ON MIXED MODELS, REML AND BLUP

Shayle R. Searle

Biometrics Unit and Statistics Center
Cornell University, Ithaca, NY 14853, U.S.A.

BU-1256-M August 1994

ABSTRACT

These notes briefly review some of the main features of the mixed model of analysis of variance, of restricted maximum likelihood (REML) estimation of variance components, and of best linear unbiased prediction (BLUP) of realized but unobservable random effects in mixed models.

1. THREE KINDS OF MODELS

1.1 Fixed effects models

A customary model equation for data from a completely randomized design of, say, n observations in each of a classes is

$$y_{ij} = \mu + \alpha_i + e_{ij} \,, \tag{1}$$

with y_{ij} being the j'th observation in class i, for $i = 1, \cdots, a$ and $j = 1, \cdots, n$. The μ in (1) represents a general mean, α_i is the effect on the datum of its being in class i, and e_{ij} is a residual error defined initially as $e_{ij} = y_{ij} - E(y_{ij})$ for $E(y_{ij}) = \mu + \alpha_i$, where E represents expectation over repeated sampling. In this context μ and the α_is are thought of as fixed, unknown constants, and are called fixed effects. The residual errors e_{ij} are deemed to be random variables: by definition they have zero mean, $E(e_{ij}) = 0$, and we attribute to them homogeneous variance, $\text{var}(e_{ij}) = \sigma_e^2 \; \forall$ i and j, and zero covariances, $\text{cov}(e_{ij}, e_{i'j'}) = 0$ except when $i = i'$ and $j = j'$. These are the standard basics of a traditional analysis of variance model, and because the e_{ij} in (1) are the only random terms, and the α_is are fixed effects, it is known as a fixed effects model. Its familiar sums of squares and mean squares are summarized in the analysis of variance (ANOVA) table of Table 1.

Table 1: Analysis of variance

Term	Degrees of Freedom	Sum of Squares	Mean Square
Classes	a − 1	$SSA = \sum_{i=1}^{a} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$MSA = SSA/(a-1)$
Residual	a(n − 1)	$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2$	$MSE = SSE/[a(n-1)]$
Total, corrected	an − 1	$SST_m = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})^2$	

The prime utility of this table is, on assuming normality (i.e., the e_{ij}s have a normal distribution – and hence the y_{ij}s do, too), that under the hypothesis that the α_is are all equal, the F-statistic F = MSA/MSE follows a Fisher's F-distribution and so can be used to test that hypothesis. A second use for the calculation summarized in the ANOVA table is that E(MSE) = σ_e^2, from which we take an estimate of σ_e^2 as

$$\hat{\sigma}_e^2 = \text{MSE} \, . \tag{2}$$

Furthermore, the best linear unbiased estimator (BLUE) of $\alpha_i - \alpha_{i'}$ is

$$\text{BLUE}(\alpha_i - \alpha_{i'}) = \bar{y}_{i\cdot} - \bar{y}_{i'\cdot} \, . \tag{3}$$

1.2 Random effects models

In the preceding discussion we think of the α_is as constants (i.e., fixed effects) arising from the fact that the a classes from which the data have come are a set of particular classes that have been specifically chosen for study. For example, they might be fertilizers in an agricultural experiment, or drugs in a clinical trial, or different brands of cotton thread in a towel-making factory – and so on. In contrast to this there are situations where the classes have not been specifically chosen but can well be considered as a random sample from some population of classes. For example, the data might be dairy cow milk yields where the cows are daughters of a sample of bulls purchased for possible use in artificial insemination. In this case the classes are bulls – and their corresponding α_is in equation (1) are then random variables. Likewise the data might be responses to a drug administered by a variety of clinics – and the clinics could be considered a random sample. Again the α_is would be random

variables: more carefully put, the α_is in the data would be realized values of unobservable random variables. They are called random effects.

The model equation for either of these examples is exactly the same as (1), but definition of the α_is is different. They are assumed to be random, with zero expectation, uniform variance, and zero covariances with each other and with error terms. Thus we take

$$\begin{aligned} E(\alpha_i) = 0 \quad \text{and} \quad var(\alpha_i) = \sigma_\alpha^2 \quad \forall\ i \\ cov(\alpha_i, \alpha_{i'}) = 0 \quad \forall\ i \neq i' \quad \text{and} \quad cov(\alpha_i, e_{i'j}) \quad \forall\ i, i' \text{ and } j\ . \end{aligned} \tag{4}$$

Then in (1), each α_i and e_{ij} is random and μ is the only fixed effect. This situation is called a random effects model, or just a random model.

