That BLUP Is a Good Thing: The Estimation of Random Effects

G. K. Robinson

Abstract. In animal breeding, Best Linear Unbiased Prediction, or BLUP, is a technique for estimating genetic merits. In general, it is a method of estimating random effects. It can be used to derive the Kalman filter, the method of Kriging used for ore reserve estimation, credibility theory used to work out insurance premiums, and Hoadley’s quality measurement plan used to estimate a quality index. It can be used for removing noise from images and for small-area estimation. This paper presents the theory of BLUP, some examples of its application and its relevance to the foundations of statistics.

Understanding of procedures for estimating random effects should help people to understand some complicated and controversial issues about fixed and random effects models and also help to bridge the apparent gulf between the Bayesian and Classical schools of thought.

Key words and phrases: Best linear unbiased prediction (BLUP), estimation of random effects, fixed versus random effects, foundations of statistics, likelihood, selection index, Kalman filtering, parametric empirical Bayes methods, small-area estimation, credibility theory, ranking and selection.

1. INTRODUCTION

The acronym BLUP stands for “Best Linear Unbiased Prediction” and is in common usage in animal breeding. It is a method of estimating random effects.

The context of BLUP is the linear model

\[ y = X\beta + Zu + e \] (1.1)

where \( y \) is a vector of \( n \) observable random variables, \( \beta \) is a vector of \( p \) unknown parameters having fixed values (fixed effects), \( X \) and \( Z \) are known matrices, and \( u \) and \( e \) are vectors of \( q \) and \( n \), respectively, unobservable random variables (random effects) such that \( E(u) = 0, E(e) = 0 \) and

\[ \text{Var}[u, e] = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \sigma^2 \]

where \( G \) and \( R \) are known positive definite matrices and \( \sigma^2 \) is a positive constant.

At times, we will discuss the estimation of dispersion parameters and will use \( \theta \) to denote a vector of dispersion parameters on which the matrices \( G \) and \( R \) depend. Generally, it will be assumed that the variance-covariance structure is known except perhaps for the single parameter \( \sigma^2 \).

BLUP estimates of the realized values of the random variables \( u \) are linear in the sense that they are linear functions of the data, \( y \); unbiased in the sense that the average value of the estimate is equal to the average value of the quantity being estimated; best in the sense that they have minimum mean squared error within the class of linear unbiased estimators; and predictors to distinguish them from estimators of fixed effects. A convention has somehow developed that estimators of random effects are called predictors while estimators of fixed effects are called estimators. As discussed in Section 7.1, I prefer to use the term “estimators” for both fixed and random effects.

Mathematically, the BLUP estimates \( \hat{\beta} \) of \( \beta \) and \( \hat{u} \) of \( u \) are defined as solutions to the following simultaneous equations which were given by Henderson (1950), although in summation rather than matrix form:

\[ X^T R^{-1} X \hat{\beta} + X^T R^{-1} Z \hat{u} = X^T R^{-1} y \]

\[ Z^T R^{-1} X \hat{\beta} + (Z^T R^{-1} Z + G^{-1}) \hat{u} = Z^T R^{-1} y. \] (1.2)

These equations have sometimes been called
"mixed model equations," and $\hat{\beta}$ and $\hat{u}$ referred to as "mixed model solutions." Note that as $G^{-1}$ tends to the zero matrix these equations tend formally to the generalized least-squares equations for estimating $\beta$ and $u$ when the components of $u$ are regarded as fixed effects.

Henderson (1975) showed that provided $X$ is of full rank, $p$, the variance-covariance matrix of estimation errors is

$$
E\left[\begin{bmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{bmatrix} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{bmatrix}^T\right] = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1} \sigma^2.
$$

That BLUP estimates generally differ from the generalized least squares estimates that would be used if $u$ were regarded as fixed is illustrated by the following example.

**Example.** A simple example of model (1.1) is that of first lactation yields of dairy cows with sire additive genetic merits being treated as random effects ($u$) and herd effects being treated as fixed effects ($\hat{\beta}$). The matrix $Rh^2$ is the variance-covariance matrix of the vector $e$ of departures from a model in which yield was explicable entirely by sire effects and herd effects. The matrix $R$ will be taken to be the identity matrix. Assume that the matrix $G$ is a known multiple of the identity matrix, say $0.1I$. This would be a reasonable assumption provided that the sires were unrelated and provided that the variance ratio had been estimated previously.

Suppose that we had records as follows.

<table>
<thead>
<tr>
<th>Herd</th>
<th>Sire</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>110</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>100</td>
</tr>
</tbody>
</table>

Then the entities in equation (1.1) are

$$
y = (110, 100, 110, 100, 110, 110, 100, 100)^T,
$$

$$
\beta = (h_1, h_2, h_3)^T
$$

where $h_i$ is the environmental effect of the $i$th herd,

$$
u = (s_A, s_B, s_C, s_D)^T.
$$

where $s_j$ is the effect of the $j$th sire on his daughters' lactation yields,

$$
X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}
$$

and

$$
Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.
$$

Equation (1.2) gives us

$$
\begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 \\ 0 & 4 & 0 & 2 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 1 & 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \hat{h}_3 \\ s_A \\ s_B \\ s_C \\ s_D \end{bmatrix} = \begin{bmatrix} 210 \\ 310 \\ 420 \\ 110 \\ 110 \\ 220 \\ 500 \end{bmatrix},
$$

which has solution

$$
\hat{\beta} = (105.64, 104.28, 105.46)^T,
$$

$$
\hat{u} = (0.40, 0.52, 0.76, -1.67)^T.
$$

If sire effects were treated as fixed, then equation (1.2) would be changed by omission of $G^{-1}$. This means that the last four diagonal elements in the left-hand-side matrix of equation (1.3) would be reduced by 10. The matrix equation for $\hat{\beta}$ and $\hat{u}$ would no longer be of full rank, but a solution can


be obtained by setting, arbitrarily, \( s_D = 0 \). This solution is
\[
\hat{\beta} = (100, 100, 100)^T, \\
\hat{u} = (10, 10, 10, 0)^T.
\]

The solution given by (1.5) is the least-squares solution with which most statisticians are well acquainted. Intuitively, each of the sires other than \( D \) has daughters that yield 10 units more on average than the daughters of sire \( D \) to which they can be directly compared.

The BLUP solution given by (1.4) takes into account the information that sire effects have less variation than the variance of lactation yields from daughters of a single sire. The extent to which a sire's estimated genetic merit is regressed toward the mean depends on the amount of information available concerning that sire. For instance, sire \( C \) is estimated to be better than sires \( A \) and \( B \) because more is known about him—the lactation yields of his daughters are the same (110) as those of sires \( A \) and \( B \).

The variance-covariance matrix of the estimates from the mixed model is \( \sigma^2 \) times the inverse of the left-hand-side matrix in equation (1.3). The diagonal elements of the inverse matrix for \( s_A, s_B, s_C \) and \( s_D \) are 0.0954, 0.0941, 0.0916 and 0.0833, respectively. Since the merit of a sire about which nothing was known would have a variance of \( 0.1 \sigma^2 \), there has been little gain in precision of sire effect estimates due to data on lactations of daughters.

**Remarks.** In this example, numbers with few significant digits have been used in order to make the example easier to follow. Consequently, variance parameters should not be estimated from the given data. In practical situations, the variance ratio that was taken to be 0.1 or the variance \( \sigma^2 \) may need to be estimated from the same data as is used to estimate sire genetic merits.

My introduction to the estimation of random effects was as statistician for the Australian Dairy Herd Improvement Scheme in mid-1980. This means that I think first of the estimation of genetic merits of dairy cattle when I think of estimating random effects. Readers might like to allow for this point of view.

### 2. Objectives

In a discussion at the Royal Statistical Society, Dawid (1976) remarked

A constant theme in the search for justification for what statisticians do. To read the textbooks, one might get the distorted idea that 'Student' proposed his \( t \)-test because it was the Uniformly Most Powerful Unbiased test for a Normal mean, but it would be more accurate to say that the concept of UMPU gains much of its appeal because it produces the \( t \)-test, and everyone knows that the \( t \)-test is a good thing.

The words "a good thing" in the title of this paper are to be interpreted as coming from this quotation. I wish to argue that the BLUP method for estimating random effects is "a good thing" just as Student's \( t \)-test is "a good thing."

I believe that the Classical school of thought in statistical inference should accept estimation of random effects as a legitimate activity. This theme will be developed in Section 4.3, which gives a classical justification for BLUP, and in Section 6, which lists applications. If estimation of random effects were accepted as legitimate by the Classical school, then the Bayesian and Classical schools of thought in statistics would differ less than much current rhetoric suggests.

Another objective is to encourage communication between people who deal with the various applications where random effects are estimated. The 50th Anniversary Conference, Iowa State Statistical Laboratory, encouraged such communication. See Harville (1984). Much theory has been developed separately in each of several areas of application and further theoretical work in each area might be assisted by looking at other fields. The computing problems associated with estimating random effects might also be alleviated by learning about methods used in other areas of application. See also Kackar and Harville (1984) and Robinson and Jones (1987) on the computational problems of estimating standard errors.

Another objective of this paper is to ask people to question the meanings of some fundamental statistical ideas. These include unbiasedness, likelihood, and the distinction between fixed and random effects. They will be discussed further in Section 7.

### 3. Structure

Section 1 of this paper has introduced BLUP and the estimation of random effects without justifying the mathematical formulae used. Section 4 presents some basic theory on estimation of random effects assuming \( \theta \) is known. This shows that BLUP can be derived in many different ways and is robust with respect to philosophy of statistics. Section 5 discusses the relationship between estimation of random effects and other theoretical ideas. Its
purposes is to show that understanding the estimation of random effects can help with the understanding of other theory. Section 6 reviews applications involving estimation of random effects. It shows that many groups of people are estimating random effects and that it makes sense. Section 7 reviews some fundamental ideas about statistics, suggesting that an understanding of estimation of random effects should influence our approaches and attitudes.

4. DERIVATIONS OF BLUP

Four derivations of BLUP are given below. Those in Sections 4.1 and 4.2 require the assumption of normality. Those in Sections 4.3 and 4.4 do not require normality as they only use first and second moments.

4.1 Henderson’s Justification

Henderson (1950) described the BLUP estimates (1.2) as being “joint maximum likelihood estimates.” Henderson (1973, page 16) explained that his derivation had actually been to assume that $u$ and $e$ are normally distributed and to maximize the joint density of $y$ and $u$ with respect to $\beta$ and $u$. He suggests that this should not be called “maximum likelihood” because the function being maximized is not a likelihood.

The joint density of $y$ and $u$ is

\[
(2\pi\sigma^2)^{-\frac{1}{2}n-1} \left[ \det \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right]^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta - Zu)^T \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{bmatrix} y - X\beta - Zu \end{bmatrix} \right\}.
\]

(4.1)

To maximize this with respect to $\beta$ and $u$ requires minimizing

\[
(y - X\beta - Zu)^T G^{-1} \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} (y - X\beta - Zu) + u^T G^{-1} u + (y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu).
\]

Differentiating this with respect to $\beta$ and $u$ using the usual rules for vector differentiation of scalar functions and equating the derivatives to zero gives Henderson’s mixed model equations (1.2).

4.2 Bayesian Derivation

A Bayesian derivation of BLUP is straightforward. Regard $\beta$ as a parameter with a uniform, improper prior distribution and $u$ as a parameter which has a prior distribution that has mean zero and variance $G\sigma^2$, independent of $\beta$. Given $\beta$ and $u$, the density of $y$ is

\[
(2\pi\sigma^2)^{-\frac{1}{2}n-1} \left[ \det (R) \right]^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta - Zu)^T R^{-1} \begin{bmatrix} y - X\beta - Zu \end{bmatrix} \right\}.
\]

The prior density is

\[
(2\pi\sigma^2)^{-\frac{1}{2}} \left[ \det (G) \right]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (u^T G^{-1} u) \right\}.
\]

Therefore the posterior density for $\beta$ and $u$ is proportional to expression (4.1), and so the posterior mode is given by the BLUP estimates.

Dempfe (1977) gave a Bayesian presentation along these lines. Lindley and Smith (1972) presented a derivation which is equivalent to this.

It is generally true that Bayesian procedures are not affected by a stopping rule, provided that the stopping rule depends only on the data included in the analysis. This can be of substantial consolation in some applications. In estimating genetic merits of animals, mating and culling decisions depend on available information. Henderson (1965) investigated the conditions under which BLUP estimates are unbiased despite selection. In geostatistics, decisions about where to drill are based on data available at the time.

4.3 Within the Classical School

The simplest case. The simplest case of estimation of random effects is in the estimation of residuals from a simple normal model.

Suppose that $n$ observations are taken from a normal population which has mean $\mu$ and variance $\sigma^2$ known to be 1. If the observations are $X_1, X_2, \ldots, X_n$ with mean $\bar{X}$, then it is common to estimate the parameter $\mu$ by $\bar{X}$. Other estimates may be used if robustness in some sense is required, but we will here assume that $\bar{X}$ is the most desirable estimator.

The model could be written in the form

$X_i = \mu + \varepsilon_i$,

where $\varepsilon_i$ is the error associated with the $i$th observation and comes from a standard normal distribution. These errors are also called residuals, being what is left of the observational data after the deterministic component is removed. Now, we might wish to ask: “What is the best estimate of $\varepsilon_i$?” The obvious estimate of the residual $\varepsilon_i$ is

$\hat{\varepsilon}_i = X_i - \bar{X}$. 
Properties of the residuals as estimators of the unknown errors are the following.

1. They are linear in the data.
2. They are unbiased in the sense that
   \[ E[\hat{\varepsilon}_i] = E[e_i] . \]
   Note however that \( E[\hat{\varepsilon}_i | e_i] \) is nearer to zero than is \( e_i \). In some circumstances, people tend to expect ‘unbiased’ to be interpretable as meaning that the expectation of an estimate of a random effect given the true value of the random effect is equal to that random effect. This is not the case.
3. They have minimum mean square error amongst the class of linear unbiased estimators.

In this very simple case, none of this is very interesting or enlightening; but it is noticeable that there is a situation where estimation of random effects is standard practice. We need to consider situations involving more than one source of variation before anything nontrivial happens. However, it does suggest that it is reasonable to ask that estimates of random effects be linear, unbiased and minimum mean square error.

The general case. Henderson (1963) showed using Lagrange multipliers that BLUP estimates of linear combinations of fixed and random effects are the estimates that satisfy the classical requirements of being linear, unbiased and minimum mean square error. Harville (1976) showed, further, that the Gauss–Markov theorem could be extended to cases when matrices \( G \) and \( R \) are of less than full rank. See also Ishii (1969, Example 2, pages 482–487).

A more intuitive approach to showing that the BLUP estimates have minimum mean squared error within the class of linear unbiased estimates was given by Harville (1990). First, note that
\[ E[yy^T] = X\beta^T X^T + ZGZT\sigma^2 + R\sigma^2 \]
\[ E[uy^T] = GZT\sigma^2 \]
and that, as Henderson, Kempthorne, Searle and von Krosigk (1959) showed, an alternative form for the BLUP estimates is
\[ \hat{\beta} = \left( X^T(R + ZGZ^T)^{-1} X \right)^{-1} X^T(ZGZ^T + R)^{-1} y \]
\[ \hat{u} = (Z^T R^{-1} Z + G^{-1})^{-1} \left[ Z^T R^{-1} - Z^T R^{-1} X \right]^{-1} X^T (R + ZGZ^T)^{-1} \left[ X^T(R + ZGZ^T)^{-1} X \right]^{-1} y. \]
Linear unbiased estimates of zero are of the form \( a^T y \), where \( a \) satisfies \( X^T a = 0 \). They are uncorrelated with the errors of BLUP estimates, since
\[ E[(\hat{\beta} - \beta)^T y^T a] \]
\[ = E[(X^T(R + ZGZ^T)^{-1} X)^{-1} X^T(ZGZ^T + R)^{-1} \cdot E[yy^T] a - \beta E[y^T] a \]
\[ = \beta \beta^T X^T a + \left( X^T(R + ZGZ^T)^{-1} X \right)^{-1} X^T \sigma^2 \]
\[ = 0 \]
and
\[ E[(\hat{u} - u)y^T a] \]
\[ = (Z^T R^{-1} Z + G^{-1})^{-1} \left[ Z^T R^{-1} - Z^T R^{-1} X \right]^{-1} X^T \]
\[ \cdot \left( (R + ZGZ^T)^{-1} \cdot E[yy^T] a - E[u^T y] a \right) \]
\[ = (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} (ZGZ^T + R) \sigma^2 \]
\[ = GZT\sigma^2 \]
\[ = 0. \]

Any linear unbiased estimator of a linear combination \( b^T \beta + c^T u \) of fixed and random effects must be of the form \( b^T \hat{\beta} + c^T \hat{u} + a^T y \), where \( X^T a = 0 \). Its variance-covariance matrix of estimation errors is
\[ E[\{ b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) + a^T y \}] \]
\[ = E[\{ b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) \}] \]
\[ = E[\{ b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) \}] + E[a^T yy^T] a \]
\[ + E[b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u)] y^T a \]
\[ = E[\{ b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) \}] \]
\[ \cdot (b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) + a^T y)] \]
\[ = E[\{ b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) \}] \]
\[ \cdot (b^T(\hat{\beta} - \beta) + c^T(\hat{u} - u) + a^T y)] \]
\[ + E[a^T yy^T] a. \]

Now \( E[a^T yy^T] a = E[a^T y(a^T y)^T] \) is a symmetric positive semidefinite matrix, so this variance-covariance matrix of estimation errors exceeds that for the BLUP estimate \( b^T \hat{\beta} + c^T \hat{u} \).

