The Annals of Statistics
1991, Vol. 19, No. 4, 1748–1770

BAYESIAN PREDICTION IN LINEAR MODELS: APPLICATIONS TO SMALL AREA ESTIMATION[1]

BY GAURI SANKAR DATTA AND MALAY GHOSH

University of Georgia and University of Florida

This paper introduces a hierarchical Bayes (HB) approach for prediction in general mixed linear models. The results find application in small area estimation. Our model unifies and extends a number of models previously considered in this area. Computational formulas for obtaining the Bayes predictors and their standard errors are given in the general case. The methods are applied to two actual data sets. Also, in a special case, the HB predictors are shown to possess some interesting frequentist properties.

1. Introduction. It has been some time now that the government agencies in the United States, Canada and elsewhere have recognized the importance of small area estimation. Estimation of this type is particularly well suited in a setting that involves several areas (or strata) with a small number of samples available from each individual stratum. The estimates of the parameters of interest (like the mean, variance, etc.) for these areas can profitably "borrow strength" from other neighboring areas.

The appropriateness of model-based inference for small area estimation is widely recognized. We may refer to Fay and Herriot (1979), Ghosh and Meeden (1986), Ghosh and Lahiri (1987), Battese, Harter and Fuller (1988), Prasad and Rao (1990), Choudhry and Rao (1988), Royall (1978) and Lui and Cumberland (1989), among others. The methods that have usually been proposed use either a variance components approach or an empirical Bayes (EB) approach, although the distinction between the two is often superfluous [Harville (1988, 1990)]. Both these procedures use certain mixed linear models for prediction purposes. First, assuming the variance components are known, certain best linear unbiased predictors (BLUPs) or EB predictors are obtained for the unknown parameters of interest. Then the unknown variance components are estimated typically by Henderson's method of fitting of constants or the restricted maximum likelihood (REML) method, and the resulting estimated BLUPs (also referred to as empirical BLUPs) are used for final prediction.

Received March, 1989; revised November 1990.

[1]The second author's research was partially supported by NSF Grants DMS-87-01814 and DMS-89-01334. Part of this work was completed when the second author was an ASA Senior Research Fellow at the Bureau of the Census and the Bureau of Labor Statistics, while the first author was an ASA/NSF/Census Research Associate.

AMS 1980 *subject classifications*. 62D05, 62F11, 62F15, 62J99.

Key words and phrases. Hierarchical Bayes, empirical Bayes, mixed linear models, best linear unbiased prediction, best unbiased prediction, small area estimation, nested error regression model, random regression coefficients model, two-stage sampling, elliptically symmetric distributions.

Although the above approach is usually quite satisfactory for point prediction, it is very difficult to estimate the standard errors associated with these predictors. This is primarily due to the lack of closed-form expressions for the mean squared errors (MSEs) of the estimated BLUPs. Kackar and Harville (1984) suggested an approximation to the MSEs [see also Harville (1985, 1988, 1990) and Harville and Jeske (1989)]. Prasad and Rao (1990) proposed estimates of these approximate MSEs in three specific mixed linear models. The work of Prasad and Rao (1990) suggests that their approximations work well when the number of small areas is sufficiently large. It is not clear though how these approximations fare for a small or even a moderately large number of strata.

Ghosh and Lahiri (1989) proposed a hierarchical Bayes (HB) procedure as an alternative to the estimated BLUP or the EB procedure. In a HB procedure, if one uses the posterior mean for estimating the parameter of interest, then a natural estimate of the standard error associated with this estimator is the posterior s.d. The estimate, though often complicated, can be found exactly via numerical integration without any approximation.

The model considered by Ghosh and Lahiri (1989) was, however, only a special case of the so-called "nested error regression model." A similar model was considered by Stroud (1987), but his general analysis was performed only for the balanced case, that is, when the number of samples was the same for each stratum. Battese, Harter and Fuller (1988) first considered the nested error regression model in the context of small area estimation and performed a variance components analysis.

The objective of this article is to present a unified Bayesian prediction theory for mixed linear models with particular emphasis on small area estimation. A general Bayesian normal theory model is presented in Section 2 which can be regarded as an extension of the HB ideas of Lindley and Smith (1972) to prediction. Most of the models considered by earlier authors can be regarded as special cases of our model, and certain specific illustrations are provided. Also, in this section, we have provided in a very general framework the posterior distribution as well as the resulting posterior means and variances of the unobserved population units given the sampled units. The proof of the main result of this section is given in the Appendix. For nonnormal HB analysis, one may refer to Albert (1988) or Morris (1988).

In Section 3, we discuss the computational issues related to the estimation of parameters of interest with particular emphasis on the estimation of population means simultaneously for several small areas. Closed-form expressions cannot usually be obtained for the posterior means and s.d.'s of such parameters, and numerical integration becomes a necessity. For very high dimensional integrals, direct numerical integration is often unreliable, and sometimes even impossible to execute, and some of the recently advocated Monte Carlo integration techniques may be of help. We shall indicate in Section 3 how the Gibbs sampling technique introduced by Geman and Geman (1984), and more recently popularized by Gelfand and Smith (1990), works in some important special cases of our general framework. The related substitu-

tion sampling algorithm of Tanner and Wong (1987) and the traditional importance sampling technique will also be discussed very briefly.

However, in small dimensions, it is often easier to perform direct numerical integration than to use any Monte Carlo numerical integration method. For instance, if the integrand is a very complicated function and cannot be approximated very accurately by a simple smooth function, the importance sampling technique can at best result in a slow convergence of the desired integral. The Gibbs sampling is usually very slow, and for evaluation of small dimensional integrals, any simplicity of this approach cannot adequately compensate for the enormous computing time needed for the method's successful execution.

For the sake of illustration of our methods, we have thus used in Section 4 direct numerical integration methods for data analysis. Two examples are considered in this paper. The first example given in Section 4.1 requires numerical evaluation of two-dimensional integrals, while the second given in Section 4.2 requires evaluation of one-dimensional integrals. The data set considered in the first example pertains to the Patterns of Care Studies, a study involving the quality of treatment received by cancer patients having radiation therapy as the primary mode of treatment. The present data form a subset of a much larger data set analyzed in Calvin and Sedransk (1991). We have considered a stratified finite population from which samples are drawn in two stages using simple random sampling at each stage. The HB estimator of the population mean is compared with an EB estimator proposed in Ghosh and Lahiri (1988), a design unbiased estimator given in Cochran (1977), page 303, an expansion estimator, a ratio type estimator and another estimator proposed in Royall (1976). The HB estimator has the smallest average mean squared error among these six and the improvement over all but the EB estimator is quite substantial.

The second example is related to the prediction of areas under corn and soybeans for 12 counties in North Central Iowa. The problem was originally considered by Battese, Harter and Fuller (1988) using a variance components method. We have used this example to illustrate how a naive EB approach can sometimes grossly underestimate the associated standard error of an EB estimator. In this particular example, the posterior s.d.'s as obtained by us are slightly smaller than the ones of Battese, Harter and Fuller.

In Section 5, we have considered a special case of the general HB model and have provided the posterior distribution of the unobserved population units given the sampled units. In this special case, the HB predictors of the linear parameters of interest are shown to be the best within the class of all linear unbiased predictors under the assumption of finiteness of second moments. For a class of spherically symmetric distributions including but not limited to the normal, the HB predictors are shown to be optimal within the class of all unbiased predictors. Optimality properties of this type extend the earlier work of Henderson (1963) and others on the prediction of real-valued parameters to the prediction of vector-valued parameters. The proof of the main result of Section 2 is deferred to the Appendix.

2. The description and analysis of the HB model. Consider the following Bayesian model:

(A) Conditional on $\mathbf{b} = (b_1, \ldots, b_p)^T$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t)^T$ and r, let

$$\mathbf{Y} \sim N\big(\mathbf{Xb}, r^{-1}(\boldsymbol{\Psi} + \mathbf{ZD}(\boldsymbol{\lambda})\mathbf{Z}^T)\big),$$

where $\mathbf{Y}$ is $N \times 1$.

(B) $\mathbf{B}$, R and $\boldsymbol{\Lambda}$ have a certain joint prior distribution proper or improper.

Stage (A) of the model can be identified as a general mixed linear model. To see this, write

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zv} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{e}$ and $\mathbf{v}$ are mutually independent, with $\mathbf{e} \sim N(\mathbf{0}, r^{-1}\boldsymbol{\Psi})$ and $\mathbf{v} \sim N(\mathbf{0}, r^{-1}\mathbf{D}(\boldsymbol{\lambda}))$, where $\mathbf{e}$ is $N \times 1$ and $\mathbf{v}$ is $q \times 1$; in the above $\mathbf{X}$ $(N \times p)$ and $\mathbf{Z}$ $(N \times q)$ are known design matrices, $\boldsymbol{\Psi}$ is a known positive definite (p.d.) matrix, while $\mathbf{D}(\boldsymbol{\lambda})$ $(q \times q)$ is a p.d. matrix which is structurally known except possibly for some unknown $\boldsymbol{\lambda}$. In the examples to follow, $\boldsymbol{\lambda}$ involves the ratios of the variance components. Sometimes we will denote $\mathbf{D}(\boldsymbol{\lambda})$ by $\mathbf{D}$ when $\boldsymbol{\lambda}$ is known.

In the context of small area estimation, partition $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{e}$, and rewrite the model given in (2.1) as

$$\begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix}\mathbf{b} + \begin{pmatrix} \mathbf{Z}^{(1)} \\ \mathbf{Z}^{(2)} \end{pmatrix}\mathbf{v} + \begin{pmatrix} \mathbf{e}^{(1)} \\ \mathbf{e}^{(2)} \end{pmatrix}, \tag{2.2}$$

where $\mathbf{Y}^{(1)}$ and $\mathbf{e}^{(1)}$ are $n \times 1$, $\mathbf{X}^{(1)}$ is $n \times p$ and $\mathbf{Z}^{(1)}$ is $n \times q$. Also, $\mathbf{Y}^{(2)}$ and $\mathbf{e}^{(2)}$ are $(N - n) \times 1$, $\mathbf{X}^{(2)}$ is $(N - n) \times p$ and $\mathbf{Z}^{(2)}$ is $(N - n) \times q$. We assume for simplicity that $\operatorname{rank}(\mathbf{X}^{(1)}) = p$.

