Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

AREA-LEVEL MODELS USING DATA FROM MULTIPLE SURVEYS

Sharon L. Lohr and Lynn M. R. Ybarra[1]

ABSTRACT

Small area estimation methods typically combine direct estimates from a survey with predictions from a regression model to obtain estimates of population quantities which have reduced mean squared error. In this paper, we consider effects of errors or missing values in the covariates used in the regression model. Such situations occur when the covariates come from another survey or from an administrative source with incomplete data. We present and develop properties of models that allow survey and administrative data to be used as auxiliary information for estimating quantities of interest in a primary survey. The methodology accounts for the survey designs and missing data in the error structure.

KEY WORDS: Best linear unbiased prediction; Measurement error; Multivariate mixed effects model.

1. INTRODUCTION

National surveys such as the U.S. Current Population Survey (CPS) or the U.S. National Crime Victimization Survey (NCVS) give accurate estimates of poverty or criminal victimization at the national level. These surveys do not, however, contain sufficient sample sizes to give reliable estimates by themselves for “small areas” such as states, counties or minority groups, or to provide detailed information about events such as domestic violence that affect only a small part of the population. Current methods for estimating poverty in counties incorporate auxiliary administrative information from sources such as tax records and food stamp programs as explanatory variables in a regression equation; the predicted value of the regression is combined with a direct estimate of poverty from the CPS to estimate the county poverty rate (Citro and Kalton, 1999). If the regression model gives accurate predictions, the mean squared error of the resulting small area estimate is smaller than that for the direct estimate of county poverty from the CPS. Properties of the small area estimates such as bias and mean squared error are derived conditionally on the auxiliary information: This approach assumes that the auxiliary data are available for all areas and are measured without error.

In many situations, however, auxiliary information is available that can help in the estimation, but that information is not exact. Auxiliary information may be available from another survey, or may come from an administrative source in which imputation has been used to fill in missing values. In both of these cases, the auxiliary information is measured with error—sampling and nonsampling error for survey data, and imputation error for incomplete administrative data. We give four examples of situations where auxiliary information may have errors.

Estimates of income and poverty at the state and county level currently use the CPS, which samples approximately 60,000 households each year, to obtain a direct estimate of poverty in the small areas. If funded by Congress, however, the proposed American Community Survey (ACS) will sample 3 million households each year. For most small areas, the ACS is expected to give more precise estimates of quantities it measures; it has been suggested as a source of auxiliary information for improving accuracy of estimates from the CPS. The ACS still contains sampling error for many small areas, however, and that error should be accounted for in any error bounds reported for the estimates.

[1]Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona 85287-1804 USA (sharon.lohr@asu.edu; lynn.ybarra@asu.edu)

The U.S. Bureau of the Census is interested in estimates of health insurance coverage in states, with initial focus on coverage of low-income children. These estimates are used to allocate funds for the State Children's Health Insurance Program. The CPS is used for the primary estimate of health insurance coverage. One research question (Campbell et al., 2002) involves whether to use auxiliary data that may be provided by some states and that are of variable quality.

Another example involves estimating violent crime victimization rates in each state, or the total amount of medical expenses incurred in each state as a result of violent crime. The U.S. National Crime Victimization Survey (NCVS) provides this information, but has insufficient sample sizes to provide accurate estimates for every state. The FBI Uniform Crime Reports (UCR), which provides statistics on crimes reported to police agencies, is an excellent source of auxiliary information. Although the UCR underestimates the amount of crime and its costs to society, victimization rates from the UCR are positively correlated with victimization rates from the NCVS. Reporting to the UCR is voluntary, however, and the UCR data set has many holes; in addition, data reported by some police agencies may be inaccurate.