4.4 Goldberger’s Derivation

Goldberger (1962) considered a linear model
\[ y = X\beta + \varepsilon, \]
where the disturbance $\varepsilon$ satisfies $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \Omega$. Given a new (observable) vector $x_\ast$ of regressors and (unobservable) prediction disturbance $\varepsilon_\ast$, which is correlated with the disturbances for the data already obtained, satisfying

$$
E(\varepsilon_\ast) = 0 \\
E(\varepsilon_\ast^\top \varepsilon_\ast) = w^\top \Omega w.
$$

Goldberger’s equation (3.12) tells us that the best linear unbiased predictor of the future observation $y_\ast = x_\ast^\top \beta + \varepsilon_\ast$ is

$$
x_\ast^\top (X^\top \Omega^{-1}X)^{-1} X^\top \Omega^{-1} y + w^\top \Omega^{-1} y \\
- w^\top \Omega^{-1} X (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y.
$$

For our model, $\varepsilon = Zu + e$, so

$$
\Omega = (ZGZ^T + R)\sigma^2.
$$

To estimate $x_\ast^\top \beta + z_\ast^\top u$ take

$$
\varepsilon_\ast = z_\ast^\top u
$$

and hence

$$
w^\top = E[z_\ast^\top u (Zu + e)^\top] = z_\ast^\top GZ^T \sigma^2.
$$

So Goldberger’s derivation tells us that the best linear unbiased predictor of $x_\ast^\top \beta + z_\ast^\top u$ is

$$
x_\ast^\top \left[ X^\top (ZGZ^T + R)^{-1} X \right]^{-1} X^\top (ZGZ^T + R)^{-1} y \\
+ z_\ast^\top GZ^T (ZGZ^T + R)^{-1} y \\
- z_\ast^\top GZ^T (ZGZ^T + R)^{-1} \\
\cdot X \left[ X^\top (ZGZ^T + R)^{-1} X \right]^{-1} \\
\cdot X^\top (ZGZ^T + R)^{-1} y,
$$

which is

$$
x_\ast^\top \hat{\beta} + z_\ast^\top GZ^T (ZGZ^T + R)^{-1} (y - \hat{\beta}).
$$

Using the matrix identity (5.2), below, and the second of the simultaneous equations (1.2), this is

$$
x_\ast^\top \hat{\beta} + z_\ast^\top (Z^\top R^{-1} Z + G^{-1})^{-1} Z^\top R^{-1} (y - \hat{\beta}) \\
= x_\ast^\top \hat{\beta} + z_\ast^\top \hat{u}.
$$

Thus Goldberger’s predictor is the same as that given by (1.2).

To the best of my knowledge, Goldberger was the first to use the term “best linear unbiased predictor” and Henderson started using the acronym BLUP in 1973.

Goldberger’s derivation seems unobjectionable from a Classical viewpoint. His emphasis is on prediction, but his formulae still apply generally to prediction of a future observation, $y_\ast$, which is perfectly correlated with a past disturbance, so they do apply to estimation of random effects.

5. LINKS WITH OTHER STATISTICAL THEORY

5.1 Recovery of Inter-Block Information

Henderson, Kempthorne, Searle and von Krosigk (1959, page 196) showed that the BLUP estimate $\hat{\beta}$ is identical with the generalized least-squares estimate of $\beta$ that would be obtained after recovery of inter-block information if the random effects $u$ were block effects. They eliminated $\hat{u}$ from equation (1.2), giving

$$
X^\top R^{-1} X \hat{\beta} \\
- X^\top R^{-1} Z (Z^\top R^{-1} Z + G^{-1})^{-1} Z^\top R^{-1} X \hat{\beta} \\
= X^\top R^{-1} y - X^\top R^{-1} Z (Z^\top R^{-1} Z + G^{-1})^{-1} Z^\top R^{-1} y.
$$

Now using a matrix identity which is commonly used in this subject area

$$
(R + ZGZ^T)^{-1} = R^{-1} - R^{-1} Z (Z^\top R^{-1} Z + G^{-1})^{-1} Z^\top R^{-1},
$$

gives

$$
X^\top (R + ZGZ^T)^{-1} \hat{X} = X^\top (R + ZGZ^T)^{-1} y.
$$

Hence

$$
\hat{\beta} = \left[ X^\top (R + ZGZ^T)^{-1} X \right]^{-1} \left[ X^\top (R + ZGZ^T)^{-1} y, \right.
$$

which can be seen to be the generalized least squares estimate of $\beta$, since the variance-covariance matrix of the random effects is

$$
E[(Zu + e)(Zu + e)^\top] = R + ZGZ^T.
$$

The matrix identity (5.1) is a particular case of

$$
(A + UBV) \left[ A^{-1} - A^{-1} U \\
\cdot (I + BVA^{-1} U)^{-1} BVA^{-1} \right] \\
= I + UBVA^{-1} - U(I + BVA^{-1} U)^{-1} BVA^{-1} \\
- UBVA^{-1} U(I + BVA^{-1} U)^{-1} BVA^{-1} = I
$$

of which the history and many variants, generalizations and special cases are discussed by Henderson.
and Searle (1981). In this paper, we will also use the matrix equality

\[ (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} = GZ^T (R + ZGZ^T)^{-1}, \]

which can be derived from (5.1).

Recovery of interblock information is most commonly discussed for experimental data from incomplete block designs. Recovery of the interblock information improves the efficiency of the estimates of the fixed effects, but the estimates based only on intra-block information are often considered to be satisfactory.

For unbalanced data, estimates of fixed effects based only on intra-block information can sometimes be quite unsatisfactory. An example of this is presented below. It was discussed in Henderson, Kempthorne, Searle and von Krosigk (1959).

**EXAMPLE.** Fifty cows produce an average of 100 kilograms of butterfat in their first lactations. Forty of these cows survive to complete their second lactations. The average first lactation butterfat yield of these 40 cows is 110 kilograms and the average second lactation butterfat yield is 140 kilograms. Estimate the average difference between lactations in the absence of culling!

One answer is \((140 - 100)\) kg = 40 kg, using all cows. Another answer is \((140 - 110)\) kg = 30 kg, using only the cows which completed second lactations. This is the estimate that uses only intra-cow information. If the true correlation between first and second lactations can be taken to be 1/2, then recovery of inter-cow information gives the estimate 35 kg.

Intuitively, the cows culled are likely to be worse than average. Therefore the cows completing second lactations are likely to be better than average; so the first answer is likely to be too large. The cows not culled are unlikely to be as good as they appear to be because the data has been selected. An extreme case to illustrate this is that if first lactation yield and second lactation yield were uncorrelated, then culling on first lactation would not increase average second lactation yield but it would increase the first lactation yield of cows completing a second lactation by rejecting some data. The second answer is likely to be too small because the 110 is an overestimate.

In this example, the effect of lactation parity is being regarded as a fixed effect (treatment) and cow effects are being regarded as random effects (blocks). Thinking of the BLUP estimate of the lactation parity effect as being a Bayesian estimate, the likelihood principle tells us that the estimate does not need to be modified if culling has taken place, provided that culling decisions were based only on the data included in the analysis. Within the Classical framework, Henderson (1975) showed that selection and culling which is based on linear combinations \((L^T y)\) of the data \(y\) do not affect the optimality of the BLUP estimates provided that \(L^T X = 0\). An aspect of this formalization that I cannot understand is the meaning of selection based on the \(L\) matrix. In examples given in Henderson (1973, 1975, 1984), the numbering of the random effects is always such that best is first but ranking is not a linear function of the data. The problem has been of interest to Henderson throughout his career, but his work involving the \(L\) matrix is not widely understood or accepted.

Estimates of environmental and genetic trends from dairying data tend to suffer from biases similar to this, unless the inter-cow information is recovered. Of course, the resulting estimates are sensitive to the values used for dispersion parameters, such as the correlation in the example above, but to not recover information is equivalent to using extreme values for dispersion parameters, and is worse.

### 5.2 Random Effects Models

Mathematically, it is easy to see from equations (1.2) and (5.2) that when there are no fixed effects the BLUP estimates of the random effects are given by

\[ (Z^T R^{-1} Z + G^{-1}) \hat{u} = Z^T R^{-1} y \]

and have variance-covariance matrix

\[ (Z^T R^{-1} Z + G^{-1})^{-1} \sigma^2. \]

In animal breeding, BLUP for random effects models is known as the selection index. See Smith (1936) and Hazel (1943). Lush (1949) referred to it as “most probable producing ability.”

Henderson (1963) describes BLUP as a form of selection index. In our notation, if \( \beta \) were known, then \( y - X\beta = Zu + e \) follows a random effects model, so the best estimate of \( u \) is

\[ (Z^T R^{-1} Z)^{-1} Z^T R^{-1} (y - X\beta). \]

If we replace \( \beta \) by the estimate

\[ \hat{\beta} = \left[ X^T (R + ZGZ^T)^{-1} X \right]^{-1} X^T (R + ZGZ^T)^{-1} y, \]

then the resulting estimate of \( u \) is the BLUP estimate, as shown by Henderson (1963).
Box and Tiao (1968) presented a detailed derivation of BLUP estimates for random effects models within a Bayesian framework. Dempfle (1977) gave the following Bayesian derivation, which I find intuitively helpful. The idea is due to Robertson (1955) and is only applicable when the matrix $Z$ is of full rank.

The generalized least squares estimate of $u$ is

$$\hat{u}_1 = (Z^TR^{-1}Z)^{-1}Z^TR^{-1}y$$

and has precision

$$\text{Var}(\hat{u}_1 - u) = (Z^TR^{-1}Z)^{-1}.$$ 

The prior estimate of $u$ is

$$\hat{u}_2 = 0$$

and has precision

$$\text{Var}(\hat{u}_2 - u) = G.$$ 

The best estimate of $u$ gives these two estimates weight in inverse proportion to their precision and is

$$\hat{u} = \left[ \text{Var}(\hat{u}_1 - u)^{-1} + \text{Var}(\hat{u}_2 - u)^{-1} \right]^{-1}$$

$$\cdot \left[ \text{Var}(\hat{u}_1 - u)^{-1}\hat{u}_1 + \text{Var}(\hat{u}_2 - u)^{-1}\hat{u}_2 \right]$$

$$= [Z^TR^{-1}Z + G^{-1}]^{-1}Z^TR^{-1}y.$$ 

A straightforward classical derivation is to use standard results on the multivariate normal distribution (e.g., Searle, 1971, page 47). Since $u$ and $y$ have zero means and variance-covariance matrix

$$\left( \begin{array}{cc} G & GZ^T \\ ZG & ZG^T + R \end{array} \right) \sigma^2,$$

the distribution of $u$ given $y$ has mean

$$GZ^T(ZG^T + R)^{-1}y$$

(5.5)

and variance

$$\left[ G - GZ^T(ZG^T + R)^{-1}ZG \right] \sigma^2$$

$$= (Z^TR^{-1}Z + G^{-1})^{-1} \sigma^2$$

in agreement with (5.3) and (5.4). This derivation shows that, when there are no fixed effects to be estimated simultaneously, the theory of estimating random effects follows the theory of correlation very closely.

Ideas about correlation are quite old. Pearson (1896, page 261) wrote

The fundamental theorems of correlation were for the first time and almost exhaustively discussed by BRAVAIS ('Analyse Mathématique sur les probabilités des erreurs de situation d'un point'. Mémoires par divers Savants, T. IX., Paris, 1846, pp 255–332) nearly half a century ago. He deals with the correlation of two and three variables ... GALTON ... introduced an improved notation ...

Random effects models are also related to the idea of regression to the mean attributed by Davis (1986) to Galton. The best estimate of a characteristic of an offspring given the characteristics of the parents is regressed towards the population mean from the parental average.

**Example.** Suppose that true intelligence quotient (IQ) is normally distributed with mean 100 and standard deviation 15. Two tests are available. Both tests give scores that are normally distributed with mean the true IQ. The first test score has standard deviation 10 given true IQ, while the second test score has standard deviation 5. A person scoring 130 on the first test would be estimated to have a true IQ of 120.8 and a person scoring 130 on the second test would be estimated to have a true IQ of 127. Features of these estimates worth noting are as follows.

- They are shrunk towards the overall mean (100) from the data. The amount of shrinkage is greater when the data point is less informative.
- They are biased given true IQ. This is obvious since the raw scores are unbiased and the estimates are nontrivial linear functions of the raw scores.
- They have zero average bias when averaged over the distribution of possible true IQs.
- The expected value of true IQ given the data is equal to the BLUP estimate of IQ, by (5.5).

This example is far from new. In the discussion to Lindley and Smith (1972), Novick suggested that Kelley (1927) was familiar with the basic ideas of shrinkage estimators. Henderson (1973, page 15) explained that considering this example was crucial in his development of BLUP in 1949.

### 5.3 Fixed Effects Models and Admissibility

Fixed effects models are, of course, a particular case of mixed models. Stein's (1956) demonstration that the sample mean is inadmissible for the mean of a multidimensional normal population of known variance when the dimensionality is at least three has led to some theoretical work that I believe to be
of little practical value. This work is characterized by a tendency to combine unrelated estimation problems. BLUP helps us to know when to combine estimation problems. Situations where estimation problems ought to be combined are when the parameters to be estimated can be regarded as coming from some distribution. Equivalently, they are "exchangeable," or are "random effects." I agree with the view expressed by E. F. Harding in discussion of Lindley and Smith (1972) that estimates of the characteristics of butterflies in Brazil, ball bearings in Birmingham, and brussels sprouts in Belgium ought not to be related to each other.

5.4 Estimation of Variance Parameters

The estimation of variance parameters is a very extensive topic. See, for instance, Khuri and Sahai (1985). The comments below concentrate on one method of estimating variance parameters, REML, which can be interpreted as either Classical or Bayesian.

For balanced experimental data, the analysis of variance provides estimates which are often considered acceptable. Sometimes estimates of variance components are negative, in which case they are taken to be zero.

For unbalanced data, REML is the method of estimating variance components that seems to have the best credentials from a Classical viewpoint. See Robinson (1987) for a recent discussion with examples. It was expounded by Patterson and Thompson (1971). They called it "modified maximum likelihood." Some people now refer to it as "restricted maximum likelihood,"-while others use the term "residual maximum likelihood."

Thompson (1973) generalized REML to the multivariate case and showed that it may be used even when the data available has been selected in certain ways.

Consider the problem of estimating $\theta$ for the linear model given by equation (1.1). Bayesian statisticians would, in principle, start with a joint prior distribution for $\theta$, $\beta$ and $\mu$. If a uniform prior distribution is used for $\beta$, then the posterior mode gives a point estimate of $\theta$ and the BLUP estimates $\hat{\beta}$ and $\hat{\mu}$ given that $\theta$. These are not ideal estimators. Bayesian statisticians would prefer to estimate $\theta$ by integrating over $\beta$ and $\mu$ rather than merely looking at the posterior for $\hat{\beta}$ and $\hat{\mu}$.

Harville (1974) showed that REML is equivalent to marginalizing the likelihood over the fixed effect parameters, so practical approximate Bayesian procedures for estimating dispersion parameters can use REML to approximately integrate over the fixed and random effects.

The Classical concept of modified profile likelihood due to Barndorff-Nielsen (1983) can also be regarded as an approximate Bayesian technique in which second derivatives of the posterior density are used to approximately integrate out nuisance parameters by assuming normal distributions with the given second derivatives for the nuisance parameters. Such approximate integration gives a multiplicative factor that is the exponential of the $-\frac{1}{2}$ power of the determinant of the observed information matrix for the nuisance parameters.

5.5 Estimation of Outliers

When fitting models with only one variance component, it is common practice to compute residuals in order to look for outliers. If it appears that some data points are outliers, then they may be ignored in some circumstances, they may be highlighted in other circumstances, or appropriately robust methods of analysis may be used for estimating parameters of interest.

When fitting models with two or more variance components, BLUP estimates of the realized values of random effects are the natural generalization of the concept of residuals. Outlying values of some random effects will mean that groups of data points are likely to fail to fit the model, but looking at the estimates of the random effects provides a more sensitive test for such outliers than merely looking at residuals. Fellner (1986) discusses this topic with some examples.

5.6 Estimation of Fixed and Random Effects when the Dispersion Parameters Must be Estimated

For most of this paper, estimation of random effects is considered under the assumption that $\theta$ is known. All approaches seem to agree as to the best estimates of random effects in this case. There is less consensus about what to do when $\theta$ must be estimated from the data.

Student's $t$-test differs from simple use of the normal distribution in that it takes the uncertainty of estimating the variance of a normal distribution into account when considering hypotheses about the mean. It is natural to suspect that BLUP estimates or their estimated precision would need to be modified when $\theta$ must be estimated.

