Chapter 4

Linear Mixed Models

4.1 Model Formulation

The linear mixed model (LMM) is very flexible and capable of fitting a large variety of datasets. It is widely used for repeated measures data or longitudinal studies where data are grouped. The form of the LMM that we use is that of Laird and Ware (1982) which can be considered an extension of the classical linear model. The literature on linear mixed models will often refer to the collection of data that forms a profile as a "cluster" or "subject", depending on the particular application. We use the term "profile" throughout but note that applications of the methods and analysis presented here apply if the data are represented by clusters or subjects. The LMM allows us to account for the correlation within profiles and to consider the profiles as a random sample from a common population distribution, which may be more realistic in many applications. A good introduction to the LMM can be found in Verbeke and Molenberghs (2000) or Schabenberger and Pierce (2002).

If we have m profiles of data, each of which has n_i measurements, where i refers to the i^{th} profile, we can fit a separate linear model to each profile. The model fit in matrix form

is given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, \ldots, m, \tag{4.1}$$

where $\mathbf{y}_i$ is a n_i by 1 vector of responses for profile i, $\mathbf{X}_i$ is a n_i by p matrix of the regressor variables associated with the fixed effects, $\boldsymbol{\beta}_i$ is the p by 1 parameter vector of fixed effects for the i^{th} profile, and $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \mathbf{R}_i)$ is the n_i by 1 vector of errors where $\mathbf{R}_i$ is a n_i by n_i positive definite matrix. If the errors are assumed to be independent, then $\mathbf{R}_i = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix and the estimates of the parameters can be easily obtained via least squares (LS) methods. The estimated parameter vector is given by

$$\widehat{\boldsymbol{\beta}}_{i,LS} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y_i} \text{ for } i = 1, 2, \ldots, m. \tag{4.2}$$

In contrast, the LMM has random effects in addition to the fixed effects of the classical linear model and is given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \text{ for } i = 1, 2, \ldots, m, \tag{4.3}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects that is the same for all profiles, $\mathbf{Z}_i$ corresponds to a n_i by q matrix of the predictor variables with random effects, $\mathbf{b}_i \sim MN(\mathbf{0}, \mathbf{D})$ is a q by 1 vector of random effects for the i^{th} cluster where $\mathbf{D}$ is a q by q positive definite matrix. Because we have written (4.3) in terms of each of the individual profiles, we refer to this particular model formulation as the "unstacked" form.

The model in (4.3) is flexible enough to allow the errors to be independent or correlated. If correlated, $\mathbf{R}_i$ is often assumed to be a simple form such as compound symmetry (CS) or autoregressive (AR) in order to reduce the number of covariance parameters that need to be estimated. For more details on the various types of correlated error structures that can be assumed for $\mathbf{R}_i$, see Littell et al. (1996) or Schabenberger and Pierce (2002). Similar

structure can be imposed on $\mathbf{D}$, but here we restrict $\mathbf{D}$ to be a diagonal matrix. Thus the random effects are assumed to be uncorrelated with each other.

In addition, we assume that $cov(\boldsymbol{\epsilon}_i, \mathbf{b}_i) = \emptyset$, where $\emptyset$ is a n by q matrix of zeros, which means that the random effects and the random errors are uncorrelated, resulting in the conditional model given by

$$\mathbf{y}_i|\mathbf{b}_i \sim MN(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i). \tag{4.4}$$

Furthermore we assume that $\mathbf{Z}_i$ is either a subset of or equal to the $\mathbf{X}_i$ matrix, so any columns in the $\mathbf{Z}_i$ matrix are also contained in the $\mathbf{X}_i$ matrix and thus $p \geq q$. The case where $\mathbf{Z}_i = \mathbf{X}_i$ is referred to as the random coefficients model (Demidenko, 2004) because all the fixed effects have a corresponding random effect. This restriction of $\mathbf{Z}_i$ being contained within $\mathbf{X}_i$ does not eliminate any of the forms of this model that are in common practice. For examples of cases where this restriction is used see Waternaux, Laird, and Ware (1989), Lesaffre, Asefa, and Verbeke (1999), Longford (2001), or Xu (2003).

The corresponding marginal model is given by

$$\mathbf{y}_i \sim MN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i) \text{ for } i = 1, 2, \ldots, m, \tag{4.5}$$

where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$ is a n_i by n_i positive definite matrix.

The model in (4.3) allows for two levels of correlation for the measurements within a profile. The first results from the random effects which cause all the measurements within a profile to be correlated to each other. The second results from the within-profile variance-covariance matrix, $\mathbf{R}_i$. Vonesh and Chinchilli (1997, p. 256) noted that in some applications it makes sense to consider both levels and give some references where both levels are used. Chi and Reinsel (1989) also recommended the use of both levels of correlation where needed.

