Tatra Mt. Math. Publ. **39** (2008), 1–15

TATRA MOUNTAINS *Mathematical Publications*

ACCOUNTING FOR THE ESTIMATION OF VARIANCES AND COVARIANCES IN PREDICTION UNDER A GENERAL LINEAR MODEL: AN OVERVIEW

David A. Harville

ABSTRACT. The problem considered is essentially that of predicting a linear combination of the fixed and/or random effects of a linear mixed-effects model. Applications are widespread; they include small-area estimation, the estimation (or prediction) of breeding values, the estimation of treatment contrasts (from the results of a comparative experiment), and the analysis of longitudinal data. The best linear unbiased predictor (BLUP) depends on functions of variance components and/or other such parameters. In practice, the values of these functions are typically unknown, and resort is made to the predictor (the so-called empirical BLUP) obtained from the BLUP by replacing the "true" values of the functions with even translation-invariant estimators (such as the REML estimators). This paper provides an overview of various results on the empirical BLUP (and includes a few extensions). The focus is on the mean squared error (MSE) of the empirical BLUP and on the approximation and estimation of the MSE. Some attention is given to prediction intervals.

1. Introduction

The estimation of a treatment contrast or a small-area mean or the estimation or prediction of an animal's breeding value is often based on a mixed-effects linear model in which

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1.1}$$

where $\mathbf{y}$ is an $N \times 1$ observable random vector, $\mathbf{X}$ is an $N \times P$ known matrix of rank P^*, $\boldsymbol{\beta}$ is a $P \times 1$ vector of unknown parameters (called fixed effects), $\mathbf{Z}$ is an $N \times K$ known matrix, and $\mathbf{u}$ and $\mathbf{e}$ are unobservable random column vectors (of

2000 Mathematics Subject Classification: Primary 62J07, 62J10, 62F10.
Keywords: mixed linear model, BLUP, BLUE, empirical predictor, REML, mean squared error, two-stage estimation, prediction interval.

random effects and "errors") with $\mathrm{E}(\mathbf{u}) = \mathbf{0}$, $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, and $\mathrm{cov}(\mathbf{u}, \mathbf{e}) = \mathbf{0}$, with $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{R}$ for some strictly positive parameter σ and some known positive definite matrix $\mathbf{R}$, and with $\mathrm{var}(\mathbf{u}) = \sigma^2\mathbf{D}$ for some nonnegative definite matrix $\mathbf{D}$ that is functionally dependent on one or more unknown parameters. Typically, each of the quantities of interest is expressible in the form of a linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta} + \boldsymbol{\delta}'\mathbf{u}$ of the model's fixed and/or random effects.

It is instructive to regard the problem of predicting a linear combination of the fixed and/or random effects of model (1.1) as a special case 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 $\mathbf{y}$. In the general prediction problem, it is assumed that $\mathrm{E}(w) = \boldsymbol{\lambda}'\boldsymbol{\beta}$, $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{var}(w) = v_w$, $\mathrm{cov}(\mathbf{y}, w) = \mathbf{v}_{yw}$, and $\mathrm{var}(\mathbf{y}) = \mathbf{V}_y$. Here, v_w and the elements of $\mathbf{v}_{yw}$ and $\mathbf{V}_y$ are specified functions of an unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_C)'$, whose value is restricted to a known set Ω, and $\mathbf{V}_y$ is assumed to be nonsingular (for all $\boldsymbol{\theta} \in \Omega$). At times, $v_w(\boldsymbol{\theta})$, $\mathbf{v}_{yw}(\boldsymbol{\theta})$, and $\mathbf{V}_y(\boldsymbol{\theta})$ are written for v_w, $\mathbf{v}_{yw}$, and $\mathbf{V}_y$, respectively. It is assumed that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable or, equivalently, that $\boldsymbol{\lambda} = \mathbf{X}'\mathbf{k}$ for some vector $\mathbf{k}$.

Let $t(\mathbf{y})$ represent a (point) predictor of w. The difference $t(\mathbf{y}) - w$ is called the *prediction error*, and $\mathrm{E}\{[t(\mathbf{y}) - w]^2\}$ is termed the mean squared error (MSE) of $t(\mathbf{y})$. (Unless otherwise indicated, expectations are with respect to the joint distribution of w and $\mathbf{y}$.) The predictor $t(\mathbf{y})$ is said to be *unbiased* if $\mathrm{E}[t(\mathbf{y}) - w] = 0$—refer, for example, to Bibby and Toutenburg (1977), Goldberger (1962), and Henderson (1963).

If the joint distribution of w and $\mathbf{y}$ were known, then the conditional mean $\mathrm{E}(w|\,\mathbf{y})$ could serve as a predictor of w. It would have minimum MSE among all predictors of w and would be unbiased, as is well-known and as is easily verified.

If $\boldsymbol{\lambda}'\boldsymbol{\beta}$, $\mathbf{X}\boldsymbol{\beta}$, $\mathbf{v}_{yw}$, and $\mathbf{V}_y$ were known, then a possible predictor would be

$$\boldsymbol{\lambda}'\boldsymbol{\beta} + \mathbf{v}'_{yw}\mathbf{V}_y^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \tau + \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{y}, \tag{1.2}$$

where $\tau = (\boldsymbol{\lambda}' - \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{X})\boldsymbol{\beta}$. It would have minimum MSE among all linear predictors of w (e.g., Rao 1973, sec. 4a.11) and would be unbiased. It and its MSE coincide with what Hartigan (1969) refers to as the linear expectation and linear variance of w given $\mathbf{y}$.

If $\mathbf{v}_{yw}$ and $\mathbf{V}_y$ (but not $\boldsymbol{\lambda}'\boldsymbol{\beta}$ and $\mathbf{X}\boldsymbol{\beta}$) were known (or, more generally, were known up to a constant of proportionality), then a possible predictor, say $\tilde{w}$, could be obtained from the quantity (1.2) by replacing τ by what would be the best (minimum MSE) linear unbiased estimator (BLUE) of τ, which is

$$\tilde{\tau} = (\boldsymbol{\lambda}' - \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{y}.$$

By definition,

$$\tilde{w} = \tilde{\tau} + \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{y} = \mathbf{h}'\mathbf{y},$$

where $\mathbf{h}' = \left[(\boldsymbol{\lambda}' - \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-}\mathbf{X}' + \mathbf{v}'_{yw}\right]\mathbf{V}_y^{-1}$. The quantity $\tilde{w}$ would (if $\mathbf{v}_{yw}$ and $\mathbf{V}_y$ were known) be the best (minimum MSE) linear unbiased predictor (BLUP) of w (Goldberger 1962; Henderson 1963)—refer, for example, to Harville (1991, sec. 1) for a proof.

Let us write $\tilde{w}(\boldsymbol{\theta})$ or $\tilde{w}(\mathbf{y};\boldsymbol{\theta})$ for $\tilde{w}$. It is common practice to obtain a predictor of w by replacing $\boldsymbol{\theta}$ in $\tilde{w}(\boldsymbol{\theta})$ with an estimator of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$. Let us suppose that $\hat{\boldsymbol{\theta}}$ is an even translation-invariant estimator such as Patterson and Thompson's (1971) restricted maximum likelihood (REML) estimator. [A possibly vector-valued function $\mathbf{f}(\mathbf{y})$ of $\mathbf{y}$ is said to be translation-invariant if $\mathbf{f}(\mathbf{y} + \mathbf{X}\mathbf{b}) = \mathbf{f}(\mathbf{y})$ for every vector $\mathbf{b}$ (and every value of $\mathbf{y}$); a possibly vector-valued function $\mathbf{g}(\mathbf{t})$ of a vector $\mathbf{t}$ is said to be even if $\mathbf{g}(-\mathbf{t}) = \mathbf{g}(\mathbf{t})$ for every $\mathbf{t}$ in the domain of $\mathbf{g}$ (and is said to be odd if $\mathbf{g}(-\mathbf{t}) = -\mathbf{g}(\mathbf{t})$ for every such $\mathbf{t}$).]

