Survey Methodology, June 1992
Vol. 18, No. 1, pp. 75–94
Statistics Canada

REML Estimation in Empirical Bayes Smoothing of Census Undercount

NOEL CRESSIE[1]

ABSTRACT

One way to assess the undercount at subnational levels (*e.g.* the state level) is to obtain sample data from a post-enumeration survey, and then smooth those data based on a linear model of explanatory variables. The relative importance of sampling-error variances to corresponding model-error variances determines the amount of smoothing. Maximum likelihood estimation can lead to oversmoothing, so making the assessment of undercount over-reliant on the linear model. Restricted maximum likelihood (REML) estimators do not suffer from this drawback. Empirical Bayes prediction of undercount based on REML will be presented in this article, and will be compared to maximum likelihood and a method of moments by both simulation and example. Large-sample distributional properties of the REML estimators allow accurate mean squared prediction errors of the REML-based smoothers to be computed.

KEY WORDS: Linear model; Maximum likelihood; Restricted maximum likelihood; Variance components.

1. INTRODUCTION

Although a census attempts to carry out a complete enumeration of the population, for various reasons the final tallies are inaccurate. Census personnel, from its director down to the thousands of temporary enumerators, are part of a mammoth task whose accuracy relies on everyone doing their jobs to perfection.

Moreover, events that are beyond human control (*e.g.* weather, natural disaster) must stay within expected limits. Clearly, in a country the size of the U.S.A. (in terms of both population and geography), many opportunities arise to give an imperfect census count. But size is not the only problem; **heterogeneity** of both population and geography gives a **differentially** imperfect count.

The inaccuracies are typically expressed in terms of undercount, so that a negative value implies an overcount. Suppose the U.S.A. is divided into $i = 1, \ldots, n$ areas (*e.g.* states, including Washington DC). In the i-th area, let T_i be the true (unknown) count and C_i be the census count. Then the undercount, expressed as a percentage of the true count, is defined as,

$$U_i \equiv \{(T_i - C_i)/T_i\}100. \tag{1.1}$$

The problem of **differential** undercount is a serious one when census counts are used to apportion political power and revenue to areas and subareas. (Further discussion of these issues can be found in Ericksen and Kadane 1985, Freedman and Navidi 1986 and Cressie 1988). States like California, Texas, and New York would gain much from adjusting for undercount, *i.e.* from replacing C_i with F_iC_i, where F_i is an **adjustment factor**.

The correct adjustment to use is,

$$F_i = T_i/C_i, \tag{1.2}$$

[1] Noel Cressie, Department of Statistics, Iowa State University, Ames, IA, U.S.A. 50011.

which is related to undercount by,

$$F_i = \{1 - U_i/100\}^{-1}.$$

As it stands, (1.2) is not helpful for adjustment, since the true count T_i is unknown. To obtain extra information that will allow F_i to be estimated, the U. S. Census Bureau conducts a post-enumeration survey (PES) that determines whether people in the PES were or were not counted in the census (*e.g.* Wolter 1986). The survey consists of several hundred thousand households, yielding "raw" adjustment factors $\{Y_i : i = 1, \ldots, n\}$ that are in need of smoothing.

Assume that, given F_i,

$$Y_i \sim \mathrm{Gau}(F_i, \delta_i^2), \tag{1.3}$$

i.e. Y_i has, conditional on F_i, a Gaussian distribution with mean F_i and variance δ_i^2. Adding the further assumption of independence, one obtains,

$$\underset{\sim}{Y} \sim \mathrm{Gau}(\underset{\sim}{F}, \Delta), \tag{1.4}$$

where $\underset{\sim}{Y} \equiv (Y_1, \ldots, Y_n)'$, $\underset{\sim}{F} \equiv (F_1, \ldots, F_n)'$, and Δ is the $n \times n$ diagonal matrix $\mathrm{diag}\{\delta_1^2, \ldots, \delta_n^2\}$.

Now assume that,

$$\underset{\sim}{F} \sim \mathrm{Gau}(X\underset{\sim}{\beta}, \Gamma(\tau^2)), \tag{1.5}$$

where X is an $n \times p$ matrix of explanatory variables, $\underset{\sim}{\beta}$ is a $p \times 1$ vector of (unknown) coefficients of the linear model, $\Gamma(\tau^2)$ is an $n \times n$ diagonal matrix:

$$\Gamma(\tau^2) \equiv \tau^2 D \tag{1.6}$$

and $D \equiv \mathrm{diag}\{1/C_1, \ldots, 1/C_n\}$. The heteroskedastic model (1.5) and (1.6) is discussed at considerable length in Cressie (1990). It is intuitively sensible that the adjustment factor, for an area whose population is large, has a smaller variance; Cressie (1989) provides both a Bayesian and a frequentist justification for this intuition.

Another way to write the model (1.4) and (1.5) is:

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\nu} + \underset{\sim}{\epsilon}, \tag{1.7}$$

where the $n \times 1$ vectors $\underset{\sim}{\nu}$ and $\underset{\sim}{\epsilon}$ are statistically independent, $\underset{\sim}{\nu} \sim \mathrm{Gau}(\underset{\sim}{0}, \Gamma(\tau^2))$, and $\underset{\sim}{\epsilon} \sim \mathrm{Gau}(\underset{\sim}{0}, \Delta)$. Now, assuming that $\delta_1^2, \ldots, \delta_n^2$ are calculated using sampling-variance formulas appropriate for the PES sampling frame, the only parameters left to estimate are $\underset{\sim}{\beta}$ and τ^2. Thus, the two variance components Δ and $\Gamma(\tau^2)$ only contribute one unknown parameter, namely τ^2. It is worth noting that the methods developed in this article can be easily generalized beyond this simple variance-components problem. The general linear model is considered in Section 3.

In Section 2, the Bayes predictor and the empirical Bayes predictor of $\underset{\sim}{F}$ will be given. Estimation of $\underset{\sim}{\beta}$ is straightforward, but there are several possible ways τ^2 could be estimated. Section 3 presents maximum likelihood (m.l.), method-of-moments, and restricted maximum likelihood (REML) approaches. The effect of estimation of τ^2, on mean squared prediction errors, is investigated in Section 4. Section 5 compares the approaches by simulation and by example, and Section 6 presents conclusions and a discussion.

2. EMPIRICAL BAYES PREDICTION

In this article, the true population of any small area is considered to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. Therefore, statistical models for undercount are **conditional** on the observed census counts. The model (1.4), (1.5), and (1.6) has been introduced in Section 1, and will be assumed throughout Sections 2, 3, and 4.

Using a matrix analogue of squared-error loss, the optimal predictor is $E(\underset{\sim}{F} \mid \underset{\sim}{Y})$ (Cressie 1990), which is,

$$\underset{\sim}{p}^*(\underset{\sim}{Y}) \equiv \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}X\underset{\sim}{\beta} \qquad (2.1)$$

and the mean-squared-prediction-error matrix is,

$$E\{(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))'\} = \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\}\Gamma(\tau^2). \qquad (2.2)$$

For the loss matrix, $L(\underset{\sim}{F},\underset{\sim}{p}) \equiv (\underset{\sim}{F} - \underset{\sim}{p})(\underset{\sim}{F} - \underset{\sim}{p})'$, (2.1) is easily seen to be a **Bayes** predictor of $\underset{\sim}{F}$. In reality, $\underset{\sim}{\beta}$ and τ^2 are unknown and so (2.1) is not a statistic (*i.e.* is not a function only of the data). The proper Bayesian approach would be to put further priors and hyperpriors on all unknown parameters. (This solution to the conundrum of unknown parameters is sometimes called hierarchical Bayes, and demands a prior knowledge of process variability that many scientists do not feel they have. Nevertheless, noninformative priors and hyperpriors, particularly, often yield sensible estimators.) Often the posterior distributions are analytically intractable. Should the model and prior be specified according to their conditional distributions, the Gibbs sampler could be used to obtain, numerically, all required marginal and joint distributions (*e.g.* Gelfand and Smith 1990).

An alternative approach, the one taken in this article, is to treat all parameters, except $\underset{\sim}{F}$, as fixed but unknown, and to use the data $\underset{\sim}{Y}$ to estimate them. This approach is called **empirical Bayes**. Although a parametric (conjugate) prior is assumed in this article, one could also work with a nonparametric prior (*e.g.* Laird and Louis 1987).