Many survey designs in the U.S. are now being integrated to allow combination of estimates. The U.S. National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) share the same primary sampling units (psu's): the psu's selected for NHIS are used as a sampling frame for NHANES. NHIS is a stratified multistage probability sample of about 100,000 persons (40,000 households) per year. The design is described in detail in Botman et al. (2000). NHANES conducts medical examinations of participants, however, and the mobile examination unit can only visit 15 psu's per year (about 5000 persons), as opposed to 358 psu's for NHIS. Because of the small sample size, NHANES data are usually accumulated over time in order to produce estimates. State and local estimates from NHANES have low precision. The NHIS data provide better estimates of quantities measured at some localities, but the data come from an interview rather than an examination: for example, in NHANES, prevalence of diabetes is estimated using the results of the medical exams, while in NHIS it is estimated using the results of questionnaires. We would expect, though, that the questionnaire results would be highly correlated with the medical examination results, and thus that the NHIS would provide high-quality auxiliary information for use with NHANES data for improved small area estimation.

Fay and Herriot (1979) first studied improved estimation in small areas using known vectors of covariate means. Since then, many other models have been studied. Prasad and Rao (1990) put many of these estimators in a unified framework and derived second-order approximations to the mean squared errors of the estimators. Schaible (1996) described indirect small area estimators used by U.S. government agencies. Rao (2003) gave a detailed account of work done in small area estimation to date.

Suppose there are m areas of interest (for example, m = 50 if states are small areas). We are interested in a characteristic Y_i of area i. For some (or all) areas, we have data from the primary survey. Let y_i be an unbiased estimator of Y_i from the survey, with sampling variance $V(y_i) = \psi_i$. Administrative data for area i, $\mathbf{A}_i$, is assumed to be measured without error. We consider the p-vector $\mathbf{x}_i$ to be measurements from the auxiliary data source, which may have sampling error and/or bias. Each vector $\mathbf{x}_i$ estimates the true characteristic for that area, $\mathbf{X}_i$. We assume that $E(\mathbf{x}_i) = \mathbf{X}_i + \mathbf{b}_i$ and $V(\mathbf{x}_i) = \mathbf{\Sigma}_i$. Note that if $\mathbf{X}_i$ is measured exactly, the bias $\mathbf{b}_i$ and covariance matrix $\mathbf{\Sigma}_i$ are both zero.

Often the characteristic of interest will be a mean or proportion: For the crime example, Y_i will represent the proportion of persons who are victims of crime in state i. The NCVS gives a direct estimate y_i; the auxiliary data $\mathbf{x}_i$ come from the UCR.

The goal in this paper is to use the auxiliary data $\mathbf{x}_i$ to improve estimation of the characteristic of interest Y_i. Lohr and Prasad (2001) developed unit-level models when the auxiliary data come from another survey; in those models, however, one must be able to match individuals' data from the primary and auxiliary survey. The area-level models of this paper require only that one be able to link the information at the area level. In Section 2, we describe consequences of using the Fay-Herriot (1979) model when the $\mathbf{x}_i$'s are measured with error, and discuss an empirical Bayes approach. In Section 3, we extend the multivariate Fay-Herriot model to incorporate the error in estimation. Section 4 presents measurement error models that incorporate the uncertainty in $\mathbf{x}_i$ into the estimator; Section 5 contains concluding remarks.

2. FAY-HERRIOT MODEL AND EMPIRICAL BAYES ESTIMATION

The Fay-Herriot (1979) model leads to the best linear unbiased predictor (BLUP) of Y_i. If y_i and Y_i are assumed to be normally distributed, the Fay-Herriot estimator can be motivated by a Bayesian approach (see Rao, 2003, chapter 9). In that case $y_i \mid Y_i,\ \psi_i \sim N(Y_i,\ \psi_i)$; the regression model for the population quantity is given by

$$Y_i \mid \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \alpha, \beta \sim N(\mathbf{A}_i^T \alpha + \mathbf{X}_i^T \beta, \sigma_v^2). \tag{2.1}$$

It is assumed that the quantities (y_i, Y_i) are independent across the areas, conditionally on $\mathbf{A}_i$, $\mathbf{X}_i$, α, β and σ_v^2. With this model, the posterior distribution of Y_i is

$$Y_i \mid y_i, \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \alpha, \beta, \psi_i \sim N[\gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T \alpha + \mathbf{X}_i^T \beta),\ \psi_i \gamma_i^*] \tag{2.2}$$

where $\gamma_i^* = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. Note that the posterior mean is the BLUP for Y_i when all of the quantities that are conditioned on are known; the mean squared error (MSE) of the BLUP is the posterior variance.

