Chapter 2

Linear Mixed Models (LMMs)

2.1 Introduction

There are a number of books that cover the details of linear mixed models, including McCulloch and Searle (2000); Searle et al. (1992); Verbeke and Molenberghs (2000) so I will not attempt to cover the topic in detail here. However, I do want to point out several facets of linear mixed models and their estimation that are relevant to generalized linear mixed models and establish some basic notation. Again, I begin with an example, this one quite simple.

2.2 Example: Propranolol and hypertension

Below (Table 2.1) are data from an early, double-blind trial of the effect of a drug, Propranolol, on hypertension. Blood pressure was measured after administration of the drug and a placebo both in the upright and recumbent positions. There are two main questions of interest. First, does Propranolol have the same influence in recumbent and upright positions (i.e, is there a lack of interaction) and second, if the answer to the first question is yes, is it effective?

If we let Y_{ijk} denote the blood pressure measurement on the kth individual, ith position and jth drug condition, then the standard model for such an analysis is

$$\begin{aligned} Y_{ijk}|p_k &\sim \text{indep. } \mathcal{N}(\mu_{ijk}, \sigma^2), \\ \mu_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + p_k, \end{aligned} \tag{2.1}$$

where the vertical bar between Y_{ijk} and p_k indicates that the specification is conditional on the p_k. To this we add the important assumption that the person effects, p_k, follow a distribution:

$$p_k \sim \text{i.i.d. } \mathcal{N}(0, \sigma_p^2). \tag{2.2}$$

The mere declaration of the p_k as random variables, as contrasted with treating them as fixed, unknown parameters, induces a correlation between measurements

TABLE 2.1.
Data for the Propranolol/hypertension example

	Blood Pressure (mmHg)				
	Recumbent		Upright		
Patient	**Placebo**	**Drug**	**Placebo**	**Drug**	**Average**
1	96	71	73	87	81.75
2	96	85	104	76	90.25
3	92	89	83	90	88.50
4	97	110	101	85	98.25
5	104	85	112	94	98.75
6	100	73	101	93	91.75
7	93	81	88	85	86.75
Ave	96.86	84.86	94.57	87.14	90.86

taken on the same person. This is easily shown by using an identity similar to iterated expectations, namely

$$\text{Cov}(X, Y) = \text{E}[\text{Cov}(X, Y|Z)] + \text{Cov}(\text{E}[X|Z], E[Y|Z]). \tag{2.3}$$

In the present context we use (2.3) by conditioning on p_k as follows:

$$\begin{aligned} \text{Cov}(Y_{ijk}, Y_{i'j'k}) &= \text{E}[\text{Cov}(Y_{ijk}, Y_{i'j'k}|p_k)] + \text{Cov}(\mu_{ijk}, \mu_{i'j'k}) \\ &= 0 + \text{Cov}(\mu_{ijk}, \mu_{i'j'k}) \\ &= \text{Cov}(p_k, p_k) \\ &= \sigma_p^2. \end{aligned} \tag{2.4}$$

The second equality in (2.4) is true since the Y_{ijk} are assumed to be conditionally independent and the third follows since the only random quantities in the conditional mean are the p_k. Similar calculations give the variance of Y_{ijk} or $Y_{i'j'k}$ as $\sigma_p^2 + \sigma^2$ so that the correlation is

$$\text{Corr}(Y_{ijk}, Y_{i'j'k}) = \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2}, \tag{2.5}$$

the well-known intraclass correlation result (Snedecor and Cochran, 1989).

2.3 Fixed versus random factors

This distinction, between treating a term in the model as a random variable as opposed to a fixed, unknown constant is so crucial to the remaining development that it is worth some emphasis. We make the following definition

DEFINITION. If a distribution is assumed for the levels of a factor it is a *random factor*. If the values are fixed, unknown constants it is a *fixed factor*.

Declaring a factor to be random has several ramifications

- *Scope of inference.* Inferences can be made on a statistical basis to the population from which the levels of the random factor are assumed to have been selected.
- *Incorporation of correlation in the model.* Observations that share the same level of the random effect are being modeled as correlated.
- *Accuracy of estimates.* Using random factors involves making extra assumptions but gives more accurate estimates.
- *Estimation method.* Different estimation methods must be used as compared to regular regression or analysis of variance.

