propriate loss function) across the range of small areas. Such studies depend on "target values" for the parameter of interest for each small area, and generally accepted values of these target values are rarely, if ever, available (if they were, then there would be no need for indirect estimates). Thus, evaluation studies tend to produce conflicting and ambiguous results and leave all concerned less than completely satisfied. A good case in point are the many problems associated with use of a synthetic estimator to adjuste for state population undercounts in the 1990 census.

Having emphasized some of the problems associated with applications of indirect estimators, we should also mention the obvious fact that these estimation methods provide practitioners with many useful tools. Challenging research issues concerning the estimation of meaningful measures of error remain; without such measures, we must be cautious regarding inferences and actions based on these estimators. Nevertheless, in many applications, these methods provide us with an attractive alternative to the use of high variance direct estimates or, in some cases, no estimates at all.

Comment

Avinash C. Singh

The review paper of Ghosh and Rao fills a very important gap by giving a comprehensive and coherent picture of various developments in small area estimation over the last twenty years. This area is fascinating for at least three reasons: (1) there is a great demand for small area statistics by both government and private sectors for purposes of planning and policy analysis; (2) the small area problem provides a fertile ground for theoretical and applied research; and (3) the problem has attracted the attention of both Bayesians and frequentists because both approaches arise naturally and often seem to give similar results.

The main theme of my discussion is to compare and contrast the Bayesian and frequentist solutions to the problem of small area estimation. Why is it that for this problem the two approaches to statistical inference seem to converge in many practical examples including the one considered by Ghosh and Rao; that is, they provide similar results for both point estimates and the corresponding measures of uncertainty? Can we make some general statements about the similarity between the two approaches for small area estimation? How do their frequentist properties compare? Questions about the frequentist properties of some empirical Bayes methods are also raised by Ghosh and Rao in Section 5.2. Although the task of making exact comparisons is a difficult one, it is possible to make asymptotic comparisons for large m—the number of small areas. This will be the focus of my discussion.

Avinash C. Singh is Senior Methodologist, Methods Development and Analysis Section, Social Survey Methods Division, Statistics Canada, Ottawa K1A 0T6. He is also Adjunct Research Professor, Department of Mathematics and Statistics, Carleton University, Ottawa K1S 5B6.

1. MODEL REFORMULATION

As discussed in the review paper of Robinson (1991), understanding of procedures for estimating fixed and random effects helps to bridge the apparent gulf between the Bayesian and frequentist schools of thought. The present discussion will also strengthen this point. First, it will be convenient for our purposes to reformulate the model with fixed and random effects for small area estimation. Now, the general mixed linear model is given by

$$y = X\beta + Z\nu + \epsilon \tag{1}$$

where y is the n-vector of element-level data; X and Z are known matrices of orders $n \times p$ and $n \times m$, respectively, with rank $(X) = p$; β is a p-vector of fixed effects; ν is a m-vector of small area specific random effects and ϵ is a n-vector of random errors independent of ν such that $\nu \sim \text{WS}(0, G), \epsilon \sim \text{WS}(0, R)$. The abbreviation "WS" stands for "wide sense"; that is, the distribution is specified only up to the first two moments. The covariance matrices G and R depend on some parameters λ called variance components. For the reformulation of (1), we will regard the fixed effects β as random with mean 0 and covariance matrix $\sigma_\beta^2 I$ where $\sigma_\beta^2 \to \infty$. Thus, the limiting prior distribution of β is uniform (improper) which is commonly assumed in the Bayesian approach. The reformulation is useful for computational convenience as well as for making connections

between the Bayesian and frequentist approaches. Writing $\alpha = (\beta^T, \nu^T)^T$ and $F = (X, Z)$, we have the reformulated model,

$$y = F\alpha + \varepsilon, \quad \alpha = \alpha^0 + \xi, \tag{2}$$

where $\alpha^0 = 0$, $\xi = (\beta^T, \nu^T)^T \sim \text{WS}(0, \Gamma), \Gamma) = \text{diag}(\sigma_\beta^2 I, G)$ and ξ is independent of ε.