Let $\mathbf{z} = \mathbf{L}'\mathbf{y}$, where $\mathbf{L}$ is an $N \times (N - P^*)$ matrix such that $\mathbf{L}'\mathbf{X} = \mathbf{0}$ and $\operatorname{rank}(\mathbf{L}) = N - P^*$. The vector $\mathbf{z}$ satisfies the definition of a maximal invariant (with respect to transformations of the general form $T(\mathbf{y}) = \mathbf{y} + \mathbf{X}\mathbf{b}$), and hence a possibly vector-valued function $\mathbf{f}(\mathbf{y})$ is translation-invariant if and only if it depends on $\mathbf{y}$ only through the value of $\mathbf{z}$ (e.g., Lehmann 1986, sec. 6.2). Accordingly, $\hat{\boldsymbol{\theta}}$ is expressible as an (even) function, say $\hat{\boldsymbol{\theta}}(\mathbf{z})$, of $\mathbf{z}$. Substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in $\tilde{w}(\boldsymbol{\theta}) = \tilde{w}(\mathbf{y};\boldsymbol{\theta})$ gives the so-called empirical BLUP $\hat{w}_{\text{EBLUP}} = \tilde{w}(\hat{\boldsymbol{\theta}}) = \tilde{w}(\mathbf{y};\hat{\boldsymbol{\theta}})$.

The empirical BLUP can be regarded as a special case of a predictor of the form

$$\int_{\Omega} \tilde{w}(\mathbf{y};\boldsymbol{\omega})\ d\mathcal{P}(\boldsymbol{\omega};\mathbf{z}), \tag{1.3}$$

where $\mathcal{P}(\cdot;\mathbf{z})$ is a probability distribution defined on Ω that may depend on $\mathbf{z}$ and is such that $\mathcal{P}(\cdot;-\mathbf{z}) = \mathcal{P}(\cdot;\mathbf{z})$. When $\mathcal{P}(\cdot;\mathbf{z})$ is a degenerate probability distribution that assigns probability one to the single point $\hat{\boldsymbol{\theta}}(\mathbf{z})$, the "integrated BLUP" (1.3) reduces to the empirical BLUP $\hat{w}_{\text{EBLUP}}$. Nondegenerate choices for $\mathcal{P}(\cdot;\mathbf{z})$ include posterior distributions derived from various proper or improper prior distributions via a Bayesian analysis of the observable random vector $\mathbf{z}$.

The objective in what follows is essentially that of providing an overview of various results on the prediction of w; many of these results are specific to the empirical BLUP. The emphasis is on the expected value of the predictor, on the MSE of prediction (and on its approximation and estimation), and on pivotal quantities (for obtaining prediction intervals).

2. Prediction error: Symmetry of distribution

Let $\hat{w} = \hat{w}(\mathbf{y})$ represent a predictor of w. Suppose that $\hat{w}$ is location-equivariant in the sense that $\hat{w}(\mathbf{y}+\mathbf{X}\mathbf{b}) = \hat{w}(\mathbf{y})+\boldsymbol{\lambda}'\mathbf{b}$ for every $P\times 1$ vector $\mathbf{b}$ and every value of $\mathbf{y}$. Suppose further that $\hat{w}(\mathbf{y})$ is an odd function of $\mathbf{y}$. And observe that the empirical BLUP $\hat{w}_{\text{EBLUP}}$ is an odd location-equivariant predictor and so is the integrated BLUP (1.3).

Let us consider the distribution of the prediction error $\hat{w}-w$. In doing so, we will frequently need to make reference to one or the other of the following two alternative assumptions about the joint distribution of $\mathbf{y}$ and w:

N. The joint distribution of $\mathbf{y}$ and w is multivariate normal (MVN).

S. The joint distribution of $\mathbf{y}$ and w is symmetric [in the sense that the joint distribution of $-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ and $-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})$ is the same as that of $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ and $w-\boldsymbol{\lambda}'\boldsymbol{\beta}$.

Kackar and Harville (1984) observed that, under Assumption S, the distribution of $\hat{w}-w$ is symmetric about 0. They noted that this result follows from Wolfe's (1973) Theorem 2.1—in applying Wolfe's Theorem 2.1, take $\mathbf{Z} = \big[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})', w-\boldsymbol{\lambda}'\boldsymbol{\beta}\big]'$, $g(\mathbf{Z}) = -\mathbf{Z}$, and $U(\mathbf{Z}) = \hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})$, and observe that $\hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-(w-\boldsymbol{\lambda}'\boldsymbol{\beta}) = \hat{w}(\mathbf{y})-w$. Alternatively, the symmetry of the distribution of $\hat{w}-w$ can be verified directly: it suffices to observe that $\hat{w}(\mathbf{y})-w = \hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})$ and that

$$\begin{aligned}\hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-(w-\boldsymbol{\lambda}'\boldsymbol{\beta}) &\sim \hat{w}\big[-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big]-\big[-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})\big]\\ &= -\hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\big[-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})\big]\\ &= -\big[\hat{w}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-(w-\boldsymbol{\lambda}'\boldsymbol{\beta})\big].\end{aligned}$$

A stronger result is obtainable. Take $\mathbf{s}$ to be any vector of even functions of $\mathbf{z}$—$\hat{\boldsymbol{\theta}}$ is one such vector. Then, under Assumption S, the joint distribution of $-(\hat{w}-w)$ and $\mathbf{s}$ is the same as that of $\hat{w}-w$ and $\mathbf{s}$, and hence the conditional distribution of $\hat{w}-w$ given $\mathbf{s}$ is symmetric about 0. This result was presented (for the special case where $\mathbf{s}$ is $\hat{\boldsymbol{\theta}}$) by Kackar and Harville (1984) and, as pointed out by them, it can be deduced from Wolfe's (1973) Theorem 2.2. In the special case where $v_w = 0$ (and hence where $w = \boldsymbol{\lambda}'\boldsymbol{\beta}$ with probability 1), it is essentially the same as a result given by Seely and Hogg (1982, sec. 2).

The result that (under Assumption S) the distribution of $\hat{w}-w$ is symmetric about 0 extended a result obtained earlier by Kackar and Harville (1981) in the special case where $\hat{w}$ is the empirical BLUP $\hat{w}_{\text{EBLUP}}$. The latter result can in turn be regarded as an extension of results obtained even earlier in the further special case where $v_w = 0$ (i.e., where $w = \boldsymbol{\lambda}'\boldsymbol{\beta}$ with probability 1). The first such result appears to have been that of Kakwani (1967); it was followed by the results of Fuller and Battese (1973), Guilkey and Schmidt (1973), and Hartley and Jayatillake (1973) (among others).