In the Propranolol example, we are almost certainly willing to assume that the p_k follow a distribution. It is unlikely that we would be satisfied in drawing conclusions only about the seven subjects in the experiment. We will almost certainly be willing to assume that the patients in our study can be regarded as a random sample from some larger population of patients (appropriately defined). We may give more pause to the specific assumption of a normal distribution, and we return to this topic in later chapters.

If we are willing to make the random sampling assumption then the effects in the model associated with the patients (the p_k) can also be regarded as a random sample. Further, a main reason for conducting an experiment such as this one by measuring the same subject under all four conditions is to gain the advantage of making within-subject comparisons. This is a more precise comparison because it exploits the within-subject correlation of the measurements.

Thus the key step is to answer the question: "Am I willing to assume the effects come from a distribution?" If the answer is yes, the factor is considered a random factor.

2.4 Estimation and prediction

Traditional analysis of the Propranolol example would be by analysis of variance with F-tests for the effects of position, drug and their interaction. This is a perfectly good analysis for a simple, balanced situation. However, as soon as the structure of the random effects becomes complicated and/or the data are unbalanced, then the traditional methods become approximate and inefficient. From a theoretical point of view, the statistics captured in the usual ANOVA table, on which all further calculations are based (such as the F-test), are no longer sufficient statistics. This makes methods such as maximum likelihood more attractive.

Indeed, this is exactly what more modern statistical procedures, such as SAS Proc MIXED, do. They calculate the likelihood or variants (such as restricted maximum likelihood) and base estimation and tests on those calculations. This has the significant advantage of being able to handle quite complicated correlated data structures as well as unbalanced data. Of course, it leads to much more difficult statistical computing.

a. A more general formulation

Before describing selected results concerning linear mixed models I consider a more general form of the linear mixed model. The model (2.1) has terms describing both fixed $(\mu, \alpha_i, \beta_j, (\alpha\beta)_{ij})$ and random (p_k) factors, which enter the model in the same fashion. If we let $\mathbf{X}$ represent the model matrix for the fixed effects, $\boldsymbol{\beta}$ the fixed effects parameters, $\mathbf{Z}$ the model matrix for the random effects, and $\mathbf{u}$ the random effects, then we can write a more generic version of the mixed model as

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \end{aligned} \tag{2.6}$$

where, rather obviously, $\mathbf{R}$ represents the variance-covariance matrix of $\mathbf{Y}$ conditional on $\mathbf{u}$ and $\mathbf{D}$ is the variance-covariance matrix of $\mathbf{u}$. For the Propranolol example, $\mathbf{X}$ and $\mathbf{Z}$ would be indicator matrices of zeros and ones, while $\mathbf{R} = \mathbf{I}\sigma^2$ and $\mathbf{D} = \mathbf{I}\sigma_p^2$.

b. Means and variances

From (2.6) and the iterated expectation identities, it is straightforward to calculate the mean and variance-covariance matrix of $\mathbf{Y}$. The mean is given by

$$\begin{aligned} \mathrm{E}[\mathbf{Y}] &= \mathrm{E}\left[\mathrm{E}[\mathbf{Y}|\mathbf{u}]\right] \\ &= \mathrm{E}\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\right] \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathrm{E}[\mathbf{u}] \\ &= \mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{2.7}$$

The variance-covariance matrix is given by

$$\begin{aligned} \mathrm{Cov}(\mathbf{Y}) &= \mathrm{E}\left[\mathrm{Cov}(\mathbf{Y}|\mathbf{u})\right] + \mathrm{Cov}\left(\mathrm{E}[\mathbf{Y}|\mathbf{u}]\right) \\ &= \mathbf{R} + \mathrm{cov}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) \\ &= \mathbf{R} + \mathrm{Cov}(\mathbf{Z}\mathbf{u}) \\ &= \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}'. \end{aligned} \tag{2.8}$$

Therefore $\mathbf{V} \equiv \mathrm{Cov}(\mathbf{Y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$.