The problem of interest is estimation (or prediction) of $L^T\alpha$ for some $(p+m)$-vector L. In the context of small areas, the vector L can be chosen appropriately to denote the superpopulation mean θ_i of each small area i. Note that if for each small area, population size is large and the sampling fraction is negligible, the estimation of finite population means is essentially equivalent to that of superpopulation means.

An important feature of the above reformulation [equation (2)] is that for known variance components λ, it provides a common model for both frequentist and Bayesian approaches. Not only does it provide a common starting point, both approaches yield identical estimates and the corresponding measures of uncertainty. Since the parameter of interest is inherently random in nature due to finiteness of the small area population, it is very appealing to have a unified formulation which gives identical results. However, for unknown λ, there is some divergence between the two approaches (see Section 3). First, we will consider the case of known λ.

2. CASE OF KNOWN VARIANCE COMPONENTS (λ KNOWN)

In this section, we show that when distributions are specified only in a wide sense, the Gauss-Markov theory (in the frequentist case) and the linear Bayes theory (in the Bayesian case) coincide. Under the frequentist approach for model (1), the objective is to find the best linear unbiased predictor (BLUP) of $\alpha = (\beta^T, \nu^T)^T$; that is, $\hat{\alpha} = a_0 + Ay$ is chosen to minimize

$$E\|a_0 + Ay - \alpha\|^2 \tag{3}$$

over all vectors a_0 and matrix A of appropriate dimensions. Here β is regarded as fixed and the expectation in (3) is with respect to y and ν. On the other hand, under the Bayesian approach, the objective is to find the (unbiased) linear Bayes estimate (LBE) of α as the prior information is specified in a wide sense only. The fixed effect β is assumed to have a uniform, improper prior distribution. Thus, the LBE $\tilde{\alpha} = A\alpha^0 + B(y - F\alpha^0)$ is obtained by minimizing

$$E\|A\alpha^0 + B(y - F\alpha^0) - \alpha\|^2 \tag{4}$$

over all matrices A and B of appropriate dimensions. Note that the chosen form of the linear estimator $\tilde{\alpha}$ is intuitive and is equivalent to the general form of a linear estimator under the condition of unbiasedness. Also note that the expectation in (4) is with respect to y, ν and also β. Now, the BLUP $\hat{\alpha}$ and its MSE coincide with the LBE $\tilde{\alpha}$ and its Bayes risk, respectively. This follows from the results of Sallas and Harville (1981) and Zehnwirth (1988). Sallas and Harville establish that the BLUP $\hat{\alpha}$ and its MSE can be obtained respectively as limits of BLUPs and MSEs of α defined by the reformulated model (2) as $\sigma_\beta^2 \to \infty$. Zehnwirth (in the context of Kalman filtering) shows that the BLUP of α under model (2) is indeed the LBE and that MSE of BLUP equals the Bayes risk of LBE. Therefore, the LBE $\tilde{\alpha}$ which is the limit of LBEs as $\sigma_\beta^2 \to \infty$ coincides with the BLUP $\hat{\alpha}$ and the same is true of their measures of uncertainty. The corresponding expressions can be obtained as

$$\hat{\alpha} = \tilde{\alpha} = \lim_{\sigma_\beta^2 \to \infty} [\alpha^0 + \Gamma)F^T(F\Gamma)F^T + R)^{-1}(y - F\alpha^0)] \tag{5}$$

and

$$\begin{aligned} \text{MSE}(\hat{\alpha}) &= \text{Bayes Risk } (\tilde{\alpha}) \\ &= \lim_{\sigma_\beta^2 \to \infty} [I - \Gamma F^T(F\Gamma)F^T + R)^{-1}F]\Gamma. \end{aligned} \tag{6}$$

See Sallas and Harville (1981) for closed form expressions of the above limits. An expedient way to get the expressions in (5) and (6) is to think of them respectively as the posterior mean and variance of α under normality. Notice that under normality, the posterior mean is linear and the posterior variance does not depend on y. Therefore, the usual Bayes theory under normality also coincides with the linear Bayes theory when the prior distribution is specified in a wide sense only.