Now consider what happens if an estimate $\hat{\mathbf{X}}_i$ is substituted for the population quantity $\mathbf{X}_i$. In practice, $\mathbf{x}_i$ might be used for $\hat{\mathbf{X}}_i$. Let $\mathbf{C}_i = \text{MSE}(\hat{\mathbf{X}}_i)$ and let $\tilde{Y}_i^* = \gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta)$ be the substitution estimator of Y_i. Then $\text{MSE}(\tilde{Y}_i^*) = \gamma_i^* \psi_i + (1-\gamma_i^*)^2 \beta^T \mathbf{C}_i \beta$. Note that if the matrix $\mathbf{C}_i$ is large, the mean squared error of the substitution estimator can be larger than ψ_i, the variance of the estimator using only the survey data. Thus if the auxiliary information is inaccurate, its use can result in a less accurate estimate than if no auxiliary information were used. The relative weight, γ_i^*, given to the direct estimator, y_i, in the substitution estimator may be too small since the relative weight does not account for all of the error in the predicted value from the regression. In addition, if it is pretended that there is no error in estimating $\mathbf{X}_i$, the posterior variance in (2.2) is smaller than the true MSE.

We can correct the MSE by incorporating the error in estimating $\mathbf{X}_i$ into the model in (2.1). Suppose that $\mathbf{x}_i$ is also normally distributed: $\mathbf{x}_i \mid \mathbf{X}_i, \mathbf{\Sigma}_i \sim N(\mathbf{X}_i, \mathbf{\Sigma}_i)$ and that $\mathbf{X}_i \mid \mu_X, \mathbf{\Sigma} \sim N(\mu_X, \mathbf{\Sigma})$. Then the posterior distribution of $\mathbf{X}_i$ is

$$\mathbf{X}_i \mid \mathbf{x}_i, \mu_x, \Sigma_i, \Sigma \sim N[\mathbf{c}_i,\ \mathbf{D}_i], \tag{2.3}$$

where $\mathbf{c}_i = \mathbf{\Sigma}\,(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\,\mathbf{x}_i + \mathbf{\Sigma}_i\,(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\,\mu_X$ and $\mathbf{D}_i = \mathbf{\Sigma}\,(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\,\mathbf{\Sigma}_i = \mathbf{\Sigma}_i\,(\mathbf{\Sigma}_i + \mathbf{\Sigma})^{-1}\,\mathbf{\Sigma}$. Combining (2.2) and (2.3), the posterior distribution of Y_i has mean $\gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T \alpha + \mathbf{c}_i^T \beta)$ and variance $\gamma_i^* \psi_i + (1-\gamma_i^*)^2 \beta^T \mathbf{D}_i \beta$. With the additional assumptions on the distribution of the auxiliary survey data, the posterior variance is correct for the MSE. The relative weight γ_i^*, however, still does not account for the error in estimating $\mathbf{X}_i$; it is possible for the posterior variance to be larger than ψ_i so that incorporating the auxiliary $\mathbf{x}$ information may result in a decrease in precision. The models in Sections 3 and 4 adjust the relative weights so that this problem will not occur.

The above discussion assumed that the regression parameters and covariance matrices are known. These must in general be estimated from the data; in that case, under appropriate regularity conditions, the MSE of the resulting small area estimator is the posterior variance above plus lower order terms. The extra terms in the MSE due to estimating the parameters are given in Ybarra (2003).

3. MULTIVARIATE FAY-HERRIOT MODEL

Fay (1987) and Datta et al. (1991) developed a Fay-Herriot-type model for a multivariate response, and showed that it often results in more efficient estimators for a small area quantity of interest than the univariate Fay-Herriot model. Datta et al. (1991) were interested in estimating Y_i, the median income of four-person households in state *i*. The direct estimate y_i was from the CPS. The auxiliary information, $\mathbf{x}_i$ = (3/4) (median income of five-person households)

+ (1/4) (median income of three-person households) also came from the CPS. They found that using a multivariate model reduced the MSE of the estimator of Y_i.