In the above $\mathbf{Y}^{(1)}$ is the vector of sampled units from m small areas, while $\mathbf{Y}^{(2)}$ is the vector of unsampled units. It is possible to partition $\mathbf{Y}^{(1)T}$ into $\mathbf{Y}^{(1)T} = (\mathbf{Y}_1^{(1)T}, \ldots, \mathbf{Y}_m^{(1)T})$, where $\mathbf{Y}_i^{(1)}(n_i \times 1)$ is the vector of sampled units for the ith small area. Similarly, $\mathbf{Y}^{(2)T}$ can be partitioned as $\mathbf{Y}^{(2)T} = (\mathbf{Y}_1^{(2)T}, \ldots, \mathbf{Y}_m^{(2)T})$, where $\mathbf{Y}_i^{(2)}((N_i - n_i) \times 1)$ is the vector of unsampled units for the ith small area.

Following the model-based approach in survey sampling, one of the primary objectives of this paper is to find the conditional (predictive) distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. The analysis will be done in two stages. In the latter part of this section, we derive the predictive distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$ putting independent uniform prior distributions on $\mathbf{B}$ and gamma distributions on $R, \Lambda_1 R, \ldots, \Lambda_t R$.

Before finding the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$, we identify some of the existing models introduced for small area estimation by several authors as special cases of (2.2). In what follows, we shall use the notation $\mathbf{I}_u$ for an identity matrix of order u, $\mathbf{1}_u$ for a u-component column vector with each element equal to 1 and $\mathbf{J}_u = \mathbf{1}_u\mathbf{1}_u^T$. Also, let $\operatorname{col}_{1 \le i \le p}(\mathbf{B}_i)$ denote the matrix $(\mathbf{B}_1^T, \ldots, \mathbf{B}_p^T)^T$ and let $\oplus_{i=1}^p \mathbf{A}_i$ denote the matrix $\begin{bmatrix} \mathbf{A}_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{A}_p \end{bmatrix}$.

First, consider the nested error regression model

$$Y_{ij} = \mathbf{x}_{ij}^T\mathbf{b} + v_i + e_{ij}, \qquad j = 1, \ldots, N_i,\ i = 1, \ldots, m. \tag{2.3}$$

The model was considered by Battese, Harter and Fuller (1988). They assumed the v_i's and e_{ij}'s to be mutually independent with v_i's iid $\mathrm{N}(0, (\lambda r)^{-1})$, and e_{ij}'s iid $N(0, r^{-1})$. In this case, $\mathbf{X}^{(1)} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{1 \le j \le n_i}(\mathbf{x}_{ij}^T))$, $\mathbf{X}^{(2)} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{n_i+1 \le j \le N_i}(\mathbf{x}_{ij}^T))$, $\mathbf{Z}^{(1)} = \oplus_{i=1}^m \mathbf{1}_{n_i}$ and $\mathbf{Z}^{(2)} = \oplus_{i=1}^m \mathbf{1}_{N_i - n_i}$, $\mathbf{\Psi} = \mathbf{I}_N$, $t = 1$, $\boldsymbol{\lambda} = \lambda$ and $\mathbf{D}(\boldsymbol{\lambda}) = \lambda^{-1}\mathbf{I}_m$. In the further special case of Ghosh and Lahiri (1989), $\mathbf{x}_{ij} = \mathbf{x}_i$ for every $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Note that $\lambda = V(e_{ij})/V(v_i)$, a ratio of the variance components.

The random regression coefficients model of Dempster, Rubin and Tsutakawa (1981) [see also Prasad and Rao (1990)] is also a special case of ours. In this setup, $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{\Psi}$ and $\mathbf{D}(\boldsymbol{\lambda})$ are the same as in the nested error regression model, but

$$\mathbf{Z}^{(1)} = \bigoplus_{i=1}^{m} \left[\mathrm{col}_{1 \le j \le n_i} \mathbf{x}_{ij}^T\right], \qquad \mathbf{Z}^{(2)} = \bigotimes_{i=1}^{m} \left[\mathrm{col}_{n_i+1 \le j \le N_i} \mathbf{x}_{ij}^T\right].$$

The models given in Choudhry and Rao (1988) are special cases of our general model as well.

It is possible also to include certain cross-classification models as special cases of our general linear model. For example, suppose there are m small areas labeled $1, \ldots, m$. Within each small area, units are further classified into c subgroups (socioeconomic class, age, etc.) labeled $1, \ldots, c$. The cell sizes N_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, c$, are assumed to be known. Let Y_{ijk}, $k = 1, \ldots, N_{ij}$, denote the measurement on the kth individual in the (i, j)th cell. Conditional on $\mathbf{b}$, r and $\boldsymbol{\lambda}$, suppose

$$\begin{aligned} Y_{ijk} = \mathbf{x}_{ij}^T\mathbf{b} + \tau_i + \eta_j + \gamma_{ij} + e_{ijk}, \\ k = 1, \ldots, N_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, c, \end{aligned} \tag{2.4}$$

with τ_i's, η_j's, γ_{ij}'s and e_{ijk}'s mutually independent with e_{ijk}'s iid $N(0, r^{-1})$, γ_{ij}'s iid $N(0, (\lambda_3 r)^{-1})$, η_j's iid $N(0, (\lambda_2 r)^{-1})$ and τ_i's iid $N(0, (\lambda_1 r)^{-1})$. Special cases of this model have been considered by several authors. Lui and Cumberland (1989) [also Royall (1978)] considered a model where τ_i's and γ_{ij}'s are degenerate at zeros. Also, they assumed the variance ratio λ_2 to be known in deriving their estimators and did not address the issue of unknown λ_2 appropriately.

Next we show that the two-stage sampling model with covariates and m strata is a special case of our general linear model. Suppose that the ith stratum contains L_i primary units. Suppose also that the jth primary unit within the ith stratum contains N_{ij} subunits. Let Y_{ijk} denote the value of the characteristic of interest for the kth subunit within the jth primary unit from the ith stratum ($k = 1, \ldots, N_{ij}$, $j = 1, \ldots, L_i$, $i = 1, \ldots, m$). From the ith stratum, a sample of l_i primary units is taken. For the jth selected primary unit within the ith stratum, a sample of n_{ij} subunits are selected. Without

loss of generality, the sample values are denoted by Y_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, l_i$, $i = 1, \ldots, m$.

Assume conditional on $\mathbf{b}$, r and $\boldsymbol{\lambda}$:

$$Y_{ijk} = \mathbf{x}_{ij}^T \mathbf{b} + \xi_i + \eta_{ij} + e_{ijk}, \qquad k = 1, \ldots, N_{ij},\ j = 1, \ldots, L_i,\ i = 1, \ldots, m, \tag{2.5}$$

where ξ_i's, η_{ij}'s and e_{ijk}'s are mutually independent with ξ_i's iid $N(0, (\lambda_1 r)^{-1})$, η_{ij}'s iid $N(0, (\lambda_2 r)^{-1})$, e_{ijk}'s iid $N(0, r^{-1})$. Let

$$\mathbf{Y}^{(1)} = \operatorname*{col}_{1 \le i \le m} \left[\operatorname*{col}_{1 \le j \le l_i} \left\{ \operatorname*{col}_{1 \le k \le n_{ij}} (Y_{ijk}) \right\} \right],$$

$$\mathbf{Y}^{(2)} = \operatorname*{col}_{1 \le i \le m} \left[\operatorname*{col}_{1 \le j \le L_i} \left\{ \operatorname*{col}_{u_{ij} \le k \le N_{ij}} (Y_{ijk}) \right\} \right],$$

$$u_{ij} = 1 + n_{ij} I_{[j \le l_i]},\ i = 1, \ldots, m.$$

$$\mathbf{v} = \left(\mathbf{s}^T \mathbf{w}_1^T \mathbf{w}_2^T\right)^T, \qquad \mathbf{s} = \operatorname*{col}_{1 \le i \le m} (\xi_i), \qquad \mathbf{w}_1 = \operatorname*{col}_{1 \le i \le m} \left(\operatorname*{col}_{1 \le j \le l_i} (\eta_{ij}) \right)$$

and

$$\mathbf{w}_2 = \operatorname*{col}_{1 \le i \le m} \left(\operatorname*{col}_{l_i + 1 \le j \le L_i} (\eta_{ij}) \right).$$

Also, let $\mathbf{e}^{(i)}$ be defined similarly as $\mathbf{Y}^{(i)}$, $i = 1, 2$. Then (2.5) can be written as (2.2) with appropriately defined $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Note that here $t = 2$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, $\boldsymbol{\Psi} = \mathbf{I}_N$, $\mathbf{D}(\boldsymbol{\lambda}) = \operatorname{Diag}(\lambda_1^{-1} \mathbf{I}_m, \lambda_2^{-1} \mathbf{I}_L.)$ with $N = \sum_{i=1}^m \sum_{j=1}^{L_i} N_{ij}$. The ideas can be extended directly to multistage sampling. We may mention here that Bayesian analysis for two-stage sampling was introduced first by Scott and Smith (1969) in a much simpler framework. A multistage analog of their work was provided by Malec and Sedransk (1985).