Kackar and Harville (1981) showed that estimates $\hat{\beta}$ and $\hat{\mu}$ remain unbiased when $\theta$ must be estimated provided that the estimates of components of $\theta$ are translation-unbiased and are even functions of the data vector. This suggests that the BLUP point estimates will generally not need to be modified. We do in principle, however, need to modify the estimated precision of the BLUP esti-
mates. In practice, this difficulty is sometimes ignored or handled by conservative interpretation of calculations based on the best point estimate of the dispersion parameters.


From a Bayesian point of view, the estimation of dispersion parameters taking the uncertainty in the fixed and random effects into account and the estimation of fixed and random effects taking uncertainty in the dispersion parameters into account are not different in principle, but different types of approximation are likely to be appropriate. A practical approximate Bayesian procedure for estimating the precision of fixed and random effects is to approximate the Bayesian posterior distribution for $\theta$ by a discrete distribution over a small number of possibilities and to use BLUP at each of these possibilities thereby calculating a mixture of distributions as posterior for the fixed and random effects.

5.7 Empirical Bayes Methods

Empirical Bayes methods are concerned first with estimating distributions from which random effects have been generated. Once a distribution of random effects has been estimated, this distribution is used to estimate realized values of random effects using Bayes’ Theorem. If the distribution of random effects is Gaussian, then empirical Bayes methods would use BLUP estimates to estimate fixed and random effects.

When parametric assumptions are made about the distribution of the random effects then the statistical methods employed are described as parametric empirical Bayes. BLUP is equivalent to one of the techniques of parametric empirical Bayes methodology.

5.8 Ranking and Selection

BLUP was originally developed for ranking and selection in the contexts of animal breeding and genetics. It is an appropriate technique when the ideal ranking or selection criteria involve unobservable characteristics that may be regarded as random effects.

Much theoretical work on ranking and selection seems to have been done in ignorance of BLUP. This work tries to control the probability that ranking or selection decisions will be made correctly, when the probability is to be calculated before the data are available. In contrast, BLUP is concerned with correct selection between random effects given the data, as in Berger and Deely (1988).

Stein (1945) presented a procedure that obtains an interval estimate of preassigned length for the mean of a normal distribution when the population variance is unknown. It achieves this by sampling in two stages and ignoring the information about the population variance contained in the sample variance for the second sample. It is not a satisfactory technique for practical application because it concentrates on initial precision rather than final precision, it does not achieve its nominal coverage probability conditional on the ancillary statistic $s_i^2 / s_2^2$ where $s_i^2$ is the sample variance in the $i$th stage of sampling, and it violates the likelihood principle. Stein’s procedure is a foundation stone for some theoretical work on ranking and selection that I believe to be misguided. BLUP would be a better foundation stone.

6. APPLICATIONS

There are a number of situations in which estimation of random effects is precisely what is required. Morris (1983) has discussed several of them. A brief discussion of some of them should help to establish the point that estimation of random effects is a legitimate activity, even if not very common.

There are four features which are common to most of the applications.

1. The data and the random effects to be estimated are often multivariate.
2. Computational issues are often extremely important.
3. Sparsity or approximate sparsity of partial correlation matrices is more important than sparsity or approximate sparsity of correlation matrices.
4. Random effects selected on the basis of the estimates made are often of particular interest.

People having difficulty with one of these features for any of the applications should consider how the feature is handled for the other applications.

6.1 Estimation of Merit of Individuals

Efron and Morris (1977) discussed estimating the batting abilities of 18 baseball players. The true batting abilities can be regarded as random effects drawn from a distribution. To estimate the mean and variance of the distribution of batting abilities
is of some interest, but the main problem is estimating the random effects.

In plant variety trials, it is sometimes realistic to regard the varieties as random effects since they have been generated by processes that are random at the chromosome level. The objective of variety trials is generally to find the best varieties or to estimate the yield (or some other characteristic) of all varieties, not to estimate the parameters of the distribution from which the varieties are a sample.

### 6.2 Selection Index

In quantitative genetics, the selection index provides a way of ranking plants or animals given measurements on several traits on the individuals to be ranked and their relatives.

The selection index can be seen to be a particular case of the BLUP estimates of random effects

\[ \hat{\mu} = (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} y \]

\[ = GZ^T (R + ZGZ^T)^{-1} y \]

using (5.2). The model being fitted is

\[ y = Zu + e, \]

where \( u \) is a vector of additive genetic merits of animals with variance-covariance matrix \( G \) and \( e \) is a vector of other influences including nonadditive genetic merits, permanent environmental effects and measurement errors. In this context \( u \) is often referred to as “genotype,” although it only gives the so-called “additive” part of genetic merit, which is the average merit of all potential offspring for a population of potential mates. The matrix \( G \) is called the genetic variance-covariance matrix. The measurable quantities, \( y \), are referred to as “phenotype,” and \( \text{Var}(y) = R + ZGZ^T \) is called the phenotypic variance-covariance matrix.

Consider a situation where a trait is measured with equal precision on an animal and on one of its parents. Suppose that the variance of genotype is \( \sigma^2 \) and the variance from other sources is \( \sigma^2(1 - h^2)/h^2 \) so that the heritability (ratio of genotypic variance to total variance) is \( h^2 \). Taking

\[ y = \begin{bmatrix} \text{measurement on animal} \\ \text{measurement on parent} \end{bmatrix}, \]

\[ Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \]

\[ G = \sigma^2 \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}, \]

the genotypes of the two animals having correlation \( \frac{1}{2} \), and

\[ R = \frac{1 - h^2}{h^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

gives

\[ \hat{\mu} = \frac{h^2}{4 - h^4} \begin{bmatrix} 4 - h^2 & 2 - 2h^2 \\ 2 - 2h^2 & 4 - h^2 \end{bmatrix} \]

in agreement with a formula from Turner and Young (1969, page 148). This tells us how to weight the two pieces of data in order to best estimate the genetic merit of the animal. It provides the best index combining the two pieces of information on which to base selection decisions.

It is often easier to work with the BLUP form of the selection index because the matrix \( G^{-1} \) is generally sparser than \( G \), as shown by Henderson (1976). Essentially, this sparsity arises because \( G^{-1} \) is a matrix of partial correlations, and the partial correlations of genetic merits of pairs of animals that are not mates or parent and offspring given the genetic merits of all their mates, parents and offspring is zero, as argued by Robinson (1986).

### 6.3 Estimation of Quality

In his many seminars on quality and management, Deming often refers to the distinction between “analytic” and “enumerative” studies as explained in Deming (1950, Chapter 7). One way of explaining this distinction is that analytic studies are concerned with estimating fixed effects while enumerative studies are concerned with estimating random effects.

Average quality over a particular time period is a random effect. The quality measurement plan (QMP) of Hoadley (1986) regards the true quality index for a period of production as having come from a probability distribution. The QMP estimates the true quality index. It is an example of BLUP where distributions other than normal are assumed.

An example of quality estimation that illustrates the distinction between estimation of fixed effects and estimation of random effects is sampling and testing of a shipload of coal or iron ore. The average grade of a large number of samples taken from a conveyor belt as the ship is loaded provides a fairly precise estimate of the average grade of the shipload. However, it may not provide a precise estimate of the average grade of all shiploads, particularly if there is substantial long-term variation in grade.

If the grade of ore as a function of cumulative tonnage, \( X(t) \), is regarded as a stochastic process,
then the average grade of the shipload is a random effect. It is \( \int_A^B X(t) \, dt / (B - A) \) for some constants \( A \) and \( B \) that delimit the shipload. The average grade of all shiploads is a fixed effect. It is the mean of a stochastic process.

Deming would regard the sampling and testing as being an "analytic" study insofar as it is studying the mean of all shiploads and as being an "enumerative" study in so far as it is studying a single shipload.


It should be noted that many sampling standards are designed to estimate the population mean of a process (an analytic study—a fixed effect), not the mean of a lot (an enumerative study—a random effect), but they fail to make the distinction clear. Often, the mean of a lot can be estimated much more precisely than the process mean with given data.

6.4 Time Series and Kalman Filtering

Some time series problems are concerned with estimating fixed parameters associated with time series. Many other time series problems may usefully be regarded as problems of estimating random effects.

The observed value of a time series is the sum of signal and noise that may be regarded as random effects that differ in their spectra. Smoothing, filtering and prediction problems all involve estimation of the random effects that form the signal.

To illustrate the link between estimation of random effects and time series, we consider the Kalman filter, which is used for estimating the current value of the signal in a time series. The estimates obtained by Kalman filtering and by BLUP must be the same because of their optimality properties. However, it seems worthwhile to consider an approach to Kalman filtering directly from BLUP. See also Broemeling and Diaz (1985), Harrison and Stevens (1976) and Sallas and Harville (1981). Incidentally, Sorenson (1970) suggests that the Kalman filter is not entirely due to Kalman.

We follow the terminology of Meinhold and Singpurwala (1983) who explained the Kalman filter in a Bayesian framework.

Suppose that unobservable vector-valued random variables \( u_t \) are related by

\[
    u_t = G_t u_{t-1} + w_t
\]

for \( t = 1, 2, \ldots, n \) with \( u_0 = 0 \). An observable vector-valued random variable \( y_t \) is related to the \( u_t \) by

\[
    y_t = F_t u_t + v_t.
\]

Assume that \( G_t \) and \( F_t \) are known. Further, suppose that the \( w_t \) and \( v_t \) are independent and normally distributed with zero means and known variances \( W_t \) and \( V_t \).

This is an example of a random effects model. The variance-covariance matrix for the random effects is not simple. Denoting it by

\[
    G = (g_{ij})
\]

where

\[
    g_{ij} = \text{Cov}(u_i, u_j) = E[u_i u_j^T],
\]

it is defined recursively by the following equations:

\[
    g_{11} = W_1
\]

\[
    g_{ij} = G_{i} g_{i-1, j} \quad \text{for } i > j
\]

\[
    g_{ij} = g_{i+1, j}^T G_{j}^T \quad \text{for } i < j
\]

\[
    g_{ii} = G_{i} g_{i-1, i} + W_i = g_{i-1, i}^T G_{i}^T + W_i.
\]

The matrix \( G^{-1} \) is simpler than \( G \). It is block tridiagonal essentially because the partial correlations of pairs of nonadjacent \( u_t \), given an intermediate \( u_t \), are always zero. It is

\[
    G^{-1} = (g^{ij})
\]

where

\[
    g^{nn} = W_{n+1}^{-1},
\]

for \( i < n, \)

\[
    g^{ii} = W_{i}^{-1} + G_{i+1}^T W_{i+1}^{-1} G_{i+1}
\]

\[
    g^{i+1i} = -G_{i+1}^T W_{i+1}^{-1}
\]

and all other \( g^{ij} \) are zero matrices.

The matrix algebra required to show equivalence of BLUP estimates with the Kalman-Bucy algorithm is not trivial, but the statistical theory is simple. The Kalman-Bucy algorithm is computationally efficient because computations for time \( t \) make use of results obtained for time \( t - 1 \). This use of the immediate past depends crucially on the simplicity of the matrix \( G^{-1} \), which is due to the Markovian nature of the process generating the \( u_t \).

A general principle illustrated here is that partial correlation matrices are more important computationally than correlation matrices.

Note that, being equivalent to BLUP estimation
of random effects, Kalman filtering is not maximum likelihood.

6.5 Removing Noise from Images

Besag (1986) reviewed many methods of attempting to restore images that have been corrupted by noise. The simplest models for continuously variable intensity over a grid are auto-normal models in which the point intensities are normally distributed and have nonzero partial correlations with intensities at neighboring points on the grid. Within our framework, the true image is a random effect that we wish to estimate. What Besag calls the “maximum a posteriori” estimate of the true image is equivalent to BLUP for these models.

As in other applications being discussed, the objective of primary interest is to estimate random effects.

6.6 Geostatistics

The method of geological reserve estimation known as Kriging is essentially the same as BLUP. It was developed independently from BLUP. A difference from the estimation of genetic merits is that variance parameters are estimated more frequently than in genetic applications.

The random effect that is estimated is the pattern of grade of ore as a function of position in two- or three-dimensional space. Estimation of the random effect is of much more immediate interest to a company that is considering a mining operation than is estimation of the parameters of the distribution from which the pattern of grade is a sample.

There is much more to ore reserve estimation than Kriging. A concise summary of some of the difficulties written by Matheron appears in the Foreword to Journel and Huijbregts (1978). One very substantial difficulty is that the economic feasibility of mining a deposit is based on a small number of drill-holes, but decisions about which blocks of ore will be processed and which will be dumped as waste will be made using data from blast-holes that are much more numerous than the drill-holes. The average grade of the ore that will be selected for processing is much harder to predict than is the overall average grade. This difficulty is most acute when only a small proportion of the rock to be mined will be processed.

6.7 Credibility Theory

Credibility theory is a collection of ideas used by actuaries to work out insurance premiums. As an example, consider setting workers’ compensation premiums for industrial companies that have different numbers of employees and different safety records. This example was discussed by Mowbray (1914).

In terminology like that of the present paper, let \( u_i \) denote the true fair premium for company number \( i \). The \( u_i \) may be regarded as random effects and the BLUP estimates of them will be useful. If a company has an extensive claims record, then its estimated premium will depend almost entirely on its own record. If a company has little claims record, then the estimated premium for that company will differ little from the average premium for all companies.

The classical approach to credibility is to assume that there is a credibility factor, denoted \( Z(t) \), where \( t \) is the size of the risk class, such that the estimated fair premium is

\[
Z(t) y + \{1 - Z(t)\} m',
\]

where \( y \) is the fair premium as estimated using data for the single risk class only, and \( m' \) is the average fair premium over all risk classes. The function \( Z(t) \) is chosen in a somewhat arbitrary manner which need not concern us here. (See Hickman, 1975.)

The Bayesian approach to credibility is to use conjugate prior distributions for the \( u_i \). The development of the Bayesian approach is attributed to Bailey (1945, 1950) and Mayerson (1965). See Kahn (1982) for a brief review.

6.8 Small-Area Estimation

Small-area estimation involves using direct survey information from areas of individual interest together with information on similar or related areas. It has been found that more precise estimates can be made than if information on the other areas had been ignored. One approach is to regard the quantities of interest in the individual areas as being random effects which are to be estimated. See Battese, Harter and Fuller (1988), Prasad and Rao (1986), Fuller and Harter (1987) and the other papers in the same volume from the May 1985 International Symposium on Small Area Statistics for further information about this field of application.

7. DISCUSSION

7.1 Prediction or Estimation?

Henderson originally used the term “predictor” rather than “estimator” in order to evade criticism of BLUP. Henderson (1984, page 37) expressed
doubts about the appropriateness of the terminology:

Which is the more logical concept, prediction of a random variable or estimation of the realized value of a random variable? If we have an animal already born, it seems reasonable to describe the evaluation of its breeding value as an estimation problem. On the other hand, if we are interested in evaluating the potential breeding value of a mating between two potential parents, this would be a problem in prediction. If we are interested in future records, the problem is clearly one of prediction.

From conversations with him, I believe that he accepts the weaknesses of the terminology:

- in some applications, the thing being estimated has already occurred, and
- BLUP is a predictor only in the same way as most estimates are predictors—if they were not relevant to something which might happen in the future then they would not be of interest.

It has become common practice to “estimate” fixed effects and to “predict” random effects. I believe that Henderson would be content to use terms such as “estimate” and “estimate of realized value.” I prefer simple terminology—to use “estimation” of both fixed and random effects.

7.2 Unbiasedness and Shrinkage

The BLUP estimator $\hat{u}$ of $u$ is sometimes said to be “unbiased” because it satisfies

$$E[\hat{u}] = E[u].$$

It is also described as a “shrinkage” estimator because its components are less spread than the generalized least-squares estimates that would be used if the components of $u$ were regarded as fixed effects. These two descriptions seem to be in conflict with one another.

The difficulty is that the use of the term “unbiased” is very different from the condition

$$E[\hat{u} | u] = u \quad \text{for all } u,$$

which is what many people intuitively expect, particularly in circumstances where the random processes generating the random effects (genetic merits of dairy bulls or ore grades in a deposit, for instance) occur prior to other processes generating noise in the data. Whenever I mean “unbiased” in the sense that $E[\hat{u}] = E[u]$ I try to make this very clear. Perhaps some new terminology would be a good idea.

7.3 Fixed and Random Effects

Eisenhart (1947) distinguished between two uses for analysis of variance: (1) detection and estimation of fixed (constant) relations among the means of subsets of the universe of objects concerned; and (2) detection and estimation of components of (random) variation associated with a composite population. He suggested two parallel sets of questions to help distinguish fixed and random effects, and these tend to imply that estimation of random effects is not sensible because effects of interest must be treated as fixed.

1. “Are the conclusions to be confined to the things actually studied (the animals, or the plots); to the immediate sources of these things (the herds, or the fields); or expanded to apply to more general populations (the species, or the farmland of the state)?”

2. “In complete repetitions of the experiment would the same things be studied again (the same animals, or the same plots); would new samples be drawn from the identical sources (new samples of animal from the same herds, or new experimental arrangements on the same fields); or would new samples be drawn from more general populations (new samples of animals from new herds, or new experimental arrangements on new fields)?”

