Bayesian and Non-Bayesian Procedures for Small-Area Estimation: Relationships and Computational Aspects

Frederick L. Hulting and David A. Harville
Department of Statistics, Iowa State University, Ames, IA 50011

1 Problem Formulation

1.1 Introduction

The topic of small area estimation has received much attention in the recent statistics literature. Small-area estimation is concerned with using sample data from a population, scattered over a large domain, to make inferences about the average, or total, of some quantity in subdomains of that population. One objective is to provide the best possible estimates for areas which contain few, if any, sampling units.

The problem of small-area estimation can, in many instances, be formulated as a special case of the general problem of predicting the realization of a random variable w based on the value of an observable random vector $\underline{y}$, where $\underline{y}$ follows a mixed linear model with a single set of random effects and w is a linear combination of fixed and random effects. A primary purpose of the present paper is to review and discuss some recent results on this general prediction problem and the special case of small area estimation. Our coverage includes both frequentist and Bayesian approaches. In particular, we discuss the relationships between the approaches, examine the computations required to implement the various approaches in the case of unbalanced data, and compare, via a Monte Carlo study, the frequentist properties of the Bayesian and non-Bayesian procedures.

1.2 Small-Area Estimation

Let y be the variable of interest, and a the number of "small areas" under consideration. One model employed in small-area estimation is

$$y_{ij} = \underline{x}'_{ij}\underline{\beta} + \underline{k}'_{ij}\underline{v} + e_{ij}, \qquad (1)$$

$(i = 1, \ldots, a;\ j = 1, \ldots, n_i)$. Here, y_{ij} represents the value of y for the jth unit sampled from the ith small area, $\underline{x}'_{ij}$ and $\underline{k}'_{ij}$ represent the values of two vectors $\underline{x}'$ and $\underline{k}'$ of explanatory or indicator variables, and $\underline{\beta} = (\beta_1, \ldots, \beta_b)'$ is a vector of unknown parameters. Further, the elements of the vector $\underline{v} = (v_1, \ldots, v_a)'$ — one element for each small area — are regarded as uncorrelated random variables having mean zero and common variance σ_v^2, while the e_{ij}'s are random variables that are uncorrelated with each other and with the v_i's and have mean zero and common variance σ_e^2.

The quantity of interest is usually the population mean $\bar{Y}_i$ of y for the ith small area, which may be expressed as

$$\bar{\underline{x}}'_i\underline{\beta} + \bar{\underline{k}}'_i\underline{v} + \bar{\underline{e}}_i, \qquad (2)$$

where $\bar{\underline{x}}'_i$ and $\bar{\underline{k}}'_i$ are the population means of $\underline{x}'$ and $\underline{k}'$ for the ith small-area and where $\bar{\underline{e}}_i = \bar{Y}_i - \bar{\underline{x}}'_i\underline{\beta} - \bar{\underline{k}}'_i\underline{v}$. Following Prasad and Rao (1986), we assume that the population size N_i for the ith area is large, in which case $\bar{\underline{e}}_i \approx 0$. Often, the vector $\underline{k}_{ij}$ is chosen to be the vector whose i^{th} element equals one and remaining elements equal zero, in which case $\underline{k}'_{ij}\underline{v} = v_i$ and $\bar{\underline{x}}'_i\underline{\beta} + \bar{\underline{k}}'_i\underline{v} = \bar{\underline{x}}'_i\underline{\beta} + v_i$ is interpretable as the small-area mean.

Models of the form (1) were adopted by, e.g., Fay and Herriot (1979), Stroud (1987), Dempster and Rhagunathan (1987), and Dempster, Rubin and Tsutakawa (1981).

Crop-Area Example . Battese et al. (1988) discuss the application of small-area methods to the prediction of county crop areas using satellite information. Their data consists of two different determinations of areas under corn and soybeans for a sample of segments from 12 north-central Iowa counties (a segment is about 250 hectares): (1) the number of hectares reported by farm operators, and (2) the number of pixels classified as corn or soybeans (as determined from LANDSAT readings).

Here we focus on the estimation of mean hectares of corn per segment for the 12 counties represented in the sample. Let y_{ij} be the reported hectares of corn in the jth segment of the ith county, and let x_{1ij} and x_{2ij} represent the number of satellite pixels in a sample segment classified as corn and soybeans, respectively. Battese et al. (1988) propose the nested error regression model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i + e_{ij}, \qquad (3)$$

which is a special case of model (1). The data, along with more details about the sample, are given by Battese et al. (1988).

1.3 General Prediction Problem

Let $\underline{y}$ represent an $n \times 1$ observable random vector. Suppose that $\underline{y}$ follows the mixed-effects linear model

$$\underline{y} = \underline{X}\underline{\beta} + \underline{Z}\underline{s} + \underline{e} \ , \qquad (4)$$

where $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters and where $\underline{s}$ and $\underline{e}$ are unobservable, statistically independent random vectors of dimensions $m \times 1$ and $n \times 1$ that

are distributed as MVN$(0, \sigma_s^2 I)$, and MVN$(0, \sigma_e^2 I)$, respectively. Here, the variance components σ_s^2 and σ_e^2 are unknown parameters satisfying $\sigma_s^2 \geq 0$ and $\sigma_e^2 > 0$, and X and Z are given matrices. Define $p^* = \text{rank}(X)$, $r = \text{rank}(X, Z) - \text{rank}(X)$, and $f = n - \text{rank}(X, Z)$. We assume that $r > 0$ and $f > 0$.