That $\hat{w} - w$ is distributed symmetrically about 0 implies that $\hat{w}$ is unbiased provided the expected value of $\hat{w}$ exists. Jiang (1999, 2000) established the existence of the expected value of $\hat{w}_{\text{EBLUP}}$ for a class of linear mixed-effects models with two variance components and subsequently for a much broader class of linear mixed-effects models. Previously, existence had been established in some relatively simple special cases (in which $v_w = 0$) by Weiler and Culpin (1970) and Fuller and Battese (1973).

3. Mean squared error of prediction

The prediction error of the odd location-equivariant predictor $\hat{w}(\mathbf{y})$ can be decomposed in accordance with the following identity:

$$\hat{w}(\mathbf{y}) - w = \big[\tilde{w}(\mathbf{y};\boldsymbol{\theta}) - w\big] + \big[\hat{w}(\mathbf{y}) - \tilde{w}(\mathbf{y};\boldsymbol{\theta})\big]. \qquad (3.1)$$

The first component $\tilde{w}(\mathbf{y};\boldsymbol{\theta}) - w$ represents the prediction error that would be incurred if $\boldsymbol{\theta}$ were known and the BLUP were used to predict w. The second component $\hat{w}(\mathbf{y}) - \tilde{w}(\mathbf{y};\boldsymbol{\theta})$ represents the effect on the prediction error of $\boldsymbol{\theta}$ being unknown. In the special case of the empirical BLUP $\hat{w}_{\text{EBLUP}}$, the second component in the representation (3.1) is expressible in the form

$$\tilde{w}(\mathbf{y};\hat{\boldsymbol{\theta}}) - \tilde{w}(\mathbf{y};\boldsymbol{\theta}) = \Big\{\ell\big[\hat{\boldsymbol{\theta}}(\mathbf{z});\boldsymbol{\theta}\big]\Big\}'\mathbf{z}\,, \qquad (3.2)$$

where $\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})$ is an $(N-P^*)\times 1$ vector whose elements are functionally dependent on $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$—refer, for example, to Harville (1985) or Harville and Jeske (1992) for an explicit expression for $\ell(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})$.

The second component $\hat{w}(\mathbf{y}) - \tilde{w}(\mathbf{y};\boldsymbol{\theta})$ of the prediction error depends on $\mathbf{y}$ only through the value of $\mathbf{z}$ and is an odd function of $\mathbf{z}$. That this is so in the special case of the empirical BLUP $\hat{w}_{\text{EBLUP}}$ is evident from expression (3.2). That it is so in general is evident upon observing that $\hat{w}(\mathbf{y}) - \tilde{w}(\mathbf{y};\boldsymbol{\theta})$ is an odd translation-invariant function of $\mathbf{y}$ [and upon recalling that $\mathbf{z}$ is a maximal invariant for transformations of $\mathbf{y}$ of the general form $T(\mathbf{y}) = \mathbf{y} + \mathbf{X}\mathbf{b}$].

Under Assumption S, the joint distribution of the two components of the prediction error is symmetric about the (2-dimensional) null vector both unconditionally and conditionally on the vector $\mathbf{s}$ of even functions of $\mathbf{z}$, as is verifiable by, for example, making use of Wolfe's (1973) Theorems 2.1 and 2.2. Regardless of the form of the joint distribution of w and $\mathbf{y}$, the first component of the prediction error is such that its (unconditional) expected value is 0 and its (unconditional) variance is

$$\begin{aligned}&\operatorname{var}\bigl[\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w\bigr]\\&\quad=v_w-\mathbf{v}_{yw}'\mathbf{V}_y^{-1}\mathbf{v}_{yw}+\bigl(\boldsymbol{\lambda}-\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{v}_{yw}\bigr)'\bigl(\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X}\bigr)^{-}\bigl(\boldsymbol{\lambda}-\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{v}_{yw}\bigr)\end{aligned}\tag{3.3}$$

and is such that

$$\operatorname{cov}\bigl[\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w,\,\mathbf{z}\bigr]=\mathbf{h}'\mathbf{V}_y\mathbf{L}-\mathbf{v}_{yw}'\mathbf{L}=\mathbf{0}.\tag{3.4}$$

Now, suppose that we adopt Assumption N. Then, the distribution of the first component $\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w$ is MVN [with mean 0 and variance (3.3)]. Moreover, it follows from result (3.4) that $\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w$ is statistically independent of $\mathbf{z}$. Thus, the distribution of $\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w$ conditional on the vector $\mathbf{s}$ (of even functions of $\mathbf{z}$) is the same as its unconditional distribution, and [recalling that $\hat{w}(\mathbf{y})-\tilde{w}(\mathbf{y}:\boldsymbol{\theta})$ depends on $\mathbf{y}$ only through the value of $\mathbf{z}$] the two components of the prediction error are distributed independently, both unconditionally and conditionally on $\mathbf{s}$. These results extend to an arbitrary odd location-equivariant predictor results obtained by, for example, Kackar and Harville (1984) and Harville (1985) for the empirical BLUP $\hat{w}_{\text{EBLUP}}$ and for predictors of the form (1.3).

The conditional (given $\mathbf{s}$) and unconditional expected values of the second component of the prediction error equal 0 [provided the expected value of $\hat{w}(\mathbf{y})$ exists]. The MSE of $\hat{w}(\mathbf{y})$ is $m(\boldsymbol{\theta})=\mathrm{E}\bigl\{[\hat{w}(\mathbf{y})-w]^2\bigr\}$. Corresponding to decomposition (3.1) of the prediction error, we have the following decomposition of the MSE:

$$m(\boldsymbol{\theta})=m_1(\boldsymbol{\theta})+m_2(\boldsymbol{\theta}),\tag{3.5}$$

where $m_1(\boldsymbol{\theta})=\operatorname{var}\bigl[\tilde{w}(\mathbf{y};\boldsymbol{\theta})-w\bigr]$ and $m_2(\boldsymbol{\theta})=\mathrm{E}\bigl\{[\hat{w}(\mathbf{y})-\tilde{w}(\mathbf{y};\boldsymbol{\theta})]^2\bigr\}$. Jiang (2000) showed that, in the special case of the empirical BLUP, $m(\boldsymbol{\theta})$ is finite [and hence $m_2(\boldsymbol{\theta})$ is finite] for a broad class of linear mixed-effects models.

As an immediate consequence of identity (3.5), we have that $m(\boldsymbol{\theta})\geq m_1(\boldsymbol{\theta})$, in agreement with earlier results by, for example, Khatri and Shah (1981), Kackar and Harville (1984), Eaton (1985), and Harville (1985). Note that [assuming the existence of the expected value of $\hat{w}(\mathbf{y})$] $m_2(\boldsymbol{\theta})=\operatorname{var}\bigl[\hat{w}(\mathbf{y})-\tilde{w}(\mathbf{y};\boldsymbol{\theta})\bigr]$. Except for very simple special cases, $m_2(\boldsymbol{\theta})$ is not expressible in "closed form." Corresponding to the decomposition (3.5) for the unconditional MSE is the decomposition for the conditional (on $\mathbf{s}$) MSE given by

$$m^*(\boldsymbol{\theta};\mathbf{s})=m_1(\boldsymbol{\theta})+m_2^*(\boldsymbol{\theta};\mathbf{s}),\tag{3.6}$$

where

$$m^*(\boldsymbol{\theta};\mathbf{s})=\mathrm{E}\bigl\{[\hat{w}(\mathbf{y})-w]^2\,|\,\mathbf{s}\bigr\}$$

and where

$$m_2^*(\boldsymbol{\theta};\mathbf{s})\ \text{is}\ \mathrm{E}\bigl\{[\hat{w}(\mathbf{y})-\tilde{w}(\mathbf{y};\boldsymbol{\theta})]^2\,|\,\mathbf{s}\bigr\}\ \text{or (equivalently)}\ \operatorname{var}\bigl[\hat{w}(\mathbf{y})-\tilde{w}(\mathbf{y};\boldsymbol{\theta})\,|\,\mathbf{s}\bigr].$$