Next, in this section, we provide the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. The following nomenclature will be used to label certain known distributions. A random variable Z is said to have a Gamma(α, β) distribution if it has pdf $f(z) = [\exp(-\alpha z) \alpha^\beta z^{\beta - 1} / \Gamma(\beta)] I_{[z > 0]}$. A random vector $\mathbf{T} = (T_1, \ldots, T_p)^T$ is said to have a multivariate t-distribution with location parameter $\boldsymbol{\mu}$, scale parameter $\boldsymbol{\Phi}$ and degrees of freedom ν if it has pdf

$$g(\mathbf{t}) \propto |\boldsymbol{\Phi}|^{-1/2} \left[\nu + (\mathbf{t} - \boldsymbol{\mu})^T \boldsymbol{\Phi}^{-1} (\mathbf{t} - \boldsymbol{\mu}) \right]^{-(\nu + p)/2} \tag{2.6}$$

[see Zellner (1971) page 383, or Press (1972) page 136]. Assume $\nu > 2$. Then $E(\mathbf{T}) = \boldsymbol{\mu}$, $V(\mathbf{T}) = (\nu/(\nu - 2))\boldsymbol{\Phi}$.

We assume condition (A) given at the beginning of this section. In stage (B) of the model, it is assumed that

$$\mathbf{B}, R, \Lambda_1 R, \ldots, \Lambda_t R \text{ are independently distributed} \tag{2.7}$$

with $\mathbf{B} \sim$ uniform(R^p), $R \sim$ Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$, $a_0 \ge 0$, $g_0 \ge 0$, $\Lambda_i R \sim$ Gamma$(\frac{1}{2}a_i, \frac{1}{2}g_i)$, $i = 1, \ldots, t$, with $a_i > 0$, $g_i \ge 0$, $i = 1, \ldots, t$. In this way, some improper gamma distributions are included as a possibility in our prior.

Before stating the main result of this section we need to introduce additional notation. We write $\Sigma \equiv \Sigma(\lambda) = \Psi + \mathbf{Z}\mathbf{D}(\lambda)\mathbf{Z}^T$, partition Σ into $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and define $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

Also, let

$$\mathbf{K} = \Sigma_{11}^{-1} - \Sigma_{11}^{-1}\mathbf{X}^{(1)}\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}, \tag{2.8}$$

$$\mathbf{M} = \Sigma_{21}\mathbf{K} + \mathbf{X}^{(2)}\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}, \tag{2.9}$$

$$\begin{aligned}\mathbf{G} = \Sigma_{22.1} &+ \left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1} \\ &\times\left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{T}.\end{aligned} \tag{2.10}$$

The posterior distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ is given in the following theorem in two steps.

THEOREM 1. *Consider the model given in* (2.1) [*or* (2.2)] *and* (2.7). *Assume that* $n + \sum_{i=0}^{t} g_i - p > 2$. *Then, conditional on* $\Lambda = \lambda$ *and* $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$, $\mathbf{Y}^{(2)}$ *is distributed as multivariate-t with degrees of freedom* $n + \sum_{i=0}^{t} g_i - p$, *location parameter* $\mathbf{M}\mathbf{y}^{(1)}$ *and scale parameter*

$$\left(n + \sum_{i=0}^{t} g_i - p\right)^{-1}\left[a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right]\mathbf{G}.$$

Also, the conditional distribution of Λ *given* $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ *has pdf*

$$\begin{aligned}f(\lambda|\mathbf{y}^{(1)}) \propto |\Sigma_{11}|^{-1/2}&\left|\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right|^{-1/2}\left[\prod_{i=1}^{t}\lambda_i^{gi/2-1}\right] \\ &\times\left[a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right]^{-[n+\sum_{i=0}^{t} g_i - p]/2}.\end{aligned} \tag{2.11}$$

The proof of Theorem 1 is deferred to the Appendix. Using the moments of a multivariate-t distribution, it follows now that if $n + \sum_{i=0}^{t} g_i > p + 2$, then

$$E\left[\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}\right] = E_{\cdot}(\mathbf{M}|\mathbf{y}^{(1)})\mathbf{y}^{(1)}, \tag{2.12}$$

$$\begin{aligned}V\left[\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}\right] = V(\mathbf{M}\mathbf{y}^{(1)}|\mathbf{y}^{(1)}) &+ \left(n + \sum_{i=0}^{t} g_i - p - 2\right)^{-1} \\ &\times E\left[\left\{a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right\}\mathbf{G}\middle|\mathbf{y}^{(1)}\right].\end{aligned} \tag{2.13}$$

Using (2.12) and (2.13), it is possible to find the posterior means and variances of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \mathbf{A}\mathbf{Y}^{(1)} + \mathbf{C}\mathbf{Y}^{(2)}$, where $\mathbf{A}$ and $\mathbf{C}$ are known matrices. The Bayes estimate of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ under any quadratic loss is its posterior

mean, and is given by

$$\mathbf{e}_B(\mathbf{y}^{(1)}) = \left[\mathbf{A} + \mathbf{C}E(\mathbf{M}|\mathbf{y}^{(1)})\right]\mathbf{y}^{(1)}, \tag{2.14}$$

using (2.12). Similarly, using (2.13), one may obtain

$$V\left[\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{y}^{(1)}\right] = \mathbf{C}V(\mathbf{Y}^{(2)}|\mathbf{y}^{(1)})\mathbf{C}^T. \tag{2.15}$$

Note that when $\mathbf{A} = \oplus_{i=1}^{m}\mathbf{1}_{n_i}^T$ and $\mathbf{C} = \oplus_{i=1}^{m}\mathbf{1}_{N_i-n_i}^{\mathrm{T}}$, $\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ reduces to the vector of population totals for the m small areas. Computational issues related to the simultaneous estimation of several small area totals will be addressed in Section 3.

3. Numerical computations. It is evident from Theorem 1 that the conditional distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)}$ cannot usually be obtained analytically because of the complicated posterior pdf of $\boldsymbol{\Lambda}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ [see (2.11)]. As mentioned in the Introduction, Monte Carlo numerical integration is a distinct possibility, particularly when the dimension of $\boldsymbol{\lambda}$ is large. One may think of the importance sampling method as a natural candidate for such purposes. To implement such a procedure, we write $f(\boldsymbol{\lambda}|\mathbf{y}^{(1)})$ given in (2.11) as $f(\boldsymbol{\lambda}|\mathbf{y}^{(1)}) = ck(\boldsymbol{\lambda}, \mathbf{y}^{(1)})$, where the norming constant c has to be numerically evaluated. Now, for any real-valued function $h(\boldsymbol{\lambda})$,

$$\int_0^\infty \cdots \int_0^\infty h(\boldsymbol{\lambda})\, f(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\, d\boldsymbol{\lambda}$$

$$= \frac{\int_0^\infty \cdots \int_0^\infty h(\boldsymbol{\lambda})\left\{k(\boldsymbol{\lambda}, \mathbf{y}^{(1)})/g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\right\}g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\, d\boldsymbol{\lambda}}{\int_0^\infty \cdots \int_0^\infty \left\{k(\boldsymbol{\lambda}, \mathbf{y}^{(1)})/g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\right\}g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\, d\boldsymbol{\lambda}},$$

where $g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})$ is some "standard" pdf from which a random sample can easily be generated. Hence $\int_0^\infty \cdots \int_0^\infty h(\boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\mathbf{y}^{(1)})\, d\boldsymbol{\lambda}$ can be approximated by

$$\frac{\sum_{i=1}^{s} h(\boldsymbol{\lambda}^{(i)})\left\{k(\boldsymbol{\lambda}^{(i)}, \mathbf{y}^{(1)})/g(\boldsymbol{\lambda}^{(i)}|\mathbf{y}^{(1)})\right\}}{\sum_{i=1}^{s} k(\boldsymbol{\lambda}^{(i)}, \mathbf{y}^{(1)})/g(\boldsymbol{\lambda}^{(i)}|\mathbf{y}^{(1)})},$$

where the number of replicates is very large, and $\boldsymbol{\lambda}^{(i)}$'s are generated from $g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})$.

Unfortunately, finding $g(\boldsymbol{\lambda}|\mathbf{y}^{(1)})$ in the present context can be quite formidable. Even when $\boldsymbol{\lambda}$ is one-dimensional, $f(\boldsymbol{\lambda}|\mathbf{y}^{(1)})$ may turn out to be multimodal, and thus defy any simple approximation. One such example appears in Ghosh and Rao (1991). In such circumstances, it is natural to seek other Monte Carlo integration methods.

The recently advertised Gibbs sampler bears some interesting promise, at least in the special case when $\boldsymbol{\Psi} = \mathbf{I}_N$ and $\mathbf{D}(\boldsymbol{\lambda}) = \mathrm{Diag}(\lambda_1^{-1}\mathbf{I}_{q_1}, \ldots, \lambda_t^{-1}\mathbf{I}_{q_t})$, where $\sum_{i=1}^{t} q_i = q$. We shall write $W_i = R\Lambda_i$, and correspondingly $w_i = r\lambda_i$, $i = 1, \ldots, t$. We assign a uniform (R^p) prior for $\mathbf{B}$, a Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$ prior for R and Gamma$(\frac{1}{2}a_i, \frac{1}{2}g_i)$ priors for the W_i's, where $\mathbf{B}, R, W_1, \ldots, W_t$ are all independently distributed.

