on both sides of (5.9), we obtain

$$\begin{aligned}\widehat{D}_{\text{REML}} = \sum_{i=1}^{N} \Big[& \widehat{D} Z_i^T \widehat{\Sigma}_i^{-1} (Y_i - X_i\widehat{\beta})(Y_i - X_i\widehat{\beta})^T \widehat{\Sigma}_i^{-1} Z_i \widehat{D} \\ & + \widehat{D} - \widehat{D} Z_i^T \widehat{\Sigma}_i^{-1} Z_i \widehat{D} \\ & + \widehat{D} Z_i^T \widehat{\Sigma}_i^{-1} X_i V X_i^T \widehat{\Sigma}_i^{-1} Z_i D \Big] \Big/ N,\end{aligned} \tag{5.12}$$

where $V = \text{var}(\beta | X_1, \ldots, Y_N)$ and $\widehat{\Sigma}_i$ and $\widehat{D}$ on the RHS are evaluated at $\widehat{D}_{\text{REML}}$ and $\widehat{\sigma}^2_{\text{REML}}$. Since the first three terms are identical to the expression for D_{ML}, and the last term is always positive, it is clear that the REML estimate is generally larger than the ML.

REMARK. We note a connection between the LMM and the LMCD. Suppose we are in the "true" missing data setting in the LMCD, where

$$Y_i = X_i\beta + e_i, \text{var}(e_i) = \Sigma_i,$$

and Σ is an unstructured $n \times n$ covariance matrix for person with complete data. How do these models relate? We can view the unstructured Σ as a special case of the LMM with $\sigma^2 = 0$ by taking $\underset{n_i \times n}{Z_i} = I_i$ to be the matrix of zeros and ones indicating missingness and $D_{n\times n} = \Sigma_{n\times n}$. Then

$$Y_i = X_i\beta + I_i b_i.$$

Here b_i is an $n \times 1$ vector of residuals for a person which might be viewed as "complete data residuals," i.e., e_i if all n observations are observed and

$$\begin{aligned}\text{var}(Y_i) &= Z_i D Z_i^T \\ &= I_i \Sigma I_i^T = \Sigma_i.\end{aligned}$$

Using this trick, we can use programs designed to fit variance component models to fit general (unstructured) Σ, but this is unnecessary as several programs are now available which fit unstructured Σ.

5.5 Estimating the Random Effects

Why estimate the random effects?

1. We might want individual growth curves. Examples include evaluating surrogate markers, and determining maximal growth rates for individuals (Donnelly *et al.*, 1995; Tsiatis *et al.*, 1996).
2. Random effects estimates can be used to establish value or cost. Examples include setting insurance premiums for small regions or small subgroups, establish breeding values for individual animals, establish optimal drilling locations in a region (Robinson, 1991; Cressie 1991).
3. Random effects estimates can be used to evaluate individual quality or performance. Examples include estimating both performance and outcomes for individual hospitals, event rates for local areas, etc.

In estimating individual effects it is sometimes advocated to treat the b_i's as fixed and use OLS to estimate an enlarged vector of fixed effects $\beta^* = (\beta, b_1, \ldots, b_N)$. This approach can be especially useful in the clustered data setting where b_i is a scalar and each n_i is substantial. However, this approach is of limited utility in a general setting for reasons discussed in Chapter 1, and because often many subjects will have small n_i and little information to estimate b_i. As we will discuss, the idea underlying random effects estimates is to "borrow" information from the whole in order to estimate individual effects.

Estimates of random effects are usually called *predictors* rather than estimates; they can be motivated in a Bayesian way (Empirical Bayes) or by extending the Gauss Markov theorem to include random effects. We take up the latter approach first. Recall that for the usual linear model

$$\underset{N\times 1}{Y} = \underset{N\times p}{X} \quad \underset{p\times 1}{\beta} + \underset{N\times 1}{e}$$

where var $(e) = \sigma^2 I$, the **B**est **L**inear **U**nbiased **E**stimator (BLUE) of any estimable contrast $C^T\beta$ satisfies

$$E(C^T\widehat{\beta}) = C^T\beta$$

and

$$\text{var}(C^T\widehat{\beta}) \leq \text{var}(C^T\beta^*)$$

for any other unbiased estimator β^*, where both $\widehat{\beta}, \beta^*$ are linear in Y, i.e., both estimates take the form AY for some A. To extend this to random effects, let B denote the vector of all random effects,

$$B^T = (b_1^T, \ldots, b_N^T)^T.$$

A BLUP is defined as a **B**est **L**inear **U**nbiased **P**redictor of

$$\lambda = C_1^T\beta + C_2^T B.$$

In this case both B and $Y = (Y_1^T, \ldots, Y_N^T)^T$ are treated as random variables. The BLUP has the same general properties, i.e., it is a linear function of the data Y, and it is unbiased and minimum variance in the class of unbiased linear predictors. It is not hard to show that for our case,

$$\widehat{b_i} = E(b_i|Y_i, \widehat{\beta}, \theta) = Z_i D\Sigma_i^{-1}(Y_i - X_i\widehat{\beta})$$

where $\widehat{\beta} = \widehat{\beta}(\Sigma(\theta)^{-1})$ is the BLUP estimator for fixed θ. Thus the BLUP estimator depends upon the unknown θ, which is typically estimated by ML or REML (Harville, 1977).