Suppose now that $\underset{\sim}{\beta}$ is unknown, but that τ^2 in (1.6) is (for the moment) known. Again, using the matrix analogue of squared-error loss, the optimal linear unbiased predictor is obtained by substituting the generalized-least-squares estimator:

$$\hat{\underset{\sim}{\beta}} \equiv \{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y}$$

into (2.1), yielding

$$\begin{aligned}\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2) &= \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\tau^2)(\Delta + \Gamma(\tau^2))^{-1}\} \\ &\quad X\{X'(\Delta + \Gamma(\tau^2))^{-1}X\}^{-1}X'(\Delta + \Gamma(\tau^2))^{-1}\underset{\sim}{Y} \equiv \Lambda(\tau^2)\underset{\sim}{Y} \qquad (2.3)\end{aligned}$$

(Cressie 1990). The mean-squared-prediction-error matrix is,

$$\begin{aligned}M_1(\tau^2) &\equiv E\{(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2))(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\tau^2))'\} \\ &= \Lambda(\tau^2)\Delta\Lambda(\tau^2)' + (\Lambda(\tau^2) - I)\Gamma(\tau^2)(\Lambda(\tau^2) - I)'. \qquad (2.4)\end{aligned}$$

More realistically, τ^2 is also unknown. An **empirical Bayes predictor** is obtained by substituting an estimator $\hat{\tau}^2$ into $\Lambda(\tau^2)$ to yield,

$$\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\, \hat{\tau}^2) = \Lambda(\hat{\tau}^2)\underset{\sim}{Y}. \tag{2.5}$$

It is easy to see that when $\hat{\tau}^2$ is the maximum likelihood estimator of τ^2, then (2.5) is the maximum likelihood estimator of the Bayes predictor.

The predictor (2.5) was suggested by Ericksen and Kadane (1985) (and criticized by Freedman and Navidi 1986). Incidentally, the form of their predictors may look different to (2.1), (2.3), and (2.5), but they are in fact identical upon using the identity: $A(A + B)^{-1}B = (A^{-1} + B^{-1})^{-1}$, where A and B are square matrices such that A, B, and $A + B$ have inverses.

By substituting $\hat{\tau}^2$ into (2.4), an estimator of the mean-squared-prediction-error matrix:

$$M_1(\hat{\tau}^2) \equiv \Lambda(\hat{\tau}^2)\Delta\Lambda(\hat{\tau}^2)' + (\Lambda(\hat{\tau}^2) - I)\Gamma(\hat{\tau}^2)(\Lambda(\hat{\tau}^2) - I)' \tag{2.6}$$

is obtained. Since (2.6) does not take into account the estimation of τ^2 in $\hat{\underset{\sim}{p}}(\underset{\sim}{Y};\, \hat{\tau}^2)$, it is likely to be a biased estimator of $E\{(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\, \hat{\tau}^2))(\underset{\sim}{F} - \hat{\underset{\sim}{p}}(\underset{\sim}{Y};\, \hat{\tau}^2))'\}$. Further discussion of this important issue is given in Section 4.

Having obtained $\hat{\underset{\sim}{\beta}}$ and $\hat{\tau}^2$, model diagnostics can be computed to check the fit of the estimated model. For example, a quantile-quantile plot, of the standardized residuals $(\Delta + \Gamma(\hat{\tau}^2))^{-1/2}(\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}})$ against expected order statistics from a unit Gaussian distribution, was used to show no obvious lack of fit of the model used in Section 5. A more complete discussion of model diagnostics is given in Section 6.

3. ESTIMATION OF VARIANCE-MATRIX PARAMETERS

In this section, the general linear model,

$$\underset{\sim}{Y} \sim \mathrm{Gau}(X\underset{\sim}{\beta},\, \Sigma(\underset{\sim}{\gamma})), \tag{3.1}$$

will be assumed, where $\underset{\sim}{\gamma}$ is a $k \times 1$ vector of variance-matrix parameters. In particular, the model given by (1.4), (1.5), and (1.6) yields,

$$\Sigma(\underset{\sim}{\gamma}) = \Delta + \Gamma(\tau^2), \tag{3.2}$$

where $\underset{\sim}{\gamma}$ consists of only one parameter, τ^2.

For $\underset{\sim}{\gamma}$ known, estimation of $\underset{\sim}{\beta}$ is straightforward:

$$\hat{\underset{\sim}{\beta}}(\underset{\sim}{\gamma}) \equiv (X'\Sigma(\underset{\sim}{\gamma})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\gamma})^{-1}\underset{\sim}{Y}. \tag{3.3}$$

More realistically, $\underset{\sim}{\gamma}$ is unknown and has to be estimated; substitution of that estimator into (3.3) then yields an estimated generalized least squares estimator of $\underset{\sim}{\beta}$. In the rest of this section, three different methods of estimating $\underset{\sim}{\gamma}$ will be considered.

3.1 Maximum Likelihood Estimation

The negative log likelihood of $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$ is:

$$\begin{aligned} L(\underset{\sim}{\beta}, \underset{\sim}{\gamma}) = {} & (n/2)\log(2\pi) + (1/2)\log(|\Sigma(\underset{\sim}{\gamma})|) + \\ & (1/2)(\underset{\sim}{Y} - X\underset{\sim}{\beta})'\Sigma(\underset{\sim}{\gamma})^{-1}(\underset{\sim}{Y} - X\underset{\sim}{\beta}). \end{aligned} \tag{3.4}$$

Minimization of this function yields maximum likelihood (m.l.) estimates $\hat{\beta}_{m\ell}$ and $\hat{\gamma}_{m\ell}$. The difficult part of this minimization involves finding $\hat{\gamma}_{m\ell}$. The Gauss-Newton (scoring) algorithm is given *inter alia* by Harville (1977) and Mardia and Marshall (1984) and is repeated here for notational completeness.

Define,

$$\Sigma_i(\gamma) \equiv \partial\Sigma(\gamma)/\partial\gamma_i;\ i = 1, \ldots, k,$$
$$\Sigma^i(\gamma) \equiv \partial\Sigma^{-1}(\gamma)/\partial\gamma_i = -\ \Sigma(\gamma)^{-1}\Sigma_i(\gamma)\Sigma(\gamma)^{-1};\ i = 1, \ldots, k, \tag{3.5}$$

the $k \times 1$ vector L_γ to have i-th element:

$$(L_\gamma)_i \equiv (½)\mathrm{tr}(\Sigma(\gamma)^{-1}\Sigma_i(\gamma)) + (½)(Y - X\beta)'\Sigma^i(\gamma)(Y - X\beta), \tag{3.6}$$

and the $k \times k$ matrix J_γ to have (i,j)-th element:

$$(J_\gamma)_{ij} \equiv (½)\mathrm{tr}(\Sigma(\gamma)^{-1}\Sigma_i(\gamma)\Sigma(\gamma)^{-1}\Sigma_j(\gamma)). \tag{3.7}$$

Then,

$$\gamma^{(\ell+1)} = \gamma^{(\ell)} - (J_\gamma^{(\ell)})^{-1}L_\gamma^{(\ell)}, \tag{3.8}$$

where $J_\gamma^{(\ell)}$ and $L_\gamma^{(\ell)}$ denotes J_γ and L_γ, respectively, evaluated at $\gamma = \gamma^{(\ell)}$ and $\beta = \hat{\beta}(\gamma^{(\ell)})$.

When γ consists of only τ^2 in (1.6), the algorithm (3.8) is particularly straightforward. In the simulations and example given in Section 5, the starting value

$$(\tau^2)^{(0)} \equiv \{1/(n - p)\}(Y - X(X'D^{-1}X)^{-1}X'D^{-1}Y)'D^{-1}$$
$$(Y - X(X'D^{-1}X)^{-1}X'D^{-1}Y), \tag{3.9}$$

was used. Then (3.8) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - \{(½)\sum_{i=1}^{n} 1/(C_i\delta_i^2 + (\tau^2)^{(\ell)})^2\}^{-1}L_\tau^{(\ell)};\ \ell = 0, 1, \ldots, \tag{3.10}$$

where

$$L_\tau^{(\ell)} = (½)\sum_{i=1}^{n} 1/(C_i\delta_i^2 + (\tau^2)^{(\ell)})$$
$$-\ (½)\{Y - X\hat{\beta}((\tau^2)^{(\ell)})\}'\mathrm{diag}\{C_i/(C_i\delta_i^2 + (\tau^2)^{(\ell)})^2\}\{Y - X\hat{\beta}((\tau^2)^{(\ell)})\}. \tag{3.11}$$

Iterating (3.8) to convergence yields the m.l. estimator $\hat{\gamma}_{m\ell}$, which upon substitution into (3.3) yields the m.l. estimator $\hat{\beta}(\hat{\gamma}_{m\ell})$. Under appropriate regularity conditions (*e.g.* Mardia and Marshall 1984) $(\hat{\beta}(\hat{\gamma}_{m\ell})', \hat{\gamma}_{m\ell}')'$ is approximately multivariate Gaussian, with mean $(\beta', \gamma')'$ and asymptotic variance matrix,

$$\begin{bmatrix} (X'\Sigma(\underset{\sim}{\gamma})^{-1}X)^{-1} & 0 \\ 0 & J_{\gamma}^{-1} \end{bmatrix}; \tag{3.12}$$

when $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the matrix (3.12) becomes,

$$\begin{bmatrix} (X'\Sigma(\tau^2)^{-1}X)^{-1} & 0 \\ 0 & \left\{(1/2)\sum_{i=1}^{n} 1/(C_i\delta_i^2 + \tau^2)^2\right\}^{-1} \end{bmatrix}. \tag{3.13}$$

In practice, estimated variances and covariances are obtained by evaluating (3.12) at the m.l. estimate $\hat{\underset{\sim}{\gamma}}_{m\ell}$.