So far as the random effects are concerned, one of the prime features of interest is their variance, σ_α^2. From the mean squares of the ANOVA in Table 1 we find that under the conditions (4) the expected values of the mean squares are

$$E(MSA) = n\sigma_\alpha^2 + \sigma_e^2 \quad \text{and} \quad E(MSE) = \sigma_e^2\ .$$

From these, just like (2), we get estimators

$$\hat{\sigma}_e^2 = MSE \qquad \text{and} \qquad \hat{\sigma}_\alpha^2 = (MSA - MSE\ )/n\ . \tag{5}$$

1.3 Mixed models

Suppose we extended the preceding example of a sample of clinics administering one drug is extended to all of them administering the same five drugs, each drug to a number of patients. Then the model equation (1) could be extended to be

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \tag{6}$$

for y_{ijk} being the response by the k'th patient who received drug j in clinic i. (We exclude interactions simply for the purpose of easy illustration of a mixed model.) As before, clinics represented by the α_is would be random, but the drug effects, the β_js, would be fixed effects. Thus we have a mixture of fixed and random effects, and this is called a mixed model.

In truth, of course, all models which have a μ and error terms are a mixture of a fixed effects, μ, and random terms, the errors. But the name mixed model is reserved for models that have a mixture of fixed and random effects other than μ and error terms.

2. A GENERAL MIXED MODEL

2.1 Formulation

The kind of extension from model equation (1) for a 1-way classification to (6) for a 2-way classification can be continued to any number of factors, fixed and random, and including interactions. But to outline some of the characteristics of mixed models in general it is advantageous to resort to matrix and vector notation, writing the model equation as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \,. \tag{7}$$

The symbols here are as follows.

$\mathbf{y}$ is the vector of data.

β is the vector of unknown fixed effects in the data.

$\mathbf{X}$ is the known model matrix corresponding to β; it is often an incidence matrix (with elements 0 or 1), but it can include columns of covariates.

$\mathbf{u}$ is a vector of unknown random effects.

$\mathbf{Z}$ is the known incidence matrix corresponding to $\mathbf{u}$.

$\mathbf{e}$ is the vector of residual error terms.

Properties usually attributed to $\mathbf{u}$ and $\mathbf{e}$ are

$$\begin{aligned} &E(\mathbf{u}) = \mathbf{0} && \text{var}(\mathbf{u}) = \mathbf{D} && \text{cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0} \\ &E(\mathbf{e}) = \mathbf{0} && \text{var}(\mathbf{e}) = \mathbf{R}, = \sigma_e^2\mathbf{I} \text{ in most cases} \,. \end{aligned} \tag{8}$$

The nature of $\mathbf{u}$ is that it is partitioned as

$$\mathbf{u}' = [\mathbf{u}_1' \; \mathbf{u}_2' \cdots \mathbf{u}_i' \cdots \mathbf{u}_r'] \tag{9}$$

to accommodate r random effects factors, with $\mathbf{u}_i'$ having as elements the q_i random effects which occur in the data for the i'th random effects factor. $\mathbf{Z}$ is also partitioned, conformably for the product $\mathbf{Zu}$ with $\mathbf{u}'$ of (9):

$$\mathbf{Z} = [\mathbf{Z}_1 \; \mathbf{Z}_2 \cdots \mathbf{Z}_i \cdots \mathbf{Z}_r] \,. \tag{10}$$

Along with this partitioning of $\mathbf{u}$, it is customary to define

$$\text{var}(\mathbf{u}_i) = \sigma_i^2\mathbf{I}_{q_i} \quad \text{and} \quad \text{cov}(\mathbf{u}_i, \mathbf{u}_{i'}') = \mathbf{0} \quad \text{for } i \neq i'$$

and so (11)

$$\mathbf{D} = \text{diag}\left\{\sigma_i^2\mathbf{I}_{q_i}\right\} \quad \text{for } i = 1, \cdots, r \,.$$

With these prescriptions we then have

$$\begin{aligned} \mathrm{E}(\mathbf{y}\,|\,\mathbf{u}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \\ \mathrm{E}(\mathbf{y}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{var}(\mathbf{y}) &= \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} = \sum_{i=i}^{r} \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i' + \mathbf{R}\,. \end{aligned} \tag{12}$$