Subsequently, let us suppose that the vector $\mathbf{s}$ of even functions of $\mathbf{z}$ is such that $\hat{\boldsymbol{\theta}}$ depends on $\mathbf{z}$ only through the value of $\mathbf{s}$. In the special case of the empirical BLUP $\hat{w}_{\mathrm{EBLUP}}$, we have that

$$m_2^*(\boldsymbol{\theta}; \mathbf{s}) = \bigl[\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr]' \operatorname{var}(\mathbf{z} \,|\, \mathbf{s}) \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) \tag{3.7}$$

and

$$m_2(\boldsymbol{\theta}) = \mathrm{E}\bigl\{\bigl[\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr]' \operatorname{var}(\mathbf{z} \,|\, \mathbf{s}) \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr\} \tag{3.8}$$

(Harville and Jeske 1992, sec. 3). Harville and Jeske (1992, sec. 5) considered expressions (3.7) and (3.8) as applied to linear mixed-effects models with two variance components, taking $\mathbf{s}$ to be a minimal sufficient statistic for $\mathbf{z}$. They obtained $\operatorname{var}(\mathbf{z} \,|\, \mathbf{s})$ in "closed form" and outlined a procedure for evaluating the expected value in expression (3.8) by numerical integration or by Monte Carlo methods.

Among the various versions of formulas (3.7) and (3.8) are those obtained by taking $\mathbf{s}$ to be $\hat{\boldsymbol{\theta}}$. It is informative to consider those versions under the following two conditions:

A. $\mathrm{E}\bigl[v_w(\hat{\boldsymbol{\theta}})\bigr] = v_w$, $\mathrm{E}\bigl[\mathbf{v}_{yw}(\hat{\boldsymbol{\theta}})\bigr] = \mathbf{v}_{yw}$, and $\mathrm{E}\bigl[\mathbf{V}_y(\hat{\boldsymbol{\theta}})\bigr] = \mathbf{V}_y$.

B. The vector $\hat{\boldsymbol{\theta}}$ is a complete sufficient statistic (for $\mathbf{z}$).

Let $\mathbf{V}_z(\boldsymbol{\theta}) = \operatorname{var}(\mathbf{z}) = \mathbf{L}'\mathbf{V}_y(\boldsymbol{\theta})\mathbf{L}$. Harville and Jeske (1992, sec. 3.2) observed that, under Conditions A and B,

$$\operatorname{var}(\mathbf{z} \,|\, \hat{\boldsymbol{\theta}}) = \mathbf{V}_z(\hat{\boldsymbol{\theta}}) \quad \text{with probability 1.}$$

And, based on that result, they suggested the following approximations for $m_2^*(\boldsymbol{\theta}; \mathbf{s})$ and $m_2(\boldsymbol{\theta})$ (when $\mathbf{s}$ is taken to be $\hat{\boldsymbol{\theta}}$):

$$m_2^*(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) \doteq \bigl[\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr]' \mathbf{V}_z(\hat{\boldsymbol{\theta}}) \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) \tag{3.9}$$

and

$$m_2(\boldsymbol{\theta}) \doteq \mathrm{E}\bigl\{\bigl[\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr]' \mathbf{V}_z(\hat{\boldsymbol{\theta}}) \ell(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\bigr\}. \tag{3.10}$$

Conditions A and B tend to be at least somewhat hypothetical (especially B). Results derived under such conditions are of interest for much the same reasons that asymptotic results may be of interest; they suggest approximations and may be helpful in discerning the circumstances under which the approximations are likely to be satisfactory. Harville and Jeske carried out a small numerical study of one-way-classification random-effects models in which approximation (3.10) performed quite well.

Kackar and Harville (1984, sec. 2) proposed an approximation to the second component of the MSE of the empirical BLUP $\hat{w}_{\mathrm{EBLUP}}$. Letting $\mathbf{d}(\boldsymbol{\theta}) = \partial\tilde{w}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, their approximation is

$$m_2(\boldsymbol{\theta}) \doteq \operatorname{tr}\bigl[\mathbf{A}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\bigr], \tag{3.11}$$

where $\mathbf{A}(\boldsymbol{\theta}) = \mathrm{var}\big[\mathbf{d}(\boldsymbol{\theta})\big]$ and where $\mathbf{B}(\boldsymbol{\theta})$ is either the MSE matrix $\mathrm{E}\big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\big]$ or some approximation to the MSE matrix. They based their approximation on the observation that a second-order Taylor-series approximation for the square of the second component of the prediction error is as follows:

$$\big[\hat{w}_{\mathrm{EBLUP}} - \tilde{w}(\boldsymbol{\theta})\big]^2 = \big[\tilde{w}(\hat{\boldsymbol{\theta}}) - \tilde{w}(\boldsymbol{\theta})\big]^2 \doteq \Big\{\big[\mathbf{d}(\boldsymbol{\theta})\big]'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\Big\}^2 \tag{3.12}$$

$$= \mathrm{tr}\Big\{\mathbf{d}(\boldsymbol{\theta})\big[\mathbf{d}(\boldsymbol{\theta})\big]'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\Big\}. \tag{3.13}$$

Prasad and Rao (1990) considered the variation on approximation (3.11) that results from replacing

$$\mathbf{d}(\boldsymbol{\theta}) \text{ by the vector } \mathbf{d}_*(\boldsymbol{\theta}) = \partial\big\{[\mathbf{v}_{yw}(\boldsymbol{\theta})]'[\mathbf{V}_y(\boldsymbol{\theta})]^{-1}\mathbf{y}\big\}/\partial\boldsymbol{\theta}.$$

Thus, the Prasad-Rao approximation is

$$m_2(\boldsymbol{\theta}) \doteq \mathrm{tr}\big[\mathbf{A}_*(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big], \tag{3.14}$$

where $\mathbf{A}_*(\boldsymbol{\theta}) = \mathrm{var}\big[\mathbf{d}_*(\boldsymbol{\theta})\big]$. Their focus was on situations like those encountered in small-area estimation where the sensitivity of $\tilde{\tau}$ [which would be the BLUE of $\tau = (\boldsymbol{\lambda}' - \mathbf{v}_{yw}'\mathbf{V}_y^{-1}\mathbf{X})\boldsymbol{\beta}$ if $\boldsymbol{\theta}$ were known] to changes in the value of $\boldsymbol{\theta}$ tends to be slight relative to that of $\mathbf{v}_{yw}'\mathbf{V}_y^{-1}\mathbf{y}$. Prasad and Rao obtained some asymptotic results that establish the order of the error of approximation (3.14) [and show it to be the same as the order of the error of approximation (3.11)]; these results are specific to linear mixed-effects models (and to choices for the coefficient vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$) of the kind typically employed in small-area estimation—refer also to Singh, Stukel, and Pfeffermann (1998), Datta and Lahiri (2000), Rao (2003), Wang and Fuller (2003), and Das, Jiang, and Rao (2004).

