A Lecture Note for Linear Mixed Models

Xia Shen*

October 2, 2012

Introduction

This lecture note introduces the *linear mixed models*, which provide a framework allowing for the most complex model structure in statistics. Its theory and applications were the last to develop in statistics, and some problems like estimating the *variance components* for non-normal models are not even solved yet. Advanced *random effects* models can have a hierarchical structure with multiple layers, which enables deep explanations for the variance components. Certainly, fitting such advanced models require smart algorithms.

Both linear regression (general linear model, LM) and generalized linear model (GLM) are *fixed effects* models. In most cases, the underlying structure is

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \tag{i}$$

where $\boldsymbol{\beta}$ is a fixed parameter with dimension p, and p is typically quite small, otherwise, the model does not have sufficient degrees of freedom to make inference for all the parameters. The need to consider random effects models arises when the number of unknown parameters is large, while the number of observations per parameter is limited. For example, we follow 200 individuals over time, where from each individual we obtain two or three measurements. To account for natural variability, we want to fit a linear growth model for each individual. Let y_{ij} be the j:th measurement of individual i, recorded at time t_{ij}, and consider the model

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij} \tag{ii}$$

where β_{0i} and β_{1i} are the intercept and slope parameters for the i:th individual. In effect these parameters are random variables, and the model is called a *random effects model*. Since the means of random effects will be treated as fixed parameters, we may also call the model a *mixed effects model* or *mixed model*. In this example we have 400 unknown parameters but only a maximum of 600 observations. Of course we can try to estimate the parameters using data from each individual separately, but the estimates will be super poor if the error variance is huge. Theoretically and empirically it has been shown that better estimates can be obtained by treating the parameters as random.

To illustrate the mixed models, the rest of this lecture note is arranged as follows. We start by extending the likelihood concept to capture the information about random parameters. Thereafter, we introduce the fairly well established area of normal linear mixed models, which can be extended to non-normal cases of mixed models.

*PhD, postdoctoral researcher, Division of Computational Genetics, Department of Clinical Sciences, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden. Lecturer, Statistics Unit, Dalarna University, Borlänge, Sweden. Personal URL: http://www.19850911.com

Simple Random Effects Models

Theoretical analyses of random effects models are usually cumbersome, and it is often impossible to get any closed solution in formulae. Although in practice, we might get over such problems using numerical methods, we here first describe the simplest random effects model, where it is possible to arrive at explicit formulae.

Data Example

The dataset given in the following table from Fears et al. (1996) shows the estrone measurement results from five menopausal women, in which 16 measurements were taken from each person.

Estrone measurements from five menopausal women

$i=1$	2	3	4	5	$i=1$	2	3	4	5
23	25	38	14	46	22	26	35	17	32
23	33	38	16	36	22	30	40	18	31
22	27	41	15	30	23	30	41	20	30
20	27	38	19	29	23	29	37	18	32
25	30	38	20	36	27	29	28	12	25
22	28	32	22	31	19	37	36	17	29
27	24	38	16	30	23	24	30	15	31
25	22	42	19	32	18	28	37	13	32

The application on these data tries to answer these questions: i) Is there significant variation between women relative to with-woman variation? ii) What is the reliability of the measurements? iii) What is each woman's mean estrone concentration? Let $y_{ij} = 10 \log_{10} x_{ij}$ where x_{ij} is the raw estrone measurement. We consider the following one-way random effects model

$$y_{ij} = \mu + u_i + e_{ij} \tag{iii}$$

where μ is a fixed overall mean parameter, and

$$u_i = \text{person effect, for } i = 1, \ldots, q = 5 \tag{iv}$$
$$e_{ij} = \text{residual effect, for } j = 1, \ldots, n = 16 \tag{v}$$

Here we assume that u_i's are iid $N(0, \sigma_u^2)$, e_{ij}'s are iid $N(0, \sigma^2)$ and they are independent.

