

UNIVERSITY OF WOLLONGONG
STAT902. Advanced Data Analysis

Mixed Models

These notes are an excerpt (Chapter 4) from the book *Semiparametric Regression* by D. Ruppert, M.P. Wand & R.J. Carroll.

©David Ruppert, M.P. Wand, R.J. Carroll 2003

Mixed Models

1.1 Introduction

Mixed models are an extension of regression models that allow for the incorporation of *random effects*. However, they also turn out to be closely related to smoothing. In fact, we will show in Section 1.9 that the penalized spline smoother exactly corresponds to the optimal predictor in a mixed model framework. This link allows for mixed model methodology and software to be used in semiparametric regression analysis, as we will demonstrate in subsequent chapters.

This chapter begins with a brief review of mixed models. Readers with detailed knowledge of mixed models could skip these sections and proceed directly to Section 1.9.

1.2 Mixed Models

Much of the early work on mixed models, in particular the special case of *variance component* models, was motivated by the analysis of data from animal breeding experiments, driven by the need to incorporate heritabilities and genetic correlations in a parsimonious fashion. They have also played an important role in establishing quality control procedures and determination of sampling designs, among other applications. Overviews of this vast topic are provided by Searle, Casella and McCulloch (1992), Vonesh and Chinchilli (1997), Pinheiro and Bates (2000), McCulloch and Searle (2001) and Verbeke and Molenberghs (2000).

A more contemporary application of mixed models is the analysis of longitudinal data sets (e.g. Laird and Ware 1982, Diggle, Heagerty,

Liang and Zeger, 2002). We will use this setting to illustrate the essence of mixed modeling.

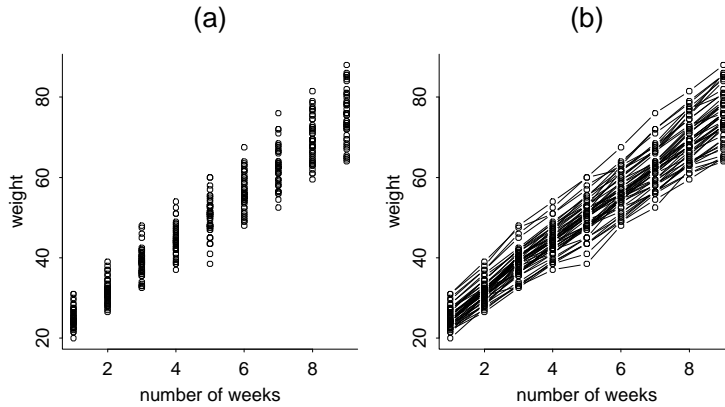


Figure 1.1: Representations of pig weight data. Panel (a) is a scatterplot of weight against week number. In (b) lines are used to connect those points pertaining to the same pig.

Figure 1.1 shows two representations of data pertaining to weight measurements of 48 pigs, for 9 successive weeks. Figure 1.1 (a) is simply a scatterplot of the weights against their corresponding week number. In Figure 1.1 (b) lines are drawn connecting those measurements that belong to the same pig. This second figure shows the longitudinal aspect of the data: there are 9 repeated measurements for each pig.

Let weight_{ij} denote the weight of pig i on week j , and let $\text{week}_j = j$ be the corresponding week number. If the data are treated *cross-sectionally*, i.e., without taking the repeated measure aspect into account, then the ordinary least squares model

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{week}_j + \varepsilon_{ij}, \quad 1 \leq i \leq 48, \quad 1 \leq j \leq 9, \quad (1.1)$$

with ε_{ij} i.i.d. $N(0, \sigma_\varepsilon^2)$, leads to a slope estimate (with estimated standard deviation) of

$$\hat{\beta}_1 = 6.21, \quad \widehat{\text{st.dev.}}(\hat{\beta}_1) = 0.0818.$$

But there are some problems with (1.1). First of all, inspection of Figure 1.1 (b) shows that the scatterplot *for each individual pig* is less variable, so one would expect that utilization of *within-pig* information would be beneficial. Related to this is the fact that (1.1) ignores the

Longitudinal data sets consist of measurements made on a set of individuals repeatedly over time.

correlation of measurements pertaining to the same pig. An analysis of the residuals shows patterns arising from this correlation, so the assumption that the ε_{ij} are independent does not hold.

An initial remedy would be to extend (1.1) to allow for a different intercept for each pig. This models the data shown in Figure 1.1 (b) as 48 parallel lines and would be written

$$\text{weight}_{ij} = \alpha_i + \beta_1 \text{week}_j + \varepsilon_{ij} \quad (1.2)$$

where α_i represents the intercept for the i th pig. This leads to a noticeably more precise estimate of β_1 , but model (1.2) is unappealing on a few counts. Firstly, it contains 49 parameters, 48 intercepts and 1 slope, which is somewhat large for such a simple data set. Secondly, it gives too much credence to the pigs used in the study. If 48 different pigs were sampled then the estimated α_i would be completely different. Normally we think of parameters being population dependent rather than sample dependent. Other longitudinal data sets have many more subjects and even fewer repeated measurements. So a model such as (1.2) is not very satisfactory when the data are longitudinal.

A remedy is to replace α_i by a *random intercept*:

$$\text{weight}_{ij} = \beta_0 + U_i + \beta_1 \text{week}_j + \varepsilon_{ij}. \quad (1.3)$$

Here

$$U_1, \dots, U_{48}$$

are treated as a random sample from, say, a $N(0, \sigma_U^2)$ distribution for some $\sigma_U^2 > 0$. The U_i term is an example of a *random effect* and has the advantage of requiring just a single parameter, σ_U^2 , which is commonly referred to as a *variance component*. Moreover, it takes into account the randomness due to other samples of pigs. For these data, one may want to consider a random slope model in which β_1 is replaced by $\beta_1 + V_i$, where V_i is a random effect that accounts for possible variability in the slopes of the growth curves. However, since we are using this example for illustrative purposes, we will assume that (1.3) is adequate.

Model (1.3) is an example of a *mixed model*. It has a *fixed component*

$$\beta_0 + \beta_1 \text{week}_j$$

and a *random component*

$$U_i \sim N(0, \sigma_U^2).$$

The next two sections describe estimation techniques for fitting (1.3). For these data they result in

$$\widehat{\beta}_1 = 6.21, \quad \widehat{\text{st.dev}}(\widehat{\beta}_1) = 0.0391$$

which is somewhat more precise than an ordinary regression model. Moreover, the random intercept U_i allows for the within-pig correlation. To appreciate this, note that the covariance between the weights of pig i , measured at two different times ($j \neq j'$), is

$$\text{cov}(\text{weight}_{ij}, \text{weight}_{ij'}) = \text{Var}(U_i) = \sigma_U^2.$$

The correlation coefficient is then

$$\text{corr}(\text{weight}_{ij}, \text{weight}_{ij'}) = \frac{\sigma_U^2}{\sigma_\varepsilon^2 + \sigma_U^2}.$$

This is estimated to be 0.775, indicating considerable within-pig correlation in this case.