In practical applications of linear mixed models, decisions have to be made as to what form to specify for $\mathbf{D}$ and $\mathbf{R}$. The random effects $\mathbf{u}$, through their variance-covariance matrix $\mathbf{D}$, are typically used to describe correlation attributable to specific entities, such as subjects, or sites, while the matrix $\mathbf{R}$ is used to describe correlation structure that remains (if any), often due to temporal or spatial correlation. So, in a model with repeated measurements over time on subjects, we might have a model with both subject random effects (to describe stable-over-time characteristics that lead to long term correlation of measurements on a subject as captured in $\mathbf{D}$) as well as an autoregressive (or other time series type error structure) to describe the additional correlation of measurements taken over time.

A common simplification for (2.6) occurs when the random effects vector describes several, independent random effects. For example, we might study the performance of students, taught in classes, within schools. So we could envision

an analysis with random effects for students, classes and schools. We might be willing to assume that within each of the vectors of effects associated with students, classes and schools the effects are independent and identically distributed and that each vector of effects is independent of one another. If we order the vector $\mathbf{u}$ by students and then classes and then schools, $\mathbf{D}$ would take the block diagonal form $\mathbf{D} = \text{diag}\{\mathbf{I}\sigma^2_{\text{student}}, \mathbf{I}\sigma^2_{\text{class}}, \mathbf{I}\sigma^2_{\text{school}}\}$. In these cases, and using r to denote the number of distinct random effects and $\mathbf{Z}_i$ and $\mathbf{u}_i$ to represent the separate model matrices and random effects, it is often convenient to rewrite (2.6) to display that fact, as follows:

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i, \mathbf{R}\right), \\ \mathbf{u}_i &\sim \text{ indep. } \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_i^2), \end{aligned} \tag{2.9}$$

c. Estimation and prediction

Some of the relevant estimation and prediction techniques are somewhat easier to describe under (2.9) and assuming $\mathbf{R} = \mathbf{I}\sigma^2$. If we take derivatives of the log likelihood for (2.9) and set them equal to zero (see Searle et al., 1992, for details) then the resulting equations are

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{Y}, \\ \text{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right) &= \mathbf{Y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Y} \qquad \text{for } i = 1, 2, \ldots, r, \end{aligned} \tag{2.10}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. The first equation can be solved explicitly for $\hat{\boldsymbol{\beta}}$, if $\mathbf{V}$ is known, but the second set of equations cannot be solved explicitly for the elements of $\mathbf{V}$ and it usually dealt with iteratively.

d. Best prediction of the random effects

As noted in Chapter 1, the best predicted value (in the sense of minimum mean squared error of prediction) of a random effect $\mathbf{u}$ given the data $\mathbf{Y}$ is $\text{E}[\mathbf{u}|\mathbf{Y}]$. A variation on the idea of a best predictor (BP) is that of a best linear unbiased predictor (BLUP). More formally:

DEFINITION. A BLUP, $\tilde{\mathbf{u}}_{blup}$, minimizes the MSE of prediction among linear unbiased predictors:

$$\text{minimize } \text{E}[(\tilde{\mathbf{u}}_{blup} - \mathbf{u})^2]$$

among $\tilde{\mathbf{u}}_{blup}$ which are linear in $\mathbf{Y}$ and for which $\text{E}[\ \tilde{\mathbf{u}}_{blup}$ - $\mathbf{u}\ ] = \mathbf{0}$.

For linear mixed models the best predictor is of the form (Searle et al., 1992)

$$\tilde{\mathbf{u}}_{bp} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.11}$$

while the best linear unbiased predictor is

$$\tilde{\mathbf{u}}_{blup} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.12}$$

We can thus see what adding on the requirements of linearity and unbiasedness gain us: the theory tell us to replace the value of $\boldsymbol{\beta}$ with its ML estimator. This resolves part of the problem of the BP depending on unknown parameters, but there is still the issue that it depends on an unknown $\mathbf{V}$.