Variance Component Estimation

The covariance between two measurements y_{ij} and y_{ik} for one particular woman i is

$$\operatorname{cov}(y_{ij}, y_{ik}) = \operatorname{cov}(\mu + u_i + e_{ij}, \mu + u_i + e_{ik}) = \sigma_u^2 \tag{vi}$$

So $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})'$ is multivariate normal with mean $\boldsymbol{\mu}$ and variance

$$\mathbf{S} = \sigma^2 \mathbf{I}_n + \sigma_u^2 \mathbf{J}_n \tag{vii}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{J}_n$ is an $n \times n$ matrix of ones. The likelihood of $\boldsymbol{\theta} = (\mu, \sigma^2, \sigma_u^2)$ is

$$L(\boldsymbol{\theta}) = -\frac{N}{2} \log |\mathbf{S}| - \frac{1}{2} \sum_i (\mathbf{y}_i - \boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \tag{viii}$$

Using some matrix algebra results (Rao, 1973, page 67), we have

$$|\mathbf{S}| = \sigma^{2(n-1)}(\sigma^2 + n\sigma_u^2) \tag{ix}$$

$$\mathbf{S}^{-1} = \frac{\mathbf{I}_n}{\sigma^2} - \frac{\sigma_u^2}{\sigma^2(\sigma^2 + n\sigma_u^2)}\mathbf{J}_n \tag{x}$$

Fixing values for (σ^2, σ_u^2), the maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}(\sigma^2, \sigma_u^2) = \frac{\sum_i \mathbf{1}'\mathbf{S}^{-1}\mathbf{y}_i}{\sum_i \mathbf{1}'\mathbf{S}^{-1}\mathbf{1}} = \bar{y} = 14.175 \tag{xi}$$

Now define the following total, person and error sum-of-squares

$$\text{SST} = \sum_{ij}(y_{ij} - \bar{y})^2 \tag{xii}$$

$$\text{SSA} = \sum_i\{\sum_j(y_{ij} - \bar{y})\}^2/n \tag{xiii}$$

$$\text{SSE} = \text{SST} - \text{SSA} \tag{xiv}$$

The *profile likelihood* for (σ^2, σ_u^2) can be shown to be

$$\log L(\sigma^2, \sigma_u^2) = -\frac{N}{2}\{(n-1)\log\sigma^2 + \log(\sigma^2 + n\sigma_u^2)\} - \frac{1}{2}\left\{\frac{\text{SSE}}{\sigma^2} + \frac{\text{SSA}}{\sigma^2 + n\sigma_u^2}\right\} \tag{xv}$$

From the likelihood we obtain the MLEs

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N(n-1)} = 0.325 \tag{xvi}$$

$$\hat{\sigma}_u^2 = (\text{SSA}/N - \hat{\sigma}^2)/n = 1.395 \tag{xvii}$$

The measurements are reliable if the correlation between two measurements for the same person is high. From the model,

$$\text{cor}(y_{ij}, y_{ik}) = \frac{\text{cov}(y_{ij}, y_{ik})}{\{\text{var}(y_{ij})\text{var}(y_{ik})\}^{1/2}} \tag{xviii}$$

$$= \frac{\sigma_u^2}{\sigma^2 + \sigma_u^2} \tag{xix}$$

This quantity is also called the *intraclass correlation*. Its estimate for the given data is 0.81. A visualization of the data and likelihoods is given in Figure 1.

Random Effects Estimation

Classical analysis of random effects models focuses on the estimation of the variance components σ^2 and σ_u^2, but recent interests are also on the estimation or *prediction* of the random effects parameters.