Let $\gamma = \sigma_s^2/\sigma_e^2$ and $\theta = (\sigma_e^2, \gamma)'$. The parameter space for θ is $\Omega = \{(\sigma_e^2, \gamma)' : \sigma_e^2 > 0, \gamma \geq 0\}$. Letting $V_y = var(y)$, we have that $V_y = \sigma_e^2 I + \sigma_s^2 ZZ' = \sigma_e^2(I + \gamma ZZ') = \sigma_e^2 V$, with $V = V(\gamma) = I + \gamma ZZ'$.

Clearly, the problem of small-area estimation can be regarded as a special case of the problem of predicting the realization of a random variable w of the form $w = \lambda'\beta + \delta's$ — we refer to the latter problem as the general prediction problem. Let $v_w = var(w)$ and $v_{yw} = cov(y, w)$, and note that $E(w) = \lambda'\beta$, $v_w = \sigma_e^2 \gamma \delta'\delta$, and $v_{yw} = \sigma_e^2 \gamma Z\delta$. Subsequently we assume that λ is such that $\lambda = X'u$ for some vector u, or equivalently that w is estimable or predictable. (Unless otherwise indicated, probabilities and expectations are defined with respect to the joint distribution of w and y.)

2 Frequentist Approaches

Traditional procedures for the general prediction problem are based on a two-stage approach to making inferences under the mixed-effects model (4): first, estimate the error variance σ_e^2 and the variance ratio γ and , second, proceed as if σ_e^2 and γ were known. In some cases, modifications are made to account for the additional variability introduced by the estimation of σ_e^2 and γ.

2.1 Point Prediction of w

Suppose that γ were known. Then, the best (minimum MSE) linear unbiased predictor (BLUP) of the value of w is $\tilde{w}(\gamma) = \lambda'\tilde{\beta} + \delta'\tilde{s}$, where $\tilde{\beta}$ is any solution to the Aitken equations $X'V^{-1}X\tilde{\beta} = X'V^{-1}y$, and $\tilde{s} = \gamma Z'V^{-1}(y - X\tilde{\beta})$. The MSE of the BLUP is $v^*(\sigma_e^2, \gamma) = \sigma_e^2 \Psi(\gamma)$, with

$$
\begin{aligned}
\Psi(\gamma) = {} & \gamma\delta'(I - \gamma Z'V^{-1}Z)\delta + \{[\lambda' - \gamma\delta'Z'V^{-1}X] \\
& \times [X'V^{-1}X]^-[\lambda - \gamma X'V^{-1}Z\delta]\} .
\end{aligned}
$$

In the case where $k'_{ij}v = v_i$, the BLUP of the small-area mean may be written as

$$
\left(\frac{\gamma}{\gamma + n_i^{-1}}\right)\bar{y}_i + \left(\bar{x}'_i - \frac{\gamma}{\gamma + n_i^{-1}}\bar{x}'_{i(s)}\right)\tilde{\beta}, \quad (5)
$$

where $\bar{x}'_{i(s)}$ and $\bar{y}_i$ are the sample averages of x' and y for the ith small area.

Except in special cases, the BLUP $\tilde{w}(\gamma)$ depends on the (unknown) value of γ. A predictor of the value of w can be obtained from $\tilde{w}(\gamma)$ by replacing γ with an estimator of γ. We refer to this predictor as the empirical BLUP.

One popular method for estimating σ_e^2 and γ is Patterson and Thompson's (1971) restricted maximum likelihood (REML) procedure. The REML estimators $\tilde{\sigma}_e^2$ and $\tilde{\gamma}$ are

obtained by maximizing the modified likelihood function $l^*(\sigma_e^2, \gamma; z)$, which is equal to

$$
\sigma_e^{2 - (n-p^*)/2}[det(V)]^{-1/2}[det(X^{*'}V^{-1}X^*)]^{-1/2}
$$
$$
\times \exp\{-(2\sigma_e^2)^{-1}(y - X\tilde{\beta})'V^{-1}(y - X\tilde{\beta})\}, \quad (6)
$$

where X^* in an $n \times p^*$ matrix whose columns consist of any p^* LIN columns of X.

The REML estimators $\tilde{\gamma}$ and $\tilde{\sigma}_e^2$ are even, translation-invariant functions of y. Subsequently, we denote by $\hat{\gamma}$ and $\hat{\sigma}_e^2$ arbitrary even, translation-invariant estimators of γ and σ_e^2. Further, we denote by $\hat{w}$ the empirical BLUP obtained from $\tilde{w}(\gamma)$ by replacing γ with $\hat{\gamma}$; that is $\hat{w} = \tilde{w}(\hat{\gamma})$. The predictor $\hat{w}$ is an unbiased predictor of the value of w, provided that the expected value of $\hat{w}$ exists (e.g., Kackar and Harville, 1984).

2.2 Estimation of the MSE of $\hat{w}$

Let $v^+(\sigma_e^2, \gamma)$ represent the MSE of $\hat{w}$. Closed-form expressions for $v^+(\sigma_e^2, \gamma)$ are available only in very simple special cases. It is common practice to approximate $v^+(\sigma_e^2, \gamma)$ by $v^*(\sigma_e^2, \gamma)$; however, this approximation is naive in that it ignores the contribution of the estimator $\hat{\gamma}$ to the variability of $\hat{w} - w$ (e.g., Khatri and Shah, 1981). Kackar and Harville (1984) approximated this contribution by $a \times b$, where $a = var[\partial\tilde{w}(\gamma)/\partial\gamma]$ and b equals or approximates the MSE $E[(\hat{\gamma} - \gamma)^2]$ of $\hat{\gamma}$. The corresponding approximation to the MSE of $\hat{w}$ is $v^\oplus(\sigma_e^2, \gamma) = v^*(\sigma_e^2, \gamma) + a \times b$. If $\hat{\gamma} \equiv \tilde{\gamma}$, then one choice for b is the large-sample variance of $\tilde{\gamma}$.