2.2 Balanced and unbalanced data

At this point we must emphasize a very important dichotomy of data: balanced and unbalanced. It is a dichotomy that cannot be universally defined mathematically, yet its general description is easily understood. Balanced data are often called equal-subclass-numbers data: they have the same number of observations in every one of the smallest (sub-most) subclasses. Unbalanced data have unequal subclass numbers, including the possibility of empty subclasses. A particular kind of unbalanced data are what can be called planned unbalanced data, such as data from experiments designed as latin squares and balanced incomplete blocks. For example, a latin square of order 3 is really a 3^3 experiment with but nine observations. Its data come from nine subclasses having one observation each and 18 subclasses have no data (see Searle *et al.*, 1992, Section 1.2b-i). Although planned unbalanced data merit special treatment akin to balanced data, we will here just think of them as part of unbalanced data in general.

The important consequence of the balanced-unbalanced dichotomy is that from balanced data, estimation is relatively straightforward.

2.3 Estimation from balanced data

Just as $\alpha_1 - \alpha_2$ of (3) is an estimable function in that simple case, so is $\boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta}$ for any non-null $\boldsymbol{\lambda}'$ in the model based on (7). Therefore, for estimating fixed effects it is appropriate to confine attention in general to estimating $\mathbf{X}\boldsymbol{\beta}$. Then the best linear unbiased estimator (BLUE) and the ordinary least squares estimator (OLSE) of $\mathbf{X}\boldsymbol{\beta}$ are the same:

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{X}^{+}\mathbf{y}\,,$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}^{+}$ is the Moore-Penrose inverse of $\mathbf{X}$.

Estimation of variance components from unbalanced data is a direct extension of (5). For as

many random effects factors as there are in the model, there will be in the usual analysis of variance of that model the same number of mean squares (or, equivalently, sum of squares) whose expected values are just linear combinations of variance components. Arraying those sums of squares in a vector $\mathbf{s}$ and the variance components in a vector $\boldsymbol{\sigma}^2$ we then have

$$\mathrm{E}(\mathbf{s}) = \mathbf{C}\boldsymbol{\sigma}^2 \,, \tag{14}$$

from which we estimate $\boldsymbol{\sigma}^2$ as

$$\hat{\boldsymbol{\sigma}}^2 = \mathbf{C}^{-1}\mathbf{s} \,. \tag{15}$$

This method of estimation is known as the ANOVA method, and for balanced data it yields estimators that have attractive properties. They are minimum variance quadratic unbiased. Under normality assumptions, they are minimum variance unbiased, and their sampling variances are available as are unbiased estimators of those sampling variances. Details of these features are given in Searle *et al.* (1992, Chapter 4).

2.4 Estimating fixed effects from unbalanced data

For unbalanced data, the BLUE and the OLSE of the fixed effects are not necessarily equal.

$$\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{X}^{+}\mathbf{y} \,, \tag{16}$$

the same formal formula as with balanced data. But the BLUE is more complicated:

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \,. \tag{17}$$

It noticeably involves $\mathbf{V}-$ and in a manner that assumes $\mathbf{V}$ to be non-singular. (Singular $\mathbf{V}$ involves even more complications – see Searle, 1994, for a recent discussion.) Five features of (15) are worthy of note.

–i. $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta})$ if and only if $\mathbf{VX} = \mathbf{XB}$ for some $\mathbf{B}$. This is a useful result because whenever $\mathbf{VX} = \mathbf{XB}$ is true one can calculate the BLUE using the OLSE of (16), which is much easier to calculate than (17) since (16) does not involve $\mathbf{V}$.

–ii. $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ of (17) demands knowing the population $\mathbf{V}$ and not just an estimate of it.

–iii. In many situations, $\mathbf{V}$ is not known, and so cannot be calculated. An obvious tactic is to replace $\mathbf{V}$ by some estimator, call it $\hat{\mathbf{V}}$. The resulting expression

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})_{\hat{\mathbf{V}}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \tag{18}$$

can be calculated, but it is not the BLUE.

–iv. The sampling variance of (15) is

$$\text{var}[\text{BLUE}(\mathbf{X}\beta)] = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}' , \tag{19}$$

but when $\mathbf{V}$ is unknown this, just like the BLUE itself, cannot be calculated.

–v. It is tempting to deduce that the sampling variance of (18) is

$$\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\,\mathbf{V}\,\hat{\mathbf{V}}^{-1}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}' . \tag{20}$$

But this is not correct because with $\hat{\mathbf{V}}^{-1}$ being an estimator of $\mathbf{V}^{-1}$ no account has been taken in (20) of the sampling variability inherent in $\hat{\mathbf{V}}$. This difficulty is given attention in Searle *et al.* (1992, Section 9.1e).