i. Prediction with a balanced data set

For a balanced data situation like that of the Propranolol data, the form of the best predictor and best linear unbiased predictor are relatively simple and informative. Consider prediction of p_k from (2.1). The best predictor is given by

$$\text{BP}(p_k) = \tilde{p}_{k,bp} = \text{E}[p_k|\bar{Y}_{\cdot\cdot k}], \tag{2.13}$$

where $\bar{Y}_{\cdot\cdot k}$ is the mean of the data for the kth person. Now p_k and $\bar{Y}_{\cdot\cdot k}$ are jointly normally distributed with a covariance given by [utilizing (2.3) again]

$$\begin{aligned} \text{Cov}(p_k, \bar{Y}_{\cdot\cdot k}) &= \text{Cov}(p_k, \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + p_k) \\ &= \text{Cov}(p_k, p_k) \\ &= \sigma_p^2. \end{aligned} \tag{2.14}$$

Since p_k and $\bar{Y}_{\cdot\cdot k}$ are jointly normal, we can use the usual formula for the conditional mean for a bivariate normal distribution to calculate the BP:

$$\begin{aligned} \text{BP}(p_k) &= \text{E}[p_k] + \text{Cov}(p_k, \bar{Y}_{\cdot\cdot k})\text{Var}^{-1}(\bar{Y}_{\cdot\cdot k})(\bar{Y}_{\cdot\cdot k} - \text{E}[\bar{Y}_{\cdot\cdot k}]) \\ &= 0 + \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}(\bar{Y}_{\cdot\cdot k} - \bar{\mu}) \\ &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}(\bar{Y}_{\cdot\cdot k} - \bar{\mu}) \end{aligned} \tag{2.15}$$

where n_k is the number of observations contributing to $\bar{Y}_{\cdot\cdot k}$, (equal to 4 in the Propanolol example) and where $\bar{\mu} \equiv \mu + \bar{\alpha}_\cdot + \bar{\beta}_\cdot + \overline{(\alpha\beta)}_{\cdot\cdot}$ and with the bars over the symbols representing averages over the subscript(s) that have been replaced with dots.

The best linear unbiased predictor is more practical, since it replaces the unknown $\bar{\mu}$ with $\bar{Y}_{\cdot\cdot\cdot}$:

$$\text{BLUP}(p_k) = \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot}). \tag{2.16}$$

In either case, the form of the predictor is interesting. It takes the naive estimator for an individual person effect, $\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot}$ and shrinks it by the factor $\sigma_p^2/(\sigma_p^2 + \sigma^2/n_k)$. How does this shrinkage factor behave? If there is large person to person variation (in relation to σ^2/n_k), suggesting that people are quite different, then the multiplier is about 1 and little shrinkage takes place. Likewise, if n_k is large, little shrinkage takes place. On the other hand, if σ_p^2 is relatively small and/or n_k is small then the shrinkage can be sizable. This is the so-called "borrowing of strength" in which information is "borrowed" either if people are similar or if the sample size is small and benefit can be had by learning from the rest of the sample.

Another viewpoint is in predicting the mean value for person k:

$$\begin{aligned} \text{BLUP}(\bar{\mu} + p_k) &= \bar{Y}_{...} + \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}(\bar{Y}_{..k} - \bar{Y}_{...}) \\ &= \left(1 - \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}\right)\bar{Y}_{...} + \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n_k}\bar{Y}_{..k} \\ &= (1-\alpha)\bar{Y}_{...} + \alpha\bar{Y}_{..k}, \end{aligned} \tag{2.17}$$

where $\alpha = \sigma_p^2/(\sigma_p^2 + \sigma^2/n_k)$. This shows that the predicted mean value for person k is a weighted average of the overall mean and the individual specific mean.

In practice it is common (Harville, 1991) to use a "plug-in" estimator and insert the estimated values of the variance and covariance parameters into the equation for the BP or BLUP. We call this the EBLUP (for estimated or empirical BLUP):

$$\text{EBLUP}(p_k) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n_k}(\bar{Y}_{..k} - \bar{Y}_{...}). \tag{2.18}$$

Here is a numerical illustration using the Propranolol example.

Numerical illustration.

$$\begin{aligned} \text{EBLUP}(p_k) &= \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n_k}(\bar{Y}_{..k} - \bar{Y}_{...}) \\ &= \frac{15.7976}{15.7976 + 85.7976/4}(81.75 - 90.86) \\ &= 0.424(-9.11) \\ &= -3.863, \end{aligned} \tag{2.19}$$

showing shrinkage of the raw estimate from -9.11 to -3.863.