The discussion of the difference between fixed and random effects given in Searle (1971) is similarly misleading in my view. On page 383, Searle writes

\[
\ldots \text{when inferences are going to be confined to the effects in the model the effects are considered fixed; and when inferences will be made about a population of effects from which those in the data are considered to be a random sample then the effects are considered as random.}
\]

Both of these people have defined their terms in such a way that estimation of random effects is not possible.

In my view, there are two questions which might need to be answered in order to decide whether the effects in a class are to be treated as fixed or random. The first of these is: Do these effects come from a probability distribution? If the effects do not come from a probability distribution then the effects should be treated as fixed.

If the effects do come from a probability distribution, then if the effects are themselves being estimated they should be treated as random. For classes
of effects which are nuisance parameters, there is a
second question to be answered: Is interclass infor-
mation to be recovered for this class of effects? If
interclass information is to be recovered then treat
the class of effects as random, otherwise fixed.

Example. When estimating the genetic merit of
dairy bulls, the herd-year-season effects that are
used to model environmental variation do come
from a probability distribution which describes
variation: that between herds, years and calving
seasons. When estimating herd-year-season effects
I believe that they should be treated as random.
However, when estimating genetic merits of dairy
bulls I believe that herd-year-season effects should
be treated as fixed because the inter-herd informa-
tion should not be recovered—Farmers’ choices of
artificial insemination bulls and their choice of
feeding regimes are somewhat related, and to re-
cover the inter-herd information would favor the
bulls that were more often chosen by farmers pro-
viding high feed intakes. Similarly, bull choices
vary between years, and yearly climatic variation
might bias estimated breeding values. The decision
not to recover inter-herd-year-season information is
based on a willingness to sacrifice some efficiency
for greater robustness relative to departures from
the assumptions of the model.

I agree with Tukey’s remark made in discussion
of Nelder (1977): “... our focus must be on ques-
tions, not models.” The choice of whether a class of
effects is to be treated as fixed or random may vary
with the question which we are trying to answer.

7.4 Generalizing Likelihood

When Fisher (1922, page 326) introduced the
idea of likelihood as the probability of data given a
hypothesis regarded as a function of the hypothesis,
I doubt that he had considered problems of estimat-
ing random effects.

For estimation of fixed effects and dispersion pa-
rameters for the random effects model (1.1) the
usual definition of likelihood seems adequate. How-
ever, if we wish to estimate the random effects, \( u \),
then the usual concept of likelihood seems to oblige
us to regard the value of \( u \) as part of the hypothe-
sis. In effect, the idea of likelihood tends to force us
to regard things that we wish to estimate as fixed
rather than random effects.

In my view, when we specify a mixed model the
dispersion of the random effects, \( u \), and the disper-
sion of the error, \( e \), ought to be given similar logical
status. How can this be achieved?

One possible attempt to resolve the problem based
on Bayesian concepts would be to regard the as-
sumed distribution of the random effects as an
objective prior distribution. It is different in logical
status from a subjective prior distribution from
which the fixed effects might be considered to have
come. One distinguishing feature is that the objec-
tive distribution of the random effects is being used
to describe variation while the subjective prior dis-
tribution of the fixed effects is being used to de-
scribe uncertainty. A second distinguishing feature
is that assumptions about the distribution of the
random effects can be tested.

To use such an objective prior distribution for
random effects should not be considered to make
one a Bayesian. Good (1965, page 8) wrote

the essential defining property of a Bayesian is
that he regards it as meaningful to talk about
the probability \( P(H \mid E) \) of hypothesis \( H \), given
evidence \( E \).

To be a Bayesian, you would need to be willing to
put a prior distribution on fixed effects as well as
random effects. Kempthorne in the discussion of
Lindley and Smith (1972, page 37) described people
who wish to estimate random effects as “legiti-
mate” Bayesians; I agree with the ideas behind
this designation, but I prefer to adhere to Good’s
use of the term “Bayesian.”

This attempt to define likelihood in a way which
allows estimation of random effects separates the
distribution of \( e \) (the likelihood), the distribution of
\( u \) (the objective prior distribution), and the uncer-
tainty about \( \beta \) and \( \theta \) (the subjective prior distribu-
tion for Bayesians) into three separate boxes. It
does not achieve the stated goal of giving similar
logical status to the dispersion of \( e \) and the disper-
sion of \( u \).

Most likely unobservables. An alternative res-
olution of the problem based on Classical concepts
is to formalize the principle behind Henderson’s
derivation of BLUP. This principle does achieve the
good of giving similar logical status to the disper-
sion of \( e \) and the dispersion of \( u \). A colleague, T.
Lwin, and I would like to suggest the name method of
most likely unobservables for it.

Given the mixed model (1.1) and having observed
the data \( y \), the method of most likely unobserv-
ables is to say that the best estimates of \( \beta \), \( u \) and \( e \)
are the ones that maximize the density of the unob-
servables \( u \) and \( e \) subject to the constraint (1.1) of
having observed the data. Because there are \( n \)
constraints and \( n \) components to the vector \( e \), the
easiest way to solve for the maximum is to set

\[
e = y - X\beta - Zu
\]

and maximize the joint density of \( u \) and \( e \) with
respect to $\beta$ and $u$. This is precisely Henderson’s derivation of BLUP discussed in Section 4.1.

For a fixed effects model, the likelihood of the data and the likelihood of the errors are equal, so the method of most likely unobservables gives estimates of the fixed effects that are the maximum likelihood estimates and gives estimates of the errors, $e$, that are the usual residuals.

7.5 On Schools of Thought

I believe that the distinction between fixed and random effects should be clarified before differences between schools of thought are considered.

Statistics is concerned with both variation and uncertainty. Classical statistics can be distinguished from Bayesian statistics by its refusal to use probability distributions to describe uncertainty. However, it is quite willing to use probability distributions to describe variation, and this should include variation between random effects. Once this is clarified, several situations in which the two schools of thought appeared to give different answers instead demonstrate close agreement between the schools.

There are a number of reasons why the estimation of random effects has been neglected to some extent by the Classical (Neyman–Pearson–Wald, Berkeley) school of thought.

1. The distinction between fixed and random effects has been often taken to be that effects are random when you are not interested in their individual values. Searle (1971), for instance, seems to suggest this. This implies that you should never be interested in estimating random effects.

2. The idea of estimating random effects seems suspiciously Bayesian to some Classical statisticians. The gulf between Bayesian and Classical statisticians seems to be like many other gulfs between schools of thought in that the adherents of each school emphasize the differences between schools rather than the similarities.

Within the Bayesian paradigm, there is little reason for distinguishing between fixed and random effects. All effects are treated as random in the sense that probability distributions used to describe uncertainty are not treated any differently from probability distributions used to describe variation.

ACKNOWLEDGMENT

I am thankful for the helpful criticism provided by the editors, C. N. Morris and the late M. H. DeGroot, two referees, colleagues too numerous to list individually—colleagues from the Herd Improvement Section of the Victorian Department of Agriculture, the Animal Genetics and Breeding Unit of the University of New England, and the Division of Mathematics and Statistics of the Commonwealth Scientific and Industrial Research Organization—and many other people with whom I have discussed these issues.

REFERENCES


THE ESTIMATION OF RANDOM EFFECTS


---

### Comment

Katherine Campbell

It has been a pleasure to read about the long history of Best Linear Unbiased Prediction, and especially about its uses in traditional statistical areas of application such as agriculture. My own experience with BLUP is in the context of ill-posed inverse problems, and I would like to discuss this paper from this point of view, where the random effects are generated by hypothesized superpopulations, in contrast with the identifiable populations considered by Robinson.

**MODEL-BASED ESTIMATION FOR ILL-POSED INVERSE PROBLEMS**

The author mentions two examples of superpopulation approaches to estimation: image restoration and geostatistics. The same ideas are also used in model-based estimation for finite populations, function approximation and many other inference problems. These problems concern inference about a reality that is in principle completely determined, but whose observation is limited by the number and/or resolution of the feasible measurements, as well as by noise. In geophysics, x-ray imaging and many other areas of science and engineering these are known as inverse problems (O’Sullivan, 1986; Tarantola, 1987).

The unknown reality we may consider to be a function \( m \) defined on some domain \( T \). The data typically consist of noisy observations on a finite number \( n \) of functionals of \( m \). We can write the data vector \( y \) in terms of a transformation \( L \) mapping \( m \) into an \( n \)-dimensional vector:

\[
y = Lm + e.
\]

In the sequel, we will assume that \( L \) is a linear transformation, i.e., that the observed functionals are linear. In particular, if the cardinality \( |T| \) of \( T \) is finite, \( L \) can be represented by an \( n \times |T| \) matrix.

BLUP arises when we embed this problem in a superpopulation model, under which \( m \) is one realization (albeit the only one of interest) of a stochastic process \( M \) indexed by \( T \). This superpopulation model has two components, corresponding to the “fixed” and “random” effects in Robinson’s discussion. The fixed effects define the mean of the superpopulation, which is here assumed to lie in a finite-dimensional subspace of functions on \( T \). We denote this subspace by \( R(F) \), the range of the linear operator \( F \) that maps a \( p \)-vector \( b \) into the function

\[
Fb = \sum b_i f_i,
\]

where \( \{f_1, \ldots, f_p\} \) is a basis for the subspace.

Any realization of \( M \) can then be written as a sum \( F\beta + u \), where \( \beta \) is an unknown vector of \( p \) fixed effects and \( u \) is a realization of a stochastic “random effects” process with mean zero and covariance \( P \). As we are interested in the realized \( m \), we need to estimate both the fixed and random effects. Among estimates that are linear functions of the data vector

\[
y = LF\beta + Lu + e,
\]

the BLUP \( \hat{m} = F\hat{\beta} + \hat{u} \) is the optimal choice: under the assumed superpopulation model \( \hat{m} \) is unbiased in the sense of Section 7.2 (i.e., \( E\hat{m} = Em \)) and it minimizes the variance of any linear functional of \( \hat{m} - m \). (To make the correspondence between equation 1 and Robinson’s equation (1.1) explicit, \( X = LF \), \( Z = L \), \( G = LPL^T \), and \( e \) is a realization of a random \( n \)-vector with mean zero and covariance \( R \). \( \hat{\beta} \) and \( \hat{u} \) are then provided by the BLUP formulas.)

---

*Katherine Campbell is a staff member of the Statistics Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.*
The following examples make the preceding discussion more concrete.

1. The term “superpopulation model” is associated with model-based inference in finite populations. A general discussion and many references are found in Cassell, Sarndal and Wretman (1977). T is here the finite set of population unit labels. Fβ is a linear model for the mean of the outcome variable m as a function of p auxiliary variables f, that are known in advance for each population unit. L is the n × |T| sampling matrix whose ith row contains a one in the tth column if the ith sampled unit is the tth population unit, and zeros elsewhere. Commonly P is modeled as a function of the auxiliary variables f with some parameters that are estimated from the data. Very often it is assumed that the outcome variable m, can be measured without error once the tth unit is sampled, and so R = 0.

2. In image analysis, T indexes the pixels. Frequently the only fixed effect is a constant, i.e., p = 1 and f₁ = 1. L is generally convolution with a known or assumed point-spread function, followed by exhaustive sampling. Although there may be |T| observations, the effective dimension of R(L) is much less than |T| because of the limited resolution of the point-spread function, and the problem is ill-conditioned. Generally, in addition, R is not zero. With large numbers of observations at their disposal, image analysts have been quite adventurous in their modeling of the random effects. A popular choice for the probabilistic structure of u in the recent literature is a Markov random field (see Marroquin, Mitter and Poggio, 1987).

3. The parallels between kriging and BLUP have been described by Cressie (1990). In geostatistical applications, T indexes a subset of two- or three-dimensional Euclidean space, and the functions fᵢ are low-order monomials in the spatial coordinates. L is again a sampling operator. R includes analytical error. Sampling error, resulting from imperfect surveying, small sample volumes and/or local heterogeneity, can be assigned to R or to a discontinuous component of P called the “nugget effect.” The continuous part of P is estimated as a function of the spatial coordinates.

4. The use of superpopulation models and BLUP for approximating the output of large computer codes as a function of many input parameters is discussed by Sacks, Welch, Mitchell and Wynn (1989). Their use in function approximation is illustrated by Blight and Ott (1975) and O’Hagan (1978). In these applications, T indexes a collection of r-vectors xᵢ = (xᵢ₁, ⋯, xᵢᵣ). The components of xᵢ are r input parameters for a computer code, or the r independent variables in the case of function approximation. Again the functions fᵢ are frequently monomials in these r variables, and L is a sampling operator. For computer codes R is zero, while curve-fitting applications may or may not include measurement error. The almost universal choice for P is a product of the form

\[ P_{x,t} = \prod_{i=1}^{r} \exp(-\rho_i |x_{i,t_i} - x_{t_i}|^\theta_i) \]

The choice of parameter values reflects assumptions about the continuity and differentiability of realizations of M but is seldom the result of estimation based on the data. Spline interpolation implies another choice, namely a generalized covariance (see below) of the form

\[ K_{x,t} = \rho |x_s - x_t|^2 \log(\theta |x_s - x_t|) \]

(see Dubrule, 1983).

**COVARIANCE MODELS**

The greatest difficulty in practical application of BLUP is the specification of G. (Specification of R is seldom problematic.) The brief discussion of Section 5.4 suggests some of the difficulties encountered; estimation of covariances is notoriously a more difficult problem than estimation of means. The covariance matrices G = LPLᵀ required by superpopulation models are nontrivial. Many areas, notably geostatistics, have developed their own methods, some rather ad hoc, to estimate this parameter.

A geostatistical innovation in this area arises from the observation that M need not possess a covariance for the BLUP to exist. Recall that in a vector-space approach to multivariate analysis, Cov M = P means that

\[ \text{Cov}\{\langle\lambda, M\rangle, \langle\nu, M\rangle\} = \langle\lambda, P\nu\rangle \]

for all vectors λ and ν. (Eaton, 1983, treats the case where |T| is finite, for which functions defined on T are just vectors in \( R^{|T|} \), which can be supplied with an inner product \( \langle\cdot, \cdot\rangle \) in the usual manner. Equation (2) in an appropriate Hilbert space of functions serves as a rigorous definition of Cov M when |T| is infinite.) In geostatistics, the BLUP is usually computed using a “generalized” covariance \( K \), for which it is sufficient that

\[ \text{Cov}\{\langle\lambda, M\rangle, \langle\nu, M\rangle\} = \langle\lambda, K\nu\rangle \]

for all λ and ν orthogonal to \( R(F) \). This enlarges somewhat the set of models available for this component of the superpopulation model, and moreover \( K \), unlike \( P \), can be estimated without correcting for the unknown fixed effects β. A popular choice
for $K$, when the fixed effects include an unknown constant, is based on the “semivariogram” function

$$\gamma(s, t) = \frac{1}{2} E((M_s - M_t)^2).$$

In practice, the BLUP is fairly robust to most aspects of the choice of $G$. However, a critical parameter is the relative size of $G$ and $R$. If $G$ is of the form $\alpha^2 \Gamma$, the resulting BLUP is moderately insensitive to the choice of $\Gamma$ but very sensitive to the parameter $\alpha$, which controls the degree of smoothing or “shrinkage” in $\hat{u}$ (Section 7.2). Cross-validation is the most common data-based method for estimating this parameter; see O’Sullivan (1986) and Woodbury (1989).

**SUPERPOPULATIONS AND RANDOM EFFECTS**

The reader may feel that the introduction of superpopulation models takes us quite far from the spirit of Robinson’s paper, wherein pains have been taken to use only the classical interpretation of probability as a description of ontological variability. In the paper, the superpopulation generating the random effects is a real population (e.g., the population of potential sires), while in the context of ill-posed inverse problems it appears that a superpopulation is introduced merely as a mechanism for imposing additional constraints on the problem so that a unique solution can be defined. In this connection several observations are in order.

First of all, whether or not the superpopulation is real does not appear to be a central philosophical problem in the acceptance of random effect estimation by the classical school. Although another realization of a real superpopulation is feasible (we could repeat the experiment with another set of sires), BLUP estimates only that realization (the set of four sires) that was actually represented in the given data. In the case of inverse problems, nature, rather than an animal breeder, provides the realization that was observed. The sticking point with respect to BLUP seems to be that, despite the fact that the realization is now fixed, we continue to model its effects as random, with different numerical results, as Robinson’s introductory example shows, than if we were to treat them as fixed.

What the superpopulation (real or imaginary) point of view makes clear is that the difference between a fixed effect and a random effect is that one belongs to the mean of the statistical model, while the other is a deviation from the mean, i.e., is described by the variance component of the statistical model. Classical statistics has no problem with this distinction in ordinary regression models, where (as in Section 4.3) such deviations are called “residuals” and may be individually estimated for diagnostic purposes, among others. No one would propose that therefore residuals should be treated as fixed effects! Similarly, BLUP preserves the distinction throughout the analysis, even after the realization of the real or implied superpopulation has been fixed and the data collected.

As used in the solution of ill-posed problems, superpopulation models often attempt, perhaps unhelpfully, to blur the distinction between the epistemological (Bayesian) and ontological (Classical) interpretations of probability. (In spirit, and also in form, superpopulation models thus come close to empirical Bayes ideas, although formal empirical Bayes techniques remain largely unexploited.) In particular, the empirical approach to the estimation of $P$ adopted in many superpopulation applications implies that this parameter reflects ontological variability, not subjective uncertainty. Empirical superpopulations for geostatistics are the subject of “deterministic geostatistics” (cf. Isaaks and Srivastava, 1988).