2.5 Estimating variance components from unbalanced data

The above discussion of replacing $\mathbf{V}$ in BLUE($\mathbf{X}\beta$) by an estimator of $\mathbf{V}$ highlights the need for estimating variance components, since from (12) we see that elements of $\mathbf{V}$ are various sums of the variance components in the model. Unfortunately the estimation of variance components from unbalanced data is enormously more complicated than from balanced. That prefix "un" causes changes far beyond what one might ever imagine for just two letters of the alphabet. This is vouched for by noting that at least eight different methods of estimation have been propounded in the literature of the last forty years: e.g., ANOVA estimation, of which Henderson's three methods are a subset, ML (maximum likelihood), REML (restricted maximum likelihood), MINQUE (minimum norm quadratic unbiased estimation) and two of its variations, I-MINQUE and MINQUE(0). All of these methods are given detailed treatment in Searle *et al.* (1992). In brief, the ANOVA method is an extension of using analysis of variance sums of squares as in (14) and (15). The difficulty is that there are no guidelines for ascertaining what sums of squares (in fact, more generally, quadratic forms of the observations) provide optimum estimators: the Henderson methods are simply three different ways of choosing some sums of squares (mostly) and one of them does not even provide a unique choice. ML is based on normality assumptions, REML is a variant of ML, and MINQUE requires no distributional assumptions but it does require using some pre-assigned (first guess) values of the variance components. I-MINQUE is an iterative variant of MINQUE (which yields REML solutions) and MINQUE(0) is

MINQUE using a particularly simple set of pre-assigned values for the σ_i^2s, namely zero for all of them except for 1.0 for σ_e^2.

This is no place to go into detail on all of these methods. Rather, we highlight major features of just REML, which is coming to be a much preferred method. (The other would be REML.)

3. REML ESTIMATION OF VARIANCE COMPONENTS

3.1 Development

Restricted maximum likelihood (REML) estimation of variance components is based on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$ as detailed in (7) – (12). It is also based on assuming that each $\mathbf{u}_i$, and $\mathbf{e}$, follow a normal distribution. Under these conditions Hartley and Rao (1967) developed equations for deriving ML (maximum likelihood) estimators, based on $\mathbf{y}$ being normally distributed with the structure set out in (7) – (12). These ML equations are far from being linear in the components of variance and have no closed-form solution. Indeed, for some years the early attempts at designing computing routines for obtaining solutions were not always successful.

An adaptation of ML was suggested by Patterson and Thompson (1971) under a title that refers to recovery of inter-block information. This adaptation is now known as REML (or marginal likelihood, in Europe). It is based on what can be described as wanting to estimate the variance components of a mixed model without having to deal with the fixed effects. This is achieved by concentrating not on the vector of data, $\mathbf{y}$, but on linear combinations of those data, $\mathbf{K}'\mathbf{y}$ with $\mathbf{K}'$ being chosen so that $\mathbf{K}'\mathbf{X} = \mathbf{0}$. Thus for

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e} \\ \mathbf{K}'\mathbf{y} &= \mathbf{K}'\mathbf{X}\boldsymbol{\beta} + \mathbf{K}'\mathbf{Zu} + \mathbf{K}'\mathbf{e} \\ &= \mathbf{K}'\mathbf{Zu} + \mathbf{K}'\mathbf{e} \quad \text{when} \quad \mathbf{K}'\mathbf{X} = \mathbf{0}\,. \end{aligned} \tag{21}$$

Moreover, as well as having $\mathbf{K}'$ satisfy $\mathbf{K}'\mathbf{X} = \mathbf{0}$, $\mathbf{K}'$ is also chosen to have full row rank, so that no element of $\mathbf{K}'\mathbf{y}$ is a linear combination of other elements of $\mathbf{K}'\mathbf{y}$. And that rank is chosen to be its maximum possible value, N – r($\mathbf{X}$), for N being the number of observations (order of $\mathbf{y}$) and r($\mathbf{X}$) being the rank of $\mathbf{X}$.