If a particular application has only the first level of correlation, so that the errors are uncorrelated but random effects are still present, the model in (4.5) reduces to $\mathbf{y}_i \sim MN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$ where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{I}$. If the application only uses the second level of correlation (that is, a fixed effects model where $\mathbf{D}$ is a null matrix and $\mathbf{R}_i$ is non-diagonal) then a time series model can often be fit to account for serial correlation among the responses. Time series models can be more restrictive because they often require equally spaced data. A LMM that uses neither of the two levels of correlation so that we have a model with uncorrelated errors with no random effects then (4.5) reduces to the general linear model in (4.1) because $\mathbf{Z}_i = \mathbf{0}$ and $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \sigma^2\mathbf{I})$.

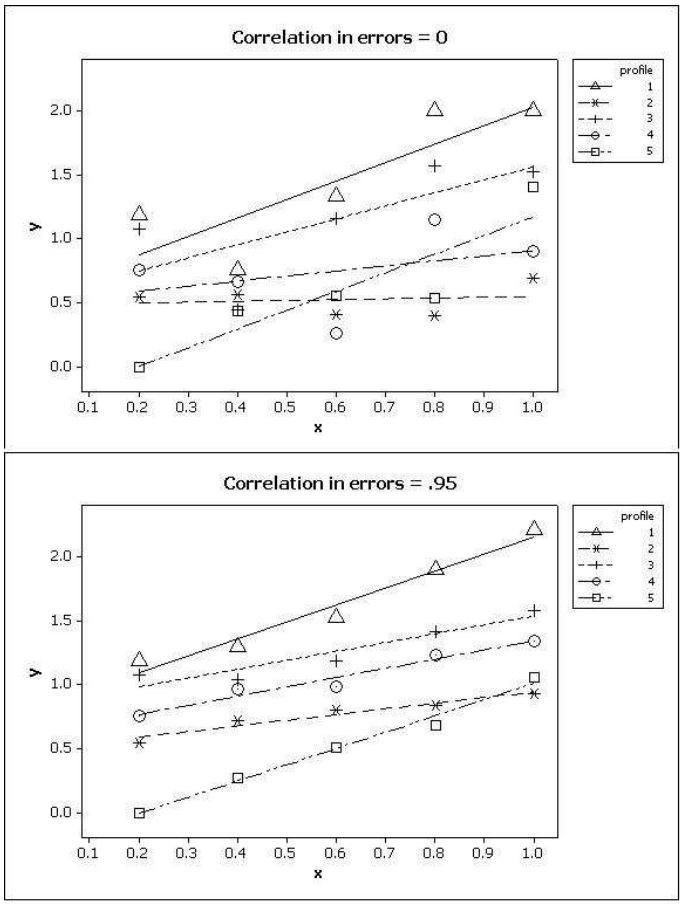
4.2 Correlation in the Errors

Autocorrelated data are very common for time ordered data, such as data representing the price of a stock over time. These time series models for autocorrelated data are usually applied to situations where there is a single profile of data. On occasion, a time series model will be fit to multiple profiles but such models often require a large number of observations per profile to ensure that the model obtained will be representative of the data. On the other hand, the LMM is usually preferable when there are multiple profiles and there are a smaller number of observations per profile which may/or may not be time ordered. With the LMM one seeks to pool information from multiple profiles in order to improve estimates and subsequent inference while with a time series model one does not usually attempt to pool information. As a result, correlated errors in the LMM may appear different when graphically displayed than a graphical display of autocorrelated time series data.

For example, consider the top panel of Figure 4.1 which shows 5 randomly generated

profiles from a LMM with no correlation in the errors. The raw data points are shown along with the simple linear regression fits for each profile. The profiles each have 5 measurements with $\boldsymbol{\beta} = [0, 1]$, $\mathbf{D} = \begin{bmatrix} .1 & 0 \\ 0 & .1 \end{bmatrix}$, $\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & .2 \\ 1 & .4 \\ 1 & .6 \\ 1 & .8 \\ 1 & 1 \end{bmatrix}$, and $\mathbf{R}_i = \sigma^2 \mathbf{I}$ where $\sigma^2 = .1$.

Figure 4.1: Randomly generated profile data along with simple linear regression fit with no correlation and correlation in the errors.



Notice that the data points appear at random on either side of the fitted profile. This is because the independent errors are just as likely to cause a point to be above the line as below the line. Now consider the same scenario but now with correlated errors following an

AR(1) structure. Thus

$$\mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix},$$

where ρ is some number between 0 and 1 and measures the strength of the correlation. Contrast with the top panel the bottom panel of Figure 4.1 which shows the profiles with correlated errors where $\rho = .95$.