**INVERSE PROBLEMS AND STATISTICS**

Although ill-posed inverse problems are certainly inference problems, this area has been neglected by Classical statisticians, apparently because probability in this context is generally thought to describe uncertainty rather than variability. Nevertheless, probabilistic regularization methods are widely used, have excellent track records and can often be given empirical interpretations. Statisticians ignore these developments at the risk of being found irrelevant by many of their colleagues in the physical sciences, where inverse problems are ubiquitous.

In particular, linear regularization methods for ill-posed inverse problems can be interpreted as BLUP under appropriate superpopulation models. I therefore welcome Robinson’s article, not only for its wide-ranging survey of BLUP history and applications, but also for its examination of the philosophical questions raised by the estimation of random effects. Better understanding of these philosophical problems may induce statisticians to reevaluate this important class of problems as an appropriate subject for statistical research.
THE ESTIMATION OF RANDOM EFFECTS

Comment

David A. Harville

The publication of Robinson's article is very timely. As related by Robinson, BLUP is a statistical methodology that has been used extensively in animal breeding, with great success. Robinson has brought to our attention the similarities between BLUP and various other methodologies, like kriging and the Kalman filter, and has noted that, while BLUP was developed via a frequentist approach to statistics, it has a Bayesian interpretation. In doing so, he has performed a valuable service to those interested in the development and application of BLUP (and the related methodologies) and to the statistics community as a whole.

The main theme of my discussion, which is developed in Sections 1, 4, and 5, is that BLUP and the related methodologies should be discussed in the common framework of a general prediction problem, that BLUP has some deficiencies that could be eliminated by a more extensive use of Bayesian ideas and that a unified, yet flexible, approach to prediction is desirable and achievable. My discussion includes (in Sections 2 and 3) some comparisons between work on BLUP and related work on empirical Bayes inference and some comments about some long-standing misconceptions regarding the use of mixed-effects linear models.

1. BLUP FOR A GENERAL PREDICTION PROBLEM

It is instructive to consider BLUP in the context of the general problem of predicting the value of an unobservable random variable $w$ based on the value of an $n \times 1$ observable random vector $y$, where the joint distribution of $w$ and $y$ has first and second moments to be denoted by $\mu_w = E(w), \mu_y = E(y), v_w = \text{var}(w), v_{yw} = \text{cov}(y, w),$ and $V_y = \text{var}(y)$. It is assumed that $\mu_y$ belongs to a known vector space $\mathcal{S}$ and that $\mu_w$ is a known linear combination of the elements of $\mu_y$, or equivalently, that $\mu_y = X\beta$ and $\mu_w = X\beta$, where $\beta$ is a $p \times 1$ vector of unknown parameters, $X$ is an $n \times p$ known matrix of rank $p^*$ (any matrix whose columns span $\mathcal{S}$), and $\lambda$ is a $p \times 1$ known vector that is expressible as $\lambda = X'k$ for some vector $k$. The quantity $v_{yw}$ and the elements of $v_{yw}$ and $V_y$ are assumed to be known functions of an unknown parameter vector $\theta = (\theta_1, \ldots, \theta_n)'$, whose value is restricted to a known set $\Omega$, and $V_y$ is assumed to be nonsingular (for all $\theta \in \Omega$). Note that, aside from the linearity of the mean structure, the primary limitation imposed by these assumptions is that $v_{uw}, v_{yw}$ and $V_y$ are unrelated to $\mu_w$ and $\mu_y$. Note also that, in the special case where $v_{uw} = 0, w = X\beta$ (with probability one), and the problem of predicting the value of $w$ is essentially equivalent to that of inference about $X\beta$.

Clearly, the problem of predicting a linear combination of the fixed and random effects in Robinson's model (1.1) can be formulated as a special case of the general prediction problem. Moreover, many of the problems considered by Robinson in his Section 6, which come from quality assurance, geostatistics, and various other fields, can likewise be formulated as special cases of the general prediction problem. For many of these problems, it may seem more natural to formulate them (directly) in terms of the general prediction problem than to follow Robinson's approach of recasting them in terms of fixed and random effects.

In introducing the BLUP—I use BLUP as an acronym for the best linear unbiased predictor as well as for best linear unbiased prediction—and in establishing its BLUPness, I prefer an approach that differs somewhat from any of those presented by Robinson. This approach, which I now describe in the context of the general prediction problem, clearly reveals the intuitive appeal of the BLUP and takes advantage of some well-known results on statistical estimation and on the MVN distribution.

Let $\tilde{w} = \mu_w + v_{yw}V_y^{-1}(y - \mu_y) = \tau + v_{yw}V_y^{-1}y$, where $\tau = \mu_w - v_{yw}V_y^{-1}\mu_y = (X'v_{yw}V_y^{-1}X)\beta$. If $\tau$ and $v_{yw}V_y^{-1}$ were known and if the joint distribution of $\tau$ and $y$ were assumed to be MVN, then the best (minimum MSE) predictor of the value of $w$ would be $E(w | y) = \tilde{w}$, as is well known. Since $\tilde{w}$ is linear in $y$, it is clear that even if the normality assumption were dropped, $\tilde{w}$ would be the best linear predictor of the value of $w$.

Now, let $\hat{\beta}$ represent any solution to the Aitken equations $X'V_y^{-1}X\hat{\beta} = X'V_y^{-1}y$. If $(X'v_{yw}V_y^{-1}X)(X'V_y^{-1}X) = X'V_y^{-1}X - v_{yw}V_y^{-1}$ were known, then the BLUE (best linear unbiased estimator) of $\tau$ would be $\hat{\tau} = (X'v_{yw}V_y^{-1}X)\hat{\beta}$, and a “natural”

David A. Harville is Professor, Department of Statistics, Iowa State University, Ames, Iowa 50011.
predictor of the value of \( w \) would be the predictor
\[
\bar{w} = \bar{\tau} + v'_{yw}V^{-1}y
\]
obtained from \( \bar{w} \) by substituting \( \bar{\tau} \) for \( \tau \). In fact, \( \bar{w} \) would be the BLUP.

To verify the BLUPness of \( \bar{w} \), observe that a predictor (of the value of \( w \)) is linear and unbiased if and only if it is expressible in the form \( r'Y + v'_{yw}V^{-1}y \), where \( r' \) is a (linear) unbiased estimator of \( \tau \), that is, where \( r'X = \lambda - v'_{yw}V^{-1}X \). Moreover, the MSE of a predictor of this form is
\[
E[(r'Y + v'_{yw}V^{-1}y - w)^2] = \text{var}(r'Y + v'_{yw}V^{-1}y - w) = \text{var}(r'Y) + \text{var}(v'_{yw}V^{-1}y - w),
\]
implying (since \( \bar{\tau} \) is the BLUE of \( \tau \)) that \( \bar{w} \) is the BLUP. A further implication is that the MSE of \( \bar{w} \) is
\[
v_{w} - v'_{yw}V^{-1}v_{yw} + (\lambda - v'_{yw}V^{-1}X)
\cdot (X'V^{-1}X)^{-1}(\lambda - X'V^{-1}v_{yw}).
\]

Robinson (in his Section 1) has chosen to define the BLUP (and to express its MSE) in terms of the mixed-model equations. Thus, his definition is specific to mixed-model prediction. I prefer to start with the more widely applicable representations (1) and (2) and to use them (together with well-known matrix identities) to derive the representations given by Robinson. The value of the latter representations is that, in the special case of mixed-model prediction, \( v_{w}, v_{yw} \) and \( V_{\gamma} \) have a relatively simple structure, and these representations indicate how (for computational purposes) to exploit that structure. In much the same way, the Kalman filter can be regarded as an algorithm for efficiently computing BLUPs in the special case of “time-series” prediction.

2. **EMPIRICAL BLUP VERSUS EMPIRICAL BAYES**

To account (in the general prediction problem) for \( \theta \) being unknown, it is common practice to adopt an even, translation-invariant estimator \( \hat{\theta} \) of \( \theta \) and, for purposes of the (point or interval) prediction of the value of \( w \), to act as though \( \hat{\theta} \) is the true value of \( \theta \). In particular, a point predictor, say \( \hat{\theta} \), can be obtained from the BLUP \( \bar{w} \) by substituting \( \hat{\theta} \) for \( \theta \). It seems natural to refer to this predictor as the **empirical BLUP**.

Empirical BLUP, as applied to mixed-effects linear models with normally distributed random effects and errors, is equivalent to PEB (parametric empirical Bayes) inference, as applied to one very important class of problems. This equivalence is discussed by Robinson in his Section 5.7, but only briefly and in rather general terms. It may be worthwhile to examine the equivalence in a relatively simple setting.

Consider first PEB inference about the means \( \mu_{i,1}, \ldots, \mu_{i,J} \) of \( I \) “groups” based on the one-way cell-mean model \( y_{ij} = \mu + \epsilon_{ij} \) \((i = 1, \ldots, I; j = 1, \ldots, J)\), where \( \epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ} \) are normally and independently distributed random variables with mean zero and common, unknown variance \( \sigma_{\epsilon}^2 \). It is instructive to describe PEB inference by relating it to HB (hierarchical Bayes) inference. HB inference about \( \mu_{i,1}, \ldots, \mu_{i,J} \) is essentially Bayesian inference based on a prior distribution that is assigned to \( \mu_{i,1}, \ldots, \mu_{i,J} \), and \( \sigma_{\epsilon}^2 \) in stages, say by (1) specifying that, conditional on \( \sigma_{\epsilon}^2 \) and “hyperparameters” \( \mu \) and \( \gamma \), the means \( \mu_{i,1}, \ldots, \mu_{i,J} \) are independently and identically distributed as \( N(\mu, \gamma \sigma_{\epsilon}^2) \), (2) assigning \( \mu \) a distribution conditional on \( \sigma_{\epsilon}^2 \) and \( \gamma \), (3) assigning \( \sigma_{\epsilon}^2 \) a distribution conditional on \( \gamma \), and (4) assigning \( \gamma \) a distribution. In PEB inference, the prior distribution might only be specified up to the values of \( \mu \), \( \gamma \), and \( \sigma_{\epsilon}^2 \), and the evaluation of point or interval estimators may be based on criteria such as PEB risk that are defined in terms of the conditional distribution of \( y_{11}, y_{12}, \ldots, y_{1J}, y_{21}, \ldots, y_{2J}, \ldots, y_{I1}, \ldots, y_{IJ} \) and \( \mu_{1,1}, \ldots, \mu_{1,J} \) given \( \mu \), \( \gamma \), and \( \sigma_{\epsilon}^2 \). The (parametric) empirical Bayesian may use frequentist methods (i.e., methods based on the conditional distribution of \( y_{11}, y_{12}, \ldots, y_{1J}, \mu_{1,1}, \ldots, \mu_{1,J}, \) \( \mu_{2,1}, \ldots, \mu_{2,J}, \ldots, \mu_{I,1}, \ldots, \mu_{I,J} \)) to obtain estimates \( \hat{\mu}, \hat{\gamma}, \) and \( \hat{\sigma}_{\epsilon}^2 \) of \( \mu \), \( \gamma \), and \( \sigma_{\epsilon}^2 \), and then, acting as though \( \hat{\mu}, \hat{\gamma}, \) and \( \hat{\sigma}_{\epsilon}^2 \) are the true values of \( \mu \), \( \gamma \), and \( \sigma_{\epsilon}^2 \), use what would, if the true values were known, be the posterior mean of \( \mu_i \) and a credible set for \( \mu_i \) as point and interval estimators. Thus, letting \( \hat{\gamma} = (1 + J\hat{\gamma})^{-1} \) \((i = 1, \ldots, I)\) and taking \( \hat{\mu} = \Sigma_i J_i \hat{\gamma}_i, \Sigma_i J_i \hat{\mu}_i \)

\[
(3) \quad \hat{\theta}_i = \hat{\mu} + (1 - \hat{\delta}_i)(\bar{y}_i - \hat{\mu}) = \hat{\delta}_i \mu + (1 - \hat{\delta}_i) \bar{y}_i.
\]
is a PEB (point) estimator of \( \mu_i \). PEB point and interval estimators may be modified to account for the estimation of \( \mu \), \( \gamma \), and \( \sigma_{\epsilon}^2 \).

For purposes of comparison, consider now the empirical BLUP approach to the prediction of the values of \( \mu + a_{1,1}, \ldots, \mu + a_{1,J} \), based on the one-way random-effects model

\[
(4) \quad y_{ij} = \mu + a_{i} + \epsilon_{ij} \quad (i = 1, \ldots, I; j = 1, \ldots, J),
\]
where \( \mu \) is an unknown parameter and where \( a_{1,1}, \ldots, a_{1,J} \) and \( \epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ} \) are normally and independently distributed random variables with
mean zero and common, unknown variances $\sigma_a^2$ and $\sigma_e^2$, respectively. Upon setting $\gamma = \sigma_a^2 / \sigma_e^2$, we find that the joint distribution of $y_{i1}, y_{i2}, \ldots, y_{iJ_i}$ and $\mu + a_1, \ldots, \mu + a_J$ under the one-way random-effects model is the same as the conditional joint distribution of $y_{i1}, y_{i2}, \ldots, y_{iJ_i}$ and $\mu_1, \ldots, \mu_J$ (given $\mu$, $\gamma$, and $\sigma_e^2$) obtained in the PEB approach to the one-way cell-mean model. Thus, the criteria adopted by the empirical BLUPer in the (point and interval) prediction of the values of $\mu + a_1, \ldots, \mu + a_J$ under the one-way random-effects model are equivalent to those adopted by an empirical Bayesian in the estimation of $\mu_1, \ldots, \mu_J$. In particular, the definitions of MSE and probability of coverage adopted by empirical BLUPers are equivalent to those adopted by empirical Bayesians. Accordingly, the (point and interval) estimators of $\mu_1, \ldots, \mu_J$ that have been proposed by empirical Bayesians tend to be similar or identical to the predictors of the values of $\mu + a_1, \ldots, \mu + a_J$ proposed by empirical BLUPers, especially the point estimators and predictors. For example, it is easy to show (say by using the mixed-model equations) that the PEB estimator $\hat{\mu}_i$ is an empirical BLUP predictor of the value of $\mu + a_i$.

While the work of empirical Bayesians has much in common with that of empirical BLUPers, there are some notable differences. In particular, in devising modifications to account for the estimation of $\gamma$ and $\sigma_e^2$, empirical Bayesians tend to pay more attention to conditional (given the values of various of the quantities $y_{i1}, y_{i2}, \ldots, y_{iJ_i}$ and $\mu_1, \ldots, \mu_J$—or of functions of those quantities—in addition to the values of $\mu$, $\gamma$, and $\sigma_e^2$) properties and relatedly to make greater use of HB ideas. Further, much of the work on PEB inference has been carried out by professional statisticians and has been theoretical in nature. This work has tended to focus on relatively simple models, like the balanced ($J_1 = \cdots = J_J$) one-way cell-mean model, since it is only these models that are tractable from a theoretical standpoint. Moreover, this work is often carried out under simplifying assumptions (e.g., an assumption that $\sigma_e^2$ is known). As a consequence, the impact of this work on statistical practice, while significant (Morris, 1983), has been restricted. By way of comparison, much of the work on empirical BLUP has been carried out by individuals in animal breeding and other fields of application, is applicable to relatively complex models (which may be intractable from a theoretical standpoint but which are often needed in practice), has focused on the problems (e.g., computational problems) encountered in applications and has had tremendous impact on statistical practice.

3. RANDOM EFFECTS

It is widely believed that the adoption of a mixed-effects linear model is appropriate only if interest centers on fixed effects and components of variance. A related and equally widespread belief is that, if inferences are to be confined to the effects of only those levels of a factor that are represented in the data, then those effects should be treated as fixed. As discussed by Robinson, both beliefs are incorrect. It seems ironic that Searle, who received a Ph.D. degree in animal breeding under the direction of C. R. Henderson, helped perpetuate these beliefs (through his highly influential book *Linear Models*). It may be worth mentioning that Searle partially redeemed himself by pointing out (on pages 461–462 of *Linear Models*) that the second part of the solution to the mixed-model equations ($\hat{a}$ in Robinson’s notation) “is, in many situations, of interest also.”

For the one-way random-effects model (4), the empirical BLUP of the value of $\mu + a_i$ is given by expression (3). If the effects $a_1, \ldots, a_J$ were treated as fixed rather than random, the empirical BLUP would be $\tilde{y}_i$. It would seem that the ultimate goal in deciding whether to treat effects as fixed or random should be to obtain the most useful methodology. Robinson suggests that effects should be treated as random only if they “come from a probability distribution.” The results of James and Stein (1961) suggest that this requirement may be overly stringent. Moreover, if strictly interpreted, it would never be satisfied. A less restrictive and perhaps more satisfactory requirement would be that the anticipated values of the effects be values that might “likely” arise in sampling a distribution.

I am not sure that the distinction between variation and uncertainty is as clear-cut as Robinson’s discussion (in his Section 7.5) would seem to indicate, nor am I convinced that “classical statistics can be distinguished from Bayesian statistics by its refusal to use probability distributions to describe uncertainty.” I believe that many classical statisticians are willing to use probability distributions to describe uncertainty anytime they feel their distributional assumptions will be acceptable to those to whom they wish to communicate their results.