Let us now specialize to the case where $\hat{\boldsymbol{\theta}}$ is the REML estimator. Denoting by $\boldsymbol{l}(\boldsymbol{\theta})$ the $C \times 1$ vector whose ith element is the partial derivative with respect to θ_i of the REML log-likelihood function, we have that $\boldsymbol{l}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ (provided $\hat{\boldsymbol{\theta}} \in \Omega$). Suppose that we take $\mathbf{B}(\boldsymbol{\theta})$ to be the large-sample variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ {so that $\big[-\mathbf{B}(\boldsymbol{\theta})\big]^{-1}$ is the expected value of the $C \times C$ matrix whose ijth element is the second-order partial derivative with respect to θ_i and θ_j of the REML log-likelihood function}. Then, upon substituting from the approximation $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \doteq \mathbf{B}(\boldsymbol{\theta})\boldsymbol{l}(\boldsymbol{\theta})$ into expression (3.12), we obtain the approximation

$$\big[\hat{w}_{\mathrm{EBLUP}} - \tilde{w}(\boldsymbol{\theta})\big]^2 \doteq \Big\{\big[\mathbf{d}(\boldsymbol{\theta})\big]'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{l}(\boldsymbol{\theta})\Big\}^2. \tag{3.15}$$

Approximation (3.15) suggests the following approximation for the second component of the MSE of the empirical BLUP $\hat{w}_{\text{EBLUP}}$:

$$m_2(\boldsymbol{\theta}) \doteq \mathrm{E}\left(\left\{\left[\mathbf{d}(\boldsymbol{\theta})\right]'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{l}(\boldsymbol{\theta})\right\}^2\right). \tag{3.16}$$

Das *et al.* (2004) obtained some asymptotic results that establish the order of the error of approximation (3.16) for various linear mixed-effects models (and choices of $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$). They further established that $\mathrm{tr}\left[\mathbf{A}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\right]$ and $\mathrm{tr}\left[\mathbf{A}_*(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\right]$ differ from $\mathrm{E}\left(\{[\mathbf{d}(\boldsymbol{\theta})]'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{l}(\boldsymbol{\theta})\}^2\right)$ only to an extent that is of the same order as the order of the error of approximation (3.16) and that, consequently, the errors of approximations (3.11) and (3.14) are of the same order as the error of approximation (3.16). Similar results were obtained by Datta and Lahiri (2000).

It is worth mentioning that the first component of the prediction error can be further decomposed as follows:

$$\tilde{w}(\mathbf{y};\boldsymbol{\theta}) - w = \left[\mathrm{E}(w\,|\,\mathbf{y}) - w\right] + \left[\tau + \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{y} - \mathrm{E}(w\,|\,\mathbf{y})\right] + (\tilde{\tau} - \tau). \tag{3.17}$$

Regardless of the form of the joint distribution of w and $\mathbf{y}$, the three components defined by identity (3.17) are uncorrelated, and each has an expected value of 0. These three components could be regarded as an "inherent component" [as in the error that would be incurred even if $\mathrm{E}(w\,|\,\mathbf{y})$ were known and were used to predict w], a "nonlinearity component", and an "unknown-means component" (a contribution to the error attributable to the estimation of τ). Corresponding to the decomposition (3.17) of $\tilde{w}(\mathbf{y};\boldsymbol{\theta}) - w$ is the following decomposition of $m_1(\boldsymbol{\theta})$:

$$m_1(\boldsymbol{\theta}) = m_{11} + m_{12} + m_{13}, \tag{3.18}$$

where

$$m_{11} = \mathrm{var}\left[\mathrm{E}(w\,|\,\mathbf{y}) - w\right] = \mathrm{E}\left[\mathrm{var}(w\,|\,\mathbf{y})\right],$$
$$m_{12} = v_w - \mathbf{v}'_{yw}\mathbf{V}_y^{-1}\mathbf{v}_{yw} - \mathrm{E}\left[\mathrm{var}(w\,|\,\mathbf{y})\right],$$

and

$$m_{13} = \left(\boldsymbol{\lambda} - \mathbf{X}'\mathbf{V}_y^{-1}\mathbf{v}_{yw}\right)'\left(\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X}\right)^-\left(\boldsymbol{\lambda} - \mathbf{X}'\mathbf{V}_y^{-1}\mathbf{v}_{yw}\right)$$

(Harville 1985).

4. Estimation of mean squared error

Let us consider the estimation of the MSE $m(\boldsymbol{\theta})$ of the predictor $\hat{w}(\mathbf{y})$ (under Assumption N). And, in doing so, let us restrict attention to the case where $\hat{w}(\mathbf{y})$ is the empirical BLUP $\hat{w}_{\text{EBLUP}} = \tilde{w}(\mathbf{y};\hat{\boldsymbol{\theta}})$.

It has been common practice to estimate $m(\boldsymbol{\theta})$ by $m_1(\hat{\boldsymbol{\theta}})$. This practice does not account for the contribution to $m(\boldsymbol{\theta})$ of $m_2(\boldsymbol{\theta})$ and, as a consequence, might be expected to result in the underestimation of $m(\boldsymbol{\theta})$. In fact, not only does $m_1(\hat{\boldsymbol{\theta}})$ tend to underestimate $m(\boldsymbol{\theta})$, but it tends to underestimate $m_1(\boldsymbol{\theta})$. In that regard, Harville and Jeske (1992) showed that if Condition A is satisfied, then

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] = m_1(\boldsymbol{\theta}) - \mathrm{E}\Big\{\big[\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\big]'\mathbf{V}_z(\hat{\boldsymbol{\theta}})\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\Big\}, \tag{4.1}$$

and that if Conditions A and B are both satisfied, then

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] = m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}). \tag{4.2}$$

Result (4.1) implies that $\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] \le m_1(\boldsymbol{\theta})$, which is an inequality given by Eaton (1985, sec. 8). Result (4.2) indicates that, under certain conditions, $m_1(\hat{\boldsymbol{\theta}})$ tends to underestimate $m_1(\boldsymbol{\theta})$ by exactly twice the amount that might have been anticipated.

Let $\mathbf{k}(\boldsymbol{\theta})$ represent the $C \times 1$ vector whose ith element is $k_i = \partial m_1(\boldsymbol{\theta})/\partial\theta_i$, and $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ the $C \times C$ matrix whose ijth element is $\lambda_{ij} = \partial^2 m_1(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$. Then, a second-order Taylor-series approximation to $m_1(\hat{\theta})$ is

$$m_1(\hat{\boldsymbol{\theta}}) \doteq m_1(\boldsymbol{\theta}) + (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\mathbf{k}(\boldsymbol{\theta}) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\boldsymbol{\Lambda}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}).$$

Upon taking expectations, we obtain the approximation

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] \doteq m_1(\boldsymbol{\theta}) + \big[\mathrm{E}(\hat{\boldsymbol{\theta}})-\boldsymbol{\theta}\big]'\mathbf{k}(\boldsymbol{\theta}) + \tfrac{1}{2}\,\mathrm{tr}\big[\boldsymbol{\Lambda}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big], \tag{4.3}$$

and, under the assumption that $\mathrm{E}(\hat{\boldsymbol{\theta}}) \doteq \boldsymbol{\theta}$, we obtain the further approximation