Because of the strong correlation and the smaller number of observations per profile, the errors tend to be similar to each other thus dampening the jagged effect of uncorrelated errors. As a result, the fitted profiles tend to appear more similar to each other when there is higher amounts of correlation in the errors. This dampening due to correlated errors will help explain our results in Chapter 5.

4.3 Data Scenarios

Profile monitoring data can be classified into several different scenarios depending on the number of observations per profile and where those observations are located within the profile. For example, all of the profiles can have measurements that are equally spaced along the profile and at the same location for all profiles. We refer to this data type as the balanced, equally spaced data. This implies that $n_i = n$ for $i = 1, 2, \ldots, m$, and that the values of the regressors (and consequently, $\mathbf{X}_i$ and $\mathbf{Z}_i$) are the same for all profiles. This does not necessarily mean that $\mathbf{X}_i = \mathbf{Z}_i$ although as mentioned in Section 4.1 it is often assumed that they are equal to each other. Balanced, unequally spaced data will occur when $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the same for all the profiles but the observations within $\mathbf{X}_i$ and $\mathbf{Z}_i$ may not necessarily be

equal distance from each other. This type of scenario would occur where more measurements for a particular profile occur in the middle than at the edges.

Unbalanced data refers to the scenario where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are not necessarily the same for all the profiles. They may not even have the same number of rows per profile, which indicates an unequal number of observations per profile.

For many control chart applications, where the profiles occur at regular time periods, the data collection is well controlled as if from a designed experiment. Thus the number of measurements per profile will often be the same and at the same locations along the profile. Thus we believe that profile monitoring applications are more likely to have balanced data

Nonetheless, we will consider both balanced and unbalanced data in our simulation studies of Chapter 5. It should be noticed that our choice of terminology for balanced and unbalanced is slightly different than that often used in the literature. The literature uses balanced and unbalanced to denote differences in sample sizes per profile. In order to simplify our comparisons, balanced and unbalanced data will both have the same number of observations for all the profiles, that is $n_i = n$ for $i = 1, 2, \ldots, m$.

4.4 Matrix Form

The model in (4.3) can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{4.6}$$

where $\mathbf{y}$ is a N by 1 stacked vector containing the responses for all the profiles with $N = \sum n_i$, $\mathbf{X}$ is a N by p stacked matrix of the $\mathbf{X}_i's$, $\boldsymbol{\beta}$ is the p by 1 vector of fixed effects, $\mathbf{Z}$ is a N by $mq = r$ block diagonal matrix such that $\mathbf{Z} = diag(\mathbf{Z}_i)$, $\mathbf{b} \sim MN(\mathbf{0}, \mathbf{B})$ is a $mq = r$ by

1 vector of random effects with $\mathbf{B} = diag(\mathbf{D})$, and $\boldsymbol{\epsilon} \sim MN(\mathbf{0}, \mathbf{R})$ is the N by 1 vector of errors with $\mathbf{R} = diag(\mathbf{R}_i)$ and $cov(\boldsymbol{\epsilon}, \mathbf{b}) = \mathbf{0}$. Note that $\mathbf{R}$ and $\mathbf{B}$ are both symmetric block diagonal matrices.

The corresponding marginal model in matrix form is $\mathbf{y} \sim MN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ where $\mathbf{V} = \mathbf{ZBZ}' + \mathbf{R} = diag(\mathbf{V}_i)$ is a N by N positive definite matrix. The matrix form of the marginal model is an alternative form of expression that we denote the stacked form that is useful to show certain results. Note that the conditional model in matrix form is $\mathbf{y}|\mathbf{b} \sim MN(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}, \mathbf{R})$.

4.5 Estimation

Under the distributional assumptions of the marginal model in (4.5), the fixed-effect parameter estimators representing the population average of all the profiles is given by $\widehat{\boldsymbol{\beta}}_{MIX}$, and the estimates of the random deviations from that population average vector are given by $\widehat{\mathbf{b}}_i$ for $i = 1, 2, \ldots, m$. If $\mathbf{V}_i$ (and consequently $\mathbf{D}$ and $\mathbf{R_i}$) are assumed known then it can be shown that

$$\widehat{\boldsymbol{\beta}}_{MIX} = \left(\sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right), \tag{4.7}$$

and the best linear unbiased predictors ("blups") are

$$\widehat{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i' \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}_{MIX} \right). \tag{4.8}$$

If we are working with the stacked form of the LMM, and we assume that $\mathbf{V}$ (and as a consequence $\mathbf{B}$ and $\mathbf{R}$) are known and if the model is correctly specified, then the fixed parameter estimators are given by

$$\widehat{\boldsymbol{\beta}}_{MIX} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \right), \tag{4.9}$$

and the vector of estimated random effects is given by

$$\widehat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right). \tag{4.10}$$