4. BAYESIAN PREDICTION

The empirical BLUP $\hat{\omega}$ has an unappealing feature, which is explainable in terms of the confinement of $\theta$ to $\Omega$. Suppose, for example, that the model is the one-way random-effects model (4), and consider the empirical BLUP $\hat{\mu}_i$ of the value of
\( \mu + a_i \). When \( \hat{\gamma} = 0 \), the empirical BLUP reduces to \( \bar{y} \), which would be the BLUP of the value of \( \mu + a_i \) if it were known with certainty that \( \gamma = 0 \). This feature seems unappealing. The true value of \( \gamma \) can be larger than zero, but no smaller. Thus, when \( \hat{\gamma} = 0 \), we may consider it “likely” that the true value of \( \gamma \) is larger than the estimated value, and therefore be reluctant to act as though the estimated value is the true value.

An appealing alternative to the empirical BLUP can be obtained via the Bayesian approach. The Bayesian approach, as applied to the general prediction problem, consists of assigning \( \beta \) and \( \theta \) a joint distribution (based on prior information), of completing the specification of the (conditional) joint distribution of \( w \) and \( y \) (given \( \beta \) and \( \theta \)) up to the values of \( \beta \) and \( \theta \), and of forming the posterior distribution of \( w \). In the event \( p^* \) equals \( p \), \( \beta \) and \( \theta \) are statistically independent, the density of \( \beta \) is proportional to a constant, and the conditional distribution of \( w \) and \( y \) given \( \beta \) and \( \theta \) is MVN, the posterior probability density function of \( w \) is

\[
(5) \quad f(w \mid y) \propto \int_\Omega g(w \mid y, \theta) l(\theta \mid y) \pi(\theta) \, d\theta,
\]

where \( g(w \mid y, \theta) \) is the probability density function of a normal distribution whose mean is the BLUP of \( w \) and whose variance is the MSE of the BLUP, \( l(\theta \mid y) \) is the likelihood function employed in REML, and \( \pi(\theta) \) is the (prior) density of \( \theta \) (e.g., Harville, 1990). The posterior mean (i.e., the mean of distribution (5)) is a weighted “average” (over \( \theta \)) of the BLUP, with weights proportional to \( l(\theta \mid y)\pi(\theta) \). Point predictors of the general form of the posterior mean provide an appealing alternative to the empirical BLUP.

The Bayesian approach is computationally intensive—so much so that its use in many mixed-model applications is at present unfeasible. Recently, there has been increased interest in the computational aspects of the Bayesian approach (e.g., Kadane, 1990). More research should be directed towards the solution of the computational problems encountered in a Bayesian approach to mixed-model prediction.

5. A UNIFIED APPROACH TO PREDICTION

Methodologies have been developed for a number of special cases of the general prediction problem, including mixed-model methodology, kriging and the Kalman filter. These methodologies tend to be identified with specific types of application such as animal breeding, geostatistics and control theory, and they have tended to evolve independently of each other, with relatively little input from professional statisticians. To facilitate the exchange of ideas and to avoid duplication of effort, it would be desirable to develop a unified, but flexible, approach to prediction that accounts for the differences among the various special cases while exploiting the similarities. The terminology used in this endeavor should be kept as free as possible from highly specialized technical jargon.

Some thoughts related to the development of a unified approach to the general prediction problem are as follows.

1. It is useful to think of a model as a family of (unconditional) joint distributions for \( w \) and \( y \) obtained by (a) starting with the assumptions implicit in the general prediction problem, (b) possibly treating \( \beta \) and \( \theta \) as random vectors having a completely specified distribution (as in the Bayesian approach) or a distribution that is specified only up to the values of unknown hyper-parameters (as in the empirical Bayes approach), and (c) possibly introducing assumptions about the form of the (conditional) joint distribution of \( w \) and \( y \) (given \( \beta \) and \( \theta \)).

2. In formulating any particular prediction problem as a special case of the general prediction problem—and in deciding on any assumptions about the form of the (conditional) joint distribution of \( w \) and \( y \) (given \( \beta \) and \( \theta \))—it may be helpful to express \( w \) and \( y \) in terms of other random variables (e.g., random effects and errors) or related to proceed in stages by introducing additional random variables and conditioning on their values.

3. A model may serve to suggest a methodology (consisting, e.g., of point and interval predictors) and also as a basis for evaluating that methodology. Typically, the methodology should be evaluated under more than one model. The properties of the methodology conditional on \( y \), on various functions of \( y \), on \( \beta \) and \( \theta \), or on other quantities—as well as its unconditional properties—may be of interest. The ultimate criterion for judging a methodology is whether it possesses the characteristics sought by potential users.

4. A distribution assigned to \( \beta \) and \( \theta \) may or may not reflect a prior opinion. When prior opinion is hard to quantify or when the results of the analysis are to be communicated to individuals with different prior opinions, it may be preferable to assign to \( \beta \) and \( \theta \) a noninformative prior distribution, in which
case prior opinion may be incorporated informally, subsequent to the analysis. In any case, the distribution assigned to $\beta$ and $\theta$ should be regarded as part of the model.

5. The posterior distribution of $w$ (i.e., the conditional distribution of $w$ given $y$) may suggest suitable point and interval predictors. With the possible exception of the term posterior distribution, which might be used in referring to any distribution that is conditional on $y$, the use of Bayesian jargon should be avoided.

Comment: The Kalman Filter and BLUP

James C. Spall

1. INTRODUCTION

Professor Robinson has given a wide-ranging account of best linear unbiased prediction with an impressive array of examples and applications. In this discussion, however, I will restrict my attention to issues regarding the Kalman filter and BLUP.

For ease of discussion, let us restate the random effects model in state-space form as given in Robinson, Section 6. The unobservable random effects (state) vector, $u_t$, evolves according to

$$u_t = G_t u_{t-1} + w_t, \quad u_0 = 0, \quad t = 1, 2, \ldots, n,$$

where $w_t$ is a noise term with mean 0 and covariance matrix $W_t$, and $G_t$ is the state transition matrix. The second equation in the model relates the state vector to the vector of observables $y_t$:

$$y_t = F_t u_t + v_t,$$

where $v_t$ is a noise term with mean 0 and covariance matrix $V_t$, and $F_t$ is the measurement matrix. Equations (1.1a, b) can be expressed in the random effects model form of Robinson by writing

$$y = Zu + e,$$

where $y = (y_1^T, y_2^T, \ldots, y_n^T)^T$, $Z = \text{block diag}(F_1, F_2, \ldots, F_n)$, $u = (u_1^T, u_2^T, \ldots, u_n^T)^T$, and $e = (v_1^T, v_2^T, \ldots, v_n^T)^T$. The covariance matrix for $u$, $G$ in the notation of Robinson, is a function of $G_t$ and $W_t$, $t = 1, 2, \ldots, n$. The structure of this covariance matrix allows for recursive algorithms of the Kalman filter/smoother form to be used to form BLUP estimates for the components of $u$. Incidentally, a slightly confusing point in Robinson, Subsection 6.4, is that it is a Kalman smoother, not filter, that produces the BLUP estimate of $u$ based on data $y$. What Robinson had in mind, I presume, is the common problem where one is interested in an estimate of $u_t$ based only on data through time $t$ (not through some later time); the Kalman filter, of course, is used for this problem. For the remainder of this discussion, I will assume that the filtering problem is the one of interest (although virtually all of the ideas would also apply in the smoothing problem).

A couple of other points are worth noting here. First, Sallas and Harville (1988) address a slightly broader problem than that considered above and by Robinson: namely the estimation of random and fixed effects via Kalman filter techniques. Second, as noted by Robinson, the Kalman filter is not entirely due to Kalman. The filter equations were essentially derived by others prior to Kalman, but it was Kalman who crystallized much of the thinking in the area and discovered several key relationships to certain systems-theoretic concepts (see Spall, 1988, for further discussion of this).

In the next two sections, I will discuss two problems that were given fairly light treatment in the Robinson paper, but that are important from the point of view of a practitioner. Section 2 describes some problems associated with constructing uncertainty bounds for the filter estimation error $\hat{u}_t - u_t$ when the noise terms have an unknown distribution (as in the general setting of Robinson, equation 1.1). Section 3 elaborates on the brief discussion of Robinson regarding uncertainty in the model parameters $\theta$.

2. UNCERTAINTY BOUNDS FOR $\hat{u}_t - u_t$ IN DISTRIBUTION-FREE SETTINGS

Robinson presents the formula for the covariance matrix of the BLUP estimation error in Section 1 of his paper, and it is well known that this covariance
matrix can be computed recursively as part of the Kalman filter (or smoother, as appropriate). However, Robinson does not present any discussion as to how this (or other) information can be used to construct probability bounds or uncertainty regions for the estimation error $\hat{u}_n - u_n$ in the distribution-free setting that is the basis for most of the paper (in the case where $w_t, v_t$ are normally distributed for all $t$, the estimation error is also normally distributed for all $t$, which makes it relatively straightforward to derive these quantities). We will outline below an approach to allow one to construct probability bounds (and hence uncertainty regions) in the distribution-free setting.

Before presenting the approach, it is worth noting that unlike many other estimators, the Kalman filter estimation error is not asymptotically normally distributed. This follows because central limit effects do not hold (as a result of the disproportionate weight given to more recent terms in the sum), as shown below. Thus standard asymptotic procedures for uncertainty region calculation (such as those that are used in maximum likelihood parameter estimation) do not apply. To establish this asymptotic nonnormality, let us consider models that are in standard uniformly completely controllable and observable (UC and UCO) form (see, e.g., Jazwinski 1970, pages 232–234; Spall and Wall, 1984; or Anderson and Moore 1979, pages 68–82). Recall from the Cramér–Lévy theorem (e.g., Feller, 1971, pages 525–526), that the sum of two nondegenerate independent random vectors is normally distributed only if both of the random vectors are normally distributed. Straightforward algebra shows that

$$
(2.1) \quad \hat{u}_n - u_n = s_{n-1} + (K_n F_n - I) w_n + K_n v_n,
$$

where $s_{n-1}$ is the weighted sum of $\{w_t, v_t\}_{t=1}^{n-1}$ and $K_n$ is the Kalman gain matrix. Since UC and UCO imply that $\alpha_0 I \leq P_n = \text{cov}(\hat{u}_n - u_n) \leq \alpha_1 I$ with $0 < \alpha_0 \leq \alpha_1 < \infty$, we know that $\|K_n F_n - I\| \geq \alpha_2$ and $\|K_n\| \geq \alpha_2$ for some $\alpha_2 > 0$ and any matrix norm (Jazwinski, 1970, pages 234–237; Deyst and Price, 1968; or Deyst, 1973). Thus by (2.1) the contributions of the $w_n$ and $v_n$ terms to $\hat{u}_n - u_n$ will be nonnegligible as $n \to \infty$. Then by the Cramér–Lévy theorem $\hat{u}_n - u_n$ is not asymptotically normal. The implications of the above have been examined in several simulation experiments with uniformly distributed noise terms. Applications of the Kolmogorov–Smirnov goodness-of-fit test (with a null hypothesis that $\hat{u}_n - u_n \sim N(0, P_n)$) yielded $p$ values of less than $10^{-6}$ for a variety of state-space parameter values, confirming that it is dangerous to assume that $\hat{u}_n - u_n$ is even approximately normal.

The approach to characterizing the uncertainty in $\hat{u}_n - u_n$ is to bound the probability $P(\hat{u}_n - u_n \in E_n)$ for rejection regions $E_n$ such that the complementary region $E_n^c$ is symmetric and convex. For example, $E_n$ might represent the points outside a $q$-dimensional spheroid, i.e., the values of $\hat{u}_n - u_n$ such that $\|\hat{u}_n - u_n\| \geq c$ for Euclidean norm $\|\cdot\|$ and some $c > 0$. Note that one easy bound is that given by Chebyshev’s inequality, e.g., $P(\|\hat{u}_n - u_n\| \geq c) \leq \frac{E\|\hat{u}_n - u_n\|^2}{c^2} = \text{trace}(P_n)/c^2$. We seek a bound that has the potential to be more precise than the Chebyshev bound. Anderson’s inequality (Anderson, 1955; Tong, 1980, page 55) provides a means to this goal.

Let us write

$$
\hat{u}_n - u_n = A_n \overline{w}_n + B_n \overline{v}_n
$$

where $\overline{w}_n = (w_T, \ldots, w_n)^T$, $\overline{v}_n = (v_T, \ldots, v_n)^T$, $A_n = (A_{n1}, A_{n2}, \ldots, A_{nn})$, $B_n = (B_{n1}, B_{n2}, \ldots, B_{nn})$, and $A_{nt}, B_{nt}$ are weighting matrices as derived from the Kalman filter and the state equation (see Spall and Wall, 1984, equation 2.2, for $B_{nt}$ and the filter contribution to $A_{nt}$). Suppose that either the $\{w_t\}$ or $\{v_t\}$ process is normally distributed and that the other sequence has an unknown symmetric, unimodal distribution (the partial assumption of normality is stronger than required for the technique, but is made here for ease of discussion). We now create a surrogate expression that will be used to form a probability that bounds $P(\hat{u}_n - u_n \in E_n)$. By the fact that UCO and UC imply that the filter is exponentially stable (Jazwinski, 1970, pages 240–242), we have that $\|A_n\| = O(e^{-c_0(n-1)})$ and $\|B_n\| = O(e^{-c_1(n-1)})$ for some $c_0, c_1 > 0$ as $n \to \infty$. To apply Anderson’s theorem, we leave the weighting matrix sequence associated with the normally distributed noise process unchanged, but modify the other sequence so that all of those weighting matrices have (at least approximately) magnitude equal to the largest magnitude matrix in the sequence. For convenience, suppose that $\{v_t\}$ is the Gaussian sequence. We then create a modified sequence $A_n^* = a_n A_n$ where $|a_n| \geq 1$ and $A_n^* = O(1)$. Then $A_n^* \overline{w}_n$ is approximately normally distributed by the Lindeberg–Feller form of the central limit theorem and so $A_n^* \overline{w}_n + B_n \overline{v}_n$ is approximately normally distributed with mean 0 and covariance matrix $A_n^* \text{block diag}[W_1, W_2, \ldots, W_q] A_n^* + B_n \text{block diag}[V_1, V_2, \ldots, V_q] B_n^T$. Since $|a_n| \geq 1$, an iterative application of Anderson’s theorem to each $w_i$ term implies

$$
(2.2a) \quad P(\hat{u}_n - u_n \in E_n) = 1 - P(A_n^* \overline{w}_n + B_n \overline{v}_n \in E_n^c) \\
\leq 1 - P(A_n^* \overline{w}_n + B_n \overline{v}_n \in E_n^c) \\
= P(A_n^* \overline{w}_n + B_n \overline{v}_n \in E_n).
$$
Similarly if \( \{ w_t \} \) has a normal distribution and \( \{ v_t \} \) has an unknown distribution, we have

\[
P(\hat{u}_n - u_n \in E_n) \leq P(A_n \hat{w}_n + B_n^* \hat{v}_n \in E_n),
\]

where the \( \{ B_n^* \} \) are chosen in the same manner as \( \{ A_n^* \} \) above, and \( A_n \hat{w}_n + B_n^* \hat{v}_n \) is approximately normally distributed with mean 0 and the obvious covariance matrix.

For bounded \( E_n^* \) (the usual case) and large \( n \), the probability bounds in (2.2a,b) may not be satisfactory since they will approach unity due to the fact that \( A_n^* \hat{w}_n \) and \( B_n^* \hat{v}_n \) are of order \( n \). However, for shorter realizations (but long enough to achieve practical central limit theorem effects) the bounds may represent an improvement over the Chebyshev inequality, as illustrated below.

For a scalar \( u \), and \( \gamma \) setting, Table 1 compares the bound of (2.2a) (using \( A_n^* = \max_t |A_{nt}| \forall t \)) with the probability values resulting from an assumption of normality for all noise terms (so \( \hat{u}_n - u_n \) is normal) and from an application of the Chebyshev inequality. All state-space parameters (\( G_t, W_t, F_t, \) and \( V_t \)) were taken to be unity and, as with Robinson’s first-lactation example, \( n = 9 \). As expected, the probability values for bound (2.2a) lie between those of the Chebyshev inequality and those resulting from the normal distribution assumption for \( \hat{u}_n - u_n \).

### Table 1

<table>
<thead>
<tr>
<th>c</th>
<th>Normal</th>
<th>Chebyshev</th>
<th>Bound (2.2a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sqrt{p_n} )</td>
<td>.32</td>
<td>1.0</td>
<td>.56</td>
</tr>
<tr>
<td>2( \sqrt{p_n} )</td>
<td>.05</td>
<td>.25</td>
<td>.24</td>
</tr>
<tr>
<td>3( \sqrt{p_n} )</td>
<td>.003</td>
<td>.11</td>
<td>.08</td>
</tr>
<tr>
<td>4( \sqrt{p_n} )</td>
<td>0</td>
<td>.06</td>
<td>.02</td>
</tr>
</tbody>
</table>

One of the beauties of the state-space formulation of the random effects model is that under the assumption of normally distributed noise terms the Kalman filter can be used for calculating the log-likelihood function and its derivatives, in addition to its usual application to state \( (u_t) \) estimation. This is achieved by writing the log-likelihood in terms of the independent “innovations” sequence \( \{ \gamma_t - F_t \hat{G}_t \hat{u}_{t-1} \}^r_{t=1} \). This process has been described in many places (e.g., Goodwin and Payne, 1977, pages 158–159; Sallas and Harville, 1988). As part of this process, Kalman filter type recursions yield the Fisher information matrix, which, when inverted, can serve as an approximation to the covariance matrix of the parameter estimation error.