1.2.1 Degrees of freedom interpretation

Consider the more general form of (1.3):

$$y_{ij} = \beta_0 + U_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m. \quad (1.4)$$

Section 1.5 describes how to fit this model. As discussed in that section, we can write the vector of fitted values as

$$\widehat{\mathbf{y}} = \mathbf{H}_0 \mathbf{y} + \mathbf{H}_U \mathbf{y} + \mathbf{H}_x \mathbf{y}$$

and, analogously to (??), define the degrees of freedom for each component of (1.4)

$$df_0 = \text{tr}(\mathbf{H}_0), \quad df_U = \text{tr}(\mathbf{H}_U) \quad \text{and} \quad df_x = \text{tr}(\mathbf{H}_x).$$

Not surprisingly,

$$df_0 = df_x = 1$$

since each of these components involve a single parameter; β_0 and β_1 respectively. When $n_1 = \dots = n_m$, it may be shown that

$$df_U = \frac{(m-1)n_1}{n_1 + \sigma_\varepsilon^2/\sigma_U^2}.$$

This shows that the effective number of parameters depends on the variance ratio

$$\sigma_\varepsilon^2/\sigma_U^2.$$

In the pig weight example,

$$df_U = 47 \left(\frac{9}{9 + \sigma_\varepsilon^2 / \sigma_U^2} \right).$$

Figure 1.2 shows this degrees of freedom plotted against $\sigma_\varepsilon^2 / \sigma_U^2$. Using restricted maximum likelihood (described in Section 1.5.4) this ratio is estimated to be $\hat{\sigma}_\varepsilon^2 / \hat{\sigma}_U^2 = 0.29$, corresponding to 46.5 degrees of freedom. We see from this that the random intercept model corresponds to a potential compromise between two possible models with fixed effects only

- (a) a single fixed intercept ($df_0 + df_U = 1$, $\sigma_U^2 = 0$). This is model (1.1).
- (b) an intercept for each pig (48 parameters. i.e., $df_0 + df_U = 48$, $\sigma_U^2 = \infty$). This is model (1.2).

In model (a), $\sigma_U = 0$ implies that $U_1, \dots, U_{48} = 0$ so all the random effects drop out of the model and we are left with a fixed effects model.

When $\sigma_U^2 = \infty$ in model (b), the interpretation is that U_1, \dots, U_{48} are no longer random but rather unknown fixed constants; the random effects have become fixed effects. Model (b) appears to have 49 parameters, $\beta_0, U_1, \dots, U_{48}$, but in this model one constraint is needed to make the parameters well-defined. Often this constraint is chosen to be $U_1 + \dots + U_{48} = 0$. Although the random intercept model is a potential compromise between (a) and (b), we see that in this example it is quite close to (b) because the random and fixed intercept models differ by only 1.5 degrees of freedom.

In this example the random intercept is closer to (b) than to (a) due to the within-pig variability being considerably lower than the between-pig variability.

There is an analogy between the \hat{U}_i 's and the $\hat{\beta}_{1k}$'s of Section ???. In both cases the estimates are shrunk in such a way that their contribution to the degrees of freedom of the fitted values is less than the number of coefficients. As will be clear by the end of this chapter, the fitting of both longitudinal data sets such as the pig weight data and nonlinear trends such as for the LIDAR data can be achieved through the same general approach.

1.3 Prediction

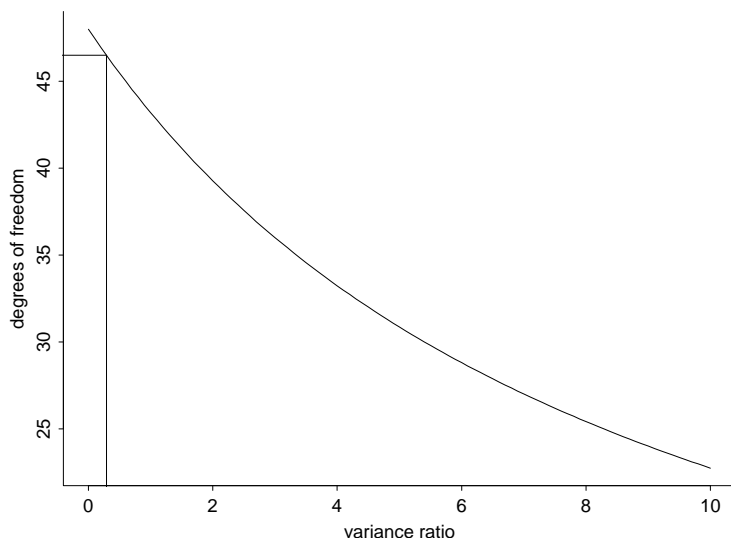


Figure 1.2: Plot of degrees of freedom for random intercepts (df_U) versus the variance ratio $\hat{\sigma}_\varepsilon^2 / \hat{\sigma}_U^2$ for the pig weight example. The lines correspond to the estimated values via REML as discussed in Section 1.5.4.

Mixed models contain fixed effects, random effects and covariance matrix parameters. For model (1.3) the fixed effects are β_0 and β_1 , the random effects are U_1, \dots, U_{48} and the covariance matrix parameters are σ_U^2 and σ_ε^2 . The *parameters* in the model are $(\beta_0, \beta_1, \sigma_U^2, \sigma_\varepsilon^2)$ and their estimation can be achieved through common statistical approaches such as maximum likelihood. This is treated for general linear mixed models in Sections 1.5.1 and 1.5.4. However, maximum likelihood does not apply to random effects. Instead we can form *predictions* of U_1, \dots, U_{48} . The difference between prediction and estimation is that, for the former, the target is random while, for the latter, it is deterministic (non-random). Some writers (e.g. Robinson, 1991; Hayes and Haslett, 1999) argue that this distinction is unnecessary and that the word “estimation” should be used for both types of targets. However, in accordance with the majority of relevant literature, we will use the classical naming convention here.

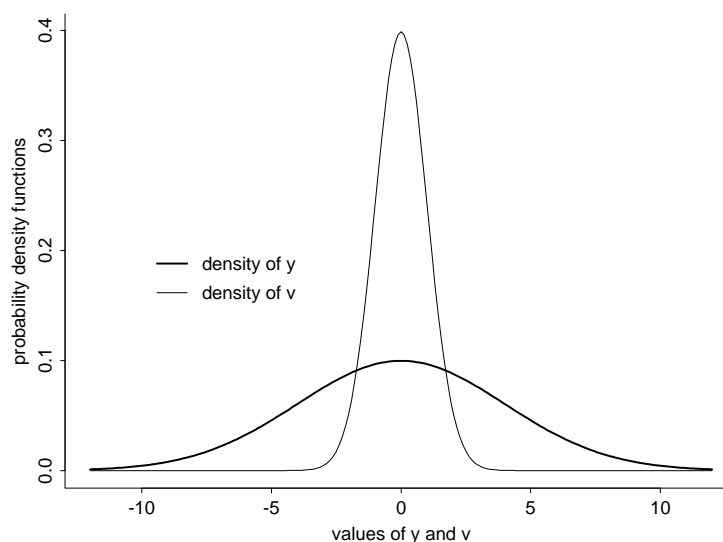
Prediction is a fundamental problem in Statistics, although its treatment in textbooks is overshadowed by estimation. An excellent synopsis of prediction is provided by Chapter 9 of McCulloch and Searle (2001). We will summarize the main points here. Figure 1.3 shows the distributions of two random variables y and v that are distributed according to

$$y = v + \varepsilon \quad \text{where} \quad \begin{bmatrix} v \\ \varepsilon \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \right).$$

Prediction is at the heart of our semiparametric and nonparametric methods.