Note that the estimator in (4.9) is a generalized least squares estimator (Vonesh and Chinchilli, 1997, p. 238) and that the expressions in (4.7) and (4.9) are equivalent. It is easy to show that $\widehat{\boldsymbol{\beta}}_{MIX}$ in (4.9) is unbiased and using (4.10) that $E(\widehat{\mathbf{b}}) = \mathbf{0}$. In addition it can be shown that

$$Var(\widehat{\boldsymbol{\beta}}_{MIX}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{4.11}$$

With the additional assumption of multivariate normality it can be shown that $\widehat{\boldsymbol{\beta}}_{MIX} \sim MN\left[\boldsymbol{\beta}, (\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\right]$ or alternatively, $\widehat{\boldsymbol{\beta}}_{MIX} \sim MN\left[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\right]$ (Schabenberger and Pierce, 2002).

Laird and Ware (1982) noted that

$$\begin{aligned}
Var(\widehat{\mathbf{b}}) &= Var\left[\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right] \\
&= Var\left[\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}\right] \\
&= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}\right] Var(\mathbf{y}) \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}\right] \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}' \\
&= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}[\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}' + \\
&\qquad \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}']\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}' \\
&= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\left[\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\right]\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}' \\
&= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}'.
\end{aligned} \tag{4.12}$$

However, as noted by Laird and Ware (1982, p. 966), Verbeke and Molenberghs (2000, p. 78), and Schabenberger and Pierce (2002, p. 431), the expression in (4.12) is not the correct expression of variability of the predictor because it ignores the variability in the random

effects, $\mathbf{b}$. Thus a more appropriate expression of variability is given by

$$\begin{aligned} Var(\widehat{\mathbf{b}} - \mathbf{b}) &= \mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B} + \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}' \\ &= \mathbf{B} - Var(\widehat{\mathbf{b}}). \end{aligned} \tag{4.13}$$

In the unstacked form for each profile (4.13) is given by

$$\begin{aligned} Var(\widehat{\mathbf{b}}_i - \mathbf{b}_i) &= \mathbf{D} - \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} + \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}\mathbf{X}_i(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}' \\ &= \mathbf{D} - Var(\widehat{\mathbf{b}}_i) \text{ for } i = 1, 2, \ldots, m. \end{aligned} \tag{4.14}$$

The above expression was derived under general conditions by Harville (1976) as an extension of the Gauss-Markov theorem to estimate linear combinations of both fixed and random effects. In Appendix A, we derive this expression in (4.13) using the variance operator and the properties of linear combinations of random variables.

In practice, $\mathbf{V}$ is not known and therefore must be estimated prior to obtaining $\widehat{\boldsymbol{\beta}}_{MIX}$ and $\widehat{\mathbf{b}}_i$. The matrix $\mathbf{V}$ can be estimated via maximum likelihood (ML) or restricted maximum likelihood (REML) and an iterative algorithm. REML is often preferred (Schabenberger and Pierce, 2002, p. 437) because it produces estimators with less bias than estimators obtained using ML. The estimates obtained from ML and REML are often very similar to each other and can sometimes be asymptotically equivalent (Demidenko, 2004, p. 146). Nonetheless, we utilize REML for all of our simulation studies.

Once the solution is obtained, $\widehat{\mathbf{V}}^{-1}$ and $\widehat{\mathbf{D}}$ are then placed in (4.7) and (4.8) to obtain the parameter estimates. If a consistent estimate of $\mathbf{V}$ is used, the distribution of $\widehat{\boldsymbol{\beta}}_{MIX}$ will be asymptotically normal (Demidenko, 2004), that is $\widehat{\boldsymbol{\beta}}_{MIX} \overset{a}{\sim} MN\left[\boldsymbol{\beta}, (\sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\right]$. The blups from (4.8) are referred to as estimated best linear unbiased predictors ("eblups") when an estimated variance-covariance matrix is used.

4.6 Properties of Blups

As noted by Verbeke and Lesaffre (1996) and Ritz (2004), the distribution of the eblups does depend on the distribution of both the $\mathbf{b}_i$'s and the $\boldsymbol{\epsilon}_i$'s. In particular, the eblups will be normally distributed as long as the random effects and errors follow a multivariate normal distribution, although the distribution of the eblups is not necessarily the same as that of the blups. Even if the random effects do not have a normal distribution, Jiang (1998) showed that the eblups will converge to their true distribution as long as both the number of profiles and number of observations per profile are increasing asymptotically (i.e. $m \to \infty$ and $n \to \infty$). In addition, the blups for different profiles will have different distributions unless $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the same for all profiles.

The blups are not independent of each other even if $\mathbf{V}$ is known because they all use the same $\widehat{\boldsymbol{\beta}}_{MIX}$ in their calculation. In fact, as shown in Appendix B, there are some cases where the blups sum to zero, thus implying correlation, because if $m-1$ of them are known, the last one is given.