As suggested by Robinson, the uncertainty in model parameters should be reflected in the uncertainty of the BLUP estimate. This may be viewed as a problem in nuisance parameter analysis, i.e., the estimation of quantities of interest in the presence of uncertainty in other terms within the model. Several authors have considered this problem in the context of state-space models. Ansley and Kohn (1986) present an expression for the asymptotic mean square error for the Kalman filter state estimate \( \hat{u}_n \) in the presence of uncertain model parameters. Spall and Garner (1990) consider how uncertainty in certain model parameters (e.g., uncertainty in a scale factor such as \( \sigma^2 \) in the model of (1.1) in Robinson) will affect the precision for estimates of other nonrandom parameters (e.g., estimates of \( \theta \) or the fixed effects \( \beta \) in the Robinson Model). Both the Ansley–Kohn and Spall– Garner approaches rely on Kalman filter type recursions, the former on differentiated state estimates and the latter on a differentiated log-likelihood function and score vector.

### 4. Concluding remarks

Although largely in the domain of the control and aerospace engineering communities until the early 1970s, the Kalman filter has now been embraced by the statistical community as the method of choice for a wide range of time series problems. The Kalman filter/state-space approach has been successfully applied in countless problems drawn from perhaps every major branch of the physical and social sciences. For these reasons, I think that the Kalman filter approach is, in the sense of Robinson, “a good thing.”
Comment

Terry Speed

Geoff Robinson is to be congratulated for writing this paper. It is lucidly written, it bridges a number of gulf's that have developed in our subject, and it is provocative. That he wrote it is clearly a Good Thing! I welcome the opportunity to say this and to make a few remarks that he might have made. I believe that these remarks will strengthen his already strong case for a much more explicit recognition of the role of BLUPs in our subject.

1. THE BAYESIAN DERIVATION

In Section 4.2 Robinson describes a Bayesian derivation, stating that the posterior mode is given by the BLUP estimates when \( \beta \) is regarded "as a parameter with a uniform, improper prior distribution and \( u \) as a parameter which has a prior distribution which has mean zero and variance \( G \sigma^2 \), independent of \( \beta \)." All this is certainly true, but it may be helpful to add that if \( \beta \) is given a proper prior (normal) distribution with mean zero and variance \( B \sigma^2 \), say, with \( u \) as before, then all of the results one could possibly want (posterior means, posterior variances, etc.) can be derived straightforwardly by the standard Bayesian formulae. Then all one has to do to derive the corresponding BLUP formulae is let \( B^{-1} \to 0 \). An identity which I have found useful, perhaps even indispensable, for carrying out this last step, is discussed in de Hoog, Speed and Williams (1990). Note that the approach just described is essentially that adopted in Dempster, Rubin and Tsutakawa (1981).

2. FORMULAE FOR \( \hat{u} \)

The only actual formulae given in the paper for \( \hat{u} \) in the general case is the rather complicated one in Section 4.3. This is a pity, because there is an obvious "plug-in" expression, namely

\[
\hat{u} = GZ^TV^{-1}(y - X\hat{\beta}),
\]

where \( V = ZGZ^T + R \). This may be viewed as the result of regressing \( u \) on \( y \), with the mean \( X\beta \) of \( y \) replaced by its obvious linear estimator.

A variant of (1) is

\[
\hat{u} = (Z^TR^{-1}Z + G^{-1})^{-1}Z^TR^{-1}(y - X\hat{\beta}).
\]

The simpler formulae (5.3) and (5.4) arising when there are no fixed effects also have more general analogues, namely

\[
(Z^T AZ + G^{-1}) \hat{u} = Z^T Ay,
\]

where \( A = R^{-1}(I - S) \), \( S = P_{\#(X)} \) being the projector onto \( \mathcal{A}(X) \) orthogonal with respect to \( (a, b) = a^TR^{-1}b \), and for the variance-covariance matrix of \( \hat{u} \):

\[
\{G^{-1} - (Z^T AZ + G^{-1})^{-1}\} \sigma^2.
\]

These expressions can be derived readily using the Bayesian approach outlined in (1) above, together with the matrix identity already referred to. I note in passing that Robinson's formulae (5.4) is in fact the variance-covariance matrix of \( \hat{u} - u \), not, as stated, of \( \hat{u} \).

3. SOLVING THE BLUP EQUATIONS

Perhaps in order to avoid messy algebra, Robinson has said little about the actual solution of the BLUP equations. I know that he has worked on this problem with some enormous data sets, and so I am hesitant to comment here. However, it does seem worthwhile to make one easy point, in order to connect this topic with another, closely related one. The obvious rearrangement of the first equation in (1.2),

\[
X^TR^{-1}X\hat{\beta} = X^TR^{-1}(y - Z\hat{u}),
\]

can be combined with either (1) or (1') above, to form the basis of an iterative solution of the BLUP equations, provided, of course, that the separate problems are readily solved. Just such a strategy is recommended more generally in Green (1985) in the context of smoothing, a topic to which I shall return.

It is also worth pointing out that (1') or (2) is to be preferred when \( G^{-1} \) has simple structure, whereas if \( G \) is simple and \( V \) is readily inverted, (1) is more useful. In many animal breeding problems it is \( G^{-1} \) which has the simpler structure, as it also does in the Kalman filter case.

4. REML AND BLUP

In Section 5.4 Robinson states "REML is the method of estimating variance components that seems to have the best credentials from a Classical

Terry Speed is Professor, Department of Statistics, University of California, Berkeley, California 94720.
viewpoint.” What he does not say, which should be of interest to readers of his paper, is that REML and BLUP are intimately connected. Indeed one view—certainly not the only one—of the REML equations for variance components is that they are simply equating observed with expected sums of squares of BLUPs. This observation goes back to the original paper by Patterson and Thompson (1971; see also Harville, 1977) and can be concisely stated within Robinson’s framework as follows.

Suppose that \( Z = [Z_1 : \cdots : Z_c] \) is blocked, corresponding to \( c \) random effects, with \( Z_i \) being \( n \times q_i \), \( i = 1, \ldots, c \), \( u = (u_1 : \cdots : u_c) \) is similarly blocked into \( q \) sets of random effects, and finally \( G = \text{diag}(G_1, \ldots, G_c) \) is diagonally blocked with \( G_i = \gamma_i I_{q_i} \), where \( q_i^2 = \gamma_i \sigma^2 \) is the variance of each independent component of the \( i \)th random effect \( u_i \). It is also convenient to denote \( e \) by \( u_0 \), put \( Z_0 = I_n \) and \( \gamma_0 = 1 \).

With this notation the REML equations take the form

\[
y^T \left( V^{-1} \bar{Q} \frac{\partial V}{\partial \sigma^2} V^{-1} \bar{Q} \right) y = \text{tr} \left( \frac{\partial V}{\partial \sigma^2} V^{-1} \bar{Q} \right)
\]

\( i = 0, \ldots, c, \) where \( Qy = X\hat{\beta} \) and \( \bar{Q} = I - Q \). (By contrast, the ML equations have no \( \bar{Q} \) term in the right-hand expression.)

Turning now to BLUPs in this context, they are (in the form (1) above)

\[
\hat{u}_i = \gamma_i Z_i^T V^{-1} \bar{Q} y
\]

\( i = 0, \ldots, c, \) and

\[
\text{var}(\hat{u}_i) = (G_i - U_i) \sigma^2
\]

\( i = 1, \ldots, c, \) where \( U_i \) is the \( i \)th diagonal block of the matrix \((Z^T AZ + G^{-1})^{-1}\). Furthermore,

\[
\text{var}(\hat{e}) = V^{-1} \bar{Q} \sigma^2 = (A - AZU Z^T A) \sigma^2
\]

where \( A = R^{-1} \bar{S} \) was defined earlier, and \( U = (Z^T AZ + G^{-1})^{-1} \). If we write \( p_i = \gamma_i^{-1} \text{tr}(U_i), i = 1, \ldots, c \), then it follows that for \( i = 1, \ldots, c \)

\[
\mathbb{E} | \hat{u}_i |^2 = (q_i - p_i) \sigma^2
\]

and

\[
\mathbb{E} | \hat{e} |^2 = \left[ (n - p) - \sum_{i=1}^{c} (q_i - p_i) \right] \sigma^2.
\]

Now the striking thing is this: the REML equations (5) can rather easily be manipulated into a form just like (8a) and (8b), with the expectation symbol \( \mathbb{E} \) omitted. Although this is not necessarily the best way to solve these equations, the repeated calculation of BLUPs and then updating the variance components is one simple iterative scheme which works quite well.

5. PENALIZED LEAST SQUARES

Suppose that we regard (1.1) as an ordinary (“fixed effects”) linear model, and that we wished to estimate \( \beta \) and \( u \) by \( R \)-weighted least squares with a “penalty” \( u^T G^{-1} u \) being added to the sum of squares term being minimized. Then we would obtain just the expression given in Section 4.1, which Henderson minimized. Such penalties are added for many reasons: to smooth, to improve the condition of the matrix to be inverted, and so on, and it has long been recognized that this is a way of making one’s linear model “quasi-Bayesian.” More precisely, it turns the standard least squares problem into a case of BLUP. This practice has a long history, dating back at least to Whittaker (1923).

6. SMOOTHING SPLINES ARE BLUPS

Continuing with the theme of the previous remark, let us see how the smoothing splines popularized by G. Wahba (see her 1990 monograph for a comprehensive exposition) are just BLUPs. This observation corrects the terminology which has been used in the spline literature for over a decade, for the Bayesian interpretation of the smoothing spline—with a partially improper prior—is just the statement heading this section.

It is simplest to deal with cubic smoothing splines on the interval \([0,1]\). If the observations are taken at \( 0 = t_1 < \cdots < t_n \leq 1 \), and are

\[
y_i = g(t_i) + \epsilon_i,
\]

\( i = 1, \ldots, n \), where \( g \) is an unknown smooth function, then the function \( g^* \) which minimizes

\[
n^{-1} \sum_{i=1}^{n} (y_i - g(t_i))^2 + \lambda \int_{0}^{1} \{g^*(u)\}^2 \, du
\]

over a suitable class \( G_0 \otimes G_1 \) of functions, has the values

\[
g^* = (g^*_0(t_i))
\]

\( i = 1, \ldots, n \),

\[
X(XX^T X)^{-1} XX^T V^{-1} y + Q_n V^{-1} \left( I - X(XX^T X)^{-1} X^T V^{-1} \right) y,
\]

where

\[
X = (t_i^{k-1}), \quad i = 1, \ldots, n; \quad k = 1, 2;
\]

\( Q_n(i,j) = Q(t_i, t_j), \quad 1 \leq i, j \leq n; \)

and

\[
Q(s,t) = \int_{0}^{1} (s-w)_+ (t-w)_+ \, dw, 0 \leq s, t \leq 1.
\]
It is easy to check that (9) is just the fitted value
\[ \hat{y} = X\hat{\beta} + Zu, \]
where, in Robinson's notation, \( \hat{\beta} \) and \( u \) are the BLUPs, \( X \) is as given above, \( Z = I_n = R \), and \( G = (n\lambda)^{-1}Q_n \).

Certainly there is more to smoothing splines than BLUPs; for example, estimates of the value of the function \( g \) at values of \( t \) other than those observed, but in many applications (9) and related expressions are all that is required.

By now it should come as no surprise to hear that the technique termed generalized maximum likelihood (GML) for estimating the smoothing parameter \( \lambda \) is no other than REML in this BLUP problem. This is readily checked by comparing formulae in this paper with ones in Wahba (1990).

With only very few changes, the identification just made to show that smoothing splines are BLUPs shows that the model robust response surface designs of Steinberg (1985) are also BLUPs. In this case the \( u \) term corresponds to sums of tensor products of orthogonal polynomials.

7. LINEAR SMOOTHERS ARE ALMOST BLUPS

There is a sense in which all linear smoothers (see Buja, Hastie and Tibshirani, 1989) are intimately related to BLUPs. A typical linear smoother \( S \) satisfies \( S^n \to T \) as \( n \to \infty \), where \( T \) is idempotent. This corresponds to a projector onto the subspace \( \mathcal{R}(X) \) in Robinson's model (1.1), and so \( Ty \) corresponds to \( X\hat{\beta} \). Thus \( (S - T)y \) corresponds to \( Zu \), and in some situations it is even possible to construct a covariance matrix \( V \) such that this correspondence is precise. Furthermore, many smoothers \( S \) have form \( S(\lambda) \), where \( \lambda \) is a parameter (bandwidth, variance ratio, smoothness penalty, etc.) that defines a family of similar smoothers. In such cases \( S(\infty) \) often has the form \( T + W \), where \( W \) is another projector, while \( S(0) = T \). Many of the problems and the formulae in the theory of linear smoothers are analogues of ones arising in the theory of BLUPs.

8. INTERVAL ESTIMATES INVOLVING BLUPS

In Section 5.6 Robinson briefly alludes to work done on estimating the precision of BLUP estimates when uncertainty in the dispersion parameter is taken into account. This general problem, and in particular the assignment of interval estimates, has attracted a lot of attention in the literature on smoothing splines (see, e.g., Nychka, 1988, for a recent review). Much concern has been given to the question of what, if any, coverage properties can be expected of a "Bayesian" posterior interval. Making interval statements about an object which is an estimate of the sum of fixed and random effects is bound to cause problems of interpretation to many people, and I would be interested to hear Geoff Robinson's comments on this point. I know that he has studied these matters closely in the past.

9. SUMMARY

In closing these few remarks, I cannot resist paraphrasing I. J. Good's memorable aphorism: "To a Bayesian, all things are Bayesian." How does "To a non-Bayesian, all things are BLUPs" sound as a summary of this fine paper?

ACKNOWLEDGMENT

The author gratefully acknowledges partial support from NSF Grant DMS-88-02378, and the hospitality of the Department of Statistics, IAS, Australian National University whilst preparing this comment.
Comment

Duane Steffey and Robert E. Kass

Dr. Robinson's well-written article provides a variety of perspectives on the general linear model and its applications. Particularly welcome are the examples of Section 6 illustrating the widespread utility of these models. We limit discussion here to three main points. First, further details are provided on approximate Bayesian methods for inference about unit-specific parameters ("random" effects). Next, we amplify Dr. Robinson’s comment on the often close agreement between Bayesian and frequentist inferences. Specifically, we give an approximation to the variance of the marginal posterior distribution of a unit-specific parameter and conjecture that the expression may also be justified on frequentist grounds as an approximation to the sampling variance of the BLUP estimator. Finally, we discuss the desirability of using all relevant information and mention some possible mechanisms for incorporating prior knowledge about animal breeding into the general linear model.

1. APPROXIMATE BAYESIAN INFERENCE

We here consider the marginal posterior distribution of a unit-specific parameter and provide a rather general variance approximation that satisfies the often-identified need (e.g., noted by Robinson in Section 5.6) to account for the uncertainty in estimating the common dispersion parameters. We begin by rewriting Robinson’s model (1.1). Switching to a formulation similar to that of Laird and Ware (1982), we consider the general linear model in which there are k experimental units and, for the ith unit,

\[ Y_i = X_i \beta + Z_i u_i + e_i. \]

Here, \( Y_i \) is an \( n_i \times 1 \) vector of responses, \( \beta \) is a \( p \times 1 \) vector of unknown population parameters and \( X_i \) is a known \( n_i \times p \) matrix linking \( \beta \) to \( Y_i \). In addition, \( u_i \) is a \( q \times 1 \) vector of unknown individu-
(y_1, \ldots, y_k) are obtained by applying the results of Kass and Steffey (1989):

\[ E(u_i \mid y) = E(u_i \mid y_i, \hat{\theta}) \]

\[ \text{VAR}(u_i \mid y) = \text{VAR}(u_i \mid y_i, \hat{\theta}) + \sum_{j,k} \delta_{jk} \delta_{jk}, \]

where \( \delta_{jk} \) is the \((j, k)\)-component of

\[ \Sigma = [-D^2 \log L(\theta) \pi(\theta) \mid_{\theta = \hat{\theta}}]^{-1} \]

and

\[ \delta_{jk} = \frac{\partial}{\partial \theta_j} E(u_i \mid y_i, \theta) \mid_{\theta = \hat{\theta}}. \]

Here, \( L(\theta) \) denotes the likelihood function, \( \pi(\theta) \) denotes the prior for \( \theta \), and \( D^2 \) denotes second-order partial differentiation with respect to the elements of \( \theta \). (For specific formulas, see Kass and Steffey, 1989; Harville, 1977.) Note the convenient interpretation of (3); the first term is the conditional posterior variance with \( \theta \) set equal to \( \hat{\theta} \), and the second term accounts for the additional uncertainty in estimating \( \theta \).