We only observe y . Based on this observation what is a good predic-

Figure 1.3: Simple illustration of prediction. The value of y is observed, while v is not. The best predictor of v is $y/5$.



tion for the value of v ? The *best predictor* (BP) of v is defined to be the \tilde{v} for which

$$E\{(\tilde{v} - v)^2\}$$

is minimized. For general (y, v) the solution is

$$\tilde{v} \equiv \text{BP}(v) = E(v|y).$$

In the current example we obtain

$$\tilde{v} = y/5.$$

This is intuitively consistent with Figure 1.3 where v is seen to be a “shrunk” version of y .

In general, if \mathbf{y} is the vector of observed data and \mathbf{v} is a random vector then best prediction corresponds to minimization of

$$E\{\|\tilde{\mathbf{v}} - \mathbf{v}\|^2\}$$

and the solution is

$$\tilde{\mathbf{v}} \equiv \text{BP}(\mathbf{v}) = E(\mathbf{v}|\mathbf{y}).$$

1.3.1 Best linear prediction

!bestlinear

The best predictor is not necessarily a linear function of \mathbf{y} . A common simplification is to restrict the family of predictors to be linear. That is

$$\tilde{\mathbf{v}} = \mathbf{A}\mathbf{y} + \mathbf{b}$$

for some matrix \mathbf{A} and vector \mathbf{b} . The solution is called the *best linear predictor* (BLP) and can be shown to be

$$\tilde{\mathbf{v}} \equiv \text{BLP}(\mathbf{v}) = \mathbf{E}(\mathbf{v}) + \mathbf{C}\mathbf{V}^{-1}\{\mathbf{y} - \mathbf{E}(\mathbf{y})\} \quad (1.5)$$

where

$$\mathbf{C} \equiv \mathbf{E}\{[\mathbf{v} - \mathbf{E}(\mathbf{v})]\{\mathbf{y} - \mathbf{E}(\mathbf{y})\}^T\} \quad \text{and} \quad \mathbf{V} \equiv \text{Cov}(\mathbf{y}).$$

If

$$\begin{bmatrix} \mathbf{v} \\ \mathbf{y} \end{bmatrix} \text{ is multivariate normal}$$

then best prediction and best linear prediction coincide. That is,

$$\text{BP}(\mathbf{v}) = \text{BLP}(\mathbf{v}) = \mathbf{E}(\mathbf{v}|\mathbf{y}) = \mathbf{E}(\mathbf{v}) + \mathbf{C}\mathbf{V}^{-1}\{\mathbf{y} - \mathbf{E}(\mathbf{y})\}.$$

1.3.2 Application to pig weight example

In model (1.3) let

$$\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_{48} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \text{weight}_{1,1} \\ \vdots \\ \text{weight}_{1,9} \\ \vdots \\ \text{weight}_{48,1} \\ \vdots \\ \text{weight}_{48,9} \end{bmatrix}.$$

Then

$$\begin{bmatrix} \mathbf{U} \\ \mathbf{y} \end{bmatrix} \text{ is multivariate normal}$$

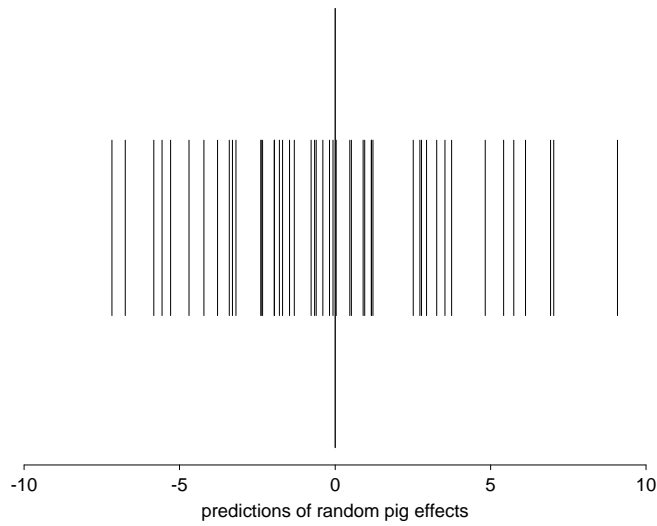
and, for given $\beta_0, \beta_1, \sigma_U^2$ and σ_ε^2 , the best predictor of U_i reduces to

$$\begin{aligned} \tilde{U}_i &= \frac{n_i \sigma_U^2}{\sigma_\varepsilon^2 + n_i \sigma_U^2} (\bar{y}_{i.} - \beta_0 - \beta_1 \bar{x}_{i.}) \\ &= \frac{9 \sigma_U^2}{\sigma_\varepsilon^2 + 9 \sigma_U^2} (\overline{\text{weight}_{i.}} - \beta_0 - \beta_1 \overline{\text{week}}) \end{aligned}$$

where $\bar{y}_i = \overline{\text{weight}_i}$ is the average weight of the i th pig and $\bar{x}_i = 5$ is the average week value. See McCulloch and Searle (2001).

Figure 1.4 shows $\tilde{U}_1, \dots, \tilde{U}_{48}$ after estimates of the variance components are plugged in. The variability in the intercepts among the 48 pigs is apparent, as is an estimated ranking of the pigs in this regard. Indeed, there is a branch of Statistics devoted to *ranking* and *selection* of subjects that has roots in animal breeding and genetics.

Figure 1.4: Predictions of U_1, \dots, U_{48} for the pig weight data. A vertical line is plotted for each \tilde{U}_i value, $1 \leq i \leq 48$.



1.4 The Linear Mixed Model (LMM)

Just like the linear model, we can generalize mixed models to arbitrary design matrices. The covariance structure of the random effects vector can also be general. The resulting general *linear mixed model* is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (1.6)$$

where

$$E \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

Note that model (1.3) is a special case of (1.6) with

$$\mathbf{y} = \begin{bmatrix} \text{weight}_{1,1} \\ \vdots \\ \text{weight}_{1,9} \\ \vdots \\ \text{weight}_{48,1} \\ \vdots \\ \text{weight}_{48,9} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & \text{week}_1 \\ \vdots & \vdots \\ 1 & \text{week}_9 \\ \vdots & \vdots \\ 1 & \text{week}_1 \\ \vdots & \vdots \\ 1 & \text{week}_9 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{9 \times 1} & \mathbf{0}_{9 \times 1} & \cdots & \mathbf{0}_{9 \times 1} \\ \mathbf{0}_{9 \times 1} & \mathbf{1}_{9 \times 1} & \cdots & \mathbf{0}_{9 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{9 \times 1} & \mathbf{0}_{9 \times 1} & \cdots & \mathbf{1}_{9 \times 1} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} U_1 \\ \vdots \\ U_{48} \end{bmatrix},$$

$$\mathbf{G} = \sigma_U^2 \mathbf{I} \quad \text{and} \quad \mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}.$$

As we will see in subsequent chapters, the general model (1.6) is extremely rich since it includes a large number of special cases that are useful in practice. The next few sections discuss statistical inference within this general framework. We will then explain how it relates to semiparametric regression modeling.