Finally, we note that the blups are examples of shrinkage estimators. They are a function of the observed data and the overall average profile given by $\mathbf{X}_i\widehat{\boldsymbol{\beta}}_{MIX}$. As a weighted average of the data and the overall profile, they are "shrunken" toward the overall profile (Verbeke and Molenberghs, 2000, pp. 80-85). The shrinkage toward the overall profile is more severe if the within profile variability is large compared to the between profile variability. As a result of the shrinkage, a histogram of the individual elements of $\widehat{\mathbf{b}}_i$ corresponding to a particular random effect has less variability than that of the random effect itself contained in $\mathbf{b}_i$. Verbeke and Lessafre (1996) give an example where data were generated such that the $\mathbf{b}_i$ followed a bimodal distribution and a histogram of the elements of $\widehat{\mathbf{b}}_i$ was unimodal.

4.7 Other Issues

Two other issues arise when fitting the LMM; zero estimates and non-convergence. Zero estimates occur when a negative estimate of a diagonal component of the variance-covariance matrix is obtained by the maximum likelihood algorithm. Because it cannot be negative, it is set to be zero, thus causing the estimated variance-covariance matrix to be singular. Setting the value to zero produces biased estimates of the variance-covariance matrix and is the multivariate analog of a similar problem that can occur with estimation of variance components in ANOVA models discussed in Searle, Casella, and McCulloch (1992, Section 3.5.c). See Verbeke and Molenberghs (2000, Section 5.6) for further discussion.

If only of some of the diagonal variance components of the $\mathbf{D}$ matrix are estimated to be zero, a T^2 statistic can still be computed. This is done by dropping out the null row and column (where all the elements are zero because we assumed that $\mathbf{D}$ is a diagonal matrix) of the variance-covariance matrix and the corresponding element of the estimated vector of coefficients. The T^2 statistic is computed using the remaining elements in the vector and variance-covariance matrix. On a rare occasion all the variance components are estimated to be negative. This situation occurs when the profiles are so similar to each other that there is no significant difference between them. If this occurs the T^2 statistic is set to be 0 indicating that none of the profiles is considered to be different from each other.

Non-convergence occurs when no estimates are obtained because of the difficulty of maximizing the likelihood. For most simple problems non-convergence is rare but is more common when the data are unbalanced, the variance components in $\mathbf{V}$ are small and/or the model has been misspecified (Verbeke and Molenberghs, 2000, Section 5.6). For all our simulation studies we tracked the frequency of non-convergence and found it to be small or zero. To reduce the frequency of non-convergence, it is often recommended to use good starting values

for the fixed parameters and components of the variance-covariance matrix. These starting values can be obtained via graphical methods (Schabenberger and Pierce, 2002).

An alternative method is to use the standard least squares estimates to obtain starting values. For example, if one needs a starting value for the fixed intercept parameter in a simple linear regression problem, separate simple linear regressions can be fit for each profile via LS. Then the average of the individual profile intercepts will serve as a good starting value for the overall intercept parameter. For the simulation studies that we performed in Chapter 5, starting values were not needed when considering in-control data because the frequency of non-convergence was low. However, starting values were needed when some of the data are out of control. For our purposes, it was not feasible to consider graphical or other methods to determine different starting values for each randomly generated dataset.

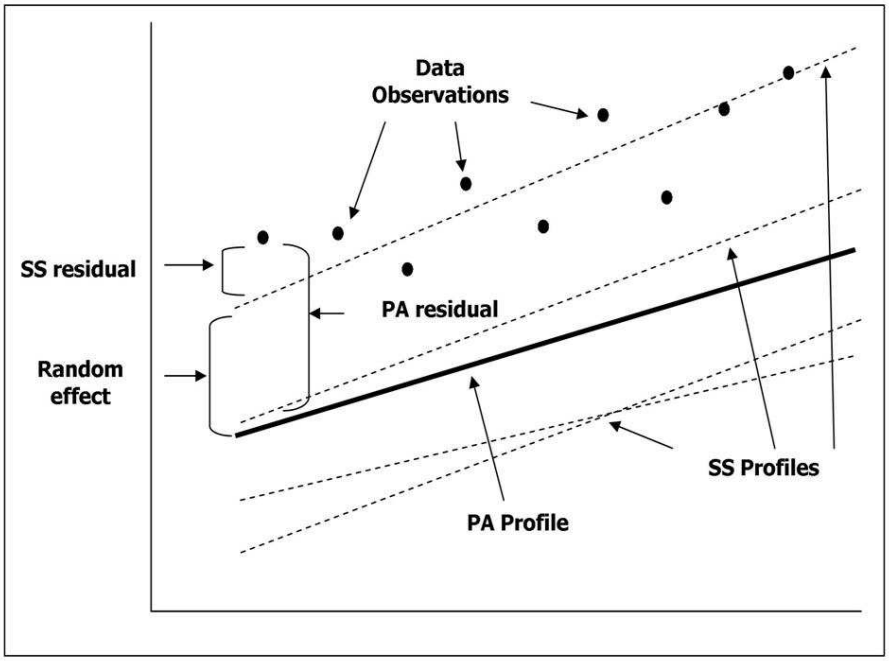
To overcome the infeasibility in situations where the non-convergence was more likely to be present, we used in our simulations the known parameter values used to generate the data as starting values of the iterative algorithm as was done by Hartford and Davidian (2000) for nonlinear models. This reduces the frequency of non-convergence just as would occur if a knowledgeable researcher were to spend a sufficient amount of time exploring, cleaning, and appropriately analyzing a single dataset.