The MSE of $\hat{w}$ is often estimated by $v^*(\hat{\sigma}_e^2, \hat{\gamma})$. Letting $\hat{a}$ and $\hat{b}$ represent the values of a and b at $\sigma_e^2 = \hat{\sigma}_e^2$ and $\gamma = \hat{\gamma}$, a more conservative estimator of the MSE is the estimator $v^\oplus(\hat{\sigma}_e^2, \hat{\gamma}) = v^*(\hat{\sigma}_e^2, \hat{\gamma}) + \hat{a} \times \hat{b}$. The asymptotic results of Prasad and Rao (1986) indicate that $v^\oplus(\hat{\sigma}_e^2, \hat{\gamma})$ tends to underestimate the MSE, though of course to a lesser extent than $v^*(\hat{\sigma}_e^2, \hat{\gamma})$, and suggest the alternative estimator $\hat{v} = v^*(\hat{\sigma}_e^2, \hat{\gamma}) + 2(\hat{a} \times \hat{b})$.

2.3 Interval Prediction of w

Approximate prediction intervals may be obtained by acting as though the quantity $t(w, y)$, defined by $t(w, y) = (\hat{w} - w)/[\hat{v}^+]^{1/2}$, is a pivotal quantity. Here, $\hat{v}^+$ represents a nonnegative function of $\hat{\sigma}_e^2$ and $\hat{\gamma}$, which is to be regarded as an estimator of the MSE of $\hat{w}$. It follows from Proposition 1 of Jeske and Harville (1988) that the distribution of $t(w, y)$ is symmetric about zero. If we approximate the upper-$\alpha/2$ point of the distribution of $t(w, y)$ by the upper-$\alpha/2$ point, say $z_{\alpha/2}$, of the standard normal distribution, we obtain from $t(w, y)$ the approximate $100(1-\alpha)\%$ prediction interval

$$
\hat{w} \pm z_{\alpha/2}[\hat{v}^+]^{1/2} . \quad (7)
$$

This interval, which is similar in form to intervals considered by Fuller (1988) in a multivariate version of the small-area problem, is likely to be satisfactory if $\hat{\gamma}$ and $\hat{v}^+$ estimate γ and $v^+(\sigma_e^2, \gamma)$ precisely. Here, we focus on the following version of interval (7): I1 : $\hat{w} \pm z_{\alpha/2}[\hat{v}]^{1/2}$, where we take $\hat{\sigma}_e^2 = \tilde{\sigma}_e^2$, and $\hat{\gamma} = \tilde{\gamma}$.

A potential improvement on interval (7) is obtained by approximating the upper-$\alpha/2$ point of the distribution of $t(w,\underline{y})$ by the upper-$\alpha/2$ point, say $t_{\alpha/2}(\hat{\nu})$ of Student's t distribution with $\hat{\nu}$ degrees of freedom, where $\hat{\nu}$ is an estimate obtained by adapting the approach of Satterthwaite (1946). By making use of this approximation, we obtain from $t(w,\underline{y})$ the approximate $100(1-\alpha)\%$ prediction interval

$$\hat{w} \pm t_{\alpha/2}(\hat{\nu})[\hat{v}^+]^{1/2} . \tag{8}$$

We consider three versions of prediction interval (8); namely I2 : $\hat{w} \pm t_{\alpha/2}(\hat{\nu}_1)[v^{\oplus}(\hat{\sigma}_e^2,\hat{\gamma})]^{1/2}$, and I3 : $\hat{w} \pm t_{\alpha/2}(\hat{\nu}_2)[\hat{v}]^{1/2}$, where $\hat{\nu}_1$, and $\hat{\nu}_2$, represent the values of $\hat{\nu}$ obtained by choosing $\hat{v}^+$ to be $v^{\oplus}(\hat{\sigma}_e^2,\hat{\gamma})$ or $\hat{v}$, respectively. In both versions, we take $\hat{\sigma}_e^2 = \tilde{\sigma}_e^2$, $\hat{\gamma} = \tilde{\gamma}$. Further details on the implementation of this approach may be found in Hulting and Harville (1989).

2.4 Computational Aspects

In this section, we give expressions for $l^*(\sigma_e^2,\gamma;\underline{z})$, $\tilde{w}(\gamma)$, $\Psi(\gamma)$, a, and b that are useful when these quantities must be calculated repeatedly for different values of γ and σ_e^2 (as is required by iterative algorithms for computing REML estimates). Following Dempster et al. (1984) and Harville and Fenech (1984), let $\underline{C} = \underline{Z}'(\underline{I} - \underline{P}_X)\underline{Z}$, $\underline{q} = \underline{Z}'(\underline{I} - \underline{P}_X)\underline{y}$, and $r = rank(\underline{C})$. Define $\Delta_1,\ldots,\Delta_r$ to be the nonzero (and hence positive) characteristic roots of $\underline{C}$, and take $\underline{R}$ to be a matrix of dimensions $m \times r$ whose columns are orthonormal characteristic vectors of $\underline{C}$ corresponding to $\Delta_1,\ldots,\Delta_r$. Furthermore, letting $\underline{D} = diag(\Delta_1,\ldots,\Delta_r)$, define $\underline{t} = (t_1,\ldots,t_r)' = \underline{D}^{-1/2}\underline{R}'\underline{q}$, and let $S_s = \sum_{i=1}^r t_i^2$, and $S_e = \underline{y}'(\underline{I} - \underline{P}_X)\underline{y} - \sum_{i=1}^r t_i^2$.