For the one-way random effects model above, the full parameter space is

$$(\boldsymbol{\theta}, \mathbf{u}) \equiv (\mu, \sigma^2, \sigma_u^2, u_1, \ldots, u_q) \tag{xx}$$

where the fixed parameter is $\boldsymbol{\theta} = (\mu, \sigma^2, \sigma_u^2)$. According to the following definition of the extended likelihood, the likelihood based on dataset $\mathbf{y}$ is

$$L(\boldsymbol{\theta}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})p_{\boldsymbol{\theta}}(\mathbf{u}) \tag{xxi}$$

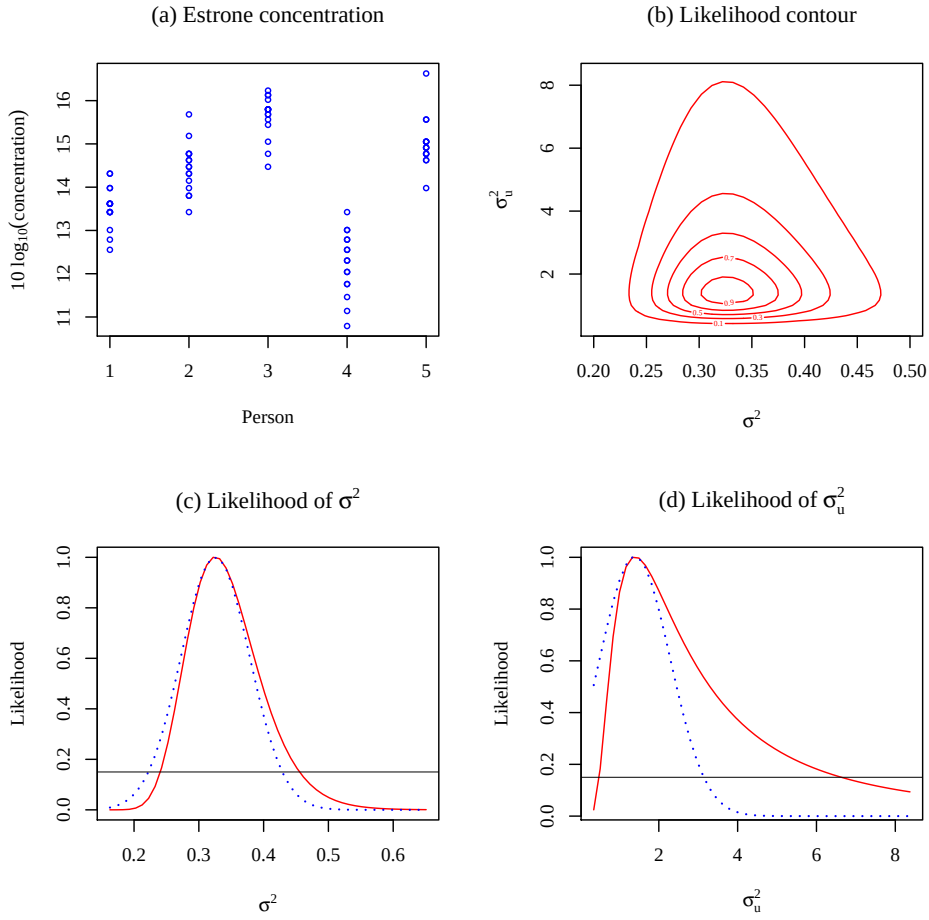

Figure 1: Likelihood analysis of a one-way random effects experiment. (a) This plot shows a significant person-to-person variability. (b) Joint likelihood of (σ^2, σ_u^2). (c) The profile likelihood of σ^2 is well approximated by the normal. (d) Poor normal approximation of the profile likelihood of σ_u^2.

Definition 1 *Assuming a statistical model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$, where $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$ is a joint density function of observed data $\mathbf{x}$ and unobserved $\mathbf{y}$ given a fixed parameter $\boldsymbol{\theta}$, the likelihood function for $\boldsymbol{\theta}$ and $\mathbf{y}$ is*

$$L(\boldsymbol{\theta}, \mathbf{y}) = p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) \tag{xxii}$$