We extend this model to allow for missing observations and measurement error, and to allow the observations to come from different sources. In the following, let $\mathbf{0}_k$ denote a k-vector with all entries 0, and let $\mathbf{I}_k$ denote the $k \times k$ identity matrix. Assume throughout this section that $\mathbf{x}_i$ is an unbiased estimator of $\mathbf{X}_i$, so that $\mathbf{b}_i = \mathbf{0}$.

Let $\mathbf{U}_i = [\mathbf{X}_i^T \;\; Y_i\,]^T$ represent the population values for each of the i areas, $i = 1, \ldots, m$. Then a model relating the population quantities for the areas to each other is $\mathbf{U}_i = \mathbf{A}_i^T\alpha + \mathbf{v}_i$, where $\mathbf{v}_i \sim N(\mathbf{0}_{p+1}, \mathbf{\Sigma}_v)$. The model covariance matrix $\mathbf{\Sigma}_v$ is partitioned as

$$\mathbf{\Sigma}_v = \begin{bmatrix} \mathbf{\Sigma}_{vxx} & \mathbf{\Sigma}_{vxy} \\ \mathbf{\Sigma}_{vxy}^T & \mathbf{\Sigma}_{vyy} \end{bmatrix}.$$

Define the vector $\mathbf{u}_i$ and the matrices $\mathbf{Z}_i$ and $\mathbf{\Psi}_i$ for three cases:

1. $\mathbf{u}_i = [\mathbf{x}_i^T, y_i\,]^T$, $\mathbf{Z}_i = \mathbf{I}_{p+1}$, $\mathbf{\Psi}_i = \text{blockdiag}(\mathbf{\Sigma}_i, \psi_i)$ if both $\mathbf{x}$ and y are observed in area i;
2. $\mathbf{u}_i = \mathbf{x}_i$, $\mathbf{Z}_i^T = [\,\mathbf{I}_p, 0]^T$, $\mathbf{\Psi}_i = \mathbf{\Sigma}_i$ if $\mathbf{x}$ is observed in area i but not y;
3. $\mathbf{u}_i = y_i$, $\mathbf{Z}_i^T = [\,\mathbf{0}_p, 1]^T$, $\mathbf{\Psi}_i = \psi_i$ if y is observed in area i but not $\mathbf{x}$.

Then the observations $\mathbf{u}_i$ follow the model $\mathbf{u}_i = \mathbf{Z}_i^T\mathbf{A}_i^T\alpha + \mathbf{Z}_i^T\mathbf{v}_i + \mathbf{e}_i$, where $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{\Psi}_i)$. The covariance matrix of $\mathbf{u}_i$ is $\mathbf{V}_i = \text{V}(\mathbf{u}_i) = \mathbf{Z}_i^T\mathbf{\Sigma}_v\mathbf{Z}_i + \mathbf{\Psi}_i$. This model fits into the block diagonal covariance structure model described in Section 6.3 of Rao (2003). The BLUP for $\mathbf{U}_i$ is then

$$\tilde{\mathbf{U}}_i = \mathbf{A}_i\tilde{\alpha} + \tilde{\mathbf{v}}_i, \tag{3.1}$$

where $\tilde{\mathbf{v}}_i = \mathbf{\Sigma}_v\mathbf{Z}_i\mathbf{V}_i^{-1}(\mathbf{u}_i - \mathbf{Z}_i^T\mathbf{A}_i\tilde{\alpha})$ and

$$\tilde{\alpha} = \left(\sum_i \mathbf{A}_i^T\mathbf{Z}_i\mathbf{V}_i^{-1}\mathbf{Z}_i^T\mathbf{A}_i\right)^{-1}\left(\sum_i \mathbf{A}_i^T\mathbf{Z}_i\mathbf{V}_i^{-1}\mathbf{u}_i\right). \tag{3.2}$$