Harville and Callanan (1989) showed that $l^*(\sigma_e^2,\gamma;\underline{z})$ is proportional to the function $l(\sigma_e^2,\gamma;\underline{z}) = (2\pi\sigma_e^2)^{-(n-p^*)/2}M(\gamma) \times \exp[-(2\sigma_e^2)^{-1}A(\gamma;t;S_e)]$, where $M(\gamma) = \{\prod_{i=1}^r(1 + \gamma\Delta_i)\}^{-1/2}$, and $A(\gamma;t;S_e) = S_e + \sum_{i=1}^r t_i^2/(1+\gamma\Delta_i)$. Note that once $\Delta_1,\ldots,\Delta_r$ and $\underline{R}$ have been computed, the additional expenditure of computing resources required to repeatedly evaluate $l(\sigma_e^2,\gamma;\underline{z})$ for different values of σ_e^2 and γ is relatively small.

Let $\underline{\zeta} = [\underline{\delta} - \underline{Z}'X(X'X)^-\underline{\lambda}]$. Analogous to the expression for $l(\sigma_e^2,\gamma;\underline{z})$ are the following expressions for $\tilde{w}(\gamma)$, $\Psi(\gamma)$, a, and b given by Harville (1989):

$$\tilde{w}(\gamma) = \underline{\lambda}'(X'X)^-X'\underline{y} + \gamma\underline{\zeta}'\underline{R}(\underline{I} + \gamma\underline{D})^{-1}\underline{D}^{1/2}\underline{t} , \tag{9}$$

$$\Psi(\gamma) = \underline{\lambda}'(X'X)^-\underline{\lambda} + \gamma\underline{\zeta}'\underline{\zeta}$$
$$-\gamma^2\underline{\zeta}'\underline{R}(\underline{I} + \gamma\underline{D})^{-1}\underline{D}\underline{R}'\underline{\zeta} , \tag{10}$$

$a = \sigma_e^2\underline{\zeta}'\underline{R}\underline{D}(\underline{I}+\gamma\underline{D})^{-3}\underline{R}'\underline{\zeta}$, and (when b is taken to be the large-sample variance of $\tilde{\gamma}$) $b = 2\{\sum_{i=1}^r \Delta_i^2/(1 + \gamma\Delta_i)^2 - (n - p^*)^{-1}[\sum_{i=1}^r \Delta_i/1 + \gamma\Delta_i]^2\}^{-1}$. Expressions for various other quantities needed to compute the end points of Intervals I2 and I3 may be found in Hulting and Harville (1989).

3 A Bayesian Approach

3.1 Formulation of Problem

As an alternative to the approach discussed in Section 2, we consider a Bayesian approach to inference about w. In the Bayesian approach, we select a prior distribution for $\underline{\beta}$ and $\underline{\theta} = (\sigma_e^2,\gamma)$ and then base our inferences on the conditional distribution of w given $\underline{y}$.

Suppose that a priori $\underline{\beta}$ and $\underline{\theta}$ are statistically independent, that $\underline{\beta} \sim MVN(\underline{\alpha},\varepsilon\underline{I})$ (where $\underline{\alpha}$ is a known $p \times 1$ vector and ε is a known positive scalar), and that $\underline{\theta}$ has a proper or improper distribution with p.d.f. equal or proportional to $\pi_2(\sigma_e^2,\gamma)$. Then, under model (4), the conditional distribution of w and $\underline{y}$ given $\underline{\theta}$ is MVN, with p.d.f $g(w,\underline{y} \mid \theta) = g_1(w \mid \underline{y},\underline{\theta}) \times g_2(\underline{y} \mid \underline{\theta})$ where $g_1(w \mid \underline{y},\underline{\theta})$ is the p.d.f. of the conditional distribution of w given $\underline{y}$ and $\underline{\theta}$, which is normal with mean $w^\varepsilon = \underline{\lambda}'\underline{\alpha} + (\underline{v}'_{yw} + \varepsilon\underline{\lambda}'X')(\underline{V}_y + \varepsilon XX')^{-1}(\underline{y} - X\underline{\alpha})$ and variance $v^\varepsilon = (v_w + \varepsilon\underline{\lambda}'\underline{\lambda}) - (\underline{v}'_{yw} + \varepsilon\underline{\lambda}'X')(\underline{V}_y + \varepsilon XX')^{-1}(\underline{v}_{yw} + \varepsilon X\underline{\lambda})$, and $g_2(\underline{y} \mid \underline{\theta})$ is the pdf of the conditional distribution of $\underline{y}$ given $\underline{\theta}$, which is MVN with mean vector $X\underline{\alpha}$ and covariance matrix $\underline{V}_y + \varepsilon XX'$. Thus, the distribution on which the Bayesian would base his/her analysis is that with p.d.f.

$$f(w \mid \underline{y}) \propto \int_\Omega g_1(w \mid \underline{y},\underline{\theta})g_2(\underline{y} \mid \underline{\theta})\pi_2(\sigma_e^2,\gamma)d\underline{\theta}. \tag{11}$$