3.2 Method-of-Moments Estimation

There is no single method-of-moments estimator of $\underset{\sim}{\gamma}$, but the general idea is to match low-order moments of data with corresponding empirical moments. If only first- and second-order moments are used, it is clear that the Gaussian assumption in (3.1) is not needed.

Let U be a positive-definite symmetric matrix. Consider the weighted regression estimator, $\hat{\underset{\sim}{\beta}}_U \equiv (X'U^{-1}X)^{-1}X'U^{-1}\underset{\sim}{Y}$, and the weighted residuals,

$$\underset{\sim}{e}_U \equiv U^{-1/2}(I - X(X'U^{-1}X)^{-1}X'U^{-1})\underset{\sim}{Y}. \tag{3.14}$$

Then, straightforward matrix algebra shows that,

$$E(\underset{\sim}{e}_U'\underset{\sim}{e}_U) = \mathrm{tr}(\Sigma(\underset{\sim}{\gamma})\Pi_U), \tag{3.15}$$

where $\Pi_U \equiv U^{-1} - U^{-1}X(X'U^{-1}X)^{-1}X'U^{-1}$. Assuming that $\Sigma(\underset{\sim}{\gamma}) = \Delta + \gamma_1\Gamma_1 + \ldots + \gamma_k\Gamma_k$, where Γ_i's are known, one obtains,

$$\sum_{i=1}^{k} \gamma_i \mathrm{tr}(\Gamma_i\Pi_U) = E(\underset{\sim}{e}_U'\underset{\sim}{e}_U) - \mathrm{tr}(\Delta\Pi_U).$$

Choice of k different U_j; $j = 1, \ldots, k$ (*e.g.* $U_1, U_1^2, \ldots, U_1^k$) yields k equations in k unknowns:

$$\sum_{i=1}^{k} \gamma_i \mathrm{tr}(\Gamma_i\Pi_{U_j}) = \underset{\sim}{e}_{U_j}'\underset{\sim}{e}_{U_j} - \mathrm{tr}(\Delta\Pi_{U_j}); j = 1, \ldots, k, \tag{3.16}$$

which can be solved for $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$. It is important to check that the solution $\hat{\underset{\sim}{\gamma}}$ is in the parameter space $\{\underset{\sim}{\gamma}: \sum_{i=1}^{k} \gamma_i\Gamma_i$ is positive-definite$\}$.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), only one matrix U in (3.16) is needed. Previous undercount predictors have based their estimate of τ^2 on $U = I$ (Ericksen and Kadane 1985;

Freedman and Navidi 1986; Ericksen, Kadane and Tukey 1989), but a small sensitivity study for the heteroskedastic model (1.6) suggested a better estimator.

Choose $U_\alpha = \Delta + \Gamma(\alpha)$ in (3.15) to mimic the model (1.7). Then, when $\alpha = \tau^2$ (the true value), Fay and Herriot (1979) show that

$$E(\underset{\sim}{e}'_{U_\alpha}\underset{\sim}{e}_{U_\alpha}) = n - p, \tag{3.17}$$

where n is the number of areas, p is the number of regressors in the matrix X (*e.g.* $p = 3$ for the selected model in Section 5), and $\underset{\sim}{e}_U$ is the standardized residual defined by (3.14). Thus, the proposed method-of-moments estimator of τ^2 is the value of α for which

$$\underset{\sim}{e}'_{U_\alpha}\underset{\sim}{e}_{U_\alpha} = n - p, \tag{3.18}$$

which can be solved using a Newton-Raphson iterative method or a simple bisection method; call the resulting estimator $\hat{\tau}^2_{mm}$.

Fay and Herriot (1979) note that the difference between $\hat{\tau}^2_{mm}$ and $\hat{\tau}^2_{m\ell}$ is manifest in how an area with small δ^2_i is weighted in the estimation procedure; $\hat{\tau}^2_{m\ell}$ gives relatively more weight to the squared residuals for such an area than does $\hat{\tau}^2_{mm}$. Based on this weighting property, and a small simulation study of bias, Cressie (1990) expressed a preference for $\hat{\tau}^2_{mm}$ over $\hat{\tau}^2_{m\ell}$. However, asymptotically, $\hat{\tau}^2_{m\ell}$ is fully efficient and has an accessible distribution theory. Lack of any (asymptotic) distributional results for $\hat{\tau}^2_{mm}$ causes its own set of problems, such as how to make inference on τ^2, and how to carry out mean-squared- prediction-error corrections in Section 4. A more satisfactory estimator, with better bias properties than the m.l. estimator, is developed below.

3.3 Restricted Maximum Likelihood Estimation

The problem is to find a suitable estimator of the variance-matrix parameters $\underset{\sim}{\gamma}$ in (3.1). The method of restricted maximum likelihood (REML), developed originally by Patterson and Thompson (1971, 1974), applies maximum likelihood to error contrasts rather than to the data themselves. (Rao (1979) calls this method MML, marginal maximum likelihood, in the context of estimation of variance components. Recently, some authors have also called it residual maximum likelihood, although they have retained the abbreviation REML.) A linear combination $\underset{\sim}{a}'\underset{\sim}{Y}$ is called an error contrast if $E(\underset{\sim}{a}'\underset{\sim}{Y}) = 0$, for all $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$; thus, $\underset{\sim}{a}'\underset{\sim}{Y}$ is an error contrast if and only if $\underset{\sim}{a}'X = \underset{\sim}{0}'$.

Let $\underset{\sim}{W} = A'\underset{\sim}{Y}$ represent a vector of $(n - p)$ linearly independent error contrasts; *i.e.* the $(n - p)$ columns of A are linearly independent and $A'X = 0$. Under the Gaussian assumption (3.1), $\underset{\sim}{W} \sim \mathrm{Gau}(\underset{\sim}{0}, A'\Sigma(\underset{\sim}{\gamma})A)$, which does not depend on $\underset{\sim}{\beta}$. Thus, the negative log likelihood function is,

$$L_W(\underset{\sim}{\gamma}) = ((n - p)/2)\log(2\pi) + (½)\log(\,|A'\Sigma(\underset{\sim}{\gamma})A|\,) + (½)\underset{\sim}{W}'(A'\Sigma(\underset{\sim}{\gamma})A)^{-1}\underset{\sim}{W}.$$

If another set of $(n - p)$ linearly independent contrasts were used to define $\underset{\sim}{W}$, the new negative log likelihood function would differ from $L_W(\underset{\sim}{\gamma})$ only by an additive constant (Harville 1974). Indeed, for the A that satisfies $AA' = I - X(X'X)^{-1}X'$ (and $A'A = I$),

$$L_W(\underset{\sim}{\gamma}) = ((n - p)/2)\log(2\pi) - (½)\log(\,|X'X|\,) + (½)\log(\,|\Sigma(\underset{\sim}{\gamma})|\,) + (½)\log(\,|X'\Sigma(\underset{\sim}{\gamma})^{-1}X|\,) + (½)\underset{\sim}{Y}'\Pi(\underset{\sim}{\gamma})\underset{\sim}{Y}, \tag{3.19}$$

where $\Pi(\underset{\sim}{\gamma}) \equiv \Sigma(\underset{\sim}{\gamma})^{-1} - \Sigma(\underset{\sim}{\gamma})^{-1}X(X'\Sigma(\underset{\sim}{\gamma})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\gamma})^{-1}$; see Harville (1974). A REML estimate of $\underset{\sim}{\gamma}$, denoted $\hat{\underset{\sim}{\gamma}}_{r\ell}$, is obtained by minimizing (3.19) with respect to $\underset{\sim}{\gamma}$. The distinction between REML and m.l. estimation becomes important when p is large relative to n. The REML method was originally proposed to estimate variance-component parameters: Numerical algorithms (Harville 1977), robust adaptations (Fellner 1986), and distribution theory (Cressie and Lahiri 1991) have been developed in this context. Kitanidis (1983) and Zimmerman (1989) give computational details for producing an iterative minimization of (3.19).