We shall write $\mathbf{v}^T = (\mathbf{v}_1^T, \ldots, \mathbf{v}_t^T)$, where $\mathbf{v}_i$ has dimension q_i. Based on the model introduced at the beginning of Section 2, the joint pdf of $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{B}, \mathbf{v}, R, W_1, \ldots, W_t$ is

$$
\begin{aligned}
&f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, r, w_1, \ldots, w_t) \\
&\quad \propto r^{n/2} \exp\Big[-\tfrac{1}{2}r\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\mathbf{b} - \mathbf{Z}^{(1)}\mathbf{v}\|^2\Big] r^{(N-n)/2} \\
&\qquad \times \exp\Big[-\tfrac{1}{2}r\|\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\mathbf{b} - \mathbf{Z}^{(2)}\mathbf{v}\|^2\Big] \\
&\qquad \times \prod_{i=1}^{t} \Big\{w_i^{q_i/2} \exp\big(-\tfrac{1}{2}w_i\|\mathbf{v}_i\|^2\big)\Big\} \exp\big(-\tfrac{1}{2}a_0 r\big) r^{g_0/2-1} \\
&\qquad \times \prod_{i=1}^{t} \big\{\exp\big(-\tfrac{1}{2}a_i w_i\big) w_i^{g_i/2-1}\big\}.
\end{aligned} \tag{3.1}
$$

Then the required conditional distributions are given by

$$
\mathbf{B}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{v}, r, w_1, \ldots, w_t \sim N\Big[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{Z}\mathbf{v}), r^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\Big], \tag{3.2}
$$

$$
\begin{aligned}
\mathbf{v}|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, r, w_1, \ldots, w_t \sim N\Bigg[&\Bigg(\mathbf{Z}^T\mathbf{Z} + \bigoplus_{l=1}^{t} r^{-1}w_l\mathbf{I}_{q_l}\Bigg)^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\mathbf{b}), \\
&r^{-1}\Bigg(\mathbf{Z}^T\mathbf{Z} + \bigoplus_{l=1}^{t} r^{-1}w_l\mathbf{I}_{q_l}\Bigg)^{-1}\Bigg],
\end{aligned} \tag{3.3}
$$

$$
\begin{aligned}
&R|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, w_1, \ldots, w_t \\
&\qquad \sim \text{Gamma}\Big(\tfrac{1}{2}\big\{\|\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{v}\|^2 + a_0\big\}, \tfrac{1}{2}(N + g_0)\Big),
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
&W_i|\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{b}, \mathbf{v}, r, w_j, j \neq i \\
&\qquad \sim \text{Gamma}\Big(\tfrac{1}{2}\big(\|\mathbf{v}_i\|^2 + a_i\big), \tfrac{1}{2}(q_i + g_i)\Big), \qquad i = 1, \ldots, t,
\end{aligned} \tag{3.5}
$$

$$
\mathbf{Y}^{(2)}|\mathbf{y}^{(1)}, \mathbf{b}, \mathbf{v}, r, w_1, \ldots, w_t \sim N\big(\mathbf{X}^{(2)}\mathbf{b} + \mathbf{Z}^{(2)}\mathbf{v}, r^{-1}\mathbf{I}_{N-n}\big). \tag{3.6}
$$

Gelfand and Smith (1990) have pointed out that it suffices to know (3.2)–(3.6) to find the joint distribution of $\mathbf{Y}^{(2)}, \mathbf{B}, \mathbf{v}, R, W_1, \ldots, W_t$ conditional on $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. Also, they have provided the recipe of finding the Monte Carlo approximation to the posterior pdf of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ on the basis of these conditional distributions. However, the procedure requires $p + q + 1 + t + N - n$ random variate generations to complete a cycle. If we run m sequences out to the ith iteration, a total of $mi\,(p + q + 1 + t + N - n)$ random variate generations are needed, and we need a great deal of total computing time. The substitution algorithm of Tanner and Wong (1987) requires even $(p + q + 1 + t + N - n)\,(p + q + t + N - n)$ random variate generations to complete a cycle in as much as other conditional distributions involving subsets of the random variables given in (3.2)–(3.6) are needed. Clearly, if the dimension of λ

is small, it is much simpler to execute direct numerical integration using one of the available packages. To carry out direct numerical integration, we have written our programs in the FORTRAN language, and have used the IMSL version 9.2 subroutine packages. A microvax computer was available for execution of our programs.

4. Data analysis. We now turn to the actual data analysis. The first set of data relates to the quality of radiation therapy care for cancer patients, while the second set of data relates to the prediction of areas under corn and soybeans for 12 counties in North Central Iowa.

4.1. *Radiation therapy data.* The data were collected with the primary objective of comparing the quality of radiation therapy for cancer patients among subpopulations of a population of facilities where radiation therapy was practiced. We have, however, used the data primarily for the comparison of several estimators of the finite population mean when two-stage sampling is performed. Our finite population of units is actually the sample units arising from a 1978 survey of patients suffering from cervical cancer. For conducting this survey, radiation therapy facilities were grouped into several strata that were thought to be relatively homogeneous in the quality of care that patients received. The five strata considered in this paper correspond to strata 1, 2, 4, 5 and 6 of Calvin and Sedransk (1991) who have provided a more detailed description of what these strata actually are. The number of facilities contained in these five strata are 10, 15, 11, 30 and 11, respectively, and are treated as primary sampling units (PSUs). Among these PSUs, we have selected a $\frac{1}{3}$ simple random sample resulting in the selection of 3, 5, 4, 10 and 4 PSUs from the five strata. From each selected PSU, with p patient records, a simple random sample of size $[\frac{1}{2}(p + 1)]$ is selected, where $[u]$ denotes the integer part of u.

The present analysis considers "pretreatment" scores for each patient. For a given patient, for each disease site, a committee of experts identified a set of services and procedures (S/P's) that were thought to be of prime importance for a complete pretreatment evaluation and for planning and monitoring therapy. The committee also assigned weights (0.5 to 4.0) to these S/P's to indicate their relative importance. Then, for each patient, a score is defined by $\sum_i W_i^* Z_i / \sum_i W_i^*$, where $Z_i = 1$ if the ith S/P is performed, while $Z_i = 0$ otherwise; W_i^* is the corresponding weight. The larger the score, the closer the patient's care conforms to acceptable standards of care.

Let Y_{ijk} denote the score for the kth patient in the jth facility within the ith stratum. Although the Y_{ijk}'s lie between 0 and 1, these are weighted averages of independent Bernoulli variables, and a normal approximation due to the CLT is not totally out of the way.

We assume the model given in (2.5) with $\mathbf{b} = \mu$, the general effect, and $\mathbf{x}_{ij} = 1$. As described in Section 2, from the ith stratum, a sample of l_i $(< L_i)$ primary units is taken, while for the jth selected primary unit within the ith stratum, a sample of n_{ij} $(< N_{ij})$ subunits are selected. We denote the sample

observations by Y_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, l_i$, $i = 1, \ldots, 5$. Also, let $\mathbf{y}^{(1)}$ be the vector of sample observations, $\bar{y}_{ij} = n_{ij}^{-1}\sum_{k=1}^{n_{ij}} y_{ijk}$, $B_{ij} = \lambda_2/(\lambda_2 + n_{ij})$, $\bar{y}_i = \sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_{ij}/\sum_{j=1}^{l_i}(1 - B_{ij})$, $\alpha_i = \lambda_1/(\lambda_1 + \lambda_2\sum_{j=1}^{l_i}(1 - B_{ij}))$, $\bar{y} = \sum_{i=1}^{5}(1 - \alpha_i)\bar{y}_i/\sum_{i=1}^{5}(1 - \alpha_i)$, $f_{ij} = (N_{ij} - n_{ij})/N_{ij}$. Then the HB predictor of $\gamma_i = \sum_{j=1}^{L_i}\sum_{k=1}^{N_{ij}} Y_{ijk}/\sum_{j=1}^{L_i} N_{ij}$, the population mean for the ith stratum, is given by

$$
\begin{aligned}
e_{\mathrm{HB}}^{i} = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1} E\Bigg[& \sum_{j=1}^{l_i} N_{ij}(1 - f_{ij}B_{ij})\bar{y}_{ij} \\
& + \left\{\left(\sum_{j=l_i+1}^{L_i} N_{ij}\right) + \sum_{j=1}^{l_i} N_{ij} f_{ij} B_{ij}\right\} \\
& \times\{(1 - \alpha_i)\bar{y}_i + \alpha_i\bar{y}\}\Big|\mathbf{y}^{(1)}\Bigg].
\end{aligned} \tag{4.1.1}
$$

The posterior pdf of Λ given in (2.11) simplifies in this case to

$$
\begin{aligned}
f(\lambda_1, \lambda_2|\mathbf{y}^{(1)}) \propto & \left(\prod_{i=1}^{m}\prod_{j=1}^{l_i} B_{ij}^{1/2}\right)\left(\prod_{i=1}^{m}\alpha_i^{1/2}\right)\left(\lambda_1\sum_{i=1}^{m}(1 - \alpha_i)\right)^{-1/2} \\
& \times\Bigg(s + a_0 + a_1\lambda_1 + a_2\lambda_2 + \sum_{i=1}^{m} K_{3i} \\
& - \left(\sum_{i=1}^{m} K_{2i}\right)^2\Big/\left(\sum_{i=1}^{m} K_{1i}\right)\Bigg)^{-(n_{\cdot\cdot}+g_0+g_1+g_2-1)/2},
\end{aligned} \tag{4.1.2}
$$

where $m = 5$, $s = \sum_{i=1}^{m}\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^2$, $K_{1i} = \lambda_1(1 - \alpha_i)$, $K_{2i} = \lambda_1(1 - \alpha_i)\bar{y}_i$, $K_{3i} = \lambda_2[\sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_{ij}^2 - (1 - \alpha_i)\sum_{j=1}^{l_i}(1 - B_{ij})\bar{y}_i^2]$ and $n_{\cdot\cdot} = \sum_{i=1}^{m}\sum_{j=1}^{l_i} n_{ij}$. In finding the HB predictor, we have used (4.1.2) with $a_0 = g_0 = g_1 = g_2 = 0$, $a_1 = a_2 = 0.0005$, and have carried out two-dimensional numerical integration.