This extended definition of likelihood agrees with Butler (1987), while Bjørnstad (1996) provided the theoretical evidence that such a likelihood carries all of the information on the unknown parameters $(\boldsymbol{\theta}, \mathbf{y})$. It can be interpreted as an augmented likelihood, since it may be written as

$$L(\boldsymbol{\theta}, \mathbf{y}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y}) p_{\boldsymbol{\theta}}(\mathbf{y}) \tag{xxiii}$$

where $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y})$ is the pure likelihood term, and $p_{\boldsymbol{\theta}}(\mathbf{y})$ is the contextual information that $\mathbf{y}$ is random. In mixed effects modeling the extended likelihood has been called h-likelihood (hierarchical likelihood) by Lee and Nelder (1996), while in smoothing literature it is known as penalized likelihood (e.g. Green and B.W., 1993).

Given $\mathbf{u}$ the outcomes y_{ij}'s are independent with mean

$$\mu_i = \mu + u_i \tag{xxiv}$$

and variance σ^2, while u_i's are iid with mean zero and variance σ_u^2. Hence

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{u}) &= -\frac{qn}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{q}\sum_{j=1}^{n}(y_{ij}-\mu-u_i)^2 \\ &\quad -\frac{q}{2}\log\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{q}u_i^2 \end{aligned} \tag{xxv}$$

Estimates of u_i can be computed by directly maximizing $\log L$ with respect to $(\boldsymbol{\theta}, \mathbf{u})$, which will be described below.

If we assume a fixed effects model, i.e. u_i's are fixed parameters, the log-likelihood of the unknown parameters is

$$\log L = -\frac{qn}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{q}\sum_{j=1}^{n}(y_{ij}-\mu-u_i)^2 \tag{xxvi}$$

which involves only the first two terms of (xxv). Using the constraint $\sum u_i = 0$ we can verify that the MLE of u_i is

$$\hat{u}_i = \bar{y}_i - \bar{y} \tag{xxvii}$$

where $\bar{y}_i$ is the average of $y_{i1}, \ldots, y_{in}$, the MLE of μ is $\bar{y}$, and the MLE of μ_i is $\bar{y}_i$ regardless of the constraint on u_i's. The constraint is not needed in the random effects model, instead, we do need to specify $E(u_i) = 0$.

Assume for the moment that $\boldsymbol{\theta}$ is known. The derivative of the log-likelihood (xxv) at a fixed value of $\boldsymbol{\theta}$ is

$$\frac{\partial \log L}{\partial u_i} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_{ij}-\mu-u_i) - \frac{u_i}{\sigma_u^2} \tag{xxviii}$$

and on setting it to zero we get

$$\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_u^2}\right)\hat{u}_i = \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_{ij}-\mu) \tag{xxix}$$

or

$$\hat{u}_i = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_u^2}\right)^{-1}\frac{n}{\sigma^2}(\bar{y}_i - \mu) \tag{xxx}$$

It is also possible to show that $\hat{u}_i$ is the conditional mean of the distribution of u_i given $\mathbf{y}$. If b_i is a fixed parameter, and we use a prior density $p_{\boldsymbol{\theta}}(\mathbf{u})$, the $\hat{u}_i$ is called a *Bayesian estimate* of u_i - This is a Bayesian interpretation of the estimate.

If the fixed parameters are known, the Fisher information for u_i is

$$I(u_i) = \frac{\partial^2 \log L}{\partial u_i^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_u^2} \tag{xxxi}$$

compared with n/σ^2 if b_i is assumed fixed. Consequently the standard error of $\hat{u}_i$ under the random effects model is **smaller** than the standard error under the fixed effects model.