In what follows, we assume that the prior distribution of $\underline{\theta}$ is a proper or improper distribution with a p.d.f. of the general form

$$\pi_2(\sigma_e^2,\gamma) = G_1(\gamma)(\sigma_e^2)^{G_2(\gamma)} \exp\{-(2\sigma_e^2)^{-1}G_3(\gamma)\}, \tag{12}$$

where $G_1(\gamma)$, $G_2(\gamma)$, and $G_3(\gamma)$ are arbitrary functions of γ such that $G_1(\gamma) > 0$, $G_2(\gamma) < (n - p^* - 4)/2$, and $G_3(\gamma) \geq 0$. For prior distributions of the general form (12), the Bayesian approach is computationally tractable. Moreover, the class of distributions of general form (12) is broad enough to cover a wide range of prior opinions. In a small-area context, Bayesian approaches to the various special cases of model (4), based on a prior distribution of the form (12), have been considered, for example, by Stroud (1987), Ghosh and Lahiri (1988), and Datta and Ghosh (1989).

3.2 Limiting Posterior $f^*(w \mid y)$

For sufficiently large values of ε, the prior distribution of $\underline{\beta}$ can be regarded as noninformative. Moreover, the results of Dempster, Rubin, and Tsutakawa (1981) and Sallas and Harville (1985) suggest that for "large" values of ε, $f(w \mid \underline{y})$ can be approximated by the p.d.f.

$$f^*(w \mid \underline{y}) = \int_\Omega g_1^*(w \mid \underline{y},\underline{\theta})h(\underline{\theta} \mid \underline{z})d\underline{\theta}.$$

Here, $g_1^*(w \mid \underline{y},\underline{\theta})$ is the p.d.f. of a normal distribution with mean $\tilde{w}(\gamma)$ and variance $v^*(\sigma_e^2,\gamma)$, and $h(\underline{\theta} \mid \underline{z}) = (1/k_1^*) \times l(\sigma_e^2,\gamma;\underline{z})\pi_2(\sigma_e^2,\gamma)$, where $k_1^* = \int_\Omega l(\sigma_e^2,\gamma;\underline{z})\pi_2(\sigma_e^2,\gamma)d\underline{\theta}$. Subsequently, we regard $f^*(w \mid \underline{y})$ as the posterior p.d.f. of w.

Let $\Gamma[\cdot]$ represent the gamma function, and, for $c > 0$, define $b(\gamma, c) = [n - p^* - 2G_2(\gamma) - c]/2$, $B_1(\gamma, c) = M(\gamma)G_1(\gamma)\Gamma[b(\gamma, c)]$, and $B_2(\gamma; \underline{t}; S_e) = (1/2)[A(\gamma; t; S_e) + G_3(\gamma)]$. Then,

$$k_1^* = (2\pi)^{(n-p^*)/2} \int_0^\infty B_1(\gamma, 2)B_2(\gamma; \underline{t}; S_e)^{-b(\gamma, 2)} d\gamma ,$$

(13)

and the posterior p.d.f. $f^*(w \mid \underline{y})$ may be written as

$$f^*(w \mid \underline{y}) = (2\pi)^{-1/2}[k_2^*]^{-1} \int_0^\infty B_1(\gamma, 1)\Psi(\gamma)^{-1/2}$$

$$\times \left\{ \frac{[w - \tilde{w}(\gamma)]}{2\Psi(\gamma)} + B_2(\gamma; \underline{t}; S_e) \right\}^{-b(\gamma, 1)} d\gamma$$

(14)

where $k_2^* = k_1^*(2\pi)^{(n-p^*)/2}$. By making use of expressions (13) and (14), $f^*(w \mid \underline{y})$ can be expressed as the ratio of two one-dimensional integrals. In general, these integrals must be evaluated by numerical integration. In carrying out the numerical integration, the two integrands must be repeatedly evaluated for different values of γ. Note that, once $\Delta_1, \ldots, \Delta_r$ and $\underline{R}$ have been computed, the repeated evaluation of the integrands requires — in light of expressions (9) and (10) for $\tilde{w}(\gamma)$ and $\Psi(\gamma)$ — very little additional computation.

Finally, we note that the posterior probability $P^*(S \mid \underline{y})$ of a set S of w-values can also be expressed as a one-dimensional integral with respect to γ, as discussed in Hulting and Harville (1989).

3.3 Characterizations of $f^*(w \mid \underline{y})$

As a practical matter, Bayesian inference requires that summaries of the posterior be computed. In this section, we describe features of $f^*(w \mid \underline{y})$ which are analogous to quantities used in the frequentist approach.

Let w_B and v_B represent the posterior mean and variance, respectively, of w; that is the mean and variance of the distribution with p.d.f. $f^*(w \mid \underline{y})$. Then, it can be shown that $w_B = \int_0^\infty \tilde{w}(\gamma)h_1(\gamma \mid \underline{z})d\gamma$, where the p.d.f. $h_1(\gamma \mid \underline{z})$, defined by $h_1(\gamma \mid \underline{z}) = \int_0^\infty h(\underline{\theta} \mid \underline{z})d\sigma_e^2$, is expressible as $h_1(\gamma \mid \underline{z}) = [k_2^*]^{-1}B_1(\gamma, 2)B_2(\gamma; \underline{t}; S_e)^{-b(\gamma, 2)}$, and $v_B = [k_1^*]^{-1} \int_0^\infty B_1(\gamma, 4)B_2(\gamma; \underline{t}; S_e)^{-b(\gamma, 4)}\{\Psi(\gamma) + [\tilde{w}(\gamma) - w_B]^2 B_2(\gamma; \underline{t}; S_e)b(\gamma, 4)\}d\gamma$. The posterior mean can be regarded as an estimator of the value of w, and can be interpreted as a weighted average (with respect to γ) of the BLUP $\tilde{w}(\gamma)$.