An alternative estimator of γ_i is due to Ghosh and Lahiri (1988) which uses estimates of B_{ij}'s and α_i's rather than assigning any prior distribution on R and Λ. The resulting EB estimate of γ_i is given by

$$
\begin{aligned}
e_{\mathrm{EB}}^{i} = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\Bigg[& \sum_{j=1}^{l_i} N_{ij}\left(1 - f_{ij}\hat{B}_{ij}\right)\bar{y}_{ij} \\
& + \left\{\sum_{j=l_i+1}^{L_i} N_{ij} + \sum_{j=1}^{l_i} N_{ij} f_{ij}\hat{B}_{ij}\right\} \\
& \times\{(1 - \hat{\alpha}_i)\bar{y}_{i*} + \hat{\alpha}_i\bar{y}_*\}\Bigg],
\end{aligned} \tag{4.1.3}
$$

where $\hat{B}_{ij} = (1 + \hat{\lambda}_2^{-1} n_{ij})^{-1}$, $\hat{\alpha}_i = \hat{\lambda}_2^{-1}(\hat{\lambda}_2^{-1} + \hat{\lambda}_1^{-1}\sum_{j=1}^{l_i}(1 - \hat{B}_{ij}))^{-1}$, $\bar{y}_{i*} = \sum_{j=1}^{l_i}(1 - \hat{B}_{ij})\bar{y}_{ij} / \sum_{j=1}^{l_i}(1 - \hat{B}_{ij})$ if $\hat{\lambda}_2^{-1} \neq 0$ and $\bar{y}_{i*} = l_i^{-1}\sum_{j=1}^{l_i}\bar{y}_{ij}$, otherwise. Similarly, $\bar{y}_* = \sum_{i=1}^{m}(1 - \hat{\alpha}_i)\bar{y}_i / \sum_{i=1}^{m}(1 - \hat{\alpha}_i)$ if $\hat{\lambda}_1^{-1} \neq 0$ and $\bar{y}_* = m^{-1}\sum_{i=1}^{m}\bar{y}_i$ otherwise. The estimators $\hat{\lambda}_1^{-1}$ and $\hat{\lambda}_2^{-1}$ are given by Ghosh and Lahiri (1988), pages 205–206.

Four other estimates of γ_i are given below. These are:

$$e_U^i = \left(\frac{L_i}{l_i}\right)\left(\sum_{j=1}^{l_i} N_{ij}\bar{y}_{ij}\right) \Big/ \left(\sum_{j=1}^{L_i} N_{ij}\right) \quad \text{(a design-unbiased estimate)}, \tag{4.1.4}$$

$$e_R^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \sum_{j=1}^{l_i}(N_{ij} - n_{ij})\bar{y}_{ij} + \left(\sum_{j=1}^{l_i} N_{ij}\bar{y}_{ij} \Big/ \sum_{j=1}^{l_i} N_{ij}\right)\left(\sum_{j=l_i+1}^{L_i} N_{ij}\right)\right] \tag{4.1.5}$$

(the ratio-type estimate),

$$e_0^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \left(\sum_{j=1}^{L_i} N_{ij} - \sum_{j=1}^{l_i} n_{ij}\right)\left(\sum_{j=1}^{l_i} n_{ij}\bar{y}_{ij}\right) \Big/ \sum_{j=1}^{l_i} n_{ij}\right] \tag{4.1.6}$$

(the expansion estimate),

$$e_{\mathrm{RO}}^i = \left(\sum_{j=1}^{L_i} N_{ij}\right)^{-1}\left[\sum_{j=1}^{l_i}\sum_{k=1}^{n_{ij}} y_{ijk} + \sum_{j=1}^{l_i}(N_{ij} - n_{ij})\bar{y}_{ij} + \left(\sum_{j=1}^{l_i} \bar{y}_{ij}/l_i\right)\left(\sum_{j=l+1}^{L_i} N_{ij}\right)\right] \tag{4.1.7}$$

(Royall's estimate).

The estimates e_R, e_0 and e_{RO} are all based on predicted values of the unobserved units on the basis of the sampled units. However, in contrast to the present model, they can possibly be justified on the basis of some other models as given for example in Royall (1976). Table 1 provides the true population means as well as the six different estimates for each stratum.

The average absolute biases of the HB estimate, the EB estimate, the design unbiased estimate, the ratio-type estimate, the expansion estimate and Royall's estimate for the given data set are given respectively by 0.03102, 0.03156, 0.12932, 0.06277, 0.06009 and 0.04844. Thus the HB estimate has a slight edge over the EB estimate and much greater edge over the others in terms of average absolute bias. Also, the total sum of squared deviations of the HB estimates from the true means is 0.0085. The corresponding figures for e_{EB},

TABLE 1
The true means γ_i's and the estimates

i	γ_i	e^i_{HB}	e^i_{EB}	e^i_U	e^i_R	e^i_0	e^i_{RO}
1	0.73326	0.79789	0.80314	0.71201	0.91849	0.92190	0.93210
2	0.76149	0.76357	0.76442	0.91002	0.77214	0.76815	0.75043
3	0.74482	0.76778	0.76844	0.78208	0.78382	0.78299	0.75043
4	0.68933	0.75057	0.74971	0.89651	0.73864	0.74003	0.71533
5	0.74549	0.74130	0.74181	0.98056	0.71313	0.72653	0.71998

e_U, e_R, e_0 and e_{RO} turn out to be 0.0091, 0.1211, 0.0391, 0.0400 and 0.0409. Thus the percentage reduction in the total sum of squared deviations for the HB estimates is 6.6 in comparison with the EB estimates, 93.0 in comparison with the design unbiased estimates, 78.3 in comparison with the ratio-type estimates, 78.8 in comparison with the expansion estimates and 79.3 in comparison with Royall's estimates. An EB point estimator is usually on par with the corresponding HB point estimator. So the small improvement of the HB estimator over the EB estimator in reducing the total sum of squared deviations is not so surprising. However, the improvement of the HB estimator over the other four estimators is indeed startling. One possible explanation for this fact is that many of the other estimators are optimal under models which do not take into account variation in the primary sampling units. Our model accounts for this extra source of variation in producing more reliable estimates.

We also mention in passing that the posterior s.d.'s associated with the HB estimates in the five strata are given respectively by 0.050, 0.036, 0.043, 0.030 and 0.039.

4.2. *Prediction of areas under corn and soybeans.* Next, we analyze a data set where the objective is to predict areas under corn and soybeans for 12 counties in North Central Iowa based on the 1978 June Enumerative Survey as well as *LANDSAT* satellite data. The data set appears in Battese, Harter and Fuller (1988) who conducted a variance components analysis for this problem. The background of this problem is as follows.

The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in 37 sample segments (each segment was about 250 hectares) of 12 counties in North Central Iowa by interviewing farm operators. Based on *LANDSAT* readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels (a term for "picture element" about 0.45 hectares) in the 12 counties. The number of segments in each country, the number of hectares of corn and soybeans (as reported in the June Enumerative Survey), the number of pixels classified as corn and soybeans for each sample segment and the county mean number of pixels classified as corn and soybeans (the total number of pixels classified as that crop divided by the number of segments in that county) are reported in

Table 1 of Battese, Harter and Fuller (1988). In order to make our results comparable to that of Battese, Harter and Fuller (1988), the second segment in Hardin County was ignored.

Battese, Harter and Fuller (1988) considered the model

$$Y_{ij} = b_0 + b_1 x_{1ij} + b_2 x_{2ij} + v_i + e_{ij}, \tag{4.2.1}$$

where i is a subscript for the county and j is a subscript for a segment within the given county ($j = 1, \ldots, N_i$, the number of segments in the ith county, $i = 1, \ldots, 12$). Here Y_{ij} is the reported number of hectares of soybeans and x_{1ij} (x_{2ij}) is the number of pixels classified as corn (soybeans) for the jth segment in the ith county. They assumed (in our notation) $E(v_i) = E(e_{ij}) = 0$, $V(v_i) = (\lambda r)^{-1}$, $V(e_{ij}) = r^{-1}$, $\operatorname{cov}(v_i, e_{ij}) = 0$, $\operatorname{cov}(v_i, v_{i'}) = 0$, $i \neq i'$, $\operatorname{cov}(e_{ij}, e_{i'j'}) = 0$ if $(i, j) \neq (i', j')$. First, assuming λ and r known, these authors obtained BLUPs of $\mu_i = b_0 + b_1 \bar{x}_{1i(p)} + b_2 \bar{x}_{2i(p)} + v_i$, $i = 1, \ldots, 12$, where $\bar{x}_{\alpha i(p)} = N_i^{-1} \sum_{j=1}^{N_i} x_{\alpha ij}$, $\alpha = 1, 2$. Then, using Henderson's method III, they obtained estimates of the variance components, and their final predictors involved the estimated variance components. [For details, see Battese, Harter and Fuller (1988).] Henderson's method being an ANOVA method could lead to negative estimates of λ^{-1}. If this were the case, Battese, Harter and Fuller set it equal to 0. This phenomenon is likely to happen, particularly when the number of small areas or strata is small.

In this particular example, we have $t = 1$, $\lambda_1 = \lambda$, $\mathbf{D}(\boldsymbol{\lambda}) = \lambda^{-1}\mathbf{I}_m$, $\boldsymbol{\Psi} = \mathbf{I}_N$. Then $\boldsymbol{\Sigma}_{11} = \operatorname{Diag}(\mathbf{I}_{n_1} + \lambda^{-1}\mathbf{J}_{n_1}, \ldots, \mathbf{I}_{n_m} + \lambda^{-1}\mathbf{J}_{n_m})$ so that $|\boldsymbol{\Sigma}_{11}| = \prod_{i=1}^{m}\{(\lambda + n_i)/\lambda\}$. Also, writing $\bar{\mathbf{x}}_i = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$, $i = 1, \ldots, m$, one gets

$$\begin{aligned}\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \sum_{i=1}^{m} n_i^2(n_i + \lambda)^{-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T \\ &= \mathbf{H}(\lambda) \quad \text{(say)}.\end{aligned} \tag{4.2.2}$$

Next, writing $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$, one gets

$$\begin{aligned}\mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 + \lambda\sum_{i=1}^{m} n_i(n_i + \lambda)^{-1}\bar{y}_i^2 \\ &\quad - \left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(y_{ij} - n_i(n_i + \lambda)^{-1}\bar{y}_i\right)\right\}^T \mathbf{H}^{-1}(\lambda) \\ &\quad \times\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(y_{ij} - n_i(n_i + \lambda)^{-1}\bar{y}_i\right)\right\} \\ &= Q_0(\lambda) \quad \text{(say)}.\end{aligned} \tag{4.2.3}$$

The conditional pdf $f(\lambda|\mathbf{y}^{(1)})$ given in (2.11) simplifies to

$$f(\lambda|\mathbf{y}^{(1)}) \propto \lambda^{(m+g_1)/2-1} \prod_{i=1}^{m} (\lambda + n_i)^{-1/2} |\mathbf{H}(\lambda)|^{-1/2} \times (a_0 + a_1\lambda + Q_0(\lambda))^{-(n+g_0+g_1-p)/2}. \tag{4.2.4}$$

The posterior means and variances of the finite population means are now obtained from (2.8)–(2.10), (2.12)–(2.13), (4.2.2)–(4.2.4) and using the formulas for iterated conditional expectations and variances.