3. CASE OF UNKNOWN VARIANCE COMPONENTS (λ UNKNOWN)

When λ is unknown, it turns out that there is some divergence between the two approaches. It is possible to get some understanding of the differences under normality. Therefore, we assume that the errors ν and ε are normal. Also, the number of small areas, m, will be assumed to be large for making asymptotic comparisons. For simplicity, we will illustrate results for the one-fold nested error regression model given by equation (4.2) of Ghosh and Rao, except that we will set $k_{ij} = 1$. Here $\lambda = (\lambda_1, \lambda_2)^T = (\sigma_\nu^2, \sigma^2)^T$ and suppose for illustration that only λ_1 is unknown. The parameters of interest are small area means $\theta_i, i = 1, \ldots m$

where $\theta_i = \overline{X}_i^T\beta + \nu_i$. If λ is known and γ_i denotes $\lambda_1(\lambda_1 + \lambda_2 n_i^{-1})^{-1}$, then the BLUP $\hat{\theta}_i$ and LBE $\tilde{\theta}_i$ (or BUP and BE respectively under normality) are obtained from (5) as

$$\hat{\theta}_i = \tilde{\theta}_i = \overline{X}_i^T\hat{\beta} + \gamma_i(\bar{y}_i - \bar{x}_i^T\hat{\beta}) \tag{7}$$

and from (6); we have, after noting that under normality the Bayes risk is same as the posterior variance (PV),

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_i) &= \mathrm{PV}(\theta_i) \\ &= \lambda_1\lambda_2 n_i^{-1}(\lambda_1 + \lambda_2 n_i^{-1})^{-1} \\ &\quad + (\overline{X}_i - \gamma_i\bar{x}_i)^T(X^TV^{-1}X)^{-1}(\overline{X} - \gamma_i\bar{x}_i) \\ &= g_1(\lambda_1) + g_2(\lambda_1), \text{ say.} \end{aligned} \tag{8}$$

Now, an EBLUP is defined by substituting a consistent estimator $\hat{\lambda}_1$ for λ_1 in $\hat{\theta}_i$ (denote by $\hat{\theta}_i(y, \hat{\lambda}_1)$) and an EB estimator is defined by substituting $\hat{\lambda}_1$ in $\tilde{\theta}_i$, to be denoted by $\tilde{\theta}_i(y, \hat{\lambda}_1)$. For facilitating comparison of the two approaches, we will assume that $\hat{\lambda}_1$ is REML. Clearly, the two estimators so defined are identical. The "naive" approximations to the corresponding measures of uncertainty obtained from (8) by substituting $\hat{\lambda}_1$ for λ_1 are also, of course, identical. The qualifying term "naive" is used to indicate that the extra variability due to estimation of λ_1 is not accounted for.

In the expression (8), the terms $g_1(\lambda_1)$ and $g_2(\lambda_1)$ are respectively $O(1)$ and $O(m^{-1})$. For finding the order of the extra term due to estimation of λ_1, first consider the frequentist approach. It can be shown by the δ-method, similar to equation (5.3) of Ghosh and Rao, that

$$\mathrm{mse}(\hat{\theta}_i(y, \hat{\lambda}_1)) = g_1(\lambda_1) + g_2(\lambda_1) + g_3(\lambda_1) + o(m^{-1}), \tag{9}$$

where $g_3(\lambda_1) = n_i^{-2}\lambda_2^2(\lambda_1 + \lambda_2 n_i^{-1})^{-3}\overline{V}(\hat{\lambda}_1)$ and $\overline{V}(\hat{\lambda}_1)$ is the asymptotic variance of $\hat{\lambda}_1$. Notice that the term $g_3(\lambda_1)$ is also $O(m^{-1})$. Substituting $\hat{\lambda}_1$ in (9), we get an estimate of MSE; but the order of bias is $O(m^{-1})$, not $o(m^{-1})$. This is so because the bias in $g_1(\hat{\lambda}_1)$ is $O(m^{-1})$, although biases in $g_2(\hat{\lambda}_1)$ and $g_3(\hat{\lambda}_1)$ are $o(m^{-1})$. To correct this, the approximation of Prasad and Rao (1990, PR for short) can be used under the assumption $E(\hat{\lambda}_1 - \lambda_1) = o(m^{-1})$ as

$$\begin{aligned} \mathrm{mse}(\hat{\theta}_i(y, \hat{\lambda}_1)) &= [g_1(\hat{\lambda}_1) + g_3(\hat{\lambda}_1)] + g_2(\hat{\lambda}_1) + g_3(\hat{\lambda}_1) \\ &= g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + 2g_3(\hat{\lambda}_1). \end{aligned} \tag{10}$$