3.2 Results

Define

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \,. \tag{22}$$

$$\mathbf{u}_0 = \mathbf{e} \quad \text{and} \quad \mathbf{Z}_0 = \mathbf{I} \,. \tag{23}$$

Then the equations that result from maximizing the likelihood of $\mathbf{K}'\mathbf{y} \sim \mathcal{N}(\mathbf{0},\ \mathbf{K}'\mathbf{V}\mathbf{K})$ are

$$\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') = \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \qquad \text{for} \quad i = 0,\ 1,\ \cdots,\ r \,, \tag{24}$$

where $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and there are r random effects factors in the mixed model.

From (12) with $\mathbf{R} = \sigma_e^2\mathbf{I}$ and using (23) we have $\mathbf{V} = \sum_{i\,=\,0}^{r} \sigma_i^2\,\mathbf{Z}_i\mathbf{Z}_i'$, and as such $\mathbf{V}$ is a population value. But in the REML equations (24) we think of $\mathbf{P}$ (through $\mathbf{V}$ and hence $\mathbf{V}^{-1}$) as being a function of the here unknown σ_i^2s; and (24) are equations that have to be solved for those σ_i^2s. Clearly, by (23), those equations are far from linear; and they have to be solved by arithmetical methods (see Section 3.3). Several features of REML estimation are worth noting.

–i. REML estimators, through being maximum likelihood estimators of non-negative parameters (i.e., variances), cannot be negative. Yet equations (24) can have negative solutions. If this occurs, the variance components corresponding to the negative solutions are estimated as zero, the associated factors are deleted from the model, and with that adjusted model (24) is recalculated.

–ii. The large sample variance-covariance matrix of the REML estimators is

$$\mathrm{var}\left(\tilde{\sigma}^2_{REML}\right) = 2\left[\left\{{}_m\ \mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\right\}_{i,\,j\,=\,0}^{\ r}\right]^{-1} \,. \tag{25}$$

The matrix being inverted here is symmetric, of order $r+1$, with each element being the trace of a product of six matrices, as shown in (25).

–iii. For balanced data, the solutions of the REML equations (24) are ANOVA estimators. This is sometimes considered to be a major merit of REML that distinguishes it from ML.

–iv. A considerable amount of algebra is involved in deriving (24) and (25), although some of the early technical reports (e.g., Searle, 1979) do go into excessive detail. A key feature of the derivation is that although there are many matrices $\mathbf{K}'$ of the form specified below (21), results (24) and (25) are invariant to whatever such $\mathbf{K}'$ is used. This in turn depends on $\mathbf{P}$ of (22) being connected to $\mathbf{K}$ by the identity

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

Searle *et al.* (1992) show all the details.

–v. ML estimation has exactly the same form of results as does REML except that in the L.H.S. of (24), and in (25), the $\mathbf{P}$ is replaced by $\mathbf{V}^{-1}$.

–vi. ML estimation includes dealing with β and yields the estimator

$$\mathbf{X}\tilde{\beta} = \mathbf{X}(\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y} \tag{26}$$

where $\tilde{\mathbf{V}}$ is the ML estimator of $\mathbf{V}$. But REML estimation does not deal with β (because $\mathbf{K}'\mathbf{X} = \mathbf{0}$). Nevertheless, having once obtained REML estimates of the variance components one would undoubtedly use them to have a REML estimate of $\mathbf{V}$ and then use that estimate as $\tilde{\mathbf{V}}$ in (26) to estimate $\mathbf{X}\beta$.

3.3 Computing

As already noted, equations (24) cannot be solved analytically. Arithmetical methods have to be used: in some applications (e.g., animal breeding, see Van Vleck, 1994) derivative-free methods are popular, but more often iteration is used, in particular the EM (expectation maximization) algorithm of Dempster *et al.* (1977); see also Searle *et al.* (1992, Section 8.3).

There is a notable connection of the iterative solution of REML equations to the MINQUE method of Rao (1971). First, for $\mathbf{P}$ of (23), observe that $\mathbf{PVP} = \mathbf{P}$. Then for the left-hand side of (24) note that

$$\begin{aligned}\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') &= \mathrm{tr}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}) = \mathrm{tr}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{V}\mathbf{P}) = \mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{V}) \\ &= \mathrm{tr}\Big(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\sum_{j=0}^{r}\mathbf{Z}_j\mathbf{Z}_j'\sigma_j^2\Big) \\ &= \sum_{j=0}^{r+1}\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\sigma_j^2 .\end{aligned}$$

Therefore (24) can be rewritten as

$$\Big\{\underset{m}{\mathrm{tr}}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\Big\}_{i,j=0}^{r}\sigma^2 = \Big\{\underset{c}{\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}}\Big\}_{i=0}^{r} . \tag{27}$$

