for Small Areas (research monograph 24). U.S. Government Printing Office, Washington, DC.

NATIONAL RESEARCH COUNCIL (1980). *Panel on Small-Area Estimates of Population and Income. Estimating Population and Income of Small Areas*. National Academy Press, Washington, DC.

NICHOL, S. (1977). A regression approach to small area estimation. Unpublished manuscript, Australian Bureau of Statistics, Canberra, Australia.

NTIS (1963). Indirect estimators in federal programs. Statistical policy working paper 21, prepared by the Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC.

PFEFFERMANN, D. and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16** 217–237.

PFEFFERMANN, D. and BARNARD, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics* **9** 73–84.

PLATEK, R. and SINGH, M. P. (1986). *Small Area Statistics: Contributed Papers*. Laboratory for Research in Statistics and Probability, Carleton Univ.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E. and SINGH, M. P. (1987). *Small Area Statistics*. Wiley, New York.

PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of mean squared errors of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171.

PURCELL, N. J. and LINACRE, S. (1976). Techniques for the estimation of small area characteristics. Unpublished manuscript.

PURCELL, N. J. and KISH, L. (1979). Estimation for small domain. *Biometrics* **35** 365–384.

PURCELL, N. J. and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Internat. Statist. Rev.* **48** 3–18.

RAO, C. R. and SHINOZAKI, N. (1978). Precision of individual estimates in simultaneous estimation of parameters. *Biometrika* **65** 23–30.

RAO, J. N. K. (1986). Synthetic estimators, SPREE and best model based predictors. In *Proceedings of the Conference on Survey Research Methods in Agriculture* 1–16. U.S. Dept. Agriculture, Washington, DC.

RAO, J. N. K. and YU, M. (1992). Small area estimation by combining time series and cross-sectional data. In *Proceedings of the Survey Research Methods Section* 1–19. Amer. Statist. Assoc., Alexandria, VA.

ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statist. Sci.* **6** 15–51.

SÄRNDAL, C. E. and HIDIROGLOU, M. A. (1989). Small domain estimation: a conditional analysis. *J. Amer. Statist. Assoc.* **84** 266–275.

SCHAIBLE, W. L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the Survey Research Methods Section* 741–746. Amer. Statist. Assoc., Washington, DC.

SCHAIBLE, W. L. (1992). Use of small area statistics in U.S. Federal Programs. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) **1** 95–114. Central Statistical Office, Warsaw.

SINGH, A. C. and MANTEL, H. J. (1991). State space composite estimation for small areas. In *Symposium 91—Spatial Issues in Statistics—Proceedings* 17–25. Statistics Canada, Ottawa.

SINGH, M. P., GAMBINO, J. and MANTEL, H. (1992). Issues and options in the provision of small area data. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) **1** 37–75. Central Statistical Office, Warsaw.

SMITH, S. K. and LEWIS, B. B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography* **17** 323–340.

SPJØTVOLL, E. and THOMSEN, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute* **2** 435–449.

STARSINIC, D. E. (1974). Development of population estimates for revenue sharing areas. Census Tract Papers, Ser. GE40, No. 10 U.S. Government Printing Office, Washington, DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*, Catalogue 91-528E. Statistics Canada, Ottawa.

STEFFEY, D. and KASS, R. E. (1991). Comment on "That BLUP is a good thing: The estimation of random effects," by G. K. Robinson. *Statist. Sci.* **6** 45–47.

STUKEL, D. (1991). *Small Area Estimation Under One and Two-Fold Nested Error Regression Model*. Ph.D. Thesis, Carleton Univ.

U.S. BUREAU OF THE CENSUS (1966). Methods of population estimation: Part I, Illustrative procedure of the Bureau's component method II. *Current Population Reports, Series P-25, No. 339*. U.S. Government Printing Office, Washington, DC.

ZIDEK, J. V. (1982). A review of methods for estimating the populations of local areas. Technical Report 82–4, Univ. British Columbia, Vancouver.