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] \doteq m_1(\boldsymbol{\theta}) + \tfrac{1}{2}\,\mathrm{tr}\big[\boldsymbol{\Lambda}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]. \tag{4.4}$$

It can be shown that

$$k_i = \begin{pmatrix}\mathbf{h}\\ -1\end{pmatrix}'\begin{pmatrix}\partial\mathbf{V}_y/\partial\theta_i & \partial\mathbf{v}_{yw}/\partial\theta_i\\ \partial\mathbf{v}_{yw}'/\partial\theta_i & \partial v_w/\partial\theta_i\end{pmatrix}\begin{pmatrix}\mathbf{h}\\ -1\end{pmatrix} \tag{4.5}$$

and, letting a_{ij} represent the ijth element of $\mathbf{A}(\boldsymbol{\theta})$, that

$$\lambda_{ij} = -2a_{ij} + \begin{pmatrix}\mathbf{h}\\ -1\end{pmatrix}'\begin{pmatrix}\partial^2\mathbf{V}_y/\partial\theta_i\partial\theta_j & \partial^2\mathbf{v}_{yw}/\partial\theta_i\partial\theta_j\\ \partial^2\mathbf{v}_{yw}'/\partial\theta_i\partial\theta_j & \partial^2 v_w/\partial\theta_i\partial\theta_j\end{pmatrix}\begin{pmatrix}\mathbf{h}\\ -1\end{pmatrix} \tag{4.6}$$

(Harville and Jeske 1992, sec. 4).

If $v_w(\boldsymbol{\theta})$, $\mathbf{v}_{yw}(\boldsymbol{\theta})$, and $\mathbf{V}_y(\boldsymbol{\theta})$ are linear in $\theta_1, \theta_2, \ldots, \theta_C$, then expression (4.6) reduces to $-2a_{ij}$, so that $\boldsymbol{\Lambda}(\boldsymbol{\theta}) = -2\mathbf{A}(\boldsymbol{\theta})$. Thus, if $v_w(\boldsymbol{\theta})$, $\mathbf{v}_{yw}(\boldsymbol{\theta})$, and $\mathbf{V}_y(\boldsymbol{\theta})$ are approximately linear in $\theta_1, \theta_2, \ldots, \theta_C$, we obtain, as an alternative to approximation (4.3), the "simplified" approximation

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] \doteq m_1(\boldsymbol{\theta}) + \big[\mathrm{E}(\hat{\boldsymbol{\theta}})-\boldsymbol{\theta}\big]'\mathbf{k}(\boldsymbol{\theta}) - \mathrm{tr}\big[\mathbf{A}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]; \tag{4.7}$$

and if in addition $\mathrm{E}(\hat{\boldsymbol{\theta}}) \doteq \boldsymbol{\theta}$ (in which case Condition A is approximately satisfied), we obtain, as an alternative to approximations (4.4) and (4.7), the even more simplified approximation

$$\mathrm{E}\big[m_1(\hat{\boldsymbol{\theta}})\big] \doteq m_1(\boldsymbol{\theta}) - \mathrm{tr}\big[\mathbf{A}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]. \tag{4.8}$$

Prasad and Rao (1990) proposed an estimator of $m(\boldsymbol{\theta})$ for use in various applications of a kind encountered in small-area estimation. The proposed estimator is

$$\hat{m}^*_{\mathrm{PR}} = m_1(\hat{\boldsymbol{\theta}}) + 2\,\mathrm{tr}\big[\mathbf{A}_*(\hat{\boldsymbol{\theta}})\mathbf{B}(\hat{\boldsymbol{\theta}})\big], \tag{4.9}$$

which can be regarded as a "simplified" version of the estimator

$$\hat{m}_{\mathrm{PR}} = m_1(\hat{\boldsymbol{\theta}}) + 2\,\mathrm{tr}\big[\mathbf{A}(\hat{\boldsymbol{\theta}})\mathbf{B}(\hat{\boldsymbol{\theta}})\big]. \tag{4.10}$$

The motivation for estimator (4.9) or (4.10) comes from results like (4.8) and (3.11) or (3.14) and the presumption that $\mathrm{tr}\big[\mathbf{A}_*(\hat{\boldsymbol{\theta}})\mathbf{B}(\hat{\boldsymbol{\theta}})\big]$ or $\mathrm{tr}\big[\mathbf{A}(\hat{\boldsymbol{\theta}})\mathbf{B}(\hat{\boldsymbol{\theta}})\big]$ is a "reasonable" estimator of $\mathrm{tr}\big[\mathbf{A}_*(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]$ or $\mathrm{tr}\big[\mathbf{A}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]$.

Some formal justification for the use of estimator (4.9) or (4.10) is provided by asymptotic results obtained (in the context of small-area estimation) by Prasad and Rao and subsequently by Datta and Lahiri (2000) and Das *et al.* (2004). Those results are for situations in which the variance components of the underlying linear mixed-effects model are estimated by methods such as REML or the method of fitting constants. They serve to justify the approximation $\mathrm{E}\big\{\mathrm{tr}[\mathbf{A}_*(\hat{\boldsymbol{\theta}})\mathbf{B}(\hat{\boldsymbol{\theta}})]\big\} \doteq \mathrm{tr}\big[\mathbf{A}_*(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})\big]$, and ultimately the approximation $\mathrm{E}(\hat{m}^*_{\mathrm{PR}}) \doteq m(\boldsymbol{\theta})$, by establishing the order of the error of the approximations. The estimator $\hat{m}^*_{\mathrm{PR}}$ is a member of classes of estimators of a more general form considered by Datta and Lahiri (2000) and Das *et al.* (2004).

Some information about the performance of estimator (4.9) or (4.10) is provided by, for example, the numerical studies of Hulting and Harville (1991), Harville and Jeske (1992), and Singh *et al.* (1998). It appears that (at least in the case of conventional linear mixed-effects models) the expected value of the estimator is reasonably close to $m(\boldsymbol{\theta})$ unless the value of $\boldsymbol{\theta}$ is close to a boundary of the parameter space Ω. When $\boldsymbol{\theta}$ is close to a boundary, $\mathrm{E}(\hat{\boldsymbol{\theta}})$ may differ significantly from $\boldsymbol{\theta}$ and [depending on the choice of $\mathbf{B}(\boldsymbol{\theta})$] $\mathbf{B}(\hat{\boldsymbol{\theta}})$ may tend to be overly "large," with the consequence that the estimator (4.9) or (4.10) may tend to exceed $m(\boldsymbol{\theta})$ by a considerable amount.

Harville and Jeske (1992) discussed various estimators of $m(\boldsymbol{\theta})$ including ones of the form

$$\hat{m}_{\mathrm{HJ}} = \hat{m}_1 + \mathrm{E}_{\mathrm{B}}\Big\{\big[\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\big]'\,\mathrm{var}(\mathbf{z}\,|\,\mathbf{s})\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\Big\}, \tag{4.11}$$

where $\hat{m}_1$ is an exactly or approximately unbiased estimator of $m_1(\boldsymbol{\theta})$ and where $\mathrm{E_B}$ is the expectation operator for the posterior distribution of $\boldsymbol{\theta}$ obtained by regarding $\mathbf{z}$ as the data vector and adopting some possibly vague prior distribution. They also discussed estimators of the somewhat more tractable form

$$\hat{m}^*_{\mathrm{HJ}} = \hat{m}_1 + \mathrm{E_B}\Big\{\big[\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\big]'\mathbf{V}_z(\hat{\boldsymbol{\theta}})\ell(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})\Big\}. \tag{4.12}$$

Moreover, for linear mixed-effects models with two variance components, they described procedures for implementing $\hat{m}_{\mathrm{HJ}}$ and $\hat{m}^*_{\mathrm{HJ}}$. And they carried out a small numerical study in which $\hat{m}_{\mathrm{HJ}}$ and $\hat{m}^*_{\mathrm{HJ}}$ were found to perform quite well.