Harville (1974) provides a Bayesian justification for REML by assuming a noninformative prior for $\underset{\sim}{\beta}$, which is statistically independent of $\underset{\sim}{\gamma}$, and showing that the marginal posterior density of $\underset{\sim}{\gamma}$ is proportional to (3.19) multiplied by the prior for $\underset{\sim}{\gamma}$. When that prior is noninformative, REML estimates correspond to marginal MAP (maximum *a posteriori*) estimates. Thus, in the situation where noninformative prior distributions for $\underset{\sim}{\beta}$ and $\underset{\sim}{\gamma}$ are independent, REML can be seen as a compromise between m.l. and Bayes estimation with squared error loss. In the case of model (1.4), (1.5) and (1.6), the latter would yield a Bayes estimate, $\int_0^\infty \tau^2 \exp\{-L_W(\tau^2)\}d\tau^2$, which can be obtained equivalently by **averaging** τ^2, weighted by the *full* likelihood, $\exp\{-L(\underset{\sim}{\beta}, \tau^2)\}$. On the other hand, m.l. yields as an estimate of τ^2 the value $\hat{\tau}^2_{m\ell}$ obtained by **maximizing** the **full** likelihood. REML averages the full likelihood over $\underset{\sim}{\beta}$ but maximizes the resulting (restricted) likelihood over τ^2.

Maximum likelihood estimation of τ^2 tends to be biased towards zero because the likelihood, as a function of τ^2, is skewed to the right. When normalized to integrate to one, the mean of such a function is generally larger than its mode (*e.g.* Groeneveld and Meeden 1977). The m.l. estimate is based on the profile of the likelihood surface of $\underset{\sim}{\beta}$ and τ^2, and this favors smaller values of τ^2. (In contrast, REML is obtained by first integrating the likelihood over $\underset{\sim}{\beta}$ and then maximizing the result over τ^2. Notice that Bayesians might advocate further integration over τ^2.)

Although the Bayesian interpretation of REML helps to explain its properties, $\hat{\underset{\sim}{\gamma}}_{r\ell}$ also has the obvious frequentist interpretation of being an estimator based on restricted information.

Minimization of (3.19) with respect to $\underset{\sim}{\gamma}$ can proceed by any of the gradient algorithms. Recall,

$$\underset{\sim}{W} = A'\underset{\sim}{Y} \tag{3.20}$$

and suppose A satisfies:

$$AA' = I - X(X'X)^{-1}X', \text{ and } A'A = I.$$

For the moment, focus all attention on the $(n - p)$ "data" $\underset{\sim}{W}$; their joint distribution depends only on $\underset{\sim}{\gamma}$, and the associated negative log (restricted) likelihood is $L_W(\underset{\sim}{\gamma})$ given by (3.19).

Define the $k \times 1$ vector $\underset{\sim}{M}_\gamma$ to have i-th element:

$$(\underset{\sim}{M}_\gamma)_i \equiv \partial L_W(\underset{\sim}{\gamma})/\partial\gamma_i = (1/2)\mathrm{tr}\{\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\} - (1/2)\underset{\sim}{Y}'\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\Pi(\underset{\sim}{\gamma})\underset{\sim}{Y}, \tag{3.21}$$

and the $k \times k$ matrix G_γ to have (i, j)-th element:

$$(G_\gamma)_{ij} \equiv E(\partial^2 L_W(\underset{\sim}{\gamma})/\partial\gamma_i\partial\gamma_j) = (1/2)\mathrm{tr}\{\Pi(\underset{\sim}{\gamma})\Sigma_i(\underset{\sim}{\gamma})\Pi(\underset{\sim}{\gamma})\Sigma_j(\underset{\sim}{\gamma})\}, \tag{3.22}$$

where $\Pi(\underset{\sim}{\gamma})$ is given below (3.19) and $\Sigma_i(\underset{\sim}{\gamma})$ is defined by (3.5). (The expressions (3.21) and (3.22) were obtained by Harville 1977.) Then, the Gauss-Newton (scoring) algorithm to find $\hat{\underset{\sim}{\gamma}}_{r\ell}$ is:

$$\underset{\sim}{\gamma}^{(\ell+1)} = \underset{\sim}{\gamma}^{(\ell)} - (G_{\gamma}^{(\ell)})^{-1}\underset{\sim}{M}_{\gamma}^{(\ell)}, \tag{3.23}$$

where $G_{\gamma}^{(\ell)}$ and $\underset{\sim}{M}_{\gamma}^{(\ell)}$ denote G_{γ} and $\underset{\sim}{M}_{\gamma}$, respectively, evaluated at $\underset{\sim}{\gamma} = \underset{\sim}{\gamma}^{(\ell)}$.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the algorithm (3.23) is particularly straightforward. In the simulations and example given in Section 5, the starting value (3.9) was used. Then (3.23) is,

$$(\tau^2)^{(\ell+1)} = (\tau^2)^{(\ell)} - (G_{\tau}^{(\ell)})^{-1}M_{\tau}^{(\ell)}, \tag{3.24}$$

where

$$M_{\tau} = (½)\mathrm{tr}\{\Pi(\tau^2)D\} - (½)\underset{\sim}{Y}'\Pi(\tau^2)D\Pi(\tau^2)\underset{\sim}{Y}, \tag{3.25}$$

$$G_{\tau} = (½)\mathrm{tr}\{\Pi(\tau^2)D\Pi(\tau^2)D\}, \tag{3.26}$$

$$\Pi(\tau^2) = \Sigma(\tau^2)^{-1} - \Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1}, \tag{3.27}$$

are evaluated at $\tau^2 = (\tau^2)^{(\ell)}$. Also, recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$ and $D = \mathrm{diag}\{1/C_1, \ldots, 1/C_n\}$.

Iterating (3.23) to convergence yields the REML estimator $\hat{\underset{\sim}{\gamma}}_{r\ell}$. It has been proved by Cressie and Lahiri (1991) that $\hat{\underset{\sim}{\gamma}}_{r\ell}$ is approximately multivariate Gaussian, with mean $\underset{\sim}{\gamma}$ and asymptotic variance matrix,

$$G_{\gamma}^{-1}. \tag{3.28}$$

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the matrix (3.28) becomes a scalar,

$$[(½)\mathrm{tr}\{\Pi(\tau^2)D\Pi(\tau^2)D\}]^{-1}. \tag{3.29}$$

In practice, estimated variances and covariances are obtained by evaluating (3.28) at $\underset{\sim}{\gamma} = \hat{\underset{\sim}{\gamma}}_{r\ell}$. Furthermore, the normalized (estimated) generalized least squares estimator, $\hat{\underset{\sim}{\beta}}(\hat{\underset{\sim}{\gamma}}_{r\ell})$ should be approximately Gaussian with asymptotic variance matrix, $(X'\Sigma(\underset{\sim}{\gamma})X)^{-1}$.

4. IMPROVED ESTIMATION OF MEAN SQUARED PREDICTION ERRORS

In what is to follow, I shall be concerned with the effect, on prediction, of estimation of $\underset{\sim}{\gamma}$ in $\Sigma(\underset{\sim}{\gamma})$ given by (3.1). Generalizing (1.5) to,

$$\underset{\sim}{F} \sim \mathrm{Gau}(X\underset{\sim}{\beta}, \Gamma(\underset{\sim}{\gamma})), \tag{4.1}$$

it is clear that

$$\Sigma(\underset{\sim}{\gamma}) = \Delta + \Gamma(\underset{\sim}{\gamma}). \tag{4.2}$$

In principle, Δ could also depend on unknown parameters (in, *e.g.* a model for sampling variances) and the results of this section are equally applicable. The optimal linear unbiased predictor is,

$$\underset{\sim}{p}(\underset{\sim}{Y};\underset{\sim}{\gamma}) = \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\underset{\sim}{Y} + \{I - \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\}$$

$$X\{X'(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}X\}^{-1}X'(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\underset{\sim}{Y} \equiv \Lambda(\underset{\sim}{\gamma})\underset{\sim}{Y}. \tag{4.3}$$

Then, the mean-squared-prediction-error matrix of $\underset{\sim}{p}(\underset{\sim}{Y};\underset{\sim}{\gamma})$, denoted $M_1(\underset{\sim}{\gamma})$, is given by,

$$M_1(\underset{\sim}{\gamma}) = \Lambda(\underset{\sim}{\gamma})\Delta\Lambda(\underset{\sim}{\gamma})' + (\Lambda(\underset{\sim}{\gamma}) - I)\Gamma(\underset{\sim}{\gamma})(\Lambda(\underset{\sim}{\gamma}) - I)'. \tag{4.4}$$

In reality, $\underset{\sim}{\gamma}$ is unknown and has to be estimated by $\hat{\underset{\sim}{\gamma}}$, say. The empirical Bayes predictor of $\underset{\sim}{F}$ is then $\underset{\sim}{p}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}})$, given by (4.3) with $\underset{\sim}{\gamma} = \hat{\underset{\sim}{\gamma}}$. In this case, $M_1(\underset{\sim}{\gamma})$ is an inappropriate measure of the predictor's precision; one should use instead,

$$M_2(\underset{\sim}{\gamma}) = E\{(\underset{\sim}{F} - \underset{\sim}{p}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}}))(\underset{\sim}{F} - \underset{\sim}{p}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}}))'\}. \tag{4.5}$$

It is the risk matrix (4.5), or an estimate of it, that should be given, along with the predictor $\underset{\sim}{p}(\underset{\sim}{Y};\hat{\underset{\sim}{\gamma}})$. However, $M_1(\hat{\underset{\sim}{\gamma}})$ is typically reported; hence, one should ask what inaccuracies result from using $M_1(\hat{\underset{\sim}{\gamma}})$ and whether a more appropriate estimator of $M_2(\underset{\sim}{\gamma})$ is available.