Comment

Noel Cressie and Mark S. Kaiser

Malay Ghosh and Jon Rao have presented us with a well written exposition of the topic of small area estimation. The past literature has been decidedly influenced by linear modeling, and we see that clearly in their paper. There has also been a tendency to judge the performance of the estimation methods by concentrating on a single, arbitrary small area. In our comment, we shall discuss what opportunities there might be to expand the class of statistical models for small area data and to consider multivariate aspects of small area estimation.

Noel Cressie is Professor of Statistics and Distinguished Professor in Liberal Arts and Sciences and Mark S. Kaiser is Assistant Professor, Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa 50011-1201.

MODELING APPROPRIATE SOURCES OF VARIABILITY

It would appear from the authors' account that the full flexibility of hierarchical modeling has not been applied to small area estimation. Two models incorporating random effects, given as equations (4.4) and (4.5), are presented in their paper. Model (4.4) is applied when both direct estimators and auxiliary data are available at the area level while model (4.5) is partitioned into sampled and unsampled units within a small area when both response and auxiliary data are available for sampling units. In either case, estimates of the area means or totals are developed. Rather than focusing on this distinction, we would like to point out the similarity of these models in the way that additional response variability is due to the random nature of model components. Using y as a vector of response data, both of the models may be considered hierarchical models of the general form,

$$(1)\ [y \mid \mu, \Sigma] = N(\mu, \Sigma) \quad \text{and} \quad [\mu \mid \beta, \Gamma] = N(X\beta, \Gamma),$$

where $[y \mid \theta]$ denotes the probability distribution of y given the parameter θ; and both Σ and Γ are positive-definite matrices. Then the marginal distribution of y is immediately (e.g., Lindley and Smith, 1972, Lemma 1)

$$(2) \qquad N(X\beta, \Sigma + \Gamma),$$

which also results from writing the models in mixed linear form (as Ghosh and Rao have chosen to do). The covariance matrix of the marginal density indicates that these types of models incorporate *sampling variability* into the distribution of y through the use of hierarchical structure. (In engineering, this approach is called state-space modeling.) What might be considered the *systematic* model component, namely $E(y) = X\beta$, is generated in the same way across all areas in model (4.4) or across all sampling units within all areas in model (4.5).

The hierarchical model described by (1) is different from the model,

$$(3)\ [y \mid \beta, \Sigma] = N(X\beta, \Sigma) \quad \text{and} \quad [\beta \mid B, \Gamma] = N(B, \Gamma),$$

for which the marginal density of y becomes

$$(4) \qquad N(XB, \Sigma + X\Gamma X^T).$$

Under the hierarchical model (3), variability in the marginal distribution of y is affected by the values of the explanatory variables observed.

A third possibility suggests itself in the situation that unit specific observations are available in each of several small areas. In this case, one might apply model (3) to each area using y_i, β_i, X_i and Σ_i to denote the dependence on area identification. The $\{\beta_i\}$ could be taken as independent and identically distributed random variables with common distribution across areas; for example, $[\beta_i \mid B, \Gamma] = N(B, \Gamma)$. Then, with assumed independence (conditional on $\{\beta_i, \Sigma_i\}$) of the $\{y_i\}$, the joint marginal of all observations is available as a product of the marginals for the m areas. More complex models allowing lack of independence for either the y_i or β_i are conceivable; and, in fact, model (3) is an example of one such. Under models of this general type, variability among observations comes not only from direct sampling variability but also from variability in the $\{\beta_i\}$ that describe the systematic relation between y and X. That such variability often exists seems a reasonable supposition. In the introductory example discussed by Ghosh and Rao, of estimating per capita income (PCI) for local administrative areas (Fay and Herriot, 1979), a regression of estimated PCI on county tax returns and housing data is assumed. The systematic relation described by such a regression may well be different for counties in different portions of a state or region, as may be the range and values of the explanatory variables used. As another example, the small areas where census undercounts are estimated can each be stratified by race. A separate regression for each race (Cressie, 1989) results in differences in regression coefficients. Finally, estimation of the distribution of regression coefficients may provide valuable information to demographers and social scientists, such as in the problem of census undercount.