Similarly,

$$\begin{aligned} \text{EBLUP}(\bar{\mu} + p_k) &= \bar{Y}_{...} + \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n_k}(\bar{Y}_{..k} - \bar{Y}_{...}) \\ &= 90.86 - 3.863 \\ &= 86.99, \end{aligned} \tag{2.20}$$

which is much closer to the overall mean of 90.86 than is the raw mean for the first patient, which is 81.75.

A primary message here is that we can declare a factor to be random but still be interested in and have the ability to obtain predictions for the specific levels of that factor.

What about standard errors for the EBPs or EBLUPs? This is somewhat difficult since, for most designs, the sampling distribution of the estimated variance components is unknown and therefore difficult to factor in to the calculations of variances and SEs (see, e.g., Kackar and Harville, 1984; Harville and Jeske, 1992).

2.5 The mixed model equations

Charles Henderson from Cornell University (an animal breeder!) made a most remarkable discovery in the 1950s (Henderson, 1953) in which he developed what have become known as the mixed model equations. This is a compact set of equations for simultaneously estimating the BLUPs and the MLE of $\boldsymbol{\beta}$. We return to the model (2.6) but with the simplification that $\mathbf{R} = \mathbf{I}\sigma^2$, namely,

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{I}\sigma^2), \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}). \end{aligned} \tag{2.21}$$

The "mixed model equations" are given by

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{D}^{-1}\sigma^2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}}_{blup} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}. \tag{2.22}$$

These can be solved for $\boldsymbol{\beta}$ and $\tilde{\mathbf{u}}_{blup}$ if the values of $\mathbf{D}$ and σ^2 are known. Of course, assuming $\mathbf{D}$ and σ^2 to be known is unrealistic and so, in practice, the mixed model equations are supplemented with an estimation equation for $\mathbf{D}$ and σ^2 and the equations are solved iteratively.

Here is more detail for the simple case where there is a single random effect with $\mathbf{D} = \mathbf{I}\sigma_u^2$. For that case, it is straightforward to show that the best linear unbiased predicted value is given by (Searle et al., 1992)

$$\tilde{\mathbf{u}}_{blup} = \sigma_u^2 \mathbf{Z}'\mathbf{P}\mathbf{y}, \tag{2.23}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}$.

The equation to solve for the MLE of σ_u^2 is (again see Searle et al., 1992)

$$\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') = \mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{Z}'\mathbf{P}\mathbf{y} \tag{2.24}$$

or equivalently

$$\begin{aligned} \sigma_u^4 \mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') &= \sigma_u^4 \mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{Z}'\mathbf{P}\mathbf{y} \\ &= \tilde{\mathbf{u}}'_{blup}\tilde{\mathbf{u}}_{blup}. \end{aligned} \tag{2.25}$$

So a subsidiary equation using the BLUPs to form a new estimate of the variance component, σ_u^2, that can be used is

$$\sigma_u^{2(m+1)} = \frac{\tilde{\mathbf{u}}'_{blup}\tilde{\mathbf{u}}_{blup}}{\sigma_u^{2(m)}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}')}, \tag{2.26}$$

where (m) indicates the round of iteration in the algorithm.

Slightly more formally then, an algorithm is:

1. Obtain starting values for the variance components.
2. Solve the MMEs (2.22) for $\boldsymbol{\beta}$ and $\tilde{\mathbf{u}}_{blup}$.
3. Solve (2.26) for σ_u^2 and a similar equation for σ^2.
4. Iterate until convergence.

TABLE 2.2.
ANOVA table for a one-way random model

	ANOVA		
Source	**d.f.**	**Mean Square**	**E[Mean Square]**
Between	$k-1$	MS(Betw)	$\sigma^2 + n\sigma_u^2$
Error	$k(n-1)$	MS(Error)	σ^2

2.6 Testing fixed effects

Testing of fixed effects in linear mixed models has been well covered in McCulloch and Searle (2000) (Section 6.4) and so I will not elaborate further here. The results for linear mixed models do not bear centrally on the discussion for generalized linear mixed models.