Numerous other procedures have been proposed for estimating $m(\boldsymbol{\theta})$, many of which have features associated with Bayesian statistics, with Monte Carlo methods, and/or with resampling methods and some of which do not require Assumption N. Refer, for example, to Hulting and Harville (1991), Harville and Jeske (1992), Singh *et al.* (1998), Jiang, Lahiri, and Wan (2002), and Hall and Maiti (2006).

5. Prediction intervals

Let us now consider the problem of obtaining a prediction interval for w (under Assumption N). Define

$$t = (\hat{w} - w)/\sqrt{\hat{m}},$$

where $\hat{w}$ is the empirical BLUP $\hat{w}_{\mathrm{EBLUP}}$ or some other odd location-equivariant predictor of w and where $\hat{m}$ is $\hat{m}_{\mathrm{PR}}$ or $\hat{m}^*_{\mathrm{PR}}$ or some other estimator of $m(\boldsymbol{\theta})$ that is expressible in the form $\hat{m} = \dot{m}(\hat{\boldsymbol{\theta}})$ [for some strictly positive function $\dot{m}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$]. A prediction interval for w is obtainable by treating t (whose distribution is symmetric about 0) as a pivotal quantity.

Based on asymptotic results, t is sometimes assigned a standard normal distribution (e.g., Singh *et al.* (1998), sec 6.2). Alternatively, as discussed by Jeske and Harville (1988) and [in the special case where $v_w = 0$ and $\dot{m}(\boldsymbol{\theta}) = m_1(\boldsymbol{\theta})$] by Giesbrecht and Burns (1985), t can be assigned a Student's t distribution with degrees of freedom, say $\hat{\nu}$, determined empirically via a Satterthwaite approximation. Specifically, $\hat{\nu}$ is taken to be an estimate of the quantity

$$\nu = 2\big[\mathrm{E}(\hat{m})\big]^2/\operatorname{var}(\hat{m}) = 2\big[\mathrm{E}(\hat{m}^2) - \operatorname{var}(\hat{m})\big]/\operatorname{var}(\hat{m})$$

obtained by equating the mean and variance of a scalar multiple of $\hat{m}$ to the mean and variance of a chi-square distribution with ν degrees of freedom.

Following Jeske and Harville and Giesbrecht and Burns, we could take

$$\hat{\nu} = 2\hat{m}^2/v_m(\hat{\boldsymbol{\theta}}),$$

where $v_m(\boldsymbol{\theta})$ equals or approximates $\text{var}(\hat{m})$. Alternatively, we could take

$$\hat{\nu} = \max\bigl\{2\bigl[\hat{m}^2/v_m(\hat{\boldsymbol{\theta}})\bigr] - 2,\ \ c\bigr\}$$

for some small positive constant c. One choice for $v_m(\boldsymbol{\theta})$ [based on a first-order Taylor-series approximation to $\dot{m}(\hat{\boldsymbol{\theta}})$] is

$$v_m(\boldsymbol{\theta}) = \bigl(\partial\dot{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\bigr)'\mathbf{B}(\boldsymbol{\theta})\bigl(\partial\dot{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\bigr);$$

another (based on a second-order Taylor-series approximation) is

$$v_m(\boldsymbol{\theta}) = \bigl(\partial\dot{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\bigr)'\mathbf{B}(\boldsymbol{\theta})\bigl(\partial\dot{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\bigr) + \tfrac{1}{2}\,\text{tr}\Bigl\{\bigl[\mathbf{B}(\boldsymbol{\theta})(\partial^2\dot{m}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}')\bigr]^2\Bigr\}.$$

Instead of assigning t a standard normal or Student's t distribution (or some other distribution), the relevant features of the distribution of t (e.g., quantiles) could be determined by simulation (with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$). This possibility was considered by Jeske and Harville.

It is sometimes desired to make simultaneous inferences for two or more, say L, unobservable random variables $w_1, w_2, \ldots, w_L$. Let $\mathbf{w} = (w_1, w_2, \ldots, w_L)'$, and take $\tilde{\mathbf{w}} = \tilde{\mathbf{w}}(\boldsymbol{\theta})$ to be the $L \times 1$ vector whose ith element is (for $\boldsymbol{\theta}$ known) the BLUP of w_i. Further, define $\hat{\mathbf{w}} = \tilde{\mathbf{w}}(\hat{\boldsymbol{\theta}})$ (so that $\hat{\mathbf{w}}$ is the $L \times 1$ vector whose ith element is the empirical BLUP of w_i).

The various results on the empirical BLUP of the unobservable random variable w can be readily extended to the vector $\hat{\mathbf{w}}$ of empirical BLUPs. For example, result (3.11) becomes

$$\mathbf{M}_2(\boldsymbol{\theta}) = \sum_{i=1}^{C}\sum_{j=1}^{C} b_{ij}(\boldsymbol{\theta})\,\text{cov}\bigl[\partial\tilde{\mathbf{w}}(\boldsymbol{\theta})/\partial\theta_i,\ \partial\tilde{\mathbf{w}}(\boldsymbol{\theta})/\partial\theta_j\bigr],$$

where $\mathbf{M}_2(\boldsymbol{\theta}) = \text{E}\bigl\{[\hat{\mathbf{w}}-\tilde{\mathbf{w}}(\boldsymbol{\theta})][\hat{\mathbf{w}}-\tilde{\mathbf{w}}(\boldsymbol{\theta})]'\bigr\}$ and where (for $i, j = 1, 2, \ldots, C$) $b_{ij}(\boldsymbol{\theta})$ is the ijth element of $\mathbf{B}(\boldsymbol{\theta})$.

Let $\hat{\mathbf{M}}$ represent an estimator of the MSE matrix $\bigl($i.e., an estimator of the matrix $\text{E}[(\hat{\mathbf{w}} - \mathbf{w})(\hat{\mathbf{w}} - \mathbf{w})']\bigr)$ of the general form $\hat{\mathbf{M}} = \dot{M}(\hat{\boldsymbol{\theta}})$ [where $\dot{M}(\boldsymbol{\theta})$ is a positive definite matrix that is functionally dependent on $\boldsymbol{\theta}$]. And define

$$F = (1/L)(\hat{\mathbf{w}} - \mathbf{w})'\hat{\mathbf{M}}^{-1}(\hat{\mathbf{w}} - \mathbf{w}).$$

Kenward and Roger (1997) considered [in the special case where $\text{var}(\mathbf{w}) = \mathbf{0}$] the use of F as a pivotal quantity for making simultaneous inferences for $w_1, w_2, \ldots, w_L$. They took $\hat{\boldsymbol{\theta}}$ to be the REML estimator of $\boldsymbol{\theta}$, took $\hat{\mathbf{M}}$ to be the matrix analogue of the estimator $\hat{m}_{\text{PR}}$, and took $\mathbf{B}(\boldsymbol{\theta})$ to be the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. And they took the distribution of the quantity F to be such that λF has Snedecor's F distribution with L and ν degrees of freedom; in doing so, they took the value of λ to be $\lambda(\hat{\boldsymbol{\theta}})$ and that of ν to be $\nu(\hat{\boldsymbol{\theta}})$, where $\lambda(\boldsymbol{\theta})$ and $\nu(\boldsymbol{\theta})$ are functions of $\boldsymbol{\theta}$ obtained by comparing the mean and variance of λF with those of the F distribution. They reported the results of a simulation study of several situations in which their approach performed well. Refer to Elston (1998) for some related results and discussion. Presumably, Kenward and Roger's formulas for the quantities $\lambda(\boldsymbol{\theta})$ and $\nu(\boldsymbol{\theta})$ (and hence their approach) could be generalized (to cover the case where $\mathrm{var}(\mathbf{w}) \neq \mathbf{0}$).