There is a difference in the modeling approach represented by (1) on the one hand, and (3) and its extensions on the other, that centers on the sources of variation in the observed responses. From a Bayesian viewpoint, this difference involves the order in which prior distributions are placed on model components. The order in which priors are assigned is pertinent, particularly in light of the fact that the data contain less information about parameters as those parameters move up in the hierarchy (Goel and DeGroot, 1981). Thus, if we have interest in the posterior distribution of β, we are well served by positioning β low in the hierarchy which leads to model (3) and its extensions rather than to model (1). Under model (1), we do not question the strength of the linear relation between y and X but are uncertain about the realization that may be observed in any particular small area. Then, a completely specified, but often uninformative, prior is placed on β as much to allow computation of a posterior distribution of μ as from genuine interest in modeling either prior or posterior distributions of β. Under model (3) and its extensions, an important source of uncertainty stems from lack of knowledge about β or $\{\beta_i\}$. Ghosh and Rao have not discussed the latter approach in small area estimation and it would be interesting and useful to see what differences might result from its application. The extension of model (3) to area-specific regression equations, in particular, offers an interesting alternative to the standard approach in that it raises the possibility of predicting the area-specific regression parameters $\{\beta_i\}$.

NONLINEAR MODELS

In an effort to increase the flexibility of small area models, it is natural to consider ways to extend the modeling concepts to nonlinear situations. One approach to nonlinear modeling that encompasses many situations is that of generalized linear models (GLMs). Ghosh and Rao mention binary and Poisson responses in Section 7 of their paper which fall into this framework. Our earlier discussion of appropriate sources of variability carries over to the GLM, and we give several models analogous to the normal models already presented. While the notation of GLMs offers flexibility in allowing nonlinear response functions, there is a concomitant reduction in flexibility for modeling lack of independence among responses. Specifically, small area responses y_i are taken to be univariate random variables, and conditional independence of these variables (conditional on parameters) is assumed throughout. Assume that y_i is distributed according to an exponential family with density (or mass) function,

$$f(y_i \mid \theta_i, \phi_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi_i) + c(y_i, \phi_i)\},$$

so that $E(y_i) = b'(\theta_i) \equiv \mu_i$ and $\text{var}(y_i) = a(\phi_i)b''(\theta_i) \equiv a(\phi_i)V(\mu_i)$. A GLM is completed by taking a known function of μ_i to be linear in a set of covariates; that is, $g(\mu_i) = x_i^T\beta \equiv \eta_i$ with $x_i = (x_{i1}, \ldots, x_{ip})^T$. One hierarchical extension of this model is to let the natural parameter θ_i be distributed according to some probability density (or mass) function $h(\theta_i \mid \lambda)$. The marginal density (or mass) function of y_i then becomes

$$(5) \qquad p(y_i \mid \lambda, \phi_i) = \int f(y_i \mid \theta_i, \phi_i)h(\theta_i \mid \lambda)d\theta_i.$$

This is the approach taken by Albert (1988) and Albert and Pepple (1989) to develop hierarchical overdispersion models. These authors take $h(\theta_i \mid \lambda)$ from a conjugate exponential family for f and then set up the GLM by linking the expected value of μ_i with a linear model as $g(E(\mu_i)) = x_i^T\beta$. This approach moves the linear model away from y to a position further up in the hierarchy and is analogous to the approach of model (1) leading exactly to that model in the case that y is normal with mean $\theta = \mu$, and g the identity mapping.