2.7 Testing random effects

When using a maximum likelihood analysis the typical tests are based on the improvement in the maximized value of the log likelihood. The difference in twice the log likelihood is compared to a chi-square distribution to test for statistical significance. For testing whether a single variance component is equal to zero the usual method must be slightly modified. Ordinarily we would take the difference in log likelihoods of the models with and without the random effect and compare that directly to a χ^2_1 critical value. The modification is to either calculate a p-value and then cut it in half, or to compare to a cutoff point with twice the nominal α level.

Why? The intuition is that testing

$$\text{H}_0 : \sigma_u^2 = 0 \quad \text{versus} \quad \text{H}_1 : \sigma_u^2 > 0 \tag{2.27}$$

is a one-sided test. The usual likelihood ratio test is inherently two-sided and must be adjusted to reflect this fact.

Specifically, consider the ANOVA and ML estimators of σ_u^2 in a balanced, one-way random model, both which are based on an ANOVA, such as that exemplified in Table 2.2:

ANOVA estimator:

$$\hat{\sigma}_u^2 = (\text{MS(Betw) - MS(Error)})/n \tag{2.28}$$

ML estimator:

$$\hat{\sigma}_u^2 = \left[\left(1 - \frac{1}{k}\right)\text{MS(Betw) - MS(Error)}\right]^+ /n, \tag{2.29}$$

where $[\cdot]^+$ denotes positive part.

If $\sigma_u^2 = 0$, then the ML estimator is often zero. In that case, the likelihood ratio

test (LRT) statistic is given by

$$\begin{aligned} LRT &= -2[\log L(\sigma_u^2 = 0) - \log L(\sigma_u^2 = \hat{\sigma}_u^2)] \\ &= -2[\log L(\sigma_u^2 = 0) - \log L(\sigma_u^2 = 0)] \\ &= 0. \end{aligned} \tag{2.30}$$

About half the time the estimate would be zero and the LRT statistic would be zero. With a point mass of approximately 0.5 at 0, the usual asymptotic distribution theory (suggesting a χ_1^2 distribution) clearly breaks down because the estimate gets "stuck" on the boundary.

So, the actual large-sample distribution under H_0 is a 50:50 mixture of a χ_1^2 and 0. Operationally, we would calculate the p-value under the assumption of χ_1^2 and cut the p-value in half! More detail can be found in Self and Liang (1987) and Stram and Lee (1994). As elaborated briefly in Chapter 9, the situation for more than a single variance component is more complicated.

a. Numerical illustration for the Propranolol data

If the model is fit with patient as a random effect it gives a value for -2 times the restricted log likelihood of 186.0517. Fit with no random effect, the value is 186.7966, with a difference of $186.7966 - 186.0517 = .7944$. This gives a p-value of $p = \Pr\{\chi_1^2 > 0.7944\} = 0.388$. Cut in half gives 0.194. Equivalently, we can compare to a chi-square cutoff of $\chi^2_{1,0.90} = 2.71$ instead of $\chi^2_{1,0.95} = 3.84$.

2.8 Generalized estimating equations

We now consider a different method of estimation, called Generalized Estimating Equations (GEEs). Again we cover only relevant results. More detail can be found in the excellent book by Diggle et al. (1994), from which this example is taken.

a. Example: milk, cows and diets

Milk was collected from 79 cows on one of three diets: barley, lupins and a mixture of both. Protein content of the milk was recorded weekly for 19 weeks after the earliest calving. Let Y_{ijt} denote the protein content of the milk from the ith cow on the j diet at time t. We consider the following model for $\mathbf{Y}_{ij}$, defined as the vector of measurements on cow i in diet j:

$$\begin{aligned} \mathbf{Y}_{ij}|\mathbf{c} &\sim \text{indep. } \mathcal{N}(\boldsymbol{\mu}_{ij}, \mathbf{R}_{ij}) \\ \mu_{ijt} &= \mu + c_{ij} + f(t) + \alpha_j, \\ \mathbf{R} &= [r_{ijt,ijt'}] = [\sigma^2 \exp(\phi|t - t'|)] \\ \mathbf{c} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_c^2), \end{aligned} \tag{2.31}$$

where $f(t)$ is a complicated function of time, which is not of interest in our discussion. This model has fixed effects of diet (α_j) and time ($f(t)$) and random cow effects (c_{ij}). Further correlation (above that introduced by the cow random effects)

is introduced by having a temporal correlation captured in $\mathbf{R}_{ij}$, which falls off exponentially as the time points become farther and farther apart. What would be an alternative to maximum likelihood as a way to form estimates for this model?

b. Weighted estimation

I return to our more general mixed model temporarily for this discussion. Suppose we believe that (2.6) holds. What consequence would it have to use ordinary least squares estimation for estimating $\boldsymbol{\beta}$? That is, how would $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ perform?