4.8 Residuals

As noted by Verbke and Molenberghs (2000, p. 151), it is not obvious which residuals to use in order to assess goodness of fit or detect outliers in the LMM. Because $\mathbf{X}_i\widehat{\boldsymbol{\beta}}_{MIX}$ represents the overall profile, we have $\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_{MIX}$ representing the deviations of the observed data from the estimated population average (PA) profile. We denote these deviations (residuals)

by the term "PA residuals". The estimated random effects can also be thought of as a type of residual and are equal to the PA residual pre-multiplied by a matrix as shown in (4.8). Thus $\mathbf{Z}_i\widehat{\mathbf{b}}_i$ represents the deviations of the subject specific (SS) profiles from the PA profile. Finally, the SS residuals, given by $\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_{MIX} - \mathbf{Z}_i\widehat{\mathbf{b}}_i$, represent the vector of deviations of the observed data from their corresponding SS profile. The relationship between the random effects, SS and PA residuals for linear profiles is shown in Figure 4.2.

Figure 4.2: Diagram illustrating the PA and SS profiles along with the random effects, PA and SS residuals.



To determine if profiles are outlying, the random effects could be used. However, because of the shrinkage that was mentioned in the previous section, the estimated random effects can be conservative. Because our focus is determination of outlying profiles, our methods will be based on the estimated random effects. We do not believe this will negatively impact our results because our method will take into account the reduced variability of $\widehat{\mathbf{b}}_i$.

To determine if observations within a profile are outliers, the SS residuals have been recommended by Verbeke and Molenberghs (2000, pp. 151-152). Alternative methods will be discussed in Section 4.10.

4.9 Checks of Goodness of Fit and Model Assumptions

When a parametric model is fit to profile data, it is important to know if the model fits the data well and if the model assumptions are met. For example, Copt and Victoria-Feser (2006) noted that ML and REML is not robust to departures from the assumption of normality of the response variable. If the model fits well and the assumptions are adequately met, then the parameter estimates obtained will be a good representation of the profile and the estimates can then be used to determine if the Phase I data are in control. Goodness-of-fit techniques and other checks of the model assumptions for the LMM such as those discussed in Verbeke and Molenberghs (2000, Chapter 4) and Demidenko (2004) can be used. If there is a not a good fit of the LMM to the data, then determining which profiles are outlying is at best a risky activity and should be used with caution. Previous literature on profile monitoring has has assumed that the right model has been fit to the data and not considered that the choice of model may be incorrect.

4.9.1 Goodness of Fit

Verbeke and Molenberghs (2000, Section 4.3) proposed to assess the goodness of fit for LMM by calculating a coefficient of multiple determination for each profile, denoted by R_i^2. They also noted that when the number of observations per profile is small, the R_i^2 values will be very high. In fact, when the number of observations per profile is 2, then $R_i^2 = 1$. As a

result, they recommended the use of scatterplots of R_i^2 vs. n to account for the number of observations per profile. An overall measure of goodness of fit can be obtained by combining the information from the individual profiles.

Xu (2003) proposed several R^2 like measures to measure the amount of variation explained by the LMM when the errors are uncorrelated. These measures focus on the amount of variation explained by the SS curve and are based on the SS residuals rather than the PA residuals. The emphasis here is to obtain and overall measure of fit for prediction purposes rather than goodness of fit of the individual profiles.

4.9.2 Existence of Random Effects

Demidenko (2004, Section 3.5) proposed a hypothesis test of $H_0 : \mathbf{D} = \mathbf{0}$ as a prior check to see if it is necessary to model the random effects. A likelihood ratio test (LRT) is proposed, however it is noted that the value of the null hypothesis lies on the boundary of the parameter space. Thus the LRT statistic does not necessarily have a χ^2 distribution but modification of the test is discussed by Demidenko (2004, Section 3.5).