A Bayesian counterpart to a prediction interval is a credible set. A set S of w-values is said to be a $100(1-\alpha)\%$ credible set if $P^*(S \mid \underline{y}) = 1 - \alpha$. The smallest $100(1-\alpha)\%$ credible set is S= $\{w : f^*(w \mid \underline{y}) \geq k_\alpha\}$, where k_α is a constant such that $P^*(S \mid \underline{y}) = 1 - \alpha$ (see e.g. Berger 1985, p. 140). This set is called the $100(1-\alpha)\%$ HPD credible set.

Consider an interval I4 : $[l, u]$, with endpoints l and u such that $f^*(l \mid \underline{y}) = f^*(u \mid \underline{y})$ and $P^*([l, u] \mid \underline{y}) = 1-\alpha$. By definition, this interval is a $100(1-\alpha)\%$ credible set, and if the posterior distribution of w is unimodal, it is the $100(1-\alpha)\%$ HPD credible set. Interval I4 must be computed iteratively via the solution of certain nonlinear equations. An

efficient computational procedure is described by Hulting and Harville (1989).

4 Related Approaches

The conditional p.d.f. $f(w \mid \underline{y})$ (and hence $f^*(w \mid \underline{y})$) also has a hierarchical Bayes (HB) interpretation. To see this, regard $\underline{s}$ as a vector of unknown parameters (fixed effects) rather than as a vector of random effects, and form the prior distribution of $\underline{\beta}$, $\underline{s}$, and σ_e^2 in three stages. In the first stage, assign $\underline{\beta}$ and $\underline{s}$ a distribution — conditional on the values of σ_e^2 and a hyperparameter γ — whose p.d.f. is $\pi_1^*(\underline{\beta}, \underline{s} \mid \sigma_e^2, \gamma)$; then, in the second stage, assign σ_e^2 a distribution — conditional on the value of γ — whose p.d.f. is $\pi_2^*(\sigma_e^2 \mid \gamma)$; and finally, in the third stage, assign γ a distribution with p.d.f. $\pi_3^*(\gamma)$. Now, suppose that apriori $\underline{\beta}$ and $\underline{s}$ are independent, and that $\underline{\beta} \sim \text{MVN}(\underline{\alpha}, \varepsilon\underline{I})$ and $\underline{s} \sim \text{MVN}(\mathbf{0}, \sigma_s^2\underline{I})$. Then, clearly, the p.d.f. of the conditional distribution for w obtained by this three-stage approach is the same as that obtained by setting $\pi_2(\sigma_e^2, \gamma) \propto \pi_2^*(\sigma_e^2 \mid \gamma)\pi_3^*(\gamma)$ in p.d.f. (11).

As in the HB approach, parametric empirical Bayes (PEB) inference procedures are based on a conditional distribution for w given $\underline{y}$ which is derived using a multistage prior on the parameters of the fixed effects model. However, in this approach, the prior distribution is specified only up to the values of σ_e^2 and γ. These unknown quantities must be estimated from the data.

If σ_e^2 and γ were known, then in the limiting case (as $\varepsilon \to \infty$ and hence where the prior distribution of $\underline{\beta}$ is noninformative), the posterior distribution of w given $\underline{y}$ and $\underline{\theta}$ would be the distribution with p.d.f. $g_1^*(w \mid \underline{y}, \underline{\theta})$, so that the posterior mean would be $\tilde{w}(\gamma)$ and the $100(1-\alpha)\%$ HPD credible set would be $\tilde{w}(\gamma) \pm z_{\alpha/2}\sigma_e[\Psi(\gamma)]^{1/2}$. A simple implementation of the PEB approach is to take the point predictor to be $\tilde{w}(\hat{\gamma})$ and the prediction interval to be $\tilde{w}(\hat{\gamma}) \pm z_{\alpha/2}\hat{\sigma}_e[\Psi(\hat{\gamma})]^{1/2}$. However, this prediction interval may not be satisfactory in that it fails to account for the additional uncertainty about the value of w that comes from not knowing σ_e^2 and γ. Morris (1983), Rubin (1982), and Laird and Louis (1987), restricting attention to a special case of the prediction problem and taking σ_e^2 to be known, discussed PEB intervals of the form $\tilde{w}(\hat{\gamma}) \pm z_{\alpha/2}\sigma_e\dot{\Psi}^{1/2}$ where $\dot{\Psi}$ is chosen — on the basis of Bayesian, bootstrap, or Monte Carlo considerations — so that $\sigma_e^2\dot{\Psi}$ is a better estimator of $\text{var}[\tilde{w}(\hat{\gamma}) - w]$ than is $\sigma_e^2\Psi(\hat{\gamma})$. To account for σ_e^2 not being known, we could, following Berger (1985, p. 172), adopt the interval $\tilde{w}(\hat{\gamma}) \pm t_{\alpha/2}(f)\bar{\sigma}_e\dot{\Psi}^{1/2}$, where $f = n - p^* - r$.