Now, under the assumptions (4.1) and (4.2) (Gaussianity is important here) and provided $\hat{\underset{\sim}{\gamma}}$ is an even and translation invariant function of the data, the results of Harville (1985) can be used to establish that $M_2(\underset{\sim}{\gamma}) - M_1(\underset{\sim}{\gamma})$ is non-negative-definite. (An estimator is even if $\hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y}) = \hat{\underset{\sim}{\gamma}}(-\underset{\sim}{Y})$ and is translation invariant if $\hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y} + X\underset{\sim}{\lambda}) = \hat{\underset{\sim}{\gamma}}(\underset{\sim}{Y})$, for any $p \times 1$ vector $\underset{\sim}{\lambda}$.) When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), the estimators $\hat{\tau}^2_{m\ell}$, $\hat{\tau}^2_{mm}$ and $\hat{\tau}^2_{r\ell}$ are all even and translation invariant. Intuitively, estimation of the unknown parameters $\underset{\sim}{\gamma}$ leads to larger mean squared prediction errors; the result above quantifies this intuition.

But, there is another potential source of bias due to the fact that $M_1(\hat{\underset{\sim}{\gamma}})$, not $M_1(\underset{\sim}{\gamma})$, is used to estimate the risk matrix. Suppose that $\hat{\underset{\sim}{\gamma}}$ is chosen to yield an unbiased estimator of the variance matrix of $(\underset{\sim}{Y}', \underset{\sim}{F}')'$, which most would agree is a desirable property. Then the results of Eaton (1985) and Zimmerman and Cressie (1991) can be used to establish that $M_1(\underset{\sim}{\gamma}) - E(M_1(\hat{\underset{\sim}{\gamma}}))$ is non-negative-definite. (The proof relies on a multivariate version of Jensen's inequality and on the fact that $\underset{\sim}{p}(\underset{\sim}{Y};\underset{\sim}{\gamma})$, which can be written as $\Lambda(\underset{\sim}{\gamma})\underset{\sim}{Y}$, minimizes the risk matrix over all linear unbiased predictors.)

Upon writing,

$$M_2(\underset{\sim}{\gamma}) - M_1(\hat{\underset{\sim}{\gamma}}) = \{M_2(\underset{\sim}{\gamma}) - M_1(\underset{\sim}{\gamma})\} + \{M_1(\underset{\sim}{\gamma}) - E(M_1(\hat{\underset{\sim}{\gamma}}))\} +$$

$$\{E(M_1(\hat{\underset{\sim}{\gamma}})) - M_1(\hat{\underset{\sim}{\gamma}})\}, \tag{4.6}$$

the results above establish that underestimation of $M_2(\underset{\sim}{\gamma})$ comes from two sources. Even if an expression for $M_2(\underset{\sim}{\gamma})$ were known, it is likely that $M_2(\hat{\underset{\sim}{\gamma}})$ would be biased for $M_2(\underset{\sim}{\gamma})$, further illustrating the inherent difficulty in estimating mean squared prediction errors.

A remedy has been suggested by Prasad and Rao (1990), based on asymptotic expansions of $M_2(\underset{\sim}{\gamma})$. Consider prediction of undercount in the i-th area, and let $[M_2(\underset{\sim}{\gamma})]_{ii}$ and $[M_1(\hat{\underset{\sim}{\gamma}})]_{ii}$ denote the (i,i)-th elements of the risk matrices $M_2(\underset{\sim}{\gamma})$ and $M_1(\hat{\underset{\sim}{\gamma}})$, respectively. Then formal application of Prasad and Rao's proposal yields the estimator of $[M_2(\underset{\sim}{\gamma})]_{ii}$,

$$[M_2(\underset{\sim}{\gamma})]^*_{ii} \equiv [M_1(\hat{\underset{\sim}{\gamma}})]_{ii} + 2\mathrm{tr}\{A_{ii}(\hat{\underset{\sim}{\gamma}})B(\hat{\underset{\sim}{\gamma}})\};\ i = 1, \ldots, n. \tag{4.7}$$

In (4.7), $A_{ii}(\underset{\sim}{\gamma})$ is a $k \times k$ matrix given by,

$$A_{ii}(\underset{\sim}{\gamma}) = \mathrm{var}\{\partial \hat{p}_i(\underset{\sim}{Y};\underset{\sim}{\gamma})/\partial \underset{\sim}{\gamma}\} \tag{4.8}$$

and $B(\underset{\sim}{\gamma})$ is a matrix that equals or approximates the $k \times k$ matrix,

$$E\{(\hat{\underset{\sim}{\gamma}} - \underset{\sim}{\gamma})(\hat{\underset{\sim}{\gamma}} - \underset{\sim}{\gamma})'\}. \tag{4.9}$$

For m.l. estimation,

$$B(\underset{\sim}{\gamma}) = J_\gamma^{-1}, \tag{4.10}$$

where J_γ is given by (3.7), and for REML estimation,

$$B(\underset{\sim}{\gamma}) = G_\gamma^{-1}, \tag{4.11}$$

where G_γ is given by (3.22).

Kass and Steffey (1989) give approximations (to the conditional variance) that are similar in spirit to (4.7), for probability distributions that are not necessarily Gaussian. However, their approach requires independent replications, which is not a feature of the distributions specified by (3.1).

Should small areas be aggregated, it is important to have an approximately unbiased estimator of all elements of $M_2(\underset{\sim}{\gamma})$. It is not difficult to generalize (4.7) to,

$$[M_2(\underset{\sim}{\gamma})]_{ij}^* = [M_1(\hat{\underset{\sim}{\gamma}})]_{ij} + 2\mathrm{tr}\{A_{ij}(\hat{\underset{\sim}{\gamma}})B(\hat{\underset{\sim}{\gamma}})\};\ i,j = 1, \ldots, n,$$

where $A_{ij}(\underset{\sim}{\gamma}) \equiv \mathrm{cov}\{\partial \hat{p}_i(\underset{\sim}{Y};\underset{\sim}{\gamma})/\partial \underset{\sim}{\gamma}, \partial \hat{p}_j(\underset{\sim}{Y};\underset{\sim}{\gamma})/\partial \underset{\sim}{\gamma}\}$. Prasad and Rao (1990) show that, to the same order of magnitude, $A_{ij}(\underset{\sim}{\gamma})$ can be replaced by $\mathrm{cov}\{\partial p_i^*(\underset{\sim}{Y})/\partial \underset{\sim}{\gamma}, \partial p_j^*(\underset{\sim}{Y})/\partial \underset{\sim}{\gamma}\}$, where $\underset{\sim}{p}^*(\underset{\sim}{Y})$ is given by (2.1); these latter derivatives can be simpler to calculate.

When $\underset{\sim}{\gamma}$ consists of only τ^2 in (1.6), calculation of $B(\underset{\sim}{\gamma})$ is straightforward; see (3.13) and (3.26). Now, consider

$$\mathrm{var}(\partial \underset{\sim}{\hat{p}}(\underset{\sim}{Y};\tau^2)/\partial\tau^2) = (\partial\Lambda(\tau^2)/\partial\tau^2)\Sigma(\tau^2)(\partial\Lambda(\tau^2)/\partial\tau^2)', \tag{4.12}$$

where $\Lambda(\tau^2)$ is given by (2.3). In terms of $\Pi(\tau^2)$ defined by (3.27), and Δ defined by (1.4),

$$\Lambda(\tau^2) = I - \Delta\Pi(\tau^2). \tag{4.13}$$

Thus, (4.12) can be calculated from (4.13) using the relationships (3.4) and (3.5). Then, $A_{ii}(\tau^2)$ given by (4.8) is the (i,i)-th element of,

$$\Delta(\partial\Pi(\tau^2)/\partial\tau^2)\Sigma(\tau^2)(\partial\Pi(\tau^2)/\partial\tau^2)'\Delta', \tag{4.14}$$

where

$$\begin{aligned}\partial\Pi(\tau^2)/\partial\tau^2 = {} & -\Sigma(\tau^2)D\Sigma(\tau^2)\{I - X(X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1}\} - \\ & \Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1}\{X'\Sigma(\tau^2)^{-1}D\Sigma(\tau^2)^{-1}X\} \\ & (X'\Sigma(\tau^2)^{-1}X)^{-1}X'\Sigma(\tau^2)^{-1} + \Sigma(\tau^2)^{-1}X(X'\Sigma(\tau^2)^{-1}X)^{-1} \\ & X'\Sigma(\tau^2)^{-1}D\Sigma(\tau^2)^{-1};\end{aligned} \tag{4.15}$$

recall that $\Sigma(\tau^2) = \Delta + \tau^2 D$, and $D = \mathrm{diag}\{1/C_1, \ldots, 1/C_n\}$.