1.5 Estimation and Prediction in LMM

We now treat, in turn, estimation of $\boldsymbol{\beta}$, prediction of \mathbf{u} and estimation of the parameters in \mathbf{G} and \mathbf{R} .

1.5.1 Estimation of fixed effects

One way to derive an estimate of $\boldsymbol{\beta}$ is to rewrite (1.6) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad \text{where} \quad \boldsymbol{\varepsilon}^* = \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}.$$

This is just a linear model with correlated errors since

$$\text{Cov}(\boldsymbol{\varepsilon}^*) \equiv \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}.$$

For given \mathbf{V} , the classical textbook estimator of $\boldsymbol{\beta}$ (e.g. Rao, 1973; Draper and Smith, 1998) is

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}. \quad (1.7)$$

and is sometimes referred to as *generalized least squares (GLS)*.

Expression (1.7) can be justified in a few ways. For \mathbf{y} having a general distribution (1.7) can be shown to be the *Best Linear Unbiased Estimator (BLUE)* for $\boldsymbol{\beta}$. Alternatively, if \mathbf{y} is multivariate normal then the right-hand side of (1.7) is both the maximum likelihood estimator and the uniformly minimum variance unbiased estimator (UMVUE). The latter is the estimator that has the best (smallest) possible variance of any unbiased estimator regardless of the parameter values.

1.5.2 Prediction of random effects

The random effects vector can be predicted via best linear prediction using (1.5). For given $\boldsymbol{\beta}$ we obtain

$$\tilde{\mathbf{u}} = \text{BLP}(\mathbf{u}) = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.8)$$

In practice $\boldsymbol{\beta}$ would be replaced by an estimator such as $\tilde{\boldsymbol{\beta}}$ in (1.7), and the parameters in \mathbf{G} and \mathbf{V} would be need to be estimated (See Section 1.6).

1.5.3 Best linear unbiased prediction (BLUP)

A more unifying way to arrive at the results of the previous two subsections is through the notion of *best linear unbiased prediction (BLUP)*. For arbitrary $n \times 1$ vectors \mathbf{s} and \mathbf{t} , this involves the determination of linear $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ to minimize the prediction error

$$E\{(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{t}^T\mathbf{Z}\tilde{\mathbf{u}} - (\mathbf{s}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{t}^T\mathbf{Z}\mathbf{u}))^2\}$$

subject to the unbiasedness condition

$$E(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{t}^T\mathbf{Z}\tilde{\mathbf{u}}) = E(\mathbf{s}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{t}^T\mathbf{Z}\mathbf{u}).$$

Then it can be shown that (e.g. Robinson, 1991; Hayes and Haslett, 1999) the solutions for $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ are

$$\begin{aligned} \text{BLUP}(\boldsymbol{\beta}) &\equiv \tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \text{and } \text{BLUP}(\mathbf{u}) &\equiv \tilde{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \end{aligned} \quad (1.9)$$

Note that BLUP for $\boldsymbol{\beta}$ is identical to the generalized least squares solution (1.7) and the BLUP for \mathbf{u} is the BLP with $\boldsymbol{\beta}$ replaced by $\text{BLUP}(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}$.

As described by Robinson (1991) there are several other ways to derive BLUP solutions. A simple, albeit somewhat *ad hoc*, way is *Henderson's justification* which makes the distributional assumptions

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}),$$

and maximizes the likelihood of the (\mathbf{y}, \mathbf{u}) over the unknowns $\boldsymbol{\beta}$ and \mathbf{u} . This leads to the criterion

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u}. \quad (1.10)$$

This shows that BLUP estimation of $(\boldsymbol{\beta}, \mathbf{u})$ involves generalized least squares with a penalty term. It is easy to show from (1.10) that the BLUP of $(\boldsymbol{\beta}, \mathbf{u})$ can also be written as

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R}^{-1} \mathbf{y} \quad (1.11)$$

where $\mathbf{C} \equiv [\mathbf{X} \ \mathbf{Z}]$ and

$$\mathbf{B} \equiv \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}.$$

The fitted values are then

$$\text{BLUP}(\mathbf{y}) = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}} = \mathbf{C}(\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R}^{-1} \mathbf{y}.$$

This “ridge regression” formulation of BLUP shows the difference between $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ explicitly.

1.5.4 Estimation of covariance matrices

There is a large and varied literature on estimation of covariance matrices in mixed models. Dictated by computational issues, the earlier literature concentrated on strategies known as *minimum norm quadratic unbiased estimation* (MINQUE) and *minimum variance quadratic unbiased estimation* (MIVQUE) (e.g. Rao, 1973). However, with the advent of better computing algorithms, maximum likelihood (ML) or *restricted maximum likelihood* (REML) have become the most common strategies for estimating the parameters in covariance matrices.

First we describe ML. As in the previous section,

$$\mathbf{V} \equiv \text{Cov}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}.$$

Then the ML estimate of \mathbf{V} is based on the model

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

Restricted maximum likelihood (REML) also goes by the names *residual maximum likelihood*, *marginal maximum likelihood* and *generalized maximum likelihood*.

The log-likelihood of \mathbf{y} under this model is

$$\ell(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2} \{n \log(2\pi) + \log |\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} \quad (1.12)$$

so the ML estimate of $(\boldsymbol{\beta}, \mathbf{V})$ is the one that maximizes the right-hand side of this expression. If one first optimizes over $\boldsymbol{\beta}$, which appears only in the last term, we obtain that for any fixed \mathbf{V} , $\ell(\boldsymbol{\beta}, \mathbf{V})$ is maximized over $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$$

which corresponds to the best linear unbiased estimator given at (1.7).