4.9.3 Normality of Random Effects

A check of the assumption of normally distributed random effects can be done by plotting a histogram of the estimated random effects, but this check will be misleading because of the shrinkage that occurs (Verbeke and Lessafre, 1996; Verbeke and Molenberghs, 2000, Section 7.8). Similarly misleading are the normal probability plot and scatter plot of random effects that were proposed by Pinheiro and Bates (2000, Section 4.3.2).

Zeger, Liang, and Albert (1988) determined that to get consistent inference for fixed

effects of the LMM, correct specification of the random effects distribution is not required. Only correct specification of the mean structure is required. However correct specification of the mean structure and the random effects distribution is required to ensure that the standard errors of the estimators are appropriate.

Lange and Ryan (1989) proposed to assess the normality of the random effects distribution through a weighted normal probability plot of the random effects. Ritz (2004) used the results of Lange and Ryan (1989) to derive a goodness of fit test to check the assumption that the blups are normally distributed. This is an overall test of normality and does not show whether or not individual profiles have been fit well by the LMM.

An early study by Butler and Louis (1992) found that the misspecification of the random effects distribution in linear mixed models does not adversely impact the inferences of fixed effects. Inferences on the fixed effects were found to be similar across different methods of obtaining those effects such as ordinary least squares (OLS), REML, or a non-parametric maximum likelihood procedure.

A similar conclusion was found by Verbeke and Lesaffre (1997a). They showed that the fixed effects and the covariance parameter estimators are consistent and asymptotically normally distributed when obtained via ML under the assumption of normality of the random effects. Consistency and asymptotic normality holds even when the random effects are not normally distributed. Their results are an extension of classical maximum likelihood theory which says that the maximum likelihood estimators are asymptotically normal as the distribution (and hence the likelihood) is correctly specified. However, their results showed that the rate of convergence does depend on the correctness of the assumed random effects distribution. Thus a nearly correct specification of the random effects distribution for a given sample size will be closer to the asymptotic normal distribution than a poorly

specified distribution. When the assumption of normality of the random effects is not a viable assumption, Verbeke and Lesaffre (1997a) recommended use of a robust sandwich type estimator in order to obtain better estimators of the standard errors for the fixed effects and components of the variance-covariance matrix.

In contrast, the estimators of the random effects are sensitive to their assumed distribution of random effects as discussed in Verbeke and Molenberghs (2000, Section 7.8.2). Thus a check of the normality assumption of the random effects distribution is recommended. If the normality assumption is not tenable an alternative method is to model the random effects distribution as a mixture of normal distributions (See Verbeke and Lesaffre, 1996 or Verbeke and Lessafre, 1997b for more details).

4.9.4 Normality of Errors

Pinheiro and Bates (2000, Section 4.3.1) checked the assumption of normally distributed errors with mean zero and the variance equal to σ^2 for the LMM. To do so, they utilized the "within-group residuals" which are equivalent to our SS residuals. They considered various graphical methods such as boxplots of the residuals by profile, a scatterplot of residuals versus the fitted values, and normal probability plots of the residuals. Their approach is very similar to the classical approach of regression residual diagnostics.

Jiang (2001) proposed a goodness of fit to check the distributional assumption of either the random effects or the errors. It would be more difficult to implement in practice because it would have to be implemented using Monte Carlo (MC) methods. In addition, the simulation study presented is limited because of its focus on the situation where there are small numbers of observations per profile.

Houseman, Ryan, and Coull (2004) proposed to assess the assumption of normally distributed errors via graphical methods. They noted that the approach of Pinheiro and Bates (2000) can be misleading because of the shrinkage of the estimated random effects. Houseman, Ryan, and Coull (2004) proposed to rotate the PA residuals ($\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}$) by premultiplying them by the Cholesky decomposition of the inverse of the estimated variance-covariance matrix. The rotated residuals can then be displayed on a Q-Q plot to determine normality.

4.9.5 Presence of Correlation

Dawson, Gennings, and Carter (1997) proposed graphical methods to determine the appropriate structure of the covariance matrix of errors for the LMM. After centering and scaling the data, they created a draftman's display and a parallel axis plot, which are complementary plots to determine the amount of correlation between successive observations and whether or not that amount is constant. Dawson, Gennings, and Carter (1997) also gave examples of the appearance of the two plots for independent data as well as data that follows an AR or CS structure. The two plots give similar results and while these plots appear to be easily interpretable for smaller dimensions, they get more unwieldy as p increases.

Demidenko (2004, Section 4.3.4) proposed a LRT to test if the autocorrelation of within profile measurements is zero. A LMM with uncorrelated errors (i.e. $\mathbf{R}_i = \sigma^2\mathbf{I}$) is nested within a LMM with an AR(1) error structure. Thus, both models can be fit and the difference in the likelihoods forms the basis of the LRT test. Demidenko (2004, Section 4.3.4) shows an example of this test and found that when the autocorrelation is small, little difference will exist for the estimates obtained by modeling the autocorrelation versus ignoring the autocorrelation. Chi and Reinsel (1989) proposed a score test of correlation of the errors.