In practice the unknown $\boldsymbol{\theta}$ is replaced by its estimate. We may use the MLE, but when the MLE is too difficult to compute other estimates may be used. Thus

$$\hat{u}_i = \left(\frac{n}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}_u^2}\right)^{-1}\frac{n}{\hat{\sigma}^2}(\bar{y}_i - \bar{y}) \tag{xxxii}$$

Comparing this with (xxvii), it is clear that the effect of the random effects assumption is to **shrink** $\hat{u}_i$ towards its zero mean, and that is why the estimate is also called a *shrinkage estimate*. The estimate of μ_i is

$$\begin{aligned}
\hat{\mu}_i &= \bar{y} - \hat{u}_i && \text{(xxxiii)}\\
&= \bar{y} + \left(\frac{n}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}_u^2}\right)^{-1} \frac{n}{\hat{\sigma}^2}(\bar{y}_i - \bar{y}) && \text{(xxxiv)}\\
&= \left(\frac{n}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}_u^2}\right)^{-1} \left(\frac{n}{\hat{\sigma}^2}\bar{y}_i + \frac{1}{\hat{\sigma}_u^2}\hat{\mu}\right) && \text{(xxxv)}\\
&= \alpha\bar{y}_i + (1-\alpha)\bar{y} && \text{(xxxvi)}
\end{aligned}$$

where

$$\alpha = \left(\frac{n}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}_u^2}\right)^{-1} \frac{n}{\hat{\sigma}^2} \qquad \text{(xxxvii)}$$

If n/σ^2 is large relative to $1/\sigma_u^2$, i.e. there is a lot of information in the data about μ_i, the α is close to one and the estimated mean is close to the sample average. The estimate is called an *empirical Bayes estimate*, as it can be thought of as implementing a Bayes estimation procedure on the mean parameter μ_i, with a normal prior that has mean μ and variance σ_u^2. It is 'empirical' since the parameter of the prior is estimated from the data.

For the estrone data above, $n = 16$, so the shrinkage parameter is

$$\alpha = 0.986 \qquad \text{(xxxviii)}$$

The sample means $\bar{y}_i$'s are

13.545 14.447 15.635 12.233 15.015

and we have the shrinkage estimates of the individual means $\hat{\mu}_i$'s as follows

13.554 14.443 15.614 12.261 15.003

They are fairly close to the sample means in this example since α is close to one.

Normal Linear Mixed Models

Let $\mathbf{y}$ be an length-N vector of outcome data, and $\mathbf{X}$ and $\mathbf{Z}$ be $N \times p$ and $N \times q$ design matrices for the fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{u}$. The standard linear mixed model specifies

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad \text{(xxxix)}$$

where $\mathbf{e}$ is $N(\mathbf{0}, \mathbf{R})$, $\mathbf{u}$ is $N(\mathbf{0}, \mathbf{G})$, and $\mathbf{u}$ and $\mathbf{e}$ are independent. The variance matrices $\mathbf{R}$ and $\mathbf{G}$ are parameterized by an unknown variance component parameter $\boldsymbol{\theta}$.

Example 1 *The one-way random effects model*

$$y_{ij} = \mu + u_i + e_{ij} \qquad \text{(xl)}$$

for $i = 1, \ldots, q$ and $j = 1, \ldots, n$, can be written in the general form (xxxix) with total data size $N = qn$ and

$$\begin{aligned}
\mathbf{X} &= \mathbf{1}_N && \text{(xli)}\\
\boldsymbol{\beta} &= \mu && \text{(xlii)}\\
\mathbf{Z} &= \left(z_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ comes from the } i\text{:th group} \\ 0, & \text{otherwise} \end{cases} \right)_{N \times q} && \text{(xliii)}\\
\mathbf{u} &= (u_1, \ldots, u_q)' && \text{(xliv)}\\
\mathbf{R} &= \sigma^2 \mathbf{I}_N && \text{(xlv)}\\
\mathbf{G} &= \sigma_u^2 \mathbf{I}_q && \text{(xlvi)}
\end{aligned}$$

$\mathbf{1}_N$ is a column vector of N ones. The variance component parameter is $\boldsymbol{\theta} = (\sigma^2, \sigma_u^2)$.