A different approach, analogous to that used in model (3), is to start with a fully specified GLM for the responses and allow β to be random. In this case, we must assign a multivariate distribution for β. For example, we might take $[\beta \mid B, \Gamma] = N(B, \Gamma)$. The marginal distribution of y_i is then

$$(6) \quad p(y_i \mid B, \Gamma, \phi_i) = \int f(y_i \mid \beta, \phi_i)h(\beta \mid B, \Gamma)d\beta.$$

In (6), we use the exponential form for y_i and the systematic specification $g(\mu_i) = x_i^T\beta$, so that $\theta_i = b'^{-1}[g^{-1}(x_i^T\beta)]$ and

$$(7) \quad \begin{aligned} &f(y_i \mid \beta, \phi_i) \\ &\quad = \exp\{[y_i b'^{-1}(g^{-1}(x_i^T\beta)) \\ &\qquad - b[b'^{-1}(g^{-1}(x_i^T\beta))]]/a(\phi_i) + c(y_i, \phi_i)\}. \end{aligned}$$

Things simplify substantially by taking g as the canonical link function $g(\cdot) = b'^{-1}(\cdot)$, giving

$$(8) \quad \begin{aligned} &f(y_i \mid \beta, \phi_i) \\ &\quad = \exp\{[y_i x_i^T\beta - b(x_i^T\beta)]/a(\phi_i) + c(y_i, \phi_i)\}. \end{aligned}$$

Using expression (8), it would be possible, at least in theory, to complete the integrations in (6). In practice, the necessary integrations might be best approached through importance sampling or, in a Bayesian analysis, the joint posterior distribution might be calculated directly through Monte Carlo resampling schemes (e.g., Smith and Roberts, 1993).

As for the analogous normal model (3), the ideas culminating in equation (8) may be extended directly to the situation of different regressions among areas. Unlike the normal situation, however, it is difficult to conceptualize the way lack of independence among responses in different small areas could be handled.

In all of these hierarchical models, ways to deal with dispersion parameters and covariance matrices become a major statistical issue. It is always possible, in theory, to find maximum likelihood estimators. Can small sample properties of maximum likelihood estimators of dispersion and covariance parameters be improved, perhaps using some appropriate analogues to REML estimation? Bayesian modeling of these (nuisance) parameters is another possibility that is becoming feasible with the recent developments in Gibbs sampling and Monte Carlo resampling schemes.

MULTIVARIATE ASPECTS

Although the problem of small area estimation is inherently multivariate, there has been a tendency to look at the performance of estimation procedures area-by-area. For example, Ghosh and Rao give the mean-squared error formula (5.5) for the ith small area. What is actually needed is the $m \times m$ mean-squared error matrix

$$(9) \qquad (\mathrm{mse}(i,j)) = E\{(\hat{\theta}^H - \theta)(\hat{\theta}^H - \theta)^T\},$$

whose diagonal elements are given by (5.5) but whose off-diagonal elements also have an important role to play.

Suppose that two small areas i and i' are combined into a new area that we denote $i \cup i'$. Now, assuming a linear model, $\theta_{i \cup i'} = w\theta_i + w'\theta_{i'}$ and $\hat{\theta}^H_{i \cup i'} = w\hat{\theta}^H_i + w'\hat{\theta}^H_{i'}$. Hence,

$$\begin{aligned} \mathrm{mse}(i \cup i', i \cup i') = w^2\, \mathrm{mse}(i,i) + (w')^2\, \mathrm{mse}(i',i') \\ +2ww'\, \mathrm{mse}(i,i'), \end{aligned}$$

which involves both diagonal and off-diagonal elements of (9). Cressie (1992) develops an approximation to (9), analogous to the univariate approximation (5.5).