On substitution into (1.12) we obtain the *profile log-likelihood* for \mathbf{V} :

$$\begin{aligned} \ell_P(\mathbf{V}) &= -\frac{1}{2} \left\{ \log |\mathbf{V}| + (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + n \log(2\pi) \right\} \\ &= -\frac{1}{2} \left[\log |\mathbf{V}| + \mathbf{y}^T \mathbf{V}^{-1} \{ \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \} \mathbf{y} \right. \\ &\quad \left. - \frac{n}{2} \log(2\pi) \right] \end{aligned} \quad (1.13)$$

ML estimates of the parameters in \mathbf{V} can be found by maximizing (1.13) over those parameters. For example, in the pig weight example,

$$\mathbf{V} = \sigma_U^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\varepsilon^2 \mathbf{I},$$

so (1.13) is a function of the variance component pair $(\sigma_U^2, \sigma_\varepsilon^2)$. Their estimation involves maximization of the bivariate function

$$\ell_P(\sigma_U^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\varepsilon^2 \mathbf{I})$$

over all $\sigma_U^2, \sigma_\varepsilon^2 \geq 0$. The answers are

$$\hat{\sigma}_{\varepsilon, \text{ML}}^2 = 4.38 \quad \text{and} \quad \hat{\sigma}_{U, \text{ML}}^2 = 14.8.$$

Derivation of the REML criterion is more complicated. It involves maximizing the likelihood of linear combinations of the elements of \mathbf{y} that do not depend on $\boldsymbol{\beta}$. Details can be found in, for example, Chapter 6 of Searle, Casella and McCulloch (1992). The resulting criterion function is the *restricted log-likelihood*

$$\ell_R(\mathbf{V}) = \ell_P(\mathbf{V}) - \frac{1}{2} \log |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|. \quad (1.14)$$

The main advantage of REML over ML is that REML takes into account the degrees of freedom for the fixed effects in the model. For example, in the special case where a random sample X_1, \dots, X_n is collected from the $N(\mu, \sigma^2)$ distribution then, with $\bar{X} = n^{-1} \sum_{i=1}^n X_i$,

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad \hat{\sigma}_{\text{REML}}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The *profile log-likelihood* of a parameter is obtained from the log-likelihood by substitution of the ML estimators of the other parameters in the model.

The $n-1$ in the denominator of $\hat{\sigma}_{\text{REML}}^2$ accounts for the estimation of μ via \bar{X} . For small sample sizes REML is expected to be more accurate than ML. But for large samples there is little difference between the two approaches.

Recently, there has also been a considerable amount of work devoted to computation of covariance matrix estimates (e.g. Lindstrom and Bates, 1988; Wolfinger, Tobias and Sall, 1994) and ensuing software development. The procedure PROC MIXED in the SAS

SAS Institute, Inc. (computing). system and the function lme() in the S-PLUS package both compute REML and ML covariance matrix estimates.

1.6 Estimated BLUP (EBLUP)

The BLUPs of $\boldsymbol{\beta}$ and \mathbf{u} given at (1.9) depend on

$$\mathbf{G} = \text{Cov}(\mathbf{u}) \quad \text{and} \quad \mathbf{R} = \text{Cov}(\boldsymbol{\varepsilon}),$$

especially through

$$\mathbf{V} = \text{Cov}(\mathbf{y}) = \mathbf{ZGZ}^T + \mathbf{R}.$$

As described in the previous section the parameters in these covariance matrices are typically estimated via ML or REML; and in practice the BLUPs are usually replaced by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y} \quad \text{and} \\ \hat{\mathbf{u}} &= \hat{\mathbf{G}} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \end{aligned}$$

where $\hat{\mathbf{G}}$ and $\hat{\mathbf{V}}$ are obtained by plugging in the ML or REML estimates of their parameters.

We will refer to $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ as *estimated* BLUPs, or EBLUPs, of $\boldsymbol{\beta}$ and \mathbf{u} . Similarly, the EBLUP of

$$\text{BLUP}\{E(\mathbf{y}|\mathbf{u})\} = \mathbf{X} \tilde{\boldsymbol{\beta}} + \mathbf{Z} \tilde{\mathbf{u}}$$

is

$$\text{EBLUP}\{E(\mathbf{y}|\mathbf{u})\} \equiv \hat{\mathbf{y}} \equiv \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{Z} \hat{\mathbf{u}}.$$

EBLUPs therefore have two sources of variability: that from estimation of $(\boldsymbol{\beta}, \mathbf{u})$, and that from estimation of \mathbf{G} and \mathbf{V} . Ideally, both should be taken into account when making inference about the quantity of interest. As described in the next section, this is a somewhat delicate matter.

1.7 Standard Error Estimation

From the BLUP expressions in Section 1.4

$$\text{Cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$$

and so the natural estimate of the standard deviation of the i th entry of the EBLUP $\hat{\beta}_i$ is

$$\widehat{\text{st.dev.}}(\hat{\beta}_i) = \sqrt{i \text{th diagonal entry of } (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}}. \quad (1.15)$$

Note that such an estimate ignores the variability due to estimation of \mathbf{V} . For larger samples this extra variability will be negligible. However, it can make a difference for smaller samples. As pointed out by McCulloch and Searle (2001, p. 258) the variance of $\hat{\beta}_i$ is largely intractable although there have been some attempts at using approximations (e.g. Kackar and Harville, 1984; Prasad and Rao, 1990) to quantify the extra variability in EBLUPs. EBLUPThe mixed model packages use (1.15) and we will usually use this estimator throughout much of this book in the hope that the samples are sufficiently large.

However, it is possible to handle these “intractable” calculations through a Bayesian approach and Markov Chain Monte Carlo methods. The details are postponed to Chapter ???. Treating the variance components as known, even though they really have been estimated, is what Bayesians call an empirical Bayes method. By taking instead a fully Bayesian approach, the extra variability in the EBLUPs due to estimation of variance components can be taken into account.

To estimate the precision of BLUPs involving \mathbf{u} we also need

$$\text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right) = \text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right).$$

Using (1.11) it can be shown that

$$\text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right) = (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \quad (1.16)$$

where, as before,

$$\mathbf{C} \equiv [\mathbf{X} \ \mathbf{Z}] \quad \text{and} \quad \mathbf{B} \equiv \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}.$$

In some contexts it is useful to estimate the *conditional* covariance matrix

$$\text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right) = \text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right).$$

From (1.11),

$$\text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right) = (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1}. \quad (1.17)$$

These results suggest the approximations

$$\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right) \simeq (\mathbf{C}^T \hat{\mathbf{R}}^{-1} \mathbf{C} + \hat{\mathbf{B}})^{-1}$$

and

$$\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right) \simeq (\mathbf{C}^T \hat{\mathbf{R}}^{-1} \mathbf{C} + \hat{\mathbf{B}})^{-1} \mathbf{C}^T \hat{\mathbf{R}}^{-1} \mathbf{C} (\mathbf{C}^T \hat{\mathbf{R}}^{-1} \mathbf{C} + \hat{\mathbf{B}})^{-1}.$$

Standard errors may also be estimated for the covariance matrix parameters, although the details are omitted. Searle *et al.* contains some details. Confidence intervals for both fixed effects and covariance matrix parameters, that is, variance components, are described by Pinheiro and Bates (2000).