The estimator of mean squared prediction error, $[M_2(\tau^2)]^*_{ii}$, is conjectured to be approximately unbiased (Prasad and Rao's 1990, results were obtained for a more specific model than is considered here). It is obtained by bringing together the relations (4.7), (4.14) and (4.10) for m.l. estimation, or (4.7), (4.14) and (4.11) for REML estimation. This estimator will be compared to the often-reported estimator $[M_1(\hat{\tau}^2)]_{ii}$, in Section 5, using 1980 U. S. Census and Post Enumeration Survey data.

5. A COMPARISON OF ESTIMATORS BY EXAMPLE AND BY SIMULATION

5.1 Example

The PEP 3-8 data from the 1980 Post Enumeration Survey, for the $n = 51$ states of the USA (including Washington, DC) are used to illustrate the empirical Bayes approach. These data are presented in Cressie (1989, Table 1, "Total" columns) and the variances $\delta_1^2, \ldots, \delta_{51}^2$ in (1.3) are obtained from Cressie's "Total" column labeled MSE$^{1/2}$ (whose squared entries will be denoted MSE$_1, \ldots,$ MSE$_{51}$). Using the relation $F_i = \{1 - U_i/100\}^{-1}$ and the δ-method, $\delta_i^2 \simeq (Y_i)^4(\text{MSE}_i)/10^4$. Eight explanatory variables, given by Ericksen, Kadane and Tukey (1989), were collapsed to the 51 states (from 66 small areas that included cities, rest of states and states). The explanatory variables are:

1. Minority percentage.
2. Crime rate.
3. Poverty percentage.
4. Percentage with language difficulty.
5. Education.
6. Housing.
7. Proportion of population in any of 16 prespecified central cities.
8. Percentage conventionally counted in the census.

To find a subset of these variables that provides a good model for undercount, I used the selection method of Ericksen, Kadane and Tukey (1989), but weighted the data proportionally to the square roots of the small areas' census counts. The variables selected were 1 (minority) and 5 (education), as well as the constant term. Henceforth, in this paper, these three variables will be the only ones considered in the linear model; *i.e.* only regression coefficients β_0, β_1 and β_5 will be fit.

Under the model (1.4), (1.5) and (1.6), the unknown parameters are $\underset{\sim}{\beta}$ and τ^2. From the scoring algorithm (3.8), the m.l. estimate of τ^2 is:

$$\hat{\tau}^2_{m\ell} = 47.32,$$

while from the scoring algorithm (3.23), the REML estimate of τ^2 is:

$$\hat{\tau}^2_{r\ell} = 58.53.$$

This illustrates a phenomenon observed from the realizations of a simulation presented below, namely, that $\hat{\tau}^2_{m\ell} < \hat{\tau}^2_{r\ell}$; an intuitive explanation is given in Section 3.3. (Parenthetically, Cressie (1990), obtained $\hat{\tau}^2_{mm} = 94.96$, but no general inequality between it, m.l., and REML is apparent.)

From the formulas in Section 3, the following estimates (with estimated standard errors in parentheses) were obtained:

m.l.	REML
$\hat{\beta}_0 = 1.03227\ (0.00708)$	$\hat{\beta}_0 = 1.03246\ (0.00724)$
$\hat{\beta}_1 = 0.0006878\ (0.0001402)$	$\hat{\beta}_1 = 0.0006941\ (0.0001436)$
$\hat{\beta}_5 = -0.001070\ (0.000231)$	$\hat{\beta}_5 = -0.001078\ (0.000236)$
$\hat{\tau}^2 = 47.32\ (32.87)$	$\hat{\tau}^2 = 58.53\ (38.1)$.

Notice that there is very little difference between the two sets of estimates, except for that of τ^2. Upon using the m.l. and REML estimates in $\hat{p}_i(\underset{\sim}{Y};\, \hat{\tau}^2)$ given by (2.5), $[M_1(\hat{\tau}^2)]_{ii}$ given by (2.6), and $[M_2(\tau^2)]^*_{ii}$ given by (4.7); $i = 1, \ldots, n$, small-area predictors and estimated root mean squared prediction errors are obtained. Table 1 shows the results for the $n = 51$ states; also shown in the table are the raw undercount data Y_i, the fitted linear model $(X\underset{\sim}{\hat{\beta}})_i$, and the weight,

$$w_i \equiv \hat{\tau}^2/(C_i\delta_i^2 + \hat{\tau}^2), \tag{5.1}$$

such that

$$\hat{p}_i(\underset{\sim}{Y};\, \hat{\tau}^2) = w_iY_i + (1 - w_i)(X\underset{\sim}{\hat{\beta}})_i;\ i = 1, \ldots, 51. \tag{5.2}$$

Notice that w_i for REML is consistently larger than w_i for m.l., which is intuitively sensible since $\hat{\tau}^2_{m\ell}$ has a notoriously large, negative bias. Thus, REML estimation of τ^2 results in less weight on the model term $(X\underset{\sim}{\hat{\beta}})_i$, but in a way so that the effect of estimation of τ^2 can be incorporated.

It is interesting to notice that one pays a price for using REML; its root mean squared prediction errors are consistently larger. This is not surprising, since we know that (asymptotically) m.l. is 100% efficient. Further, notice that the improved root mean squared prediction error, $\surd[M_2(\tau^2)]^*_{ii}$, is between 1% and 9% larger than $\surd[M_1(\hat{\tau}^2)]_{ii}$.

With regard to prediction, one can assess the importance of m.l. versus REML estimation of τ^2 by computing the weighted sum of squares,

$$\sum_{i=1}^{51} \{\hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{m\ell}) - \hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{r\ell})\}^2C_i = 15.$$

When compared to,

$$\sum_{i=1}^{51} (Y_i - 1)^2C_i = 70{,}421$$

and

$$\sum_{i=1}^{51} \{Y_i - \hat{p}_i(\underset{\sim}{Y};\hat{\tau}^2_{m\ell})\}^2 = 26{,}033,$$

Table 1: Columns, from left to right, show the 51 states according to a three-letter identifier, their raw undercounts $\{Y_i\}$, their model fits $\{(X\underset{\sim}{\hat{\beta}})_i\}$, their weights $\{w_i\}$ given by (5.1), their predictors (5.2) (headed F12), their root mean squared prediction errors $\{\surd[M_1(\hat{\tau}^2)]_{ii}\}$ (headed RMPE1), and their improved root mean squared prediction errors $\{\surd[M_2(\tau^2)]^*_{ii}\}$ (headed RMPE2). Table is given over the page.