This is exactly the form of the MINQUE equations except in (27) they have $\mathbf{P}$ replaced by a $\mathbf{P}_0$ which is $\mathbf{P}$ with every σ_i^2 replaced by some pre-assigned numerical value $\sigma_{i,0}^2$. The solution to (27) with $\mathbf{P}_0$ in

place of $\mathbf{P}$ is a MINQUE estimate. This means that a first iterate of (27) and hence of (24), is a MINQUE. Moreover, using a MINQUE estimate for a new $\mathbf{P}_0$ to get a second MINQUE estimate, i.e., iterating MINQUE (I-MINQUE) is identical to REML. We thus have

$$\text{a MINQUE} = \text{a first iterate solution of REML} \tag{28}$$

$$\text{I-MINQUE estimates} = \text{REML solutions} . \tag{29}$$

4. BLUP: BEST LINEAR UNBIASED PREDICTION

4.1 Background

In the mixed model based on $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$ we treat $\mathbf{u}$ as a vector of random variables. In reality, of course, with $\mathbf{y}$ being data, the $\mathbf{u}$ is a vector of realized (unobservable) random variables. But there are many occasions when we would like to estimate, in some sense, those realized values; e.g., student I.Q.s based on a battery of test scores. One of the first attempts at this kind of estimation was a conference paper, Henderson (1950), which aroused strong criticism for suggesting that random variables could be estimated. But Henderson persisted, because in the breeding of dairy cows to increase per-cow milk production one needed (in using artificial insemination) to estimate the genetic value of the bulls to be used as sires. And genetic values are, of course, random variables. In pursuing this kind of estimation the custom thus arose of "predicting" random variables rather than "estimating" them. And so was born the acronym of BLUP: best, linear, unbiased prediction. In truth, of course, we are estimating realized values of random variables and so a more accurate name might be BLUERVAL($\mathbf{u}$) – best, linear, unbiased realized value. Its clumsiness as a name would never dislodge the monosyllabic BLUP from its place in history! But the emphasis on estimating realized values is important.

There are many approaches to deriving BLUP($\mathbf{u}$). Searle *et al.* (1992) have six: two direct, heavily matrix-oriented approaches (Sec. 7.4c and 7.5b), a 2-stage regression method (Sec. 7.5a), a partitioning of $\mathbf{y}$ (Sec. 7.5c), Bayes (Sec. 7.5d) and Henderson's mixed model equations (Sec. 7.6). All six of these treat what I will for the moment call BLUERVAL($\mathbf{u}$) as an estimation methodology different from what is used for deriving BLUE($\mathbf{X}\beta$) in fixed models. Thus a widespread feeling has developed that BLUP is quite different from BLUE. In fact, this is not so.

4.2 BLUE in fixed effects models

Recall that in the fixed effects model with $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta},\ \mathbf{V})$ we can derive BLUE($\mathbf{X}\boldsymbol{\beta}$) quite straightforwardly by seeking $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}'\mathbf{y}$ (linear in $\mathbf{y}$) is unbiased for $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$ and has minimum variance. This leads, as in (17), to BLUE($\mathbf{X}\boldsymbol{\beta}$) $= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. In achieving this result we define unbiasedness for $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$ as

$$\mathrm{E}(\boldsymbol{\lambda}'\mathbf{y}) = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}\ , \qquad \text{and use} \qquad \mathrm{var}(\boldsymbol{\lambda}'\mathbf{y}) = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda}\ . \tag{30}$$

Note that with $\boldsymbol{\beta}$ being fixed effects, i.e., a vector of constants, expressions (30) can just as well be written as

$$\mathrm{E}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}'\mathbf{X}\boldsymbol{\beta}) = 0 \qquad \text{and} \qquad \mathrm{var}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}'\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda}\ . \tag{31}$$

In doing so, we refer to $\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}'\mathbf{X}\boldsymbol{\beta}$ as the prediction error, or estimation error: the difference of the estimator $\boldsymbol{\lambda}'\mathbf{y}$ from the thing being estimated.