1.7.1 Summary of fit to pig weights

Table 1.1 summarizes the fit of (1.3) based on REML estimation of the variance components and EBLUP.

	coef.	std.err.	z ratio
intercept	19.4	0.603	32.1
week	6.21	0.0391	159
$\hat{\sigma}_U^2 = 15.1$		$\hat{\sigma}_\varepsilon^2 = 4.39$	

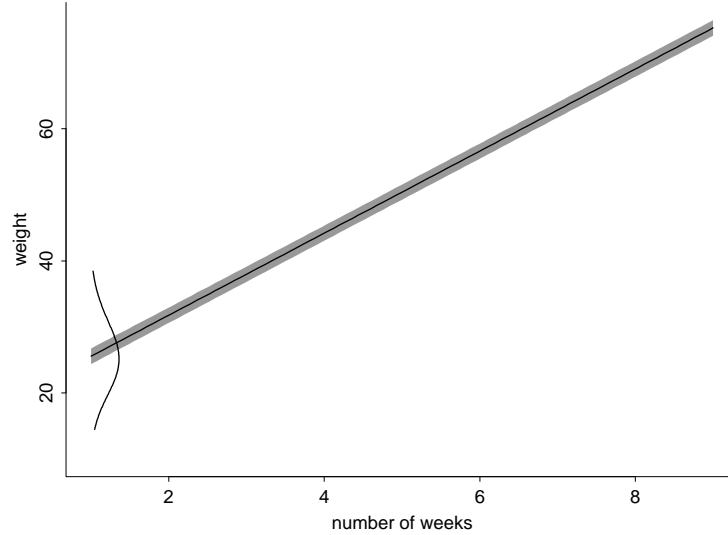
Table 1.1: Summary of REML/EBLUP fit of (1.3) for the pig weight data.

A graphical summary of the fit is shown in Figure 1.5. There is a strongly significant positive growth, but with considerable between-pig variability in the intercepts.

1.8 Hypothesis Testing

There is a rich literature on hypothesis testing in the linear mixed model framework. Verbeke and Molenberghs (1997) and Khuri, Mathew, and Sinha (1998) provide surveys. We will restrict discussion to some of the simpler methods here.

Figure 1.5: Graphical summary of the fit of REML/BLUP fit of (1.3) for the pig weight data. The line is estimated mean weight. The shaded regions corresponds to plus and minus two standard deviations. The curve at the left is a density estimate of the estimated random intercepts.



1.8.1 Normal theory tests

First consider the problem of hypothesis testing for β_i , the i th entry of $\boldsymbol{\beta}$. Ideally, this can be done through a result of the form

$$z_i \equiv \frac{\hat{\beta}_i - \beta_i}{\text{st.dev.}(\hat{\beta}_i)} \text{ approx. } N(0, 1). \quad (1.18)$$

Specifically, for the hypothesis set-up

$$\begin{aligned} H_0 : \beta_i &= 0 \\ H_1 : \beta_i &\neq 0, \end{aligned} \quad (1.19)$$

the approximate p-value is given by the tail area

$$\text{p-value} \simeq 2\{1 - \Phi(|z_{0,i}|)\}$$

where Φ is the cumulative distribution function of the standard normal distribution, and $z_{0,i}$ is given by (1.18) with $\beta_i = 0$. However, the theoretical justification of (1.18) for general mixed models is somewhat elusive due to the dependence in \mathbf{y} imposed by the random effects. Theoretical back-up for (1.18) exists in certain special cases such as those arising in analysis of variance and longitudinal data analysis (e.g. Miller, 1977). For some mixed models, including many used in the subsequent chapters of this book, justification of (1.18) remains an open problem.

1.8.2 Likelihood ratio tests

Since, as we saw in Sections 1.5.1 and 1.5.4, the parameters in mixed models can be estimated by maximum likelihood, the *likelihood ratio* procedure can be used to test hypothesis. We will first give a brief description of the likelihood ratio test procedure for general parametric models.

Let $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ be the likelihood of the parameter vector $\boldsymbol{\theta}$ based on the data in \mathbf{y} . The likelihood ratio statistic for testing a null restricted model against alternative unrestricted model is

$$\text{LR}(\mathbf{y}) = \mathcal{L}(\hat{\boldsymbol{\theta}}_0; \mathbf{y}) / \mathcal{L}(\hat{\boldsymbol{\theta}}; \mathbf{y})$$

where $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ are the maximum likelihood estimates of $\boldsymbol{\theta}$ under the null model and unrestricted model, respectively. It is more common to work with

$$-2 \log\{\text{LR}(\mathbf{y})\} = -2\{\ell(\hat{\boldsymbol{\theta}}_0; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}; \mathbf{y})\} \quad (1.20)$$

where $\ell(\boldsymbol{\theta}; \mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ is the log-likelihood.

The classical result for determining the significance of the observed value of $\ell(\boldsymbol{\theta}; \mathbf{y})$ is one that states, under H_0 ,

$$-2 \log\{\text{LR}(\mathbf{y})\} \stackrel{\text{approx.}}{\sim} \chi_\nu^2 \quad (1.21)$$

where the right-hand side is the chi-squared distribution with ν degrees of freedom and

$$\begin{aligned} \nu &= \text{number of independent parameters in unrestricted model} \\ &\quad - \text{number of independent parameters in null model.} \end{aligned}$$

For hypothesis test (1.19) $\nu = 1$ so (1.21) provides an alternative way to test this hypothesis. Since a χ_1^2 random variable is the square of a standard normal random variable we have

$$\text{p-value} \simeq 1 - \Phi(\sqrt{-2 \log \text{LR}(\mathbf{y})}). \quad (1.22)$$

Once again the dependence in \mathbf{y} means that justification of (1.21), and hence (1.22), is dependent on the type of correlation structure induced by the \mathbf{G} and \mathbf{R} matrices.

Hypothesis tests for covariance matrix parameters may also be of interest. Consider, for example, the random intercept straight line model for repeated measures regression data:

$$y_{ij} = \beta_0 + U_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m, \quad U_i \sim N(0, \sigma_U^2). \quad (1.23)$$

One may wish to determine whether the intercepts of the individuals are significantly different from one another, i.e. whether the sub-model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i \quad 1 \leq i \leq m, \quad (1.24)$$

is adequate compared with (1.23). This boils down to testing the hypotheses

$$\begin{aligned} H_0 : \sigma_U^2 &= 0 \\ H_1 : \sigma_U^2 &> 0. \end{aligned} \quad (1.25)$$

For (1.25) this would suggest that $-2 \log\{\text{LR}(\mathbf{y})\}$ be compared with percentiles from the χ_1^2 distribution. However, the theory behind (1.21) assumes that the parameter of interest is not on the boundary of its parameter space. Since the parameter space for σ_U^2 is $[0, \infty)$ this assumption is violated. In fact, under certain independence assumptions discussed below, the asymptotic distribution when H_0 is true is such that there is a 50:50 chance that

$$\hat{\sigma}_{u, \text{ML}}^2 = 0.$$

This type of behavior leads to

$$-2 \log\{\text{LR}(\mathbf{y})\} \overset{\text{approx.}}{\sim} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 \quad (1.26)$$

for (1.25), where χ_0^2 notation means a point mass at zero and the right hand side of (1.26) is shorthand for a 50:50 mixture between a χ_0^2 and a χ_1^2 distribution (Self and Liang, 1987). Use of (1.26) rather than (1.21) leads to p -values being halved.