Table 1

STATE	Y	REML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1431	1.0026	0.00439	0.00453
aka	1.0288	1.0175	0.4767	1.0229	0.00896	0.00976
arz	1.0204	1.0158	0.0742	1.0162	0.00487	0.00500
ark	0.9895	0.9962	0.1398	0.9953	0.00541	0.00562
cal	1.0307	1.0225	0.0682	1.0231	0.00322	0.00327
col	1.0033	1.0199	0.1926	1.0167	0.00473	0.00495
con	0.9886	1.0079	0.1029	1.0059	0.00435	0.00451
del	0.9938	1.0107	0.4571	1.0030	0.00739	0.00811
fla	1.0144	1.0120	0.0785	1.0122	0.00289	0.00295
gga	0.9955	1.0046	0.1639	1.0031	0.00391	0.00403
hai	1.0111	1.0105	0.2785	1.0107	0.00678	0.00730
idh	1.0125	1.0070	0.5627	1.0101	0.00531	0.00579
ill	1.0211	1.0103	0.1170	1.0116	0.00257	0.00265
ind	0.9936	1.0026	0.1413	1.0013	0.00334	0.00349
iow	0.9932	1.0033	0.1478	1.0018	0.00452	0.00475
kan	1.0056	1.0092	0.2215	1.0084	0.00466	0.00496
kty	0.9845	0.9872	0.1519	0.9868	0.00507	0.00524
lou	1.0234	1.0086	0.0263	1.0090	0.00476	0.00480
mne	1.0201	0.9992	0.3703	1.0069	0.00593	0.00645
mld	1.0242	1.0140	0.0712	1.0147	0.00406	0.00415
mas	0.9882	1.0068	0.1945	1.0032	0.00323	0.00341
mch	1.0079	1.0081	0.1601	1.0081	0.00259	0.00271
min	1.0111	1.0049	0.2793	1.0066	0.00359	0.00383
mis	1.0097	1.0086	0.1279	1.0087	0.00557	0.00575
mou	1.0080	1.0010	0.1681	1.0022	0.00350	0.00367
mon	1.0144	1.0059	0.3785	1.0091	0.00699	0.00761
neb	1.0008	1.0071	0.5117	1.0039	0.00441	0.00480
nev	1.0265	1.0151	0.2852	1.0183	0.00744	0.00802
nwh	0.9842	1.0033	0.3080	0.9974	0.00684	0.00740
nwj	1.0130	1.0105	0.0895	1.0107	0.00305	0.00314
nwm	1.0236	1.0256	0.3276	1.0249	0.00611	0.00648
nwy	1.0166	1.0119	0.0807	1.0123	0.00243	0.00247
noc	1.0118	0.9998	0.0748	1.0007	0.00421	0.00430
nod	1.0005	0.9969	0.8931	1.0001	0.00313	0.00324
oho	1.0108	1.0044	0.1273	1.0052	0.00253	0.00263
okl	0.9977	1.0018	0.1625	1.0011	0.00429	0.00451
ore	1.0027	1.0089	0.2833	1.0071	0.00434	0.00464
pen	0.9972	1.0013	0.1475	1.0007	0.00253	0.00263
rhi	1.0089	0.9939	0.4167	1.0001	0.00625	0.00678
soc	1.0632	1.0040	0.0216	1.0053	0.00555	0.00559
sod	1.0008	0.9985	0.7538	1.0002	0.00464	0.00496
ten	0.9717	0.9966	0.0755	0.9947	0.00439	0.00449
tex	1.0037	1.0149	0.0482	1.0144	0.00341	0.00345
uth	1.0040	1.0142	0.4010	1.0101	0.00524	0.00563
vmt	0.9889	1.0018	0.8232	0.9912	0.00454	0.00479
vir	1.0009	1.0058	0.1753	1.0049	0.00338	0.00354
was	1.0142	1.0121	0.1305	1.0123	0.00418	0.00434
wev	0.9942	0.9877	0.1452	0.9887	0.00603	0.00628
wis	1.0173	1.0032	0.2877	1.0073	0.00325	0.00348
wyo	1.0361	1.0127	0.3992	1.0221	0.00882	0.00963
dcl	1.0375	1.0474	0.2191	1.0452	0.01081	0.01125

Table 1 (concluded)

STATE	*Y*	ML				
		MDLFT	WGHT	F12	RMPE1	RMPE2
ala	0.9965	1.0037	0.1190	1.0028	0.00415	0.00427
aka	1.0288	1.0175	0.4241	1.0223	0.00850	0.00933
arz	1.0204	1.0157	0.0608	1.0160	0.00448	0.00459
ark	0.9895	0.9963	0.1161	0.9955	0.00506	0.00525
cal	1.0307	1.0224	0.0559	1.0228	0.00314	0.00319
col	1.0033	1.0198	0.1617	1.0171	0.00446	0.00466
con	0.9886	1.0079	0.0849	1.0063	0.00398	0.00412
del	0.9938	1.0107	0.4050	1.0039	0.00697	0.00771
fla	1.0144	1.0120	0.0644	1.0121	0.00271	0.00276
gga	0.9955	1.0046	0.1368	1.0034	0.00375	0.00385
hai	1.0111	1.0105	0.2378	1.0106	0.00629	0.00679
idh	1.0125	1.0070	0.5099	1.0098	0.00507	0.00559
ill	1.0211	1.0103	0.0967	1.0113	0.00242	0.00248
ind	0.9936	1.0026	0.1174	1.0015	0.00309	0.00323
iow	0.9932	1.0034	0.1230	1.0021	0.00418	0.00438
kan	1.0056	1.0091	0.1870	1.0085	0.00432	0.00460
kty	0.9845	0.9874	0.1264	0.9870	0.00486	0.00502
lou	1.0234	1.0086	0.0214	1.0089	0.00446	0.00449
mne	1.0201	0.9993	0.3222	1.0060	0.00557	0.00608
mld	1.0242	1.0139	0.0583	1.0145	0.00376	0.00384
mas	0.9882	1.0068	0.1634	1.0037	0.00302	0.00319
mch	1.0079	1.0081	0.1335	1.0081	0.00242	0.00252
min	1.0111	1.0049	0.2386	1.0064	0.00339	0.00362
mis	1.0097	1.0085	0.1060	1.0087	0.00526	0.00541
mou	1.0080	1.0011	0.1404	1.0021	0.00326	0.00341
mon	1.0144	1.0059	0.3299	1.0087	0.00656	0.00717
neb	1.0008	1.0071	0.4587	1.0042	0.00420	0.00461
nev	1.0265	1.0150	0.2439	1.0178	0.00692	0.00746
nwh	0.9842	1.0033	0.2646	0.9983	0.00637	0.00691
nwj	1.0130	1.0105	0.0736	1.0106	0.00283	0.00290
nwm	1.0236	1.0254	0.2826	1.0249	0.00582	0.00617
nwy	1.0166	1.0119	0.0663	1.0122	0.00231	0.00235
noc	1.0118	0.9998	0.0614	1.0005	0.00401	0.00408
nod	1.0005	0.9970	0.8710	1.0000	0.00310	0.00324
oho	1.0108	1.0045	0.1055	1.0051	0.00236	0.00245
okl	0.9977	1.0018	0.1356	1.0013	0.00396	0.00416
ore	1.0027	1.0088	0.2421	1.0074	0.00408	0.00436
pen	0.9972	1.0014	0.1227	1.0008	0.00239	0.00248
rhi	1.0089	0.9940	0.3660	0.9995	0.00591	0.00645
soc	1.0632	1.0041	0.0176	1.0051	0.00519	0.00523
sod	1.0008	0.9985	0.7122	1.0002	0.00452	0.00490
ten	0.9717	0.9967	0.0619	0.9951	0.00413	0.00422
tex	1.0037	1.0148	0.0393	1.0144	0.00329	0.00332
uth	1.0040	1.0141	0.3512	1.0105	0.00498	0.00536
vmt	0.9889	1.0019	0.7901	0.9916	0.00445	0.00477
vir	1.0009	1.0058	0.1467	1.0051	0.00317	0.00330
was	1.0142	1.0120	0.1082	1.0123	0.00391	0.00406
wev	0.9942	0.9879	0.1207	0.9886	0.00567	0.00590
wis	1.0173	1.0033	0.2461	1.0067	0.00306	0.00328
wyo	1.0361	1.0127	0.3494	1.0209	0.00829	0.00909
dcl	1.0375	1.0470	0.1849	1.0452	0.01036	0.01078

it is clear that, from a national perspective, prediction is not very sensitive to estimation methods for τ^2. (Cressie (1990) reaches the same conclusion based on a similar comparison of $\hat{\tau}^2_{m\ell}$ and $\hat{\tau}^2_{mm}$.) However, from Table 1, it is equally clear that estimated root mean squared prediction errors are considerably more sensitive.

Cressie (1990) gives expressions for the risks of adjusting using $\hat{p}(\underset{\sim}{Y};\tau^2)$ and of not adjusting. When $\hat{\tau}^2_{r\ell}$ and $\hat{\underset{\sim}{\beta}}(\hat{\tau}^2_{r\ell})$ are substituted into those expressions, the risk of adjusting is 3,253, while the risk, of not adjusting is 34,134. That is, not adjusting leads to a 949% increase in risk (provided the model defined by (1.4), (1.5) and (1.6) holds).

5.2 Simulation

To check the asymptotic distribution theory of the REML (and m.l.) estimator of τ^2, a simulation was carried out on the linear model described in Section 5.1, with parameter values:

$$\beta_0 = 1.0330, \qquad \beta_1 = 0.000712, \qquad \beta_5 = -0.000110, \qquad \tau^2 = 95.00. \tag{5.3}$$

The simulation,

$$\underset{\sim}{Y} \sim \mathrm{Gau}(X\underset{\sim}{\beta},\, \Delta + \tau^2 D), \tag{5.4}$$

where Δ is given by (1.4) the same values of $\delta_1^2, \ldots, \delta_{51}^2$, as used in Section 5.1 and Cressie in 1990, are used here and D is given by (1.6), was performed 500 times, and each time the estimates, $\hat{\tau}^2_{m\ell}$, $\hat{\tau}^2_{mm}$, and $\hat{\tau}^2_{r\ell}$ were computed. (Whenever a negative value was obtained, the estimate was set equal to zero.) The stem-and-leaf plots of the three sets of estimates are presented in Figures 1a, 1b and 1c, respectively. Notice the relatively larger number of zeros for the m.l. estimates (Figure 1a).