4.3 BLUP in mixed models

In the mixed model, based on $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, with $\mathbf{u} \sim (\mathbf{0},\ \mathbf{D})$, we can follow exactly the same procedure for developing BLUP as just described for BLUE, using in place of (31) expressions that are appropriate to mixed model estimation. The procedure is as follows: Seek $\boldsymbol{\lambda}'$ such that $\boldsymbol{\lambda}'\mathbf{y}$ is unbiased for any linear combination of $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{u}$; i.e., for $\mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} + \mathbf{t}_2'\mathbf{u}$, for any non-null $\mathbf{t}_1$ and $\mathbf{t}_2$. We use a linear combination of $\mathbf{X}\boldsymbol{\beta}$ to accommodate estimability, but since $\mathbf{u}$ represents random effects estimability is of no concern for $\mathbf{u}$ and so we use $\mathbf{t}_2'\mathbf{u}$. Now proceed exactly as in using (31) for deriving BLUE($\mathbf{X}\boldsymbol{\beta}$) only with $\mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} + \mathbf{t}_2'\mathbf{u}$ in place of $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$, i.e., the prediction error is $\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}$. Thus we seek $\boldsymbol{\lambda}$ so that

$$\mathrm{E}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}) = \mathbf{0} \qquad \text{and} \qquad \mathrm{var}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}) \equiv \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2'\mathbf{D}\mathbf{t}_2 - 2\mathbf{t}_2'\mathbf{D}\mathbf{Z}'\boldsymbol{\lambda} \tag{32}$$

is minimized with respect to $\boldsymbol{\lambda}$. The result is that what we call BLUP has the form (derived in the appendix)

$$\mathrm{BLUP}(\mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}) = \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta}^o + \mathbf{t}_2'\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o) \tag{33}$$

for

$$\boldsymbol{\beta}^o = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\ . \tag{34}$$

With (33) being true for any $\mathbf{t}_1'$ and $\mathbf{t}_2'$, putting $\mathbf{t}_1' = \mathbf{0}$ and taking $\mathbf{t}_2'$ to be successive rows of $\mathbf{I}$ gives

$$\mathrm{BLUP}(\mathbf{u}) = \mathbf{D}\mathbf{Z}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o)\ . \tag{35}$$

Likewise setting $\mathbf{t}_1$ to be successive rows of $\mathbf{I}$, and $\mathbf{t}_2 = \mathbf{0}$ gives

$$\text{BLUP}(\mathbf{X}\beta) = \mathbf{X}\beta^o = \text{BLUE}(\mathbf{X}\beta)\,. \tag{36}$$

This derivation shows with delightful clarity that BLUP is no more than a direct extension of BLUE based upon defining unbiasedness and variance in terms of prediction (estimation) error and not just in terms of the predictor. In the fixed effects model this means replacing (30) by (31) – although they are the same – but in the mixed model (31) is extended to (32). And what is additionally nice is that the algebra is very straightforward, as seen in the appendix.

4.4 Computing

As with ML estimation of fixed effects, $\mathbf{X}\beta^o$ and BLUP($\mathbf{u}$) are in terms of $\mathbf{D}$ and $\mathbf{V}$. So are expressions for variances and covariances of β^o and BLUP($\mathbf{w}$) – see Searle *et al.* (1992, Sec. 7.5d). Hence again we have the problem of having to estimate variance components to estimate $\mathbf{D}$ and $\mathbf{V}$ in $\mathbf{X}\beta^o$ and BLUP($\mathbf{u}$). Sampling properties of the estimated $\mathbf{X}\beta^o$ and BLUP($\mathbf{u}$) then have the same difficulties as occurs with the ML $\mathbf{X}\tilde{\beta}$.

One characteristic of the expressions β^o and BLUP($\mathbf{u}$) is that they are solutions to equations that Henderson established that have come to be known as the mixed model equations (MMEs). In terms of $\mathbf{D}$ and $\mathbf{R}$ of (8), which occur in $\mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}$, these equations are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \beta^o \\ \text{BLUP}(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{37}$$

β^o and BLUP($\mathbf{u}$) obtained from (37) are identical to (34) and (35), as established in Henderson *et al.* (1959) and shown in Searle *et al.* (1992, Sec. 7.6b). When $\mathbf{R} = \sigma_e^2\mathbf{I}$ and $\mathbf{D} = \text{diag}\left\{\sigma_i^2\mathbf{I}_{q_i}\right\}$, as is often the case, [see (8) and (11)],, equation (37) reduces to the simpler

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\text{diag}\left\{(\sigma_e^2/\sigma_i^2)\mathbf{I}_{q_i}\right\} \end{bmatrix} \begin{bmatrix} \beta^o \\ \text{BLUP}(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}. \tag{38}$$

The advantage of both (37) and (38) is that they have order $p + \sum_{i=1}^{r} q_i$, which is often much less than N, the order of $\mathbf{V}$, the inverse of which is needed for $\beta^o = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

References

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the E.M. algorithm. *J. Roy. Soc. Ser. B*, 39, 1-38.