REFERENCES

[1] BIBBY, J.—TOUTENBURG, H.: *Prediction and Improved Estimation in Linear Models.* Wiley, New York, 1977.

[2] DAS, K.—JIANG, J.—RAO, J. N. K.: *Mean squared error of empirical predictor*, Ann. Statist. **32** (2004), 818–840.

[3] DATTA, G. S.—LAHIRI, P.: *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems*, Statist. Sinica **10** (2000), 613–627.

[4] EATON, M. L.: *The Gauss-Markov theorem in multivariate analysis.* In: Multivariate Analysis—VI (P. R. Krishnaiah, ed.), Elsevier, Amsterdam, 1985, pp. 177–201.

[5] ELSTON, D. A.: *Estimation of denominator degrees of freedom of F-distributions for assessing wald statistics for fixed-effect factors in unbalanced mixed models*, Biometrics **54** (1998), 1085–1096.

[6] FULLER, W. A.—BATTESE, G. E.: *Transformations for estimation of linear models with nested-error structure*, J. Amer. Statist. Assoc. **68** (1973), 626–632.

[7] GIESBRECHT, F. G.—BURNS, J. C.: *Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results*, Biometrics **41** (1985), 477–486.

[8] GOLDBERGER, A. S.: *Best linear unbiased prediction in the generalized linear regression model*, J. Amer. Statist. Assoc. **57** (1962), 369–375.

[9] GUILKEY, D. K.—SCHMIDT, P.: *Estimation of seemingly unrelated regressions with vector autoregressive errors*, J. Amer. Statist. Assoc. **68** (1973), 642–647.

[10] HALL, P.—MAITI, T.: *Nonparametric estimation of mean-squared prediction error in nested-error regression models*, Ann. Statist. **34** (2006) 1733–1750.

[11] HARTIGAN, J. A.: *Linear Bayesian methods*, J. Roy. Statist. Soc., Ser. B, **31** (1969), 446–454.

[12] HARTLEY, H. O.—JAYATILLAKE, K. S. E.: *Estimation for linear models with unequal variances*, J. Amer. Statist. Assoc. **68** (1973), 189–192.

[13] HARVILLE, D. A.: *Decomposition of prediction error*, J. Amer. Statist. Assoc. **80** (1985), 132–138.

[14] HARVILLE, D. A.: *Comment on "That* BLUP *is a good thing: The estimation of random effects"*, by G. K. Robinson, Statist. Sci. **6** (1991), 35–39.

[15] HARVILLE, D. A.—JESKE, D. R.: *Mean squared error of estimation or prediction under a general linear model*, J. Amer. Statist. Assoc. **87** (1992), 724–731.

[16] HENDERSON, C. R.: *Selection index and expected genetic advance.* In: Statistical Genetics and Plant Breeding (Publication 982) (W. D. Hanson and H. F. Robinson, eds.), National Academy of Sciences, National Research Council, Washington, DC, 1963, pp. 141–163.

[17] HULTING, F. L.—HARVILLE, D. A.: *Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: Computational*

aspects, frequentist properties, and relationships, J. Amer. Statist. Assoc. **86** (1991), 557–568.

[18] JESKE, D. R.—HARVILLE, D. A.: *Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models*, Comm. Statist. Theory Methods **17** (1988), 1053–1087.

[19] JIANG, J.: *On unbiasedness of the empirical* BLUE *and* BLUP, Statist. Probab. Lett. **41** (1999), 19–24.

[20] JIANG, J.: *A matrix inequality and its statistical application*, Linear Algebra Appl. **307** (2000), 131–144.

[21] JIANG, J.—LAHIRI, P.— WAN, S.-M.: *A unified jackknife theory for empirical best prediction with M-estimation*, Ann. Statist. **30** (2002), 1782–1810.

[22] KACKAR, R. N.—HARVILLE, D. A.: *Unbiasedness of two-stage estimation and prediction procedures for mixed linear models*, Comm. Statist. Theory Methods, **10** (1981), 1249–1261.

[23] KACKAR, R. N.—HARVILLE, D. A.: *Approximations for standard errors of estimators of fixed and random effects in mixed linear models*, J. Amer. Statist. Assoc., **79** (1984), 853–862.

[24] KAKWANI, N. C.: *The unbiasedness of Zellner's seemingly unrelated regression equations estimators*, J. Amer. Statist. Assoc. **62** (1967), 141–142.

[25] KENWARD, M. G.—ROGER, J. H.: *Small sample inference for fixed effects from restricted maximum likelihood*, Biometrics **53** (1997), 983–997.

[26] KHATRI, C. G.—SHAH, K. R.: *On the unbiased estimation of fixed effects in a mixed model for growth curves*, Comm. Statist. Theory Methods **10** (1981), 401–406.

[27] LEHMANN, E. L.: *Testing Statistical Hypotheses*, (2nd ed.), Wiley, New York, 1986.

[28] PATTERSON, H. D.—THOMPSON, R.: *Recovery of inter-block information when block sizes are unequal*, Biometrika **58** (1971), 545–554.

[29] PRASAD, N. G. N.—RAO, J. N. K.: *The estimation of the mean squared error of small-area estimators*, J. Amer. Statist. Assoc. **85** (1990), 163–171.

[30] RAO, C. R.: *Linear Statistical Inference and Its Applications*, (2nd ed.), Wiley, New York, 1973.

[31] RAO, J. N. K.: *Small Area Estimation*. Wiley, Hoboken, NJ, 2003.

[32] SEELY, J.—HOGG, R. V.: *Symmetrically distributed and unbiased estimators in linear models*, Comm. Statist. Theory Methods **11** (1982), 721–729.

[33] SINGH, A. C.—STUKEL, D. M.—PFEFFERMANN, D.: *Bayesian versus frequentist measures of error in small area estimation*, J. Roy. Statist. Soc., Ser. B, **60** (1998), 377–396.

[34] WANG, J.—FULLER, W. A.: *The mean squared error of small area predictors constructed with estimated area variances*, J. Amer. Statist. Assoc. **98** 2003, 716–723.

[35] WEILER, H.—CULPIN, D.: *On the variance of weighted means*, Technometrics **12** (1970), 757–773.

[36] WOLFE, D. A.: *Some general results about uncorrelated statistics*, J. Amer. Statist. Assoc. **68** (1973), 1013–1018.

Received June 20, 2006

Mathematical Sciences Department
IBM Thomas J. Watson Research Center
P. O. Box 218
Yorktown Heights
New York 10598
U.S.A.
E-mail: harville@us.ibm.com