REMARK 1. Let $V_1(\mathbf{y}^{(1)})$ and $V_2(\mathbf{y}^{(1)})$ denote respectively the variance of the conditional expectation and expectation of the conditional variance of the finite population mean. A naive empirical Bayes procedure effectively ignores V_1 and can lead to serious underestimate of the variance. A HB procedure on the other hand rectifies this deficiency. Battese, Harter and Fuller have a frequentist approach which also incorporates the uncertainty of estimating the variance components into account.

We find the posterior means and variances of the population means for the 12 counties. Our approach eliminates the possibility of obtaining zero estimates of the variance components. The improper prior with $a_0 = a_1 = 0.005$, $g_0 = g_1 = 0$ is used for predicting areas under soybeans.

Table 2 provides the HB predictors (e_{HB}), the EB predictors (e_{EB}), the BHF predictors (e_{BHF}) and the associated standard errors s_{HB}, s_{EB} and s_{BHF}, respectively. Note that the EB predictors are obtained by replacing λ with its Henderson's Method III estimate in $E[N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}|\mathbf{y}^{(1)}, \lambda]$. Also, we provide the V_1 and V_2 values to demonstrate that V_1 can sometimes contribute significantly toward the posterior variance.

As one might anticipate, e_{HB} and e_{EB} are extremely close as point predictors; e_{BHF} differs from e_{EB} because it uses a different estimate of λ, and

TABLE 2
The predicted hectares of soybeans and standard errors

County	e_{HB}	e_{EB}	e_{BHF}	s_{HB}	s_{EB}	s_{BHF}	V_1	V_2
Cerro Gordo	78.8	78.2	77.5	11.7	11.6	12.7	7.67	128.59
Franklin	67.1	65.9	64.8	8.2	7.5	7.8	11.94	54.92
Hamilton	94.4	94.6	95.0	11.2	11.4	12.4	1.97	123.61
Hancock	100.4	100.8	101.1	6.2	6.1	6.3	1.35	37.59
Hardin	75.4	75.1	74.9	6.5	6.4	6.6	0.37	41.84
Humboldt	81.9	80.6	79.2	10.4	9.3	10.0	22.62	85.40
Kossuth	118.2	119.2	120.2	6.6	6.0	6.2	7.99	36.23
Pocahontas	113.9	113.7	113.8	7.5	7.5	7.9	0.06	55.98
Webster	110.0	109.7	109.6	6.6	6.6	6.8	0.64	43.91
Winnebago	97.3	98.0	98.7	7.7	7.5	7.9	4.11	55.70
Worth	87.8	87.2	86.6	11.1	11.1	12.1	4.06	118.17
Wright	111.9	112.4	112.9	7.7	7.6	8.0	1.62	57.48

thereby leads to slightly different predicted values. It is important to note that the difference between e_{BHF} and either e_{EB} or e_{HB} is much more pronounced than any difference between e_{HB} and e_{EB}.

The naive EB estimator, in general, underestimates the standard error in comparison with the HB estimator. With the exception of Hamilton County, s_{EB} is always smaller or equal to s_{HB}. The difference can be significant as evidenced from the figures given in Humboldt County where s_{EB} is about 10% smaller than s_{HB}.

However, s_{HB} and s_{BHF} are both very good as estimates of standard errors. In this example, while s_{BHF} is never smaller than s_{HB} by more than 6.1%, it can exceed s_{HB} by about 9.7%.

One may wonder whether the proposed HB predictors which perform so well conditionally enjoy any frequentist properties. To answer this, we undertook an extensive simulation study using the BHF model. The detailed results are not reported in this paper, but our findings indicated that the simulated mean squared errors for the HB predictors were matching those for the BHF predictors up to the fifth decimal place, while (1.96) s.d. coverage probabilities turned out to be slightly bigger for HB than BHF, both being very close to 95% under all circumstances.

5. The HB predictor in a special case. We consider in this section the special case when $\boldsymbol{\lambda}$ is known, while $\mathbf{B}$ and R are independently distributed with $\mathbf{B} \sim \text{uniform}(R^p)$ and $R \sim \text{Gamma}(\frac{1}{2}a_0, \frac{1}{2}g_0)$. We are still interested in finding the posterior distribution of $\mathbf{Y}^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$. Recall the notation $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{G}$ given in (2.8)–(2.10). Since $\boldsymbol{\lambda}$ is known in this case, we have the following Theorem 2 instead of Theorem 1.

THEOREM 2. *Assume that $n + g_0 > p + 2$. Then under the model given in* (A) *and* (B) *with $\boldsymbol{\lambda}$ known, and an independent uniform (R^p) prior for $\mathbf{B}$ and a* Gamma$(\frac{1}{2}a_0, \frac{1}{2}g_0)$ *prior for R, the conditional distribution of $Y^{(2)}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$ is multivariate-t with location parameter $\mathbf{My}^{(1)}$, scale parameter $(n + g_0 - p)^{-1}(a_0 + \mathbf{y}^{(1)T}\mathbf{Ky}^{(1)})\mathbf{G}$ and degrees of freedom $n + g_0 - p$.*

The proof of Theorem 2 is similar to the proof of the first part of Theorem 1 provided in the Appendix and is omitted. Using the properties of the multivariate-t distribution, it is now possible to obtain closed-form expressions for $E[\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}]$ and $V[\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}]$, where $\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \mathbf{AY}^{(1)} + \mathbf{CY}^{(2)}$. In particular, the Bayes estimate of $\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ under any quadratic loss is now

$$\mathbf{e}_B^*(\mathbf{y}^{(1)}) = (\mathbf{A} + \mathbf{CM})\mathbf{y}^{(1)}. \tag{5.1}$$

We may note that the posterior mean given in (5.1) does not depend on the prior distribution of R.

There are alternative ways to generate the same predictor $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ of $\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Suppose, for example, one assumes only (2.1) or (2.2) with $\mathbf{b}$ known (r may or may not be known). Then the best predictor (best linear

predictor without the normality assumption) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ in the sense of having the smallest mean squared error matrix is given by

$$E_{\theta}\left[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)}\right] = \mathbf{C}\left[\Sigma_{21}\Sigma_{11}^{-1}\mathbf{Y}^{(1)} + \left(\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)\mathbf{b}\right] + \mathbf{A}\mathbf{Y}^{(1)} \quad (\text{a.e. } \mathbf{Y}^{(1)}), \tag{5.2}$$

where $\theta = (\mathbf{b}^T, r)^T$.

[We say that $\mathbf{E} \leq \mathbf{F}$ for two symmetric matrices $\mathbf{E}$ and $\mathbf{F}$ if $\mathbf{F} - \mathbf{E}$ is nonnegative definite (n.n.d.).] If $\mathbf{b}$ is unknown, then one replaces $\mathbf{b}$ by its UMVUE (BLUE without the normality assumption)

$$\left(\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)T}\Sigma_{11}^{-1}\mathbf{Y}^{(1)}.$$

The resulting predictor of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ turns out to be $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$. In this sense, $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is also an empirical Bayes predictor of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Harville (1985, 1988, 1990) recognized this for predicting scalars.

We shall now discuss some frequentist properties of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$. First, we assume the normal model (2.1) or (2.2) with λ known. No prior distribution for $\mathbf{B}$ and R is assumed, and $\theta = (\mathbf{b}^T, r)^T$ is treated as an unknown parameter. We prove the optimality of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ within the class of all unbiased predictors of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. This result is then used to prove the optimality of $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ once again within the class of all unbiased predictors of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ for a class of spherically symmetric distributions of $\mathbf{Y}$ including but not limited to the normal distribution.

We start with the following definition.

DEFINITION 1. A predictor $\mathbf{T}(\mathbf{Y}^{(1)})$ is said to be a *best unbiased predictor* (BUP) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ if $E_{\theta}[\mathbf{T}(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ and for every predictor $\delta(\mathbf{Y}^{(1)})$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ satisfying $E_{\theta}[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all θ, $V_{\theta}[\mathbf{T}(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] \leq V_{\theta}[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})]$ for all θ provided the quantities are finite.

The following theorem is proved.