For the Bayesian approach, corrections for underestimation of $PV(\theta_i)$ due to estimation of $\hat{\lambda}_1$ can be made by using results of the asymptotic (as $m \to \infty$) hierarchical Bayes (HB) theory (cf. Kass and Steffey, 1989). This technique is justified because the HB estimator (i.e., the posterior mean of θ_i) is asymptotically equivalent to the EB estimator $\tilde{\theta}_i(y, \hat{\lambda}_1)$, the order of error in the approximation being $O(m^{-1})$. Also, the HB technique is convenient in practice because, for large m, the posterior distribution of λ_1 is independent of the choice of prior. Now, analogous to (5.11) of Ghosh and Rao [note that β is absent in the expectation operator because variability due to $\hat{\beta}$ is already accounted for in the $PV(\theta_i)$], we have the posterior variance

$$V(\theta_i|y) = E_{\lambda_1|y}V(\theta_i|y, \lambda_1) + V_{\lambda_1|y}E(\theta_i|y, \lambda_1) \tag{11a}$$

$$= E_{\lambda_1|y}(g_1(\lambda_1) + g_2(\lambda_1)) + V_{\lambda_1|y}\tilde{\theta}_i(y, \lambda_1). \tag{11b}$$

It follows from Kass and Steffey (1989) that

$$E_{\lambda_1|y}(g_1(\lambda_1) + g_2(\lambda)) = g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + O(m^{-1}), \tag{12a}$$

$$\begin{aligned} V_{\lambda_1|y}\tilde{\theta}_i(y, \lambda_1) &= d^2(\hat{\lambda}_1)\overline{V}(\hat{\lambda}_1) + o(m^{-1}) \\ &= g_3^*(\hat{\lambda}_1) + o(m^{-1}), \text{ say,} \end{aligned} \tag{12b}$$

where $d(\hat{\lambda}_1)$ is $(\partial/\partial\lambda_1)\tilde{\theta}_i(y, \lambda_1)|_{\lambda_1=\hat{\lambda}_1}$. Note that if β were known, then $g_3^*(\hat{\lambda}_1)$ simplifies to $\lambda_2^2 n_i^{-2}(\hat{\lambda}_1 + \lambda_2 n_i^{-1})^{-4}(\bar{y}_i - \bar{x}_i^T\beta)^2\overline{V}(\hat{\lambda}_1)$ which is more directly comparable to the term $g_3(\hat{\lambda}_1)$ of the frequentist approximation (9). Incidentally, for β known, $\hat{\lambda}_1$ will be the usual ML and not REML.

In the approximation (12a), the neglected term is $O(m^{-1})$. The accuracy of this approximation can be improved by including terms of order $O(m^{-1})$. By using the δ-method, Singh, Stukel and Pfeffermann (1993) obtain an improved Bayesian approximation as

$$\begin{aligned} V(\theta_i|y) &= [g_1(\hat{\lambda}_1) + g_2(\hat{\lambda}_1) + sg_3^{**}(\hat{\lambda}_1)] \\ &\quad + g_3^*(\hat{\lambda}_1) + o(m^{-1}), \end{aligned} \tag{13}$$

where $g_3^{**}(\hat{\lambda}_1)$ is $\lambda_2^2 n_i^{-2}(\hat{\lambda}_1 + \lambda_2 n_i^{-1})^{-2}(\hat{\hat{\lambda}}_1 - \hat{\lambda}_1) - g_3(\hat{\lambda}_1)$. The estimator $\hat{\hat{\lambda}}_1$ denotes an improved (over $\hat{\lambda}_1$) approximation to the posterior mean $E(\lambda_1|y)$ in the sense that $E(\lambda_1|y) = \hat{\hat{\lambda}}_1 + o(m^{-1})$ whereas $E(\lambda_1|y) = \hat{\lambda}_1 + O(m^{-1})$. Note that $\hat{\hat{\lambda}}_1$, can be obtained from the results of Tierney, Kass and Kadane (1989). The expression in (13) is a simplified version of the second order approximation of Kass and Steffey; denote it by KS-II*. Their first order (denote by KS-I) does not include the term $g_3^{**}(\hat{\lambda}_1)$. The approximation KS-II* seems more convenient for term by term comparison with the PR approximation (10) than the original second order KS-II (not considered here).