As another example, a multivariate version of the Laird and Louis (1987) bootstrap, described by the authors in Section 5.2, is straightforward to derive. Let $\theta \equiv (\theta_1, \ldots, \theta_m)$ denote the parameters of the m small areas. A large number, B, of independent bootstrap samples $\{\theta^*(b) : b = 1, \ldots, B\}$ are drawn from the estimated marginal distribution $N(X\hat{\beta}, \hat{\Sigma} + \hat{\Gamma})$; see relation (2). Estimates $\beta^*(b)$, $\Sigma^*(b)$ and $\Gamma^*(b)$ are computed from the bootstrap data $\theta^*(b)$ for each b. Then the EB bootstrap estimator and the appropriate estimated posterior variance matrix are, respectively,

$$\begin{aligned} \theta^{*EB}(\cdot) &= (1/B) \sum_{b=1}^{B} E(\theta \mid \theta^*(b), \beta^*(b), \Sigma^*(b), \Gamma^*(b)) \\ &= (1/B) \sum_{b=1}^{B} \theta^{*EB}(b), \\ V^* &= (1/B) \sum_{b=1}^{B} \mathrm{var}(\theta \mid \theta^*(b), \beta^*(b), \Sigma^*(b), \Gamma^*(b)) \\ &\quad + (1/(B-1)) \sum_{b=1}^{B} (\theta^{*EB}(b) - \theta^{*EB}(\cdot)) \\ &\qquad \cdot (\theta^{*EB}(b) - \theta^{*EB}(\cdot))^T. \end{aligned}$$

Given the geographic nature of most small area estimation problems, the question of how to aggregate is always waiting to be asked; hence, the multivariate aspects are important. The harder question of how to disagregate has been at the core of much of the debate about the adjustment of census counts. Cressie (1988) shows that adjustment based on small area estimation of both the synthetic and empirical Bayes type offers smaller risk than no adjustment even under disaggregation of the small areas. Crucial to his argument is the appropriateness of the small area model at the disaggregated level. Tukey (1983) and Wolter and Causey (1991) reach similar conclusions to Cressie; however, both articles make an assumption that when disaggregating synthetically the *true* adjustment factor is *known* at the level below which disaggregation occurs. There is no certainty that adjustment will improve counts at *all* disaggregated levels; Freedman and Navidi (1992) give a simple example to demonstrate that some adjusted counts can be worse than unadjusted counts.

CONSTRAINED ESTIMATION

In a sense, constrained estimation takes a multivariate point of view in that interest is focussed on how well the *ensemble* of the m small area estimators matches the ensemble of the m estimands. However, there is an opportunity to make the problem more explicitly multivariate.

First, we would like to fill in some of the history of constrained estimation. Tukey (1974, page 143) was aware that the ensemble of estimates gives poor information about the ensemble properties of parameters (e.g., one such property might be the population-weighted proportion of small areas whose lip-cancer rate is above .05 per thousand population years at risk). Louis (1984) addressed the problem in a normal homoscedastic model by advocating that optimal (i.e., Bayes) shrinkage estimates be modified so that their ensemble variance matches the posterior expectation of the parameters' ensemble variance. Cressie (1986, 1989) coined the term "constrained Bayes estimation" and generalized Louis' result to heteroscedastic normal models (for census undercount).

Spjøtvoll and Thomsen (1987) completely ignored the multivariate aspects of the problem by considering each area one-at-a-time. Let θ_i and $\tilde{\theta}_i$ denote the parameter and an estimator, respectively, for the ith area. Assume that both parameter and estimator are random, with first two moments finite, and that $E(\tilde{\theta}_i \mid \theta_i) = \theta_i$. They propose to estimate θ_i by

$$(10) \qquad \hat{\theta}_i = a_i\tilde{\theta}_i + b_i,$$

where a_i and b_i are solved by specifying that $E(\hat{\theta}_i) = E(\theta_i) \equiv \nu$ and $\mathrm{var}(\hat{\theta}_i) = \mathrm{var}(\theta_i) \equiv \sigma^2$. In the discussion to Spjøtvoll and Thomsen's paper, it is pointed out that the solution yields the constrained empirical Bayes estimates obtained by Cressie (1986), although no Bayes optimality criterion is invoked by the authors.