THEOREM 3. *Under the model* (2.1) *or* (2.2), $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ *is the BUP of* $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

PROOF. Write $\mathbf{H}_0 = \mathbf{A} + \mathbf{C}\Sigma_{21}\Sigma_{11}^{-1}$ and $\mathbf{U} = \mathbf{C}[\mathbf{X}^{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{X}^{(1)}]$. Then, from (5.2), $E[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)}] = \mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}$ a.e. $(\mathbf{Y}^{(1)})$. For an arbitrary predictor $\delta(\mathbf{Y}^{(1)})$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$, write

$$\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \left[\delta(\mathbf{Y}^{(1)}) - \left(H_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}\right)\right] + \left[\left(\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b}\right) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\right]. \tag{5.3}$$

Then, from (5.2) and (5.3),

$$
\begin{aligned}
&E_{\theta}\Big[\{\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^{T}\Big] \\
&\quad = E_{\theta}\Big[\{(\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)}) - \mathbf{U}\mathbf{b}\} \\
&\qquad \times \{(\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)}) - \mathbf{U}\mathbf{b}\}^{T}\Big] \qquad (5.4) \\
&\qquad + E_{\theta}\Big[\{\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\} \\
&\qquad\quad \times \{\mathbf{H}_0\mathbf{Y}^{(1)} + \mathbf{U}\mathbf{b} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^{T}\Big].
\end{aligned}
$$

Hence minimization of the left-hand side of (5.4) wrt $\delta(\mathbf{Y}^{(1)})$ amounts to the minimization of the first term in the right-hand side of (5.4) wrt $\delta(\mathbf{Y}^{(1)})$. Since $\mathbf{Y}^{(1)} \sim N(\mathbf{X}^{(1)}\mathbf{b}, r^{-1}\boldsymbol{\Sigma}_{11})$, from the classical theory of least squares it follows that the first term in the right-hand side of (5.4) is minimized wrt $\delta(\mathbf{Y}^{(1)})$ if and only if $\delta(\mathbf{Y}^{(1)}) - \mathbf{H}_0\mathbf{Y}^{(1)} = \mathbf{U}(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{Y}^{(1)}$ a.e. $(\mathbf{Y}^{(1)})$, that is, $\delta(\mathbf{Y}^{(1)}) = (\mathbf{A} + \mathbf{C}\mathbf{M})\mathbf{Y}^{(1)} = \mathbf{e}_B^*(\mathbf{Y}^{(1)})$ a.e. $(\mathbf{Y}^{(1)})$. The proof of Theorem 3 is complete. □

REMARK 2. It follows from the proof of the theorem that the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ is unique with probability 1.

REMARK 3. It is possible to generalize Theorem 2 for a more general class of distributions of $\mathbf{Y}$. Suppose that conditional on $R = r$, $\mathbf{Y} \sim N(\mathbf{X}\mathbf{b}, r^{-1}\boldsymbol{\Sigma})$, while marginally R has any proper distribution. The objective is once again to minimize the left-hand side of (5.4). We achieve this by first computing this expectation conditional on $R = r$. We may note that $E[\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})|\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}, R = r] = \mathbf{H}_0\mathbf{y}^{(1)} + \mathbf{U}\mathbf{b}$ *does not* depend on r. Hence we obtain an identity similar to (5.4) conditional on $R = r$, and as in the proof of Theorem 3, conclude that $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

Next we dispense with any distributional assumption in (2.1) and show that $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ is the BUP of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ within the class of all linear unbiased predictors. A predictor $\delta(\mathbf{Y}^{(1)})$ is said to be linear if $\delta(\mathbf{Y}^{(1)})$ has the form $\mathbf{H}\mathbf{Y}^{(1)}$ for some known $u \times n$ matrix $\mathbf{H}$. If, in addition, $E_{\theta}[\delta(\mathbf{Y}^{(1)}) - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})] = \mathbf{0}$ for all $\boldsymbol{\theta}$, we say that $\delta(\mathbf{Y}^{(1)})$ is a *linear unbiased predictor* (LUP) of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. We now introduce another definition.

DEFINITION 2. A LUP $\mathbf{P}\mathbf{Y}^{(1)}$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ is said to be a *best linear unbiased predictor* (BLUP) if for every LUP $\mathbf{H}\mathbf{Y}^{(1)}$ of $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$, $V_{\theta}(\mathbf{H}\mathbf{Y}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})) - V_{\theta}(\mathbf{P}\mathbf{Y}^{(1)} - \xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}))$ is n.n.d. for all $\boldsymbol{\theta}$.

We now prove the following theorem.

THEOREM 4. *Consider the model* (2.2) *and assume that* $E_{\theta}[\mathbf{e}] = \mathbf{0}$, $E_{\theta}[\mathbf{v}] = \mathbf{0}$, $E_{\theta}[\mathbf{e}v^T] = \mathbf{0}$, $E_{\theta}[\mathbf{e}^T\mathbf{e}] < \infty$ *and* $E_{\theta}[\mathbf{v}^T\mathbf{v}] < \infty$. *Then* $\mathbf{e}_B^*(\mathbf{Y}^{(1)})$ *is the BLUP of* $\xi(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$.

PROOF. Suppose $\mathbf{WY}^{(1)}$ is an unbiased predictor of $\boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Then $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{WY}^{(1)} - (A\mathbf{Y}^{(1)} + C\mathbf{Y}^{(2)})] = \mathbf{0}$ for all $\boldsymbol{\theta}$, which is equivalent to $(\mathbf{W} - \mathbf{A})\mathbf{X}^{(1)} = \mathbf{CX}^{(2)} = \mathbf{CMX}^{(1)}$ from (2.8) and (2.9), that is, $(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{X}^{(1)} = \mathbf{0}$. Next write

$$\begin{aligned}\mathbf{WY}^{(1)} - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) &= \mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)}) + \mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) \\ &= (\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{Y}^{(1)} + \mathbf{C}(\mathbf{MY}^{(1)} - \mathbf{Y}^{(2)}).\end{aligned} \tag{5.5}$$

Observe next that since $\mathbf{MX}^{(1)} = \mathbf{X}^{(2)}$,

$$\begin{aligned}&E_{\boldsymbol{\theta}}\Big[\mathbf{C}(\mathbf{MY}^{(1)} - \mathbf{Y}^{(2)})\{(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{Y}^{(1)}\}^T\Big] \\ &\quad = E_{\boldsymbol{\theta}}\Big[\mathbf{C}\{\mathbf{M}(\mathbf{Y}^{(1)} - E_{\boldsymbol{\theta}}(\mathbf{Y}^{(1)})) - (\mathbf{Y}^{(2)} - E_{\boldsymbol{\theta}}(\mathbf{Y}^{(2)}))\} \\ &\qquad\qquad \times \mathbf{Y}^{(1)T}(\mathbf{W} - \mathbf{A} - \mathbf{CM})^T\Big] \\ &\quad = E_{\boldsymbol{\theta}}\Big[\mathbf{C}(\mathbf{M}\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{21})(\mathbf{W} - \mathbf{A} - C\mathbf{M})^T\Big].\end{aligned} \tag{5.6}$$

But, using (2.8) and (2.9),

$$\mathbf{M}\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{21} = (\mathbf{X}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)})(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)T}. \tag{5.7}$$

Since $\mathbf{X}^{(1)T}(\mathbf{W} - \mathbf{A} - \mathbf{CM})^T = [(\mathbf{W} - \mathbf{A} - \mathbf{CM})\mathbf{X}^{(1)}]^T = \mathbf{0}$, it follows from (5.6) and (5.7) that the left-hand side of (5.6) is $\mathbf{0}$. Now, from (5.5),

$$\begin{aligned}&E_{\boldsymbol{\theta}}\Big[\{\mathbf{WY}^{(1)} - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{WY}^{(1)} - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\Big] \\ &\quad = E_{\boldsymbol{\theta}}\Big[\{\mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)})\}\{\mathbf{WY}^{(1)} - \mathbf{e}_B^*(\mathbf{Y}^{(1)})\}^T\Big] \\ &\qquad + E_{\boldsymbol{\theta}}\Big[\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\Big] \\ &\quad \geq E_{\boldsymbol{\theta}}\Big[\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}\{\mathbf{e}_B^*(\mathbf{Y}^{(1)}) - \boldsymbol{\xi}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})\}^T\Big]\end{aligned}$$

with equality if and only if $\mathbf{WY}^{(1)} = \mathbf{e}_B^*(\mathbf{Y}^{(1)})$ a.e. $(\mathbf{Y}^{(1)})$. The proof of Theorem 4 is complete. □

APPENDIX

PROOF OF THEOREM 1. Under the assumptions of the theorem, the joint pdf of $\mathbf{Y}$, $\mathbf{B}$, R and $\boldsymbol{\Lambda}$ is given by

$$\begin{aligned}&f(\mathbf{y}, \mathbf{b}, r, \boldsymbol{\lambda}) \\ &\quad \propto r^{N/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\Big[-\tfrac{1}{2}r(\mathbf{y} - \mathbf{Xb})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{Xb})\Big]\exp(-\tfrac{1}{2}a_0 r)r^{g_0/2-1} \\ &\qquad \times \exp\left(-\tfrac{1}{2}r\sum_{i=1}^{t} a_i\lambda_i\right)\prod_{i=1}^{t}(\lambda_i r)^{g_i/2-1}r^t \\ &\quad = |\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\tfrac{1}{2}r\left\{(\mathbf{y} - \mathbf{Xb})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{Xb}) + a_0 + \sum_{i=1}^{t} a_i\lambda_i\right\}\right] \\ &\qquad \times r^{(N+\sum_{i=0}^{t} g_i)2-1}\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.\end{aligned} \tag{A.1}$$

Now

$$
\begin{aligned}
&(\mathbf{y}-\mathbf{X}\mathbf{b})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})\\
&\quad=\left[\mathbf{b}-(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\right]^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})\\
&\qquad\times\left[\mathbf{b}-(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\right]+\mathbf{y}^T\mathbf{Q}\mathbf{y},
\end{aligned}
\tag{A.2}
$$

where $\mathbf{Q}=\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$. From (A.1) and (A.2), one gets the joint pdf of $\mathbf{Y}$, R and $\boldsymbol{\Lambda}$ given by

$$
\begin{aligned}
f(\mathbf{y},r,\boldsymbol{\lambda})\ &\propto\ |\boldsymbol{\Sigma}|^{-1/2}|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}|^{-1/2}r^{(N+\Sigma_{i=0}^{t}g_i-p)2-1}\\
&\quad\times\exp\left[-\tfrac{1}{2}r\left(a_0+\Sigma_{i=1}^{t}a_i\lambda_i+\mathbf{y}^T\mathbf{Q}\mathbf{y}\right)\right]\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.
\end{aligned}
\tag{A.3}
$$