The application of the PEB approach to small-area estimation has been considered by, e.g., Fay and Herriot (1979), Dempster et. al. (1981), Dempster et. al. (1984), Ghosh and Meeden (1986), and Dempster and Rhagunathan (1987). Prediction intervals I1 – I3, which were introduced in Section 2.4, have obvious PEB interpretations and are applicable to the general problem of predicting the value of w. They can be used to obtain PEB intervals for the value of w in applications not covered by previously considered special cases.

5 Frequentist Properties

For each of the point predictors $\bar{w}$, $\hat{w}$, and w_B, the distribution of the prediction error is symmetric about zero (e.g., Harville, 1985). Thus, w_B, like $\bar{w}$ and $\hat{w}$, is an unbiased predictor, provided that its expected value exists.

To evaluate properties of the point and interval predictors we carried out a Monte Carlo study. We investigated the properties of the various predictors conditional on $\tilde{\gamma} = 0$, as well as their overall properties. Five situations were considered. They correspond to one set of values for $\underline{\lambda}$, $\underline{\delta}$, $\underline{X}$, and $\underline{Z}$ in combination with the following five values of γ: $\gamma = 0$, 0.2, 0.5, 1.0, and 2.0. The set of values for $\underline{\lambda}$, $\underline{\delta}$, $\underline{X}$, and $\underline{Z}$ was that for inference about the mean amount of corn per segment in Cerro Gordo County in the crop-area example. The values of $\underline{\beta}$ and σ_e^2 were (without loss of generality) taken to be $\underline{\beta}=0$ and $\sigma_e^2=1$, and the Jeffreys (1961) prior for $\underline{\theta} = (\sigma_e^2, \gamma)'$ was used in the Bayesian analysis. For those situations where $\gamma=0$, 0.2, 0.5 or 1.0 we carried out two simulations — one conditional on $\tilde{\gamma}=0$, and the other conditional on $\tilde{\gamma} > 0$. Details may be found in Hulting and Harville (1989).

Tables 1 and 2 contain unconditional results from the Monte Carlo study. For very small values of γ, the two estimators $v^{\circledcirc}(\tilde{\sigma}_e^2,\tilde{\gamma})$, and $\hat{v}$ tend to severely overestimate the unconditional MSE $v^{+}(\sigma_e^2, \gamma)$. For the larger values of γ, $v^{\circledcirc}(\tilde{\sigma}_e^2,\tilde{\gamma})$ tends to underestimate $v^{+}(\sigma_e^2, \gamma)$, while $\hat{v}$ is nearly unbiased. Also, for the larger values of γ, the unconditional MSE of w_B is smaller than that of $\hat{w}$. However, the associated standard errors suggest these may not be real differences.

The unconditional probabilities of coverage of the Bayesian interval I4 tends to be closer to the nominal level than that of the frequentist interval I3, and the expected length of I4 is less than that of I3. The expected length of Interval I1 compares favorably with that of the Bayesian interval, but, for some values of γ, its unconditional probability of coverage is significantly less than the nominal level.

For $\gamma=0.0$, 0.2, 0.5, and 1.0, the conditional (on $\tilde{\gamma}=0$) MSEs of $\hat{w}$ were 0.032, 0.227, 0.518, and 0.959, respectively, while those of w_B were 0.053, 0.199, 0.422, and 0.680. Thus, conditional on $\tilde{\gamma}=0$, the overall performance of w_B was superior to that of $\hat{w}$.

Furthermore, conditional on $\tilde{\gamma}=0$, the probability of coverage of Intervals I2 and I3 tended to be much higher than the nominal level — when $\tilde{\gamma}=0$, the estimated degrees of freedom $\hat{\nu}_1$ and $\hat{\nu}_2$ tended to be very small and consequently the percentage points $t_{\alpha/2}(\hat{\nu}_1)$ and $t_{\alpha/2}(\hat{\nu}_2)$ tended to be very large. In particular, the conditional probability of coverage of Interval I3 was 1.000 for $\gamma=0.0$, 0.2, 0.5, and 1.0. The conditional probability of coverage of Interval I1 declined sharply with the value of γ and was only 0.488 for $\gamma=1.0$.

Overall, the conditional behavior of the Bayesian interval I4 was more sensible than that of Intervals I1 – I3. For $\gamma= 0.0$, 0.2, 0.5, and 1.0, the probabilities of coverage of Interval I4 were 1.000, 0.952, 0.852, and 0.815, respectively, and its expected lengths were 1.7, 1.8, 2.0, and 2.1. By way of comparison, the expected lengths of Interval I3 were 7.0, 7.4, 7.6 and 8.0, respectively.

Table 1. Monte Carlo Estimates of Unconditional MSEs and Expectations (and the Estimated Standard Errors of the Monte Carlo Estimates)

	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 2.0$
MSE of $\hat{w}$	0.050	0.235	0.443	0.641	0.839
	(0.001)	(0.003)	(0.007)	(0.009)	(0.012)
$E[v^{\circledcirc}(\tilde{\sigma}_e^2,\tilde{\gamma})]$	0.108	0.225	0.379	0.567	0.773
	(0.000)	(0.001)	(0.002)	(0.002)	(0.002)
$E[\hat{v}(\tilde{\sigma}_e^2,\tilde{\gamma})]$	0.135	0.264	0.431	0.630	0.840
	(0.000)	(0.001)	(0.002)	(0.002)	(0.002)
MSE of w_B	0.076	0.234	0.423	0.615	0.823
	(0.001)	(0.004)	(0.006)	(0.009)	(0.012)
$E[v_B]$	0.216	0.313	0.440	0.608	0.815
	(0.000)	(0.001)	(0.002)	(0.002)	(0.003)