If, instead, one was interested in testing the adequacy of

$$y_{ij} = \beta_0 + \varepsilon_{ij} \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m,$$

compared with (1.23) then the hypotheses would be

$$\begin{aligned} H_0 : \sigma_U^2 &= \beta_1 = 0 \\ H_1 : \sigma_U^2 &> 0 \quad \text{or} \quad \beta_1 \neq 0 \end{aligned}$$

and, under H_0 , we would have

$$-2 \log\{\text{LR}(\mathbf{y})\} \overset{\text{approx.}}{\sim} \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2.$$

More generally, if H_0 constrains one variance component and s regression coefficients to be zero then

$$-2 \log\{\text{LR}(\mathbf{y})\} \overset{\text{approx.}}{\sim} \frac{1}{2}\chi_s^2 + \frac{1}{2}\chi_{s+1}^2. \quad (1.27)$$

The notation $X \overset{\text{approx.}}{\sim} \frac{1}{2}\chi_s^2 + \frac{1}{2}\chi_t^2$ means that the random variable X has a approximate density function equal to a 50:50 mixture of the χ_s^2 and χ_t^2 densities. This is different from the density of the average of independent random variables from each of these densities.

If there are two variance components of interest then the distribution theory for $-2 \log\{\text{LR}(\mathbf{y})\}$ becomes much more complicated (e.g. Self and Liang, 1987; Stram and Lee, 1994; Silvapulle, 1996; Verbeke and Molenberghs, 1997).

The classical large-sample theory for likelihood ratio tests assumes independence of the \mathbf{y} vector under all values of the parameters, or more precisely, that the \mathbf{y} vector can be partitioned into subvectors that are independent. This assumption does not hold in general for mixed models, at least not under the alternative. However, for the simple longitudinal model (1.23) independence between subjects allows for the extension of the classical theory and validation of (1.26) and (1.27) for a large number of subjects. The idea is that these approximations assume that the number of independent observations goes to infinity, which is true if we take the subjects to be the observations and the number of subjects increases to infinity (Stram and Lee, 1994). As for the normal theory tests described in the previous section, the asymptotic distribution theory for general mixed models is more difficult and, in some instances, yet to be worked out.

One case that has been studied carefully in Crainiceanu, Ruppert, and Vogelsang (2002) and Crainiceanu and Ruppert (2002) is the balanced one-way analysis of variance with random treatment (or subject) effects. This model is given by equation (1.23) with β_1 known to equal 0 and $n_i = n$ for some fixed n and all i . If n is fixed and m tends to infinity, then the Self and Liang (1987) assumptions are met and (1.26) does hold. In contrast, if n goes to infinity with m fixed, then we have a fixed number of subjects (or treatments) and the number of observations per subject goes to infinity. In this case, the Self and Liang (1987) assumptions do not hold and (1.26) fails to hold. The asymptotic probability of zero, that is the probability attached to the χ_0^2 component, is not $\frac{1}{2}$ but rather is greater than $\frac{1}{2}$ and tends very slowly to $\frac{1}{2}$ as m goes to infinity (Crainiceanu, Ruppert, and Vogelsang, 2002). In fact, this probability is about 0.65 when m is 10 and 0.55 when m is 100. Moreover, the component that is non-zero is not χ_1^2 but something that tends to be smaller than χ_1^2 (Crainiceanu and Ruppert, 2002). Thus, the likelihood ratio test statistic tends to be *smaller* than under the Self and Liang asymptotics, so that using those asymptotics to obtain p-values gives tests that are conservative, that is, have nominal p-values larger than the true p-value. Conservative tests have smaller type 1 error probabilities than stated, which is not a problem. However, they have the disadvantage of having less power at the alternative than a test with correct type 1 error probability.

As we will see in the next section, penalized splines can be viewed as BLUPs in a certain mixed model. Approximation (1.26) is very poor when applied to penalized splines (Crainiceanu and Ruppert, 2002). In general, the asymptotics of Self and Liang (1987) do not apply to the semiparametric models we discuss in this book. This means that asymptotics cannot be used to find p-values, at least not until alternative asymptotics are derived.

As we discuss in later sections, critical values of likelihood ratio tests can be determined satisfactorily by Monte Carlo simulations. The idea is to set the values of all fixed effect and variance component parameters equal to their estimates under the null distribution and to simulate the distribution of the likelihood ratio test under the null model at the parameters and with the covariates equal to their observed values. More precisely, we simulate a large number of independent data sets, say 10,00 to 100,000, with fixed effects parameters at their estimated values and the ε 's and random effects generated according to their estimated variances, both estimations under the null hypothesis. The likelihood ratio test statistic (1.20) is calculated for each simulated data set. The p-value of the test is the proportion of simulated values of the test statistic that exceed the value at the real data.

1.8.3 Restricted likelihood ratio tests

Instead of using the likelihood function to form test statistics, one could do so using the maximum restricted log-likelihood $\ell_R(\mathbf{V})$ defined at (1.14). Since REML estimates of \mathbf{V} are unbiased, the accuracy of the test might be improved. There is some discussion on this approach in Verbeke and Molenberghs (1997). They mention that restricted likelihood can only be used to compare models with the same mean structure, that is, the same fixed effects model. The reason for this is that restricted likelihood is the likelihood of the residuals after fitting the fixed effects and so is not appropriate when there are more than one fixed effects models under consideration.

As discussed in Crainiceanu, Ruppert, and Vogelsang (2002) and Crainiceanu and Ruppert (2002), restricted likelihood ratio tests have the same complex asymptotic theory as likelihood ratio tests. For this reason, we advocate computing p-values by simulation as just discussed at the end of Section 1.8.2.

1.9 Penalized Splines as BLUPs

In Chapter ?? we considered the ordinary nonparametric regression model

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n \quad (1.28)$$

and showed how f could be estimated by penalized splines. In this section we show that this estimate can be written as the BLUP of a mixed model.

For clarity we will treat the linear case and suppose that the errors satisfy $\text{Cov}(\varepsilon) = \sigma_\varepsilon^2 \mathbf{I}$. The linear spline model for f is

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^K u_k (x_i - \kappa_k)_+. \quad (1.29)$$

Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}$$

be the coefficients of the polynomial functions and truncated line functions respectively. Corresponding to these vectors define

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_K)_+ \end{bmatrix}.$$

The penalized spline fitting criterion (??), when divided by σ_ε^2 , can then be written as

$$\frac{1}{\sigma_\varepsilon^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + (\lambda^2/\sigma_\varepsilon^2) \|\mathbf{u}\|^2.$$

Notice that this can be made to equal the BLUP criterion given at (1.10) by treating the \mathbf{u} as a set of random coefficients with

$$\text{Cov}(\mathbf{u}) = \sigma_u^2 \mathbf{I} \quad \text{where} \quad \sigma_u^2 = \sigma_\varepsilon^2 / \lambda^2.$$

Putting all of this together we have the mixed model representation of the regression spline

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix}. \quad (1.30)$$

Note that the fitted values $\tilde{\mathbf{f}}$ can be rewritten as

$$\tilde{\mathbf{f}} = \mathbf{C}(\mathbf{C}^T \mathbf{C} + \lambda^2 \mathbf{D})^{-1} \mathbf{C}^T \mathbf{y} \tag{1.31}$$

where

$$\mathbf{C} = [\mathbf{X} \ \mathbf{Z}], \quad \mathbf{D} = \text{diag}(0, 0, 1, 1, \dots, 1) \quad \text{and} \quad \lambda^2 = \sigma_\varepsilon^2 / \sigma_u^2,$$

matching (??) for $p = 1$.