It is a standard linear models exercise to show that $\hat{\boldsymbol{\beta}}_{ols}$ is unbiased:

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}_{ols}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{2.32}$$

It is also straightforward to calculate its variance as

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}) &= \mathrm{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{2.33}$$

Furthermore, $\hat{\boldsymbol{\beta}}_{ols}$ is usually fairly efficient as compared to the weighted least squares estimator, $\hat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, which has variance $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. In fact, for balanced designs, $\hat{\boldsymbol{\beta}}_{ols} = \hat{\boldsymbol{\beta}}_{wls}$ (Searle et al., 1992) so the variances are identical and $\hat{\boldsymbol{\beta}}_{ols}$ is fully efficient. For some detailed calculations see Chapter 8.

So why not just use $\hat{\boldsymbol{\beta}}_{ols}$ and standard software? The answer is that even though $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ is close to optimal, the problem is that, using standard software, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols})$ is estimated as $(\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2$, which will often be very wrong. That is, the OLS estimate isn't so bad, but the usual variance estimate is way off.

Consider the longitudinal data setting with $\mathbf{Y}_i$ distributed independently across subjects (generically interpreting "subjects"). The basic idea is to use the replication across subjects to get an empirical estimate of the variance. For this setting, where we have

$$\begin{aligned} \mathbf{Y_i} &\sim \text{indep. } \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{V}_i), \\ \boldsymbol{\mu_i} &= \mathbf{X}_i\boldsymbol{\beta}, \end{aligned} \tag{2.34}$$

$\hat{\boldsymbol{\beta}}_{ols}$ simplifies to $(\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i'\mathbf{Y}_i)$ with variance

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}) = \left(\sum_i \mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_i \mathbf{X}_i'\mathbf{V}_i\mathbf{X}_i\right)\left(\sum_i \mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \tag{2.35}$$

This can be estimated consistently (Liang and Zeger, 1986) by

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}_{ols}) = \left(\sum_i \mathbf{X}_i'\mathbf{X}_i\right)^{-1} \left(\sum_i \mathbf{X}_i'(\mathbf{X}_i - \hat{\mu}_i)(\mathbf{X}_i - \hat{\mu}_i)'\mathbf{X}_i\right) \times \left(\sum_i \mathbf{X}_i'\mathbf{X}_i\right)^{-1}, \tag{2.36}$$

the so-called "sandwich" or "robust" variance estimator.

For the milk protein data from Diggle et al. (1994), if all the animals had all 19 weeks of data, we could view the data as multivariate of dimension 19. A straightforward estimate of the covariance matrix would then be the sample variance-covariance matrix using the estimated means. In this example, as in many, there is incomplete, or missing, data. In such a case, the sample variance-covariance matrix is not as attractive to use, but (2.36) is still applicable.

The sandwich estimator has the significant advantage and is robust in the sense that it gives consistent estimates of the variance-covariance matrix, even when the variance-covariance structure is misspecified. However, Carroll et al. (1995) object to the term "robust" since that usually implies little loss of efficiency when the model is incorrect. They note that the loss of efficiency can be substantial. See Kauermann and Carroll (2001) and Drum and McCullagh (1993) for details.

2.9 Summary

This chapter has introduced the important idea of a random effect and its use to incorporate correlation into a model in a tangible and easily-understood manner. Once random effects are incorporated, the model is a correlated data model and it is not too surprising that estimation becomes more complicated. It also raises two new inferential goals. First, testing whether the random effect distribution is degenerate (so that it can be treated as a fixed effect). That is, is the source of correlation in the model represented by the random effect statistically insignificant? Second, prediction of the realized values of the random effects. Closely tied to the prediction of random effects is the idea of a shrinkage estimator. The chapter closed with a brief look at the idea of generalized estimation equations and the notion of a sandwich estimator of the variance.