The empirical Bayes strategy is to regard each b_i as a random parameter with prior $N(0, D)$. Given data Y_i, we can compute the posterior of b_i as

$$b_i|Y_i \sim N\left(\mu_{b_i}, \Sigma_{b_i}\right)$$

where

$$\mu_{b_i} = E\left(b_i|Y_i, \beta, \theta\right) = DZ_i^T\Sigma_i^{-1}\left(Y_i - X_i\beta\right)$$

and

$$\Sigma_{b_i} = \text{var}\left(b_i|Y_i, \beta, \theta\right) = D - DZ_i^T\Sigma_i^{-1}Z_iD.$$

So (if we knew β, θ) an optimal "Bayes" estimate of b_i would be μ_{b_i} (both the posterior mean and the mode). Unlike most Bayes settings, we have replication (in the form of Y_i), which enables us to estimate both θ and β from the data, so the "empirical" Bayes estimate of b_i takes the form

$$\widehat{b_i} = \mu_{b_i}|_{\widehat{\beta},\widehat{\theta}} = \widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}).$$

This is identical to the BLUP estimator when θ is estimated in the same way.

Obtaining a valid estimate for var (b_i) is a difficult problem. We do not want to use

$$\text{var}\left(b_i|Y_i, \widehat{\beta}, \widehat{\theta}\right)$$

because it underestimates the variability arising due to uncertainty about $\widehat{\beta}$. Uncertainty about θ usually has a second order effect (at least for large samples), so it is uncertainty about $\widehat{\beta}$ that is more problematic. The frequentist approach is to use

$$\text{var}\left(\widehat{b}_i - b_i|\widehat{\beta},\widehat{\theta}\right)$$

since both $\widehat{b_i}$ and b_i are random variables. The Bayesian approach is to use the marginal posterior of b_i, integrating out β. These two again yield identical variance expressions for the linear model:

$$\begin{aligned}\text{var}\left(\widehat{b}_i - b_i|Y_i,\beta,\theta\right) &= \text{var}\left(b_i|Y,\theta\right)\\ &= D - DZ_i^T\Sigma_i^{-1}Z_iD\\ &\quad +DZ_i^T\Sigma_i^{-1}X_i\left(\Sigma X_i^T\Sigma_i^{-1}X_i\right)^{-1}X_i^T\Sigma_i^{-1}Z_iD.\end{aligned}$$

This expression often used to get standard errors for EB estimates even though it does ignore error of estimation in $\widehat{\theta}$. Note that only estimates for σ^2 and D are needed to evaluate the variance expression, since β has been integrated out.

Finally, notice that the EB estimates are sometimes called shrinkage (or James–Stein) estimators. To see why, return to the two-stage model where

$$Y_i = A_iZ_i\beta + Z_ib_i + e_i$$

and

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

and assume Z_i is of full rank. After much algebra, we can show that

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

where

$$\widehat{\beta} = \widehat{\beta}\left(\Sigma(\theta)^{-1}\right), \quad \widehat{\beta}_i^{\text{OLS}} = \left(Z_i^TZ_i\right)^{-1}Z_i^TY_i,$$

and W_i is the variance ratio

$$W_i = D\left(D + \sigma^2(Z_i^TZ_i)^{-1}\right)^{-1}.$$

If $\sigma^2 = 0$, meaning that we have perfect information about β_i from Y_i, we see that $\widehat{\beta}_i^{\text{EB}} = \widehat{\beta}_i^{\text{OLS}}$. Alternatively, if $D = 0$, there is no variability in the individual random effects, and $\widehat{\beta}_i^{\text{EB}} = A_i\widehat{\beta}$.

This approach can also be used to get "smoothed" or predicted values as

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

for the observed Y_i's as well as for values of Y_{ij} that are missing. For example, if we fit straight lines (b_i is intercept and slope), then $(\widehat{\beta}, \widehat{b}_i)$ can be used to get each individual's curve and predicted values. If the mean response is estimated only at the points specified by the protocol, the empirical Bayes approach can be used to obtain estimates of any missing Y_{ij}'s.