2. APPROXIMATE VERSUS EMPIRICAL BAYES

The approximations (2) and (3) have an alternative frequentist interpretation. The approximate posterior mean (2) is commonly recognized as the parametric empirical Bayes (PEB) estimator. In addition, we believe that under appropriate regularity conditions, the approximate posterior variance (3) may also be justified fairly generally as a frequentist approximation to the sampling variance of the PEB estimator. In the context of the general linear model (1), when \( \pi(\theta) = 1 \) the expression (3) is very similar to Kackar and Harville’s (1984) approximation to the mean squared error of the PEB estimator, differing only in that Kackar and Harville use estimated expected information rather than observed information. Specifically, with

\[ \text{VAR}(G(Y_i, \hat{\theta})) = \hat{\Sigma} \{1 + O_p(n^{-1/2})\} \{1 + O_p(k^{-1/2})\}, \]

where \( \hat{\theta} \) is the MLE of \( \theta \), \( \hat{\Sigma} \) is the approximate posterior variance given in (3) and our \( O_p \) statements refer to the two-stage model—i.e., the joint distribution of \((Y_1, u_1), \ldots, (Y_k, u_k)\) for some fixed \( \theta \in \Theta \).

To indicate heuristically why we make this conjecture, we write \( V_1 = \text{VAR}(G(Y_i, \hat{\theta}) \mid Y_i = y_i) \)

and \( V_2 = \text{VAR}(G(Y_i, \hat{\theta}) \mid Y_i = y_i) \), and use the conditional variance formula

\[ \text{VAR}(G(Y_i, \hat{\theta})) = V_1 + V_2. \]

That the second term of (3) estimates \( V_2 \) can be readily seen by carrying out a Taylor series expansion and standard delta method arguments. The conditional posterior variance \( \text{VAR}(u_i \mid y_i, \hat{\theta}) \) can then be shown to estimate \( V_1 \) by linking three steps: (i) \( \text{VAR}(u_i \mid y_i, \theta) \) is approximately equal to the variance of \( \hat{u}_i \) conditional on \( u_i \), while (ii) the latter is approximately equal to the variance of the conditional posterior expectation \( \text{VAR}(G(Y_i, \theta)) \); finally, (iii) this quantity \( \text{VAR}(G(Y_i, \theta)) \) can be shown to estimate \( V_1 \). Substituting \( \hat{\theta} \) for \( \theta \) then yields the desired result. These arguments are most easily made by taking advantage of the asymptotic independence of \( Y_i \) and \( \hat{\theta} \) as \( k \to \infty \). In words, the influence of any single observation on the estimation of \( \theta \) becomes negligible as the number of experimental units increases.

Note that the conjecture (4) applies to all CIHMs, of which the linear model (1) is a special case. Hence, the approximation (3) with a vague prior such as \( \pi(\theta) = 1 \) may find usage among frequentist statisticians as a means of overcoming the mathematical complications typically encountered in estimating the precision of BLUP estimators and, more generally, of PEB estimators in hierarchical models. For related recent work on PEB inference, see Laird and Louis (1987), Carlin and Gelfand (1990), and Hill (1990).

3. PRIOR INFORMATION AND ANIMAL BREEDING

We have undertaken only a cursory review of the animal breeding literature, but we suspect that, after generations of animal breeding experiments, considerable knowledge has been accumulated about the distributions of herd and sire effects. Does the author concur in this view?

Statistical models that incorporate available prior information will typically yield inferences about quantities of interest that are more accurate than those obtained from models that ignore relevant information. Gianola and Fernando (1986) discuss Bayesian methods for estimating breeding value (the \( u \) vector) and genetic parameters (the \( \theta \) vector). They note that prior information is often available and that its use can preclude anomalies such as nonpositive definite estimated covariance matrices and “ridiculous estimates of heritability” (page 219). Also, they explain how the Bayesian approach provides a logical framework for handling
problems such as those involving sequential experiments and those in which only indirect data (from relatives) are available in predicting an individual’s breeding value.

Robinson notes (Section 4.2) that the BLUP estimates may be viewed as approximate Bayes estimates with improper uniform priors on $\beta$ and $\theta$. However, the results using proper prior distributions (even only mildly informative ones) that reflect pre-experiment knowledge can be substantially different from those obtained using flat priors; for instance, see Example 2 in Kass and Steffe (1989). For further details of Bayesian analyses with informative priors for $\beta$ and $\theta$ in the linear model (1), see Gianola and Fernando (1986), Broemeling (1985) and the references contained therein.

The difficulty in finding ways to incorporate prior information has led many applied statisticians to question the practical value of Bayesian methods. Establishing sensible methodologies has been a goal of continuing research by statistical scientists, cognitive psychologists and econometricians. While translating uncertainty into probability distributions can be challenging, the potential scientific rewards for doing so can be substantial. Some authors have advocated generating probability distributions from statements made by substantive experts as a mechanism for incorporating prior information. For example, Kadane, Dickey, Winkler, Smith and Peters (1980) present a method for specifying a conjugate prior for ($\beta, \sigma^2$) in the normal linear regression model. That procedure is based on collecting responses from substantive experts (non-statisticians) to questions about the predictive distribution of the response vector given values of the predictor variables. Such a procedure may be adaptable to the mixed effects models considered here.

Along with Kadane (1990) we would emphasize the need for more work on elicitation and would add that the need is especially great when elicitation is taken, in its broadest sense, to refer to the general process of constructing probability distributions from available background information.

We look forward to future modeling efforts that tap all sources of relevant information in order to improve inferences in the statistical problems encountered in animal breeding and many other fields of science.

Comment

Robin Thompson

Dr. Robinson's paper is valuable as it shows the links of BLUP, suggested for animal breeding applications, with methods used in other areas. An alternative way of thinking about the models used is in terms of an expectation and a variance for $y$

Robin Thompson is Head of the Biometrical Genetics Department, Edinburgh Research Station, Institute of Animal Physiology and Genetics Research, Roslin, Midlothian, EH25 9PS, Scotland.

that leads to a natural interpretation for prediction in terms of regressing future observations on present observations. I wonder why Dr. Robinson did not use such a formulation. With regard to making inferences on random estimates, can Dr. Robinson say if it is sensible to use the suggestion of most likely unobservables to construct confidence intervals for random estimates? I would also like to know which likelihood to use when testing fixed effects.
Rejoinder

G. K. Robinson

INTRODUCTION

I would like to thank the discussants for their remarks. I hope that readers will find that the discussion helps to clarify the ideas that I tried to present in my paper. Mostly, I have chosen not to use this opportunity to restate my opinion on minor points where I disagree with the discussants or where I would give different emphasis.

In this introduction I will pass quickly over a number of issues which can each be presented briefly. Issues requiring longer discussion will be laid out as separate sections.


Following comments by Harville and Speed, I think that my presentation would have been easier to understand if I had given greater emphasis to the way the linear model (1.1) would be handled if the random effects were not to be estimated. The linear model could be rewritten as

\[ y = X\beta + \varepsilon, \]

where \( \varepsilon = Zu + e \). Now \( \text{Var}(\varepsilon) = (ZGZ^T + R)^2 \) and it is convenient to denote \( ZGZ^T + R \) by \( V \). The generalized least-squares estimate

\[ \hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}y \]

is the same as the BLUP estimate as explained in Section 5.1.

As Harville and Thompson indicated, BLUP is often explained using a predictive formulation. Henderson frequently used such a formulation. (e.g., Henderson, 1973). Goldberger (1962) also used a predictive formulation. I find my presentation simpler, but I recommend that readers consider the alternative to see which they find easier to comprehend.

As pointed out by Spall, I did not clarify the distinction between smoothers and filters in my paper. His statement “it is a Kalman smoother... that produces the BLUP estimate of \( u \) based on data \( y \)” might leave readers thinking that the Kalman filter is not BLUP. In fact, the Kalman filter is the BLUP estimate of \( u \) based on the data up to time \( t, y_t \).

NOMENCLATURE

One of the major barriers to discussion in this area is the variety of nomenclature.

- I have used the term BLUP where many other people would use the term parametric empirical Bayes.
- I refer to random effects within mixed models whereas Steffey and Kass refer to unit-specific parameters within conditionally independent hierarchical models.
- Much terminology is application specific.

I do not wholeheartedly support the term BLUP because it includes the idea of predicting, and I do not believe that estimates of random effects are predictors for any greater fraction of their usage than estimates of fixed effects are predictors.

In ore reserve estimation I find it silly to speak of predicting something that happened millions of years ago. In time series, it is common to differentiate between smoothing, filtering and prediction. BLUP can be used for all three—which suggests that it is not merely prediction.

In the absence of general agreement about terminology, I would appeal for greater tolerance of other people’s terminology.

COMPUTATIONAL ISSUES

As Speed hinted at in his discussion, when I first started working on the paper I was involved in the task of designing a computating strategy for estimating the genetic merits of dairy cattle using BLUP. (My first draft of the paper was dated February 22, 1982.)

Up to that time, BLUP for large numbers of sires had been done using several different models, but BLUP for models requiring the solution of sets of simultaneous equations with equations corresponding to both male and female animals (often referred to as animal models) had only been used for small number of animals. Henderson (1975b) had proposed the model for use within single herds. The Australian Dairy Herd Improvement Scheme accepted my opinion that an animal model was computationally practical for large numbers of animals and has been using it for several years. Details of the computing strategy are given in Robinson (1986). See also Jones and Goddard (1990). A nonessential development was a method for solving the sets of up to one million simultaneous linear equations which is described in Robinson (1988).

Many other genetic evaluation schemes with large
data sets now also use animal models rather than sire-only models.

**ILL-POSED INVERSE PROBLEMS**

I agree with most of Campbell's remarks about BLUP and ill-posed inverse problems. In thinking about the distinction between uncertainty (which she refers to as the epistemological or Bayesian interpretation of probability) and variation (which she refers to as the ontological or classical interpretation of probability), I would include as variation the variation between the true images that have been or are likely to be looked at and the variation between patterns of grade in mineral deposits. The characteristic that I look for in deciding whether probability is being used to describe uncertainty or variation is that when probability is used to describe variation it should be possible to estimate the probability distribution or to test hypotheses about the probability distribution using available data.

**ILL-POSED THEORY**

While thinking about things that are ill posed, I would like to take this opportunity to indicate my disrespect for a type of theoretical work that is seldom helpful for solving real problems. The Cramér-Lévy theorem, referred to by Spall, states that if both X and Y are precisely normally distributed if both X and Y are precisely normally distributed. I regard the Cramér-Lévy theorem as an example of ill-posed theory because, although its conclusion does follow from its premises, a small departure from its premises allows a large departure from its conclusions. More specifically, it is not true that X + Y can only be approximately normally distributed if both X and Y are approximately normally distributed, as is easy to see for the case of X and Y having smooth, unimodal distributions of opposite skewness.

Spall obviously realizes that it is desirable to supplement the Cramér-Lévy theorem's conclusions. Regrettably, he only quotes the statistical significance of the departure from normality of the Kalman filter errors for his simulation experiments, not the extent of the departure.

**MOST LIKELY UNOBSERVABLES**

Thompson asked whether the method of most likely unobservables can be used to construct confidence intervals for estimates of random effects. Assuming a multivariate normal distribution as in Section 4.1, the ratio of the density of the unobservables u and e for arbitrary β and u to the density given β = ȳ and u = ȳ is \(\exp\{-Q/(2a^2)\}\), where Q is the quadratic form

\[
\begin{pmatrix}
     y - X\bar{\beta} - Zu
\end{pmatrix}
\begin{pmatrix}
    G & 0 \\
    0 & R
\end{pmatrix}
\begin{pmatrix}
     y - X\bar{\beta} - Zu
\end{pmatrix}^{-1}
\begin{pmatrix}
    \tilde{u}
\end{pmatrix}
\begin{pmatrix}
    G & 0 \\
    0 & R
\end{pmatrix}
\begin{pmatrix}
     y - X\bar{\beta} - Z\tilde{u}
\end{pmatrix}.
\]

Using equation (1.2), this can be shown to be equal to

\[
\begin{pmatrix}
    \beta - \bar{\beta} \\
    u - \tilde{u}
\end{pmatrix}
\begin{pmatrix}
    X^TV^{-1}X & X^TV^{-1}Z \\
    Z^TV^{-1}X & Z^TV^{-1}Z + G^{-1}
\end{pmatrix}
\begin{pmatrix}
    \beta - \bar{\beta} \\
    u - \tilde{u}
\end{pmatrix}.
\]

Hence using the method of most likely unobservables to construct confidence intervals is equivalent to assuming that the estimation errors have a multivariate normal distribution with the usual variance-covariance matrix which was given just below equation (1.2).

The meaning of Thompson's last question is not completely clear to me. There appear to be two likelihoods for β.

1. For the linear model \(y = X\beta + \epsilon\) the variance-covariance matrix of estimation errors is

\[
E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] = (X^TV^{-1}X)^{-1}\sigma^2
\]

and the likelihood is a multivariate normal distribution with mean \(\beta = \hat{\beta}\) and this variance-covariance matrix.

2. The method of most likely unobservables gives a likelihood for \(\beta\) and \(u\) that is a multivariate normal distribution with mean \(\beta = \hat{\beta}\), \(u = \tilde{u}\) and variance

\[
\begin{pmatrix}
    X^TV^{-1}X & X^TV^{-1}Z \\
    Z^TV^{-1}X & Z^TV^{-1}Z + G^{-1}
\end{pmatrix}^{-1}\sigma^2.
\]

These two likelihoods give the same conclusions about fixed effects, so choosing between them is not an issue. To see that they give the same conclusions, remember that the likelihood arising from the method of most likely unobservables is \(\exp\{-Q/(2\sigma^2)\}\) times the maximum at \(\beta = \hat{\beta}, u = \tilde{u}\). The derivative of the quadratic form with respect to \(u\) is

\[
2Z^TV^{-1}X(\beta - \hat{\beta}) + 2(Z^TV^{-1}Z + G^{-1})(u - \tilde{u}).
\]

The minimum of \(Q\) over \(u\) is at

\[
(u - \tilde{u}) = -(Z^TV^{-1}Z + G^{-1})^{-1}Z^TV^{-1}X(\beta - \hat{\beta})
\]

and is

\[
(\beta - \hat{\beta})^TX^TVGZ + R)^{-1}X(\beta - \hat{\beta}).
\]
Thus the maximum over \( u \) of the likelihood of the unobservables differs only by a constant factor from the likelihood for the linear model \( y = X\beta + \epsilon \).

In answer to Speed’s question in his Section 8, I believe that the coverage of the confidence intervals is the usual sort, provided that they are interpreted highlighting the probability distribution of the random effects that might have occurred.

**PRIOR INFORMATION AND ANIMAL BREEDING**

Section 3 of the discussion by Steffey and Kass seems to me to require some specific comments.

I agree that considerable knowledge about herd and sire effects is available. However, as I stated in Section 7.3 of my paper, I would prefer to treat herd-year-season effects as fixed rather than random (or, from a Bayesian perspective, to put uniform priors on them) because I am worried about the potential biases in the information contained in between herd-year-season comparisons.

The *prior information*, which Gianola and Fernando (1986) say should be used to preclude *ridiculous estimates of heritability* and other anomalies, can be put into the form of restrictions on the parameter spaces. Readers might be misled if they thought that this was a type of information that non-Bayesian statisticians would be unwilling to use.

The third paragraph is inaccurate. In Section 4.2 of my paper I find that BLUP estimates are exactly (not approximately) Bayes estimates with uniform improper prior on \( \beta \) and a point distribution (not a uniform distribution) on \( \theta \). Example 2 from Kass and Steffey (1989) illustrates that using flat or informative priors on the random effect \( (u \) in my notation) can make a substantial difference. I agree with this, but I do not agree with the implication that users of BLUP should be concerned about having implicitly used a uniform prior for \( \beta \).

**CLOSURE**

Before receiving copies of the discussion, I had wondered whether some discussion might be as intemperate as the comments by O. Kempthorne on a paper by D. V. Lindley at the Waterloo Symposium. The comments reported in Godambe and Sprott (1971, page 452) included the following:

... a former colleague, D. L. Harris, prepared a manuscript entitled “Estimation of Random Variables” in 1963. The title “bugged” some people, and the manuscript was rejected by *Biometrics* in all its wisdom. The argumentation was Bayesian, of course. Perhaps a manuscript with such a title will receive a slightly more open reception nowadays.

Kemthorne could surely not object to the openness of my paper’s reception. Having met Dewey Harris and seen the revised version of his paper (“Estimation of normally distributed random elements of certain statistical models,” Journal Paper No. J-4576, Iowa Agricultural and Home Economics Experiment Station, Ames, Project No. 1505, supported by National Science Foundation Grant G-18093), I see no evidence of lack of openness in 1963 either.

In fact, the complimentary and constructive tone of the discussion concerns me because readers might get the impression that there was general agreement with the ideas expressed in my paper. I doubt that this is the case. Two things which I believe readers should regard as controversial but which have not been raised in discussion are the following:

- In Section 5.8 I imply that much work on ranking and selection has been misguided.
- In Section 5.4 I ignore most of the methods used for estimating variance parameters.

**ADDITIONAL REFERENCES**


THE ESTIMATION OF RANDOM EFFECTS