The multivariate version of (10) is

$$\hat{\theta} = A\tilde{\theta} + b, \tag{11}$$

where A is an $m \times m$ matrix and b is an $m \times 1$ vector. Upon specifying that $E(\hat{\theta}) = E(\theta)$ and $\mathrm{var}(\hat{\theta}) = \mathrm{var}(\theta)$, Cressie (1990b, 1992) obtains a multivariate constrained estimator. In the notation of (1), $\theta = \mu$, $E(\theta) = X\beta$, $\tilde{\theta} = y$, $E(y \mid \theta) = \theta$, $\mathrm{var}(y \mid \theta) = \Sigma$, and $\mathrm{var}(y) = \Sigma + \Gamma$. Then the multivariate constrained estimator for model (1), analogous to Spjøtvoll and Thomsen's, is given by (11), where

$$A = \Gamma^{1/2}(\Sigma + \Gamma)^{-1/2} \tag{12}$$

and

$$b = \{I - \Gamma^{1/2}(\Sigma + \Gamma)^{-1/2}\}X\beta. \tag{13}$$

Notice that $\hat{\theta}$ given by (11), (12) and (13) does not shrink y towards $X\beta$ as far as the Bayes estimator θ^* does (where $A = \Gamma(\Sigma + \Gamma)^{-1}$ and $b = (I - A)X\beta$).

In an elegant paper, Ghosh (1992) derives a multivariate constrained *Bayes* estimator for model (1):

$$\theta^{@} = \{a + (1 - a)\underset{\sim}{1}\underset{\sim}{1}'/m\}\theta^*, \tag{14}$$

where

$$a = \left[\mathrm{trace}\{(I - \underset{\sim}{1}\underset{\sim}{1}'/m)V\}\left(\sum_{i=1}^{m}(\theta_i^* - \bar{\theta}^*)^2\right)^{-1} + 1\right]^{1/2},$$

$$\theta^* = E(\theta \mid y) = \{\Gamma(\Sigma + \Gamma)^{-1}\}y + (I - \Gamma(\Sigma + \Gamma)^{-1}\}X\beta,$$

$$V = \mathrm{var}(\theta \mid y) = \Gamma\{I - \Gamma(\Sigma + \Gamma)^{-1}\}\Gamma.$$

The vector $\theta^{@}$ has the property that it minimizes $E(\Sigma_{i=1}^{m}(\theta_i - t_i)^2 \mid y)$ with respect to t and subject to conditions that match first and second sample moments of t with those same moments of θ conditional on y. Cressie's proposal given by (11), (12) and (13) does not invoke any optimality conditions and so is likely to be less efficient than Ghosh's estimator (14).

Constrained Bayes estimation for more general models, such as GLMs, is presented by Ghosh (1992), although from an essentially univariate point of view. Our earlier comment, that we do not have flexible ways to model lack of independence in nonlinear, nonnormal models, is equally appropriate here.

Finally, we agree with the authors' comment about the importance of small area estimation in medical geography. A good source for recent research in this area is the May 1993 Supplement Issue of the journal *Medical Care* (Proceedings of the Fourth Biennial Regenstrief Conference, "Methods for Comparing Patterns of Care," October 27–29, 1991). We are working on incorporating spatial variation and dependence into statistical methods for these and other small area estimation problems.

ACKNOWLEDGMENT

Partial support came from the Office of Naval Research Grant N00014-93-1-0001, NSF Grant DMS-92-04521 and the National Security Agency Grant MDA904-92-H-3021.

Comment

D. Holt

The paper by Ghosh and Rao is a valuable summary of recent developments using empirical Bayes and hierarchical Bayes methods for making small area estimates. The need for methods which make provision for local variation while pooling information across areas is well established. The review is a thorough appraisal of the methods and their properties, and the numerical results reinforce earlier results which demonstrate that these methods are preferable to others such as synthetic estimation and sample size dependent estimation.

The value of these approaches is not simply in their ability to provide point estimates for each small area which, on average, have better precision. A very important additional factor is that a measure of precision (MSE) and an estimator of this can be

D. Holt is Professor, Department of Social Statistics, University of Southampton, Southampton S09 5NH, United Kingdom.