4.10 Diagnostics in LMM

Diagnostic methods to detect outliers and influential points have been proposed in LMM but they are not well developed. The need for better or more utilized diagnostics for models with random effects and/or correlated errors has been noted by a number of authors, including, Ghosh and Rao (1994), Verbeke and Molenberghs (2000), Tan, Ouwens, and Berger (2001), Houseman, Ryan, and Coull (2004) and Haslett and Dillane (2004). Diagnostic methods are needed to detect outlying profiles as well as outlying observations within profiles. The goodness-of-fit methods discussed in the previous section are more appropriate for determining if observations within a profile are similar to each other and well described by the selected model. Our focus is determining outlying profiles rather than observations within a profile. As noted by Langford and Lewis (1998), once the outlying profile is determined, it can be examined for outlying observations.

There is a wide variety of methods for determining outlying profiles in LMM. In our review of the diagnostic methods for LMM, we found four methods for determining influential or outlying data in the LMM which we label the case deletion approach, local influence approach, bootstrap approach, and the distance approach.

Banerjee and Frees (1997) extended the case deletion diagnostic approach for linear regression (Cook, 1977) to the LMM. While this extension does not allow for complete deletion of a subject, it does allow for determination of the partial influence of a subject on the estimated parameters. The measure of partial influence on the fixed parameter estimates defined by Banerjee and Frees (1997, equation 11) is a distance measure and will have a similar form as our T^2 statistic to be shown later in (5.2) and (5.5). Tan, Ouwens, and Berger (2001) considered a Cook's distance measure for the LMM and found that it does not work well in determining the correct outlying profile. Thus they proposed that Cook's distance be calcu-

lated conditional on the obtained random effects and showed that the modification improves the effectiveness of the measure.

Alternatively, Demidenko and Stukel (2005) developed methods of detecting influential profiles and/or outliers by deriving a form of leverage and Cook's distance measures for the LMM. While these measures could be computationally difficult, Demidenko and Stukel (2005) showed how the measures can be more easily computed with an updating formula if the variance-covariance matrix of the random effects were known. If the variance-covariance matrix is not known, it can be replaced by its estimate obtained via ML. Asymptotic results ensure that the diagnostic measures will still work well as long as the number of profiles is sufficiently large. Once the influential profiles are identified, they can be examined for influential observations within the profiles.

The local influence approach was proposed by Lesaffre and Verbeke (1998) and is the extension of ideas from Cook (1986) to the LMM. A local influence approach gives weights to each profile and determines the change in parameter estimates as the weights change. In contrast, a global influence approach corresponds to a method that completely removes an outlier to determine its effect, as is the case for many case deletion schemes. The local influence measure reflects how much the log likelihood changes due to a particular profile. It also has the advantage of being decomposable into interpretable components. These components relate the influential point to how well it is predicted by the model, how well the covariance structure is model, the size of the random effects, or how large the PA residuals are. A case study of the local influence approach can be found in Lesaffre, Asefa, and Verbeke (1999).

A parametric bootstrap approach was proposed by Longford (2001). Once the model has been fit to the real data, simulated datasets based on the model fit were generated and a

comparison is made of the real dataset to the simulated datasets. If the real dataset is not fit well by the model, then it will stand out when compared to the simulated datasets. Longford (2001) proposed to do this with a global influence approach where a LRT statistic is used to assess outlying profiles. However, like other one-at-a-time deletion schemes this approach will not work well when multiple outliers are similar to each other and mask each other.

The final approach determines outlying profiles based on the distance of the estimated parameter vector from the center of the group of estimated parameter vectors. It was used by Waternaux, Laird, and Ware (1989), who proposed detecting outlying profiles by using the Mahalanobis distances of the eblups. They proposed to calculate

$$T^2_{varbi,i} = \widehat{\mathbf{b}}_i \, 'Var(\widehat{\mathbf{b}}_i - \mathbf{b}_i) \, \widehat{\mathbf{b}}_i \text{ for } i = 1, 2, \ldots, m, \tag{4.15}$$

where $Var(\widehat{\mathbf{b}}_i - \mathbf{b}_i)$ is calculated from the expression in Harville (1976) or Laird and Ware (1982). Waternaux, Laird, and Ware (1989) proposed to use a Q-Q plot of the values for $T^2_{varbi,i}$ to detect outliers. We calculate the Mahalanobis distance as in (4.15) with different estimators of the variance-covariance matrix. We will evaluate the method of Waternaux, Laird, and Ware (1989) in our simulation studies of Chapter 5 to determine its efficacy.