Estimation of Fixed Parameters

From (xxxix) the marginal distribution of $\mathbf{y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}' \tag{xlvii}$$

so the log-likelihood of the fixed parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\log L(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{xlviii}$$

For fixed $\boldsymbol{\theta}$, taking the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$, we find the estimate of $\boldsymbol{\beta}$ as the solution of

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{xlix}$$

which is the well-known *generalized* or *weighted least squares* formula. The profile likelihood of $\boldsymbol{\theta}$ is

$$\log L(\boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{l}$$

where $\hat{\boldsymbol{\beta}}$ is computed in (xlix). The observed Fisher information of $\boldsymbol{\beta}$ is

$$I(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \tag{li}$$

from which the standard error for $\boldsymbol{\beta}$ can be found. There is a modified version of the profile likelihood

$$\log L_m(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \tag{lii}$$

that takes into account the estimation of $\boldsymbol{\beta}$. This matches exactly the so-called *restricted maximum likelihood (REML)*, derived by Patterson and Thompson (1971) and Harville (1974) using the marginal distribution of the error term $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$. See also Harville (1977) for further discussion on normal-based variance component estimation. Usually an iterative procedure is required to estimate the fixed parameters.

Estimation of Random Effects

The log-likelihood of all the parameters is based on the joint density of $(\mathbf{y}, \mathbf{u})$, therefore

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) = p(\mathbf{y}|\mathbf{u}) p(\mathbf{u}) \tag{liii}$$

From the mixed model specification, the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$ is normal with mean

$$E(\mathbf{y}|\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{liv}$$

and variance $\mathbf{R}$. The random effects $\mathbf{u}$ is normal with mean zero and variance $\mathbf{G}$, thus

$$\begin{aligned}\log L(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{u}) &= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \\ &\quad -\frac{1}{2}\log|\mathbf{G}| - \frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\end{aligned} \tag{lv}$$

Given the fixed parameters $(\boldsymbol{\beta},\boldsymbol{\theta})$, we take the derivative of the log-likelihood with respect to $\mathbf{u}$:

$$\frac{\partial \log L}{\partial \mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) - \mathbf{G}^{-1}\mathbf{u} \tag{lvi}$$

Setting this to zero, the estimate $\hat{\mathbf{u}}$ is the solution of

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\mathbf{u} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{lvii}$$

This estimate is also known as the *best linear unbiased predictor (BLUP)* (Robinson, 1991). In practice we replace the unknown fixed parameters by their estimates as described above. The second derivative of the log-likelihood with respect to $\mathbf{u}$ is

$$\frac{\partial^2 \log L}{\partial \mathbf{u}\partial \mathbf{u}'} = -\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{G}^{-1} \tag{lviii}$$

so the observed Fisher information is

$$I(\hat{\mathbf{u}}) = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \tag{lix}$$

Assuming the fixed effects are known, the standard errors of $\hat{\mathbf{u}}$ (also interpreted as the *prediction error* for a random parameter) can be computed as the square root of the diagonal elements of $I(\hat{\mathbf{u}})^{-1}$. Prediction intervals for $\mathbf{u}$ are usually computed using

$$\hat{\mathbf{u}}_i \pm z_{\alpha/2}\mathrm{se}(\hat{\mathbf{u}}_i) \tag{lx}$$

where $z_{\alpha/2}$ is an appropriate value from the normal table, and $\mathrm{se}(\hat{\mathbf{u}}_i)$ is the standard error of $\hat{\mathbf{u}}_i$.

Note that $I(\hat{\mathbf{u}})$ is not a function of $\mathbf{u}$. This implies that the log-likelihood of $\mathbf{u}$ alone, assuming $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are known, is **quadratic** around $\mathbf{u}$. Since $\mathbf{u}$ is random we can interpret the likelihood as a density function, so, thinking of $p(\mathbf{u})$ as the 'prior' density of $\mathbf{u}$, the 'posterior' distribution of $\mathbf{u}$ is normal with mean $\hat{\mathbf{u}}$ and variance $(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}$. This is the empirical Bayes interpretation of the formulae.