From (7), (8), (10) and (13), one can compare the Bayesian and frequentist approaches for large m

when λ is unknown. The point estimates are identical or very similar depending on the choice of $\hat{\lambda}_1$ for each approach but the associated measures of uncertainty could be quite different. In addition to the above modifications which rely on the δ-method, Singh, Stukel and Pfeffermann (1993) also obtain a modification of the asymptotic Bayes method of Hamilton (1986) which uses Monte Carlo integration (MCI) for evaluating the two terms of the posterior variance given by (11) thus avoiding computation of partial derivatives. The MCI simply entails generating λ_1-values from the approximate posterior distribution of λ_1 which is given by $N(\hat{\lambda}_1, \overline{V}(\hat{\lambda}_1))$. It is not difficult to show that the order of the neglected terms in the Hamilton (H) approximation is $O(m^{-1})$ and not $o(m^{-1})$. However, if the posterior distribution of λ_1 is approximated by $N(\hat{\hat{\lambda}}_1, \overline{V}(\hat{\lambda}_1))$, then the modified Hamilton (MH) approximation is of the desired order. Singh, Stukel and Pfeffermann (1993) report results of a Monte Carlo study on the frequentist properties of various approximations. Empirically, it is found that the KS-I approximation is biased downward, but KS-II* adds a positive term (similar to PR) and tends to be conservative. The behaviour of the MH approximation is quite similar to KS-II*, but H tends to be more biased downward than KS-I. The performance of the PR approximation is found to be best overall with respect to the frequentist properties, although other approximations provide useful alternatives. In particular, Bayesian approximations KS-II* and MH have the distinct advantage of having a dual interpretation in both frequentist and Bayesian contexts.

It may be noted that if m is quite large, then there will be hardly any difference between various approximations.

4. REMARKS

It is evident from the paper of Ghosh and Rao that great advances have been made in the field of small area estimation by both Bayesians and frequentists. It is also evident from the present discussion that there may be quite a bit of agreement between the two approaches. However, these advanced tools are not in widespread use, especially by statistical agencies conducting large scale complex surveys who face probably the greatest demand for small area statistics. Perhaps, the reason for this is the practitioner's skepticism in modelling complex survey data. Indeed, for complex surveys there is very little by way of model validation and more so for element-level modelling because of possible selection bias [see section 4 of Ghosh and Rao and a recent review by Pfeffermann (1993)]. There is no doubt that the area of model validation for complex survey data needs more research. This is also recognized by Ghosh and Rao and I would like to emphasize by noting that further work in this direction will be a very valuable contribution.

ACKNOWLEDGMENT

This research was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

Comment

Elizabeth A. Stasny

Ghosh and Rao are to be congratulated for their timely paper reviewing methods for small-area estimation. My main complaint is that a paper such as this was not available five years ago when I began working on small-area estimation problems. I particularly enjoyed the historical perspective offered in the demographics methods section of the paper; I was sorry that section was so short since much of the material described in that section is not readily available to statisticians outside of the government agencies.

As the authors noted, there is a growing demand for small-area estimates and a corresponding interest in research on procedures for producing such estimates. The widely publicized debate on adjusting the U.S. population census for the undercount to produce adjusted counts for states and large cities has made many researchers focus on small area estimation problems related to the population census. There are, however, other long-standing small-area estimation programs. One of these is the USDA's program of county-level estimation of crop and live-

Elizabeth A. Stasny is Associate Professor, Department of Statistics, 1958 Neil Avenue, 148D Cockins Hall, Ohio State University, Columbus, Ohio 43210.