Figure 1. Stem-and-leaf plots of estimated variance parameter τ^2, based on 500 simulations of (5.4): (a) maximum likelihood (Section 3.1), (b) method-of-moments (Section 3.2) and (c) restricted maximum likelihood (Section 3.3).

```
 0   000000000000001155556667
 1   0001223566667889
 2   000112356677899
 3   001112234455555779999
 4   00111111122223334444555556666777788888899999
 5   0000122233333334455566666778888899999999
 6   0000011111111222222222333334444556667777777888899999
 7   0001111111122222334444444555556666777778888889999
 8   0011122222223333333333344455566777777788889999
 9   000011122222223333345556777777788
10   00001111111233334444456677777888899
11   00011122222234444456667899
12   000111122223333336677788899
13   1223345556677999
14   0001222334445666799
15   000012223344558999
16   157899
17   001122233589
18   2568
19   145
20   7
21   2
22   88
```

Figure 1a

```
 0   000000377778
 1   011113344446679999
 2   11144455557778888
 3   00002222222233333355555666666888899999
 4   122222224466667777777999999
 5   0002222333333455577777778888
 6   0000000111333334444444446666677779999999999
 7   111122222224444455555555777778888888
 8   00000222333333335555556666666688999
 9   111222222244444466667777779999
10   000000222233333355577777888888
11   00000000111113333444444666677789999
12   1111122222224444444445777778
13   00002233666888888999
14   11122244667777999
15   002233558888
16   000001133444777799
17   122222245555788
18   00003335566899
19   26799
20   02258
21   37
22   11558
23   5
24   79
25
26
27   5
28   5
29   2
30   78
31   3
32
33   2
```

Figure 1b

The means ($\bar{X}$) and standard deviations (S) of the distributions shown in Figure 1 are:

$\hat{\tau}^2_{m\ell}$	$\hat{\tau}^2_{mm}$	$\hat{\tau}^2_{r\ell}$
$\bar{X} = 83.56$	$\bar{X} = 96.85$	$\bar{X} = 94.27.$
$S = 45.65$	$S = 57.46$	$S = 49.17.$

The means should be compared to the true value of $\tau^2 = 95.00$. The bias in $\hat{\tau}^2_{m\ell}$ is apparent; $\hat{\tau}^2_{r\ell}$ has very little bias and has a small advantage over $\hat{\tau}^2_{mm}$. With regard to standard deviations, the advantage of $\hat{\tau}^2_{r\ell}$ over $\hat{\tau}^2_{mm}$ is considerable, but it is at some disadvantage over $\hat{\tau}^2_{m\ell}$. For reasons explained in Section 3.3, that are not all statistical, bias is more of a concern than variance, and so REML estimation of τ^2 should be considered a serious alternative to m.l.

Asymptotic distribution theory for m.l. and REML can be checked from the simulations. (The method of moments is at a disadvantage in that no asymptotic distribution theory is readily available.) Substituting $\tau^2 = 95.00$ into (3.13) yields,

```
 0   00000001234777799
 1   0012334567789
 2   012225556888899
 3   112234444556889
 4   0013344445555666777888888899
 5   000011222333333444445566777788888999
 6   0001112222222334444444444566667777888899999
 7   00000011111112222222233334444555677778888899999
 8   00000001111222333455555566666777778889
 9   00011122222222333333444455555556668999999
10   00000000011122333444555566777888899
11   0001122233344455666677788888899
12   00011111122233344455567789
13   000133334555555556788
14   0001112344445667789
15   00111222344566788
16   0011122223355557999
17   011235556
18   00112566777779
19   117
20   013478
21   123
22   7
23   6
24
25   02
```

Figure 1c

$$\{\text{var}(\hat{\tau}^2_{m\ell})\}^{1/2} \simeq 48.73,$$

which should be compared to $S = 45.65$. Finally, substituting $\tau^2 = 95.00$ into (3.29) yields,

$$\{\text{var}(\hat{\tau}^2_{r\ell})\}^{1/2} \simeq 50.14,$$

which should be compared to $S = 49.17$.

The opportunity also exists to use the simulation to look at "actual" errors of prediction and to assess the performance of $M_1(\hat{\tau}^2)$ and $M_2(\tau^2)^*$. If the parameter values (5.3) were estimated from the original data, then this amounts to a parametric boostrap.

6. CONCLUSIONS AND DISCUSSION

Model-based prediction of undercount relies on careful checking of model fit. Diagnostic plots based on standardized residuals have already been suggested at the end of Section 2. The standardized BLUP residuals $\{Y_i - \hat{p}_i(\underset{\sim}{Y}; \hat{\tau}^2)\}/\{[M(\hat{\tau}^2)]_{ii}\}^{1/2}$; $i = 1, \ldots, n$, also have a role to play. They could either be used in a quantile-quantile plot (*e.g.* Cressie 1991, p. 225) or, as suggested by Calvin and Sedransk (1991), plotted against $\hat{p}_i(\underset{\sim}{Y}; \hat{\tau}^2)$; $i = 1, \ldots, n$.

One could also extend the model (1.4) to include an unknown variance-component parameter σ^2:

$$\underset{\sim}{Y} \sim \text{Gau}(\underset{\sim}{F}, \sigma^2\Delta), \tag{6.1}$$

where $\Delta = \text{diag}\{\delta_1^2, \ldots, \delta_n^2\}$. Upon fitting the more general model (6.1), (1.5) and (1.6), one could then test whether the REML estimate $\sigma_{r\ell}^2$ is significantly different from $\sigma^2 = 1$, which would provide a check on model misspecification. (In this case, REML estimation is recommended over m.l. estimation, since any bias will seriously affect inference on σ^2.)

Restricted maximum likelihood (REML) estimation of variance-matrix parameters is less likely to lead to empirical Bayes predictors that put too much weight on the regression model (1.5). The price paid is slightly larger mean squared prediction errors. Using asymptotic distribution theory for REML (which is checked by simulation), improved estimators of the mean squared prediction errors can also be obtained. Based on the model (1.4), (1.5) and (1.6), it can be concluded that there are accurate and precise ways to make inference on adjustment factors $\{F_i : i = 1, \ldots, n\}$; the predictors $\{\hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2) : i = 1, \ldots, n\}$ yield true-count and undercount predictors,

$$T_i^{\text{prd}} = \hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2)C_i \quad \text{and} \quad U_i^{\text{prd}} = 100\{1 - (\hat{p}_i(\underset{\sim}{Y};\hat{\tau}_{r\ell}^2))^{-1}\}; \quad i = 1, \ldots, n,$$

respectively. Their biases and mean-squared prediction errors can be obtained using the δ-method (*cf.* Cressie 1991, Section 3.2.2).

ACKNOWLEDGEMENTS

The research assistance of Robert Parker is gratefully acknowledged. This research was supported by Joint Statistical Agreement JSA 90-41, between the U. S. Bureau of the Census and Iowa State University. The findings and opinions expressed in this article are of the author alone, and do not necessarily reflect those of the Census Bureau.

REFERENCES

CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-48.

CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. In *Proceedings of Bureau of the Census Fourth Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 127-150.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

CRESSIE, N. (1990). Weighted smoothing of estimated undercount. In *Proceedings of Bureau of the Census 1990 Annual Research Conference*, Washington, DC: U. S. Bureau of the Census, 301-325.

CRESSIE, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

CRESSIE, N., and LAHIRI, S.N. (1991). The asymptotic distribution of REML estimators. *Statistical Laboratory Preprint 91-20*, Iowa State University, Ames, Iowa.

EATON, M.L. (1985). The Gauss-Markov Theorem in multivariate analysis. In *Multivariate Analysis – VI*, (Ed. P.R. Krishnaiah). Amsterdam: Elsevier, 177-201.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of population and housing. *Journal of the American Statistical Association*, 84, 927-944.

FAY, R.E., III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLNER, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-17.

GELFAND, A.E., and SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GROENEVELD, R.A., and MEEDEN, G.D. (1977). The mode, median and mean inequality. *American Statistician*, 31, 120-121.

HARVILLE, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.

HARVILLE, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, 80, 132-138.

KASS, R.E., and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models).*Journal of the American Statistical Association*, 84, 717-726.

KITANIDIS, P.K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19, 909-921.

LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.

MARDIA, K.V., and MARSHALL, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.

MILLER, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics*, 5, 746-762.

PATTERSON, H.D., and THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

PATTERSON, H.D., and THOMPSON, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometric Conference*. Washington, DC: Biometric Society, 197-207.

PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

RAO, C.R. (1979). MINQE theory and its relation to ML and MML estimation of variance components. *Sankhyā B*, 41, 138-153.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

ZIMMERMAN, D.L. (1989). Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21, 655-672.

ZIMMERMAN, D.L., and CRESSIE, N. (1991). Mean-squared prediction error in the spatial linear model. *Annals of the Institute of Statistical Mathematics*, 43, forthcoming.