Table 2. Monte Carlo Estimates of Unconditional Probabilities of Coverage (and Unconditional Expected Lengths)

	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 2.0$
I1	0.996	0.911	0.901	0.921	0.938
	(1.4)	(2.0)	(2.5)	(3.1)	(3.6)
I2	0.999	0.994	0.974	0.951	0.938
	(4.6)	(4.2)	(3.6)	(3.4)	(3.5)
I3	0.999	0.996	0.983	0.967	0.957
	(6.0)	(4.8)	(3.9)	(3.7)	(3.8)
I4	0.999	0.970	0.943	0.938	0.944
	(1.8)	(2.2)	(2.6)	(3.0)	(3.5)

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *J. Amer. Stat. Assoc.*, 83, 28–36.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd edition), New York: Springer-Verlag.

Datta, G. S., and Ghosh, M. (1989), "Bayesian Prediction in Mixed Linear Models: Applications to Small Area Estimation," Technical Report, Department of Statistics, University of Florida.

Dempster, A. P., and Raghunathan, T. E. (1987), "Using a Covariate for Small Area Estimation: A Common Sense Approach," in *Small Area Statistics: An International Symposium*, eds. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, New York: John Wiley.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981), "Estimation in Covariance Components Models," *J. Amer. Stat. Assoc.*, 76, 341–353.

Dempster, A. P., Selwyn, M. R., Patel, C. M., and Roth, A. J. (1984), "Statistical and Computational Aspects of Mixed Model Analysis," *Applied Statistics*, 33, 203–214.

Fay, R.E., and Herriot, R. A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *J. Amer. Stat. Assoc.*, 74, 269–277.

Fuller, W. A. (1988), "Prediction with the Multivariate Components of Variance Model," unpublished manuscript, Iowa State University, Department of Statistics.

Ghosh, M., and Lahiri, P. (1988), "A Hierarchical Bayes Approach to Small Area Estimation with Auxiliary Information," Technical Report 315, Department of Statistics, University of Florida.

Ghosh, M., and Meeden, G. (1986), "Empirical Bayes Estimation in Finite Population Sampling," *J. Amer. Stat. Assoc.*, 81, 1058–1062.

Harville, D. A. (1974), "Bayesian Inference for Variance Components Using Only Error Contrasts," *Biometrika*, 61, 383–385.

———— (1985), "Decomposition of Prediction Error," *J. Amer. Stat. Assoc.*, 80, 132–138.

———— (1989), "BLUP (Best Linear Unbiased Prediction) and Beyond," in *Advances in Statistical Methods for Genetic Improvement of Livestock*, eds. D. Gianola and K. Hammond, New York: Springer-Verlag.

Harville, D. A., and Callanan, T. P. (1989), "Computational Aspects of Likelihood-Based Inference for Variance Components," in *Advances in Statistical Methods for Genetic Improvement of Livestock*, eds. D. Gianola and K. Hammond, New York: Springer-Verlag.

Harville, D. A., and Fenech, A. P. (1985), "Confidence Intervals for a Variance Ratio, or for Heritability, in an Unbalanced Mixed Linear Model," *Biometrics*, 41, 137–152.

Hulting, F. L., and Harville, D. A. (1989), "Some Bayesian and Non-Bayesian Procedures for the Analysis of Comparative Experiments and for Small-Area Estimation: Computational Aspects, Frequentist Properties, and Relationships," Preprint 89-10, Department of Statistics, Iowa State University, Ames, Iowa.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press: London.

Jeske, D. R., and Harville, D. A. (1988), "Prediction - Interval Procedures and (Fixed-Effects) Confidence-Interval Procedures for Mixed Linear Models," *Communications in Statistics – Theory and Methods*, 17, 1053–1087.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *J. Amer. Stat. Assoc.*, 79, 853–862.

Khatri, C. G., and Shah, K. R. (1981), "On the Unbiased Estimation of Fixed Effects in a Mixed Model for Growth Curves," *Communications in Statistics – Theory and Methods*, 10, 401–406.

Laird, N., and Louis, T. A. (1987), "Empirical Bayes Confidence Intervals Based on Bootstrap Samples," *J. Amer. Stat. Assoc.*, 82, 739–750.

Morris C. N. (1983), "Parametric Empirical Bayes Confidence Intervals," in *Scientific Inference, Data Analysis, and Robustness*, New York: Academic Press, pp. 25–50.

Patterson, H. D., and Thompson, R. (1971), "Recovery of Interblock Information when Block Sizes are Unequal," *Biometrika*, 58, 545–554.

Prasad, N. G. N., and Rao, J. N. K. (1986), "On the Estimation of Mean Square Error of Small Area Predictors," in *1986 Proceedings of the Section on Survey Research Methods, American Statistical Association*, 108–116.

Rubin, D. B. (1982), "Estimation in Parallel Randomized Experiments," *Journal of Educational Statistics*, 6, 377–401.

Sallas, W. M., and Harville, D. A. (1981), "Best Linear Recursive Estimation for Mixed Linear Models," *J. Amer. Stat. Assoc.*, 76, 860–869.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110-114.

Stroud T. W. F. (1987), "Bayes and Empirical Bayes Approaches to Small Area Estimation," in *Small Area Statistics: An International Symposium*, eds. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, New York: John Wiley.