Now, integrating wrt R, one finds the pdf of $\mathbf{Y}$ and $\boldsymbol{\Lambda}$ given by

$$
\begin{aligned}
f(\mathbf{y},\boldsymbol{\lambda})\ &\propto\ |\boldsymbol{\Sigma}|^{-1/2}|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}|^{-1/2}\left(a_0+\sum_{i=1}^{t}a_i\lambda_i+\mathbf{y}^T\mathbf{Q}\mathbf{y}\right)^{-(N+\Sigma_{i=0}^{t}g_i-p)/2}\\
&\quad\times\prod_{i=1}^{t}\lambda_i^{g_i/2-1}.
\end{aligned}
\tag{A.4}
$$

Now, using a standard formula for partitioned matrices [e.g., Searle (1971), page 46], we have

$$
\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}=\mathbf{y}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}+\left(\mathbf{y}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\right)^T\boldsymbol{\Sigma}_{22.1}^{-1}\left(\mathbf{y}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\right).
\tag{A.5}
$$

Similarly,

$$
\begin{aligned}
\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}&=\mathbf{y}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\\
&\quad+\left(\mathbf{y}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\right)^T\boldsymbol{\Sigma}_{22.1}^{-1}\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)\\
&=\mathbf{t}_1^T+\mathbf{t}_2^T\quad\text{(say)},
\end{aligned}
\tag{A.6}
$$

$$
\begin{aligned}
\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}&=\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\\
&\quad+\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^T\boldsymbol{\Sigma}_{22.1}^{-1}\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right).
\end{aligned}
\tag{A.7}
$$

Using the matrix inversion formula [see Exercise 2.9, page 33 of Rao (1973)], we have from (A.7) that

$$
\begin{aligned}
&(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\\
&\quad=\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}-\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^T\\
&\qquad\times\Big\{\boldsymbol{\Sigma}_{22.1}+\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\\
&\qquad\qquad\times\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{21}^{-1}\mathbf{X}^{(1)}\right)^T\Big\}^{-1}\\
&\qquad\times\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\\
&\quad=\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}-\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^T\mathbf{G}^{-1}\\
&\qquad\times\left(\mathbf{X}^{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\right)^{-1}\quad\text{by (2.10)}\\
&\quad=\mathbf{M}_1-\mathbf{M}_2\quad\text{(say)}.
\end{aligned}
\tag{A.8}
$$

From (A.6), (A.8) and (2.8)–(2.10), we get after simplifications

$$\text{(A.9)}\quad \mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} = \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_1 - \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_1 + \mathbf{t}_2^T\mathbf{M}_1\mathbf{t}_2 - \mathbf{t}_2^T\mathbf{M}_2\mathbf{t}_2 + 2\mathbf{t}_1^T(\mathbf{M}_1 - \mathbf{M}_2)\mathbf{t}_2,$$

$$\text{(A.10)}\quad \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_1 = \mathbf{y}^{(1)T}\big(\boldsymbol{\Sigma}_{11}^{-1} - \mathbf{K}\big)\mathbf{y}^{(1)},$$

$$\text{(A.11)}\quad \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_1 = \big(\mathbf{M}\mathbf{y}^{(1)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big)^T\mathbf{G}^{-1}\big(\mathbf{M}\mathbf{y}^{(1)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big),$$

$$\text{(A.12)}\quad \mathbf{t}_2^T\mathbf{M}_1\mathbf{t}_2 = \big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big)^T\big[\boldsymbol{\Sigma}_{22.1}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{22.1}^{-1} - \boldsymbol{\Sigma}_{22.1}^{-1}\big] \times\big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big),$$

$$\text{(A.13)}\quad \mathbf{t}_2^T\mathbf{M}_2\mathbf{t}_2 = \big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big)^T \times\big[\boldsymbol{\Sigma}_{22.1}^{-1}\mathbf{G}\boldsymbol{\Sigma}_{22.1}^{-1} - 2\boldsymbol{\Sigma}_{22.1}^{-1} + \mathbf{G}^{-1}\big]\big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big),$$

$$\text{(A.14)}\quad \mathbf{t}_1^T\mathbf{M}_1\mathbf{t}_2 = \big(\mathbf{M}\mathbf{y}^{(1)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big)^T\boldsymbol{\Sigma}_{22.1}^{-1}\big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big),$$

$$\text{(A.15)}\quad \mathbf{t}_1^T\mathbf{M}_2\mathbf{t}_2 = \big(\mathbf{M}\mathbf{y}^{(1)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big)^T\big[\boldsymbol{\Sigma}_{22.1}^{-1} - \mathbf{G}^{-1}\big] \times\big(\mathbf{y}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}^{(1)}\big),$$

Using the same definition of $\mathbf{Q}$, it follows from (A.5)–(A.15) with some algebraic manipulations that

$$\text{(A.16)}\quad \mathbf{y}^T\mathbf{Q}\mathbf{y} = \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)} + \big(\mathbf{y}^{(2)} - \mathbf{M}\mathbf{y}^{(1)}\big)^T\mathbf{G}^{-1}\big(\mathbf{y}^{(2)} - \mathbf{M}\mathbf{y}^{(1)}\big).$$

Combining (A.4), (A.16) and (2.6), one gets the first part of Theorem 1.

Now to find the conditional distribution of $\boldsymbol{\Lambda}$ given $\mathbf{Y}^{(1)} = \mathbf{y}^{(1)}$, one can have as in (A.4) that the pdf of $\mathbf{Y}^{(1)}$ and $\boldsymbol{\Lambda}$ is given by

$$\text{(A.17)}\quad f(\mathbf{y}^{(1)}, \boldsymbol{\lambda}) \propto |\boldsymbol{\Sigma}_{11}|^{-1/2}\big|\mathbf{X}^{(1)T}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}^{(1)}\big|^{-1/2} \times\left(a_0 + \sum_{i=1}^{t} a_i\lambda_i + \mathbf{y}^{(1)T}\mathbf{K}\mathbf{y}^{(1)}\right)^{-(n+\sum_{i=0}^{t} g_i - p)/2} \prod_{i=1}^{t}\lambda_i^{g_i/2-1}.$$

Since $f(\boldsymbol{\lambda}|\mathbf{y}^{(1)}) \propto f(\mathbf{y}^{(1)}, \boldsymbol{\lambda})$, (2.11) follows from (A.17). □

Acknowledgments. We express our indebtedness to Professors James Calvin and Joseph Sedransk for supplying us with the set of data used for two-stage sampling. The revision has benefitted much from the very helpful comments of an Associate Editor and four referees. Thanks are due to Dr. Li-Chu Lee for carrying out some numerical computations.

REFERENCES

ALBERT, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83** 1037–1044.

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. Statist. Assoc.* **83** 28–36.

CALVIN, J. and SEDRANSK, J. (1991). The patterns of care studies. *J. Amer. Statist. Assoc.* **86** 36–48.

CHOUDHRY, G. H. and RAO, J. N. K. (1988). Evaluation of small area estimations: An empirical study. Preprint.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd. ed. Wiley, New York.

DEMPSTER, A. P., RUBIN, D. B. and TSUTAKAWA, R. K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* **76** 341–353.

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. Statist. Assoc.* **74** 269–277.

GELFAND, A. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

GHOSH, M. and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* **82** 1153–1162.

GHOSH, M. and LAHIRI, P. (1988). Bayes and empirical Bayes analysis in multistage sampling. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 195–212. Springer, New York.

GHOSH, M. and LAHIRI, P. (1989). A hierarchical Bayes approach to small area estimation with auxiliary information. In *Proceedings of the Joint Indo–U.S. Workshop on Bayesian Inference in Statistics and Econometrics*. To appear.

GHOSH, M. and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.* **81** 1058–1062.

GHOSH, M. and RAO, J. N. K. (1991). Small area estimation. Technical Report 390, Dept. Statistics, Univ. Florida.

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

HARVILLE, D. A. (1988). Mixed-model methodology: Theoretical justifications and future directions. In *Proceedings of the Statistical Computing Section* 41–49. Amer. Statist. Assoc., Alexandria, Va.

HARVILLE, D. A. (1990). BLUP and beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock*. (D. Gianola and K. L. Hammond, eds.) 239–276. Springer, New York.

HARVILLE, D. A. and JESKE, D. R. (1989). Mean squared error of estimation and prediction under a general linear model. Preprint 89-37, Statistics Laboratory, Iowa State Univ.

HENDERSON, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (W. D. Hanson and M. F. Robinson, eds.) 141–163. Publication 982, NAS-NRC, Washington, D. C.

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. Statist. Assoc.* **79** 853–862.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

LUI, K. J. and CUMBERLAND, W. G. (1989). A Bayesian approach to small domain estimation. *Journal of Official Statistics* **5** 143–156.

MALEC, D. and SEDRANSK, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *J. Amer. Statist. Assoc.* **80** 897–902.

MORRIS, C. N. (1988). Determining the accuracy of Bayesian empirical Bayes estimates in the familiar exponential families. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 251–263. Springer, New York.

PRASAD, N. G. N. and RAO, J. N. K. (1990). On the estimation of mean square error of small area predictors. *J. Amer. Statist. Assoc.* **85** 163–171.

PRESS, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston, New York.

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

ROYALL, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.* **71** 657–664.

ROYALL, R. M. (1978). Prediction models in small area estimation. In *Synthetic Estimates for Small Areas* (J. Steinberg, ed.) 63–87. NIDA Monograph Series 24. Dept. Health, Education and Welfare, Washington, D.C.

SCOTT, A. and SMITH, T. M. F. (1969). Estimation in multistage surveys. *J. Amer. Statist. Assoc.* **64** 830–840.

SEARLE, S. R. (1971). *Linear Models*. Wiley, New York.

STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics: An International Symposium*. (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 124–137. Wiley, New York.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611