Hartley, H.O. and Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

Henderson, C.R. (1950). Estimation of genetic parameters (Abstract). *Ann. Math Statist.*, 21, 309-310.

Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.N. (1959). Estimation of environmental and genetic trends in records subject to culling. *Biometrics*, 15, 192-218.

Rao, C.R. (1971). Estimation of variance and covariance components – MINQUE theory. *J. Multivar. Anal.*, 1, 445-456.

Searle, S.R. (1994). Extending some results and proofs for the singular linear model. *Linear Algebra and Its Applications*, 4th Special Issue on Linear Algebra and Statistics (in press).

Searle, S.R. (1979). Notes on variance components estimation. A detailed account of maximum likelihood and kindred methodology. Technical Report BU-673-M, Biometrics Unit, Cornell University, Ithaca, N.Y.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*, Wiley, N.Y.

Van Vleck, L.D. (1994). Experiences with derivative-free REML. *Proceedings of the 1994 Conference on the Interface of Statistics and Computing* (in press).

APPENDIX: Derivation of BLUP

For $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ with $\mathbf{u} \sim (\mathbf{0}, \mathbf{D})$ and $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, we choose to estimate $\mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} + \mathbf{t}_2'\mathbf{u}$ by $\boldsymbol{\lambda}'\mathbf{y}$, seeking $\boldsymbol{\lambda}$ such that both

$$\mathrm{E}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}) = \mathbf{0}\ , \qquad \text{i.e.,} \qquad \boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta} = \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta}\ \ \forall\ \boldsymbol{\beta}\ ,\ \Rightarrow\ \boldsymbol{\lambda}'\mathbf{X} = \mathbf{t}_1'\mathbf{X}\ ,$$

and

$$\mathrm{var}(\boldsymbol{\lambda}'\mathbf{y} - \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} - \mathbf{t}_2'\mathbf{u}) = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2'\mathbf{D}\mathbf{t}_2 - 2\mathbf{t}_2'\mathbf{D}\mathbf{Z}'\boldsymbol{\lambda}$$

is minimized with respect to $\boldsymbol{\lambda}$. Using Lagrange multipliers $2\boldsymbol{\ell}'$ we therefore minimize

$$\theta = \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2'\mathbf{D}\mathbf{t}_2 - 2\mathbf{t}_2'\mathbf{D}\mathbf{Z}'\boldsymbol{\lambda} + 2\boldsymbol{\ell}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{X}'\mathbf{t}_1)\ .$$

$$\partial\theta/\partial\boldsymbol{\lambda} = 0\ \Rightarrow\ 2\mathbf{V}\boldsymbol{\lambda} - 2\mathbf{Z}\mathbf{D}\mathbf{t}_2 + 2\mathbf{X}\boldsymbol{\ell} = \mathbf{0}\ \Rightarrow\ \boldsymbol{\lambda} = \mathbf{V}^{-1}\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \mathbf{V}^{-1}\mathbf{X}\boldsymbol{\ell}\ . \tag{A1}$$

$$\partial\theta/\partial\boldsymbol{\ell} = 0\ \Rightarrow\ \boldsymbol{\lambda}'\mathbf{X} - \mathbf{t}_1'\mathbf{X}\ , \qquad \text{i.e.,} \qquad \mathbf{X}'\boldsymbol{\lambda} = \mathbf{t}_1\ . \tag{A2}$$

Substituting (A1) into (A2) gives

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\ell} = \mathbf{X}'\mathbf{t}_1\ \Rightarrow\ \boldsymbol{\ell} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \mathbf{X}'\mathbf{t}_1)$$

and putting this back into (A1) and using that in $\boldsymbol{\lambda}'\mathbf{y}$ gives

$$\begin{aligned}\boldsymbol{\lambda}'\mathbf{y} = \mathrm{BLUP}(\mathbf{t}_1'\mathbf{X}\boldsymbol{\beta} + \mathbf{t}_2'\mathbf{u}) &= \mathbf{t}_1'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \mathbf{t}_2'\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y} \\ &= \mathbf{t}_1'\mathbf{X}\boldsymbol{\beta}^{o} + \mathbf{t}_2'\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{o})\ ,\end{aligned}$$

which is (33).