Joint Estimation of Fixed and Random Effects

Specifically, the derivative of $\log L(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{u})$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \tag{lxi}$$

Combining this with (lvi) and setting them to zero, we have

Definition 2 Henderson's Mixed Model Equation (MME) *(Henderson, 1953)*

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{lxii}$$

The estimates from this simultaneous equation are exactly those we get from (xlix) and (lvii). This classic equation suggest the possibility of estimating $\boldsymbol{\beta}$ and $\mathbf{u}$ without computing the marginal variance $\mathbf{V}$ or its inverse.

Computing Variance Components

The marginal likelihood (l) for the variance component parameter $\boldsymbol{\theta}$ is not desirable due to the terms involving $\mathbf{V}$ or $\mathbf{V}^{-1}$. Here we show an alternative formula which may be easier to compute. One can show the following identities:

$$\begin{aligned}
\mathbf{V}^{-1} &= \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1} && \text{(lxiii)}\\
\hat{\mathbf{u}} &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) && \text{(lxiv)}\\
\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) && \text{(lxv)}\\
(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) + \hat{\mathbf{u}}'\mathbf{G}^{-1}\hat{\mathbf{u}} && \text{(lxvi)}\\
|\mathbf{V}| &= |\mathbf{R}||\mathbf{G}||\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| && \text{(lxvii)}
\end{aligned}$$

Hence we can rewrite (l) as

$$\begin{aligned}
\log L(\boldsymbol{\theta}) &= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})\\
&\quad -\frac{1}{2}\log|\mathbf{G}| - \frac{1}{2}\hat{\mathbf{u}}'\mathbf{G}^{-1}\hat{\mathbf{u}} - \frac{1}{2}\log|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| && \text{(lxviii)}\\
&= \log L(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \hat{\mathbf{u}}) - \frac{1}{2}\log|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| && \text{(lxix)}
\end{aligned}$$

where $\boldsymbol{\theta}$ enters through $\mathbf{R}$, $\mathbf{G}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. The formulae for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ as functions of $\boldsymbol{\theta}$ are given by (xlix) and (lvii). Note that (lxix) is a modified profile likelihood with the extra term (so-called *REML adjustment*)

$$-\frac{1}{2}\log|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| \qquad \text{(lxx)}$$

where the matrix $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}$ is the Fisher information of $\mathbf{b}$ from (li).

Iterative Procedure

In practice, it is convenient to use a derivative free optimization routine to maximize (lxviii) where in the process we also get the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. Computationally we can view the whole estimation procedure as maximizing an objective function

$$\begin{aligned}
Q &= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})\\
&\quad -\frac{1}{2}\log|\mathbf{G}| - \frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u} - \frac{1}{2}\log|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}| && \text{(lxxi)}
\end{aligned}$$

Q is not a log-likelihood, only a device to justify the algorithm. Start with an estimate of the variance component parameter $\boldsymbol{\theta}$, then

Algorithm 1 *To fit a normal linear mixed model,*

1. *Compute $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ using MME.*
2. *Fixing $\boldsymbol{\beta}$ and $\mathbf{u}$ at $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$, maximize Q to update $\boldsymbol{\theta}$.*
3. *Iterate between 1 and 2 until convergence.*

References

J. F. Bjørnstad. On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91(434):791–806, Jun 1996.

R. Butler. A likely answer to 'what is the likelihood function?'. *Statistical decision theory and related topics IV, Vol. 1, eds. S.S. Gupta and J.O. Berger. New York: Springer Verlag.*, 1987.

T. Fears, J. Benichou, and M. Gail. A reminder of the fallibility of the Wald statistic. *American Statistician*, 50:226–227, 1996.

P. Green and S. B.W. *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman and Hall, 1993.

D. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385, 1974.

D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, Jun 1977.

C. R. Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2):226–252, Jun 1953.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678, 1996.

H. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

C. Rao. *Linear Statistical Inference and Its Applications (2nd edn.)*. New York: Wiley, 1973.

G. K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1): 15–32, Feb 1991.