Figure 1.6 shows how the mixed model approach to fitting the regression splines leads to a smooth result. In this example data are simulated according to the situation where $f(x) = \sin(3\pi x)$, $0 \leq x \leq 1$, and $\sigma_\varepsilon = 0.4$. In Figure 1.6 (a) we see the result when

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^K u_k (x_i - \kappa_k)_+ + \varepsilon_i$$

is fit using ordinary least squares. Notice that it overfits the data rather than smoothing it. The fit in Figure 1.6 (b) corresponds to treating (1.29) as a mixed model with

$$u_k \text{ i.i.d } N(0, \sigma_u^2).$$

Ordinary least squares corresponds to $\sigma_u = \infty$, where the u_k are unrestricted. Taking σ_u to be finite, in this case $\sigma_u = 3\sigma_\varepsilon = 1.2$, leads to smaller estimates of the u_k and the effect of the $(x_i - \kappa_k)_+$ being diminished. A smooth fit results.

This representation of the penalized spline as a BLUP in a mixed model is useful, since it allows smoothing to be done using mixed model methodology and software. This will be exemplified in the following chapters. It also lends itself, via e.g. Robinson (1991), to a host of other derivations including one as a *Bayesian estimator*. As explained there if $\tilde{\boldsymbol{\beta}}$ is regarded as a parameter with a uniform, improper prior then $\tilde{\mathbf{f}}$ corresponds to the posterior mode. Robinson (1991) also demonstrates how the *Kalman filter* can be used to compute BLUPs. See also the discussion by Spall (1991) and Robinson's rejoinder.

When σ_u^2 and σ_ε^2 are replaced by estimators, such as those obtained from ML and REML, the final vector of fitted values is

$$\hat{\mathbf{f}} = \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{Z} \hat{\mathbf{u}}.$$

Bayesian distribution is improper if total probability does not equal 1. This is the case if the distribution is based on the inclusion of incorporation prior beliefs about parameters of interest. Chapter ?? deals with Bayesian estimation.

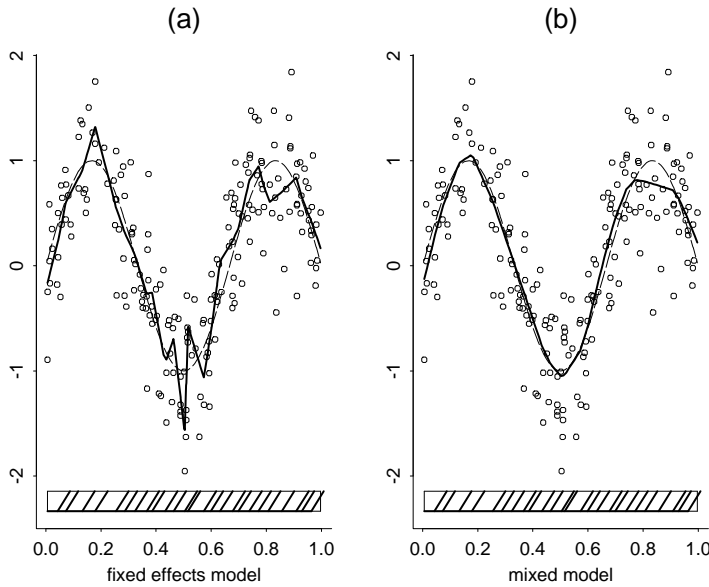


Figure 1.6: Comparison between treating the coefficients of the knots as fixed effects versus random effects. The solid curve is the estimated curve, while the dashed curve is the true function.

For arbitrary $x \in \mathbb{R}$ the estimate of $f(x)$ is

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \sum_{k=1}^K \hat{u}_k (x - \kappa_k)_+$$

where $\hat{\beta}_0$, $\hat{\beta}_1$ and \hat{u}_k , $1 \leq k \leq K$, are EBLUPs.

1.10 Bibliographical Notes

Mixed modeling is a massive and growing branch of Statistics, and here we have just summarized the aspects that are most relevant to the subsequent chapters in this book.

An excellent introduction to general design linear mixed models is Robinson (1991). REML estimation of covariance matrices is due to Patterson and Thompson (1971). An important reference for computational issues in mixed models is Harville (1977).

In recent years several books on mixed models, some with an emphasis on longitudinal data analysis, have been published. These include Searle, Casella and McCulloch (1992), Vonesh and Chinchilli (1997), Pinheiro and Bates (2000), McCulloch and Searle (2001) and Verbeke and Molenberghs (2000).

The *Kalman filter* is a fast algorithm for fitting optimal solutions to a certain class of linear statistical models. An introduction to the Kalman filter is provided by Maybeck (1979).

1.11 Summary of Formulae

Linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where

$$E \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

$$\mathbf{V} \equiv \text{Cov}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}.$$

Best linear unbiased prediction (BLUP)

$$\begin{aligned} \text{BLUP}(\boldsymbol{\beta}) &\equiv \tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \text{and } \text{BLUP}(\mathbf{u}) &\equiv \tilde{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \end{aligned}$$

Alternatively,

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = (\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C} + \mathbf{B})^{-1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}$$

where

$$\mathbf{C} = [\mathbf{X} \ \mathbf{Z}] \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}.$$

Likelihood-based estimation of \mathbf{V}

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

Profile log-likelihood for ML estimation of \mathbf{V} is

$$\ell_P(\mathbf{V}) = -\frac{1}{2} [\log |\mathbf{V}| + \mathbf{y}^T\mathbf{V}^{-1}\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\}\mathbf{y}] - \frac{n}{2} \log(2\pi).$$

Restricted profile log-likelihood for REML estimation of \mathbf{V} is

$$\ell_R(\mathbf{V}) = \ell_P(\mathbf{V}) - \frac{1}{2} \log |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|.$$

Standard error estimation

$$\text{Cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$$

$$\text{Cov} \left(\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right) = (\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C} + \mathbf{B})^{-1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C} + \mathbf{B})^{-1}.$$

BLUP representation of linear penalized spline

Scatterplot data are (x_i, y_i) , $1 \leq i \leq n$. Knots are $\kappa_1, \dots, \kappa_K$.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_K)_+ \end{bmatrix}.$$

$$\mathbf{G} = \sigma_u^2 \mathbf{I}, \quad \mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}, \quad \lambda^2 = \sigma_\varepsilon^2 / \sigma_u^2.$$