Let $\mathbf{M}_i = (\mathbf{\Sigma}_{vxx} + \mathbf{\Sigma}_i)^{-1}$ and $\kappa_i = (\mathbf{\Sigma}_{vyy} - \mathbf{\Sigma}_{vxy}{}^T\,\mathbf{M}_i\,\mathbf{\Sigma}_{vxy}) / (\mathbf{\Sigma}_{vyy} - \mathbf{\Sigma}_{vxy}{}^T\,\mathbf{M}_i\,\mathbf{\Sigma}_{vxy} + \psi_i)$. Using (3.1),

$$\tilde{Y}_{i,\text{MFH}} = \kappa_i y_i + (1-\kappa_i)\{[\mathbf{0}_p^T, 1]\mathbf{A}_i\tilde{\alpha} + \mathbf{\Sigma}_{vxy}^T\mathbf{M}_i(\mathbf{x}_i - [\mathbf{I}_p, 1]\mathbf{A}_i\tilde{\alpha})\} \tag{3.3}$$

for cases 1 and 2 above (note that for case 2, $\kappa_i = 0$). For case 3 (areas i in which $\mathbf{x}$ is not measured and hence the entries of $\mathbf{M}_i$ are 0), $\tilde{Y}_{i,\text{MFH}} = \kappa_i y_i + (1-\kappa_i)[\mathbf{0}_p^T, 1]\mathbf{A}_i\tilde{\alpha}$, and $\kappa_i = \mathbf{\Sigma}_{vyy} / (\mathbf{\Sigma}_{vyy} + \psi_i)$.

The weighting κ_i in the small area estimator in (3.3) thus depends on the variability of $\mathbf{x}_i$: κ_i is smaller, and the small area estimator depends more heavily on the direct estimator, if the variability of $\mathbf{x}_i$ is larger. If $\mathbf{X}_i$ is measured exactly (i.e., all entries of $\mathbf{\Sigma}_i$ are 0), the multivariate Fay-Herriot estimator, using assumptions of normality, coincides with the univariate Fay-Herriot estimator that incorporates the $\mathbf{x}$ information as covariates.

The MSE of the estimator in (3.3) can be obtained using standard methods: Under regularity assumptions given in Datta et al. (1991) and Ybarra (2003), we have that for cases 1 and 3, $\text{MSE}\,(\tilde{Y}_{i,\text{MFH}}) = \kappa_i\psi_i + O(m^{-1})$. For case 2, $\text{MSE}\,(\tilde{Y}_{i,\text{MFH}}) = \kappa_i^* + O(m^{-1})$, where κ_i^* is the numerator of κ_i.

In practice, $\mathbf{\Sigma}_v$ as well as α must be estimated from the data. Method of moments, maximum likelihood, or restricted maximum likelihood may be used. See Datta et al. (2001) for a comparison of the estimators of $\mathbf{\Sigma}_v$ in the univariate case.

4. MEASUREMENT ERROR MODELS

4.1 Estimation when regression parameters and variances are known

We saw in Section 2 that ignoring the error in $\mathbf{x}_i$ resulted in underestimating the mean squared error and using a non-optimal weighting. The motivation for using a measurement error model comes from the observation that omitted or inaccurate covariates can cause bias. Suppose that the model in (2.1) holds, but that the analyst fits the model without the term involving β. Since the "wrong" model is fit, estimates of the regression parameters α and the predicted values may be biased. The bias arising from omitting **X** from the covariates leads to an increase in the MSE of the predicted values. If **X** is included in the covariates, though, the error in measuring **X** must be accounted for in the estimation and mean squared error. Fuller (1987, 1990) presented a comprehensive treatment of using measurement error models for estimation of regression parameters and for prediction.

As in Section 2, let $\hat{\mathbf{X}}_i$ be an estimator of the population quantity $\mathbf{X}_i$ with $\mathbf{C}_i =$ MSE$(\hat{\mathbf{X}}_i)$. We assume that such an estimator exists for every area: If **x** is not measured in area *i*, then an empirical Bayes estimator or imputed value may be used for $\hat{\mathbf{X}}_i$. Consider the model

$$y_i = \mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta + r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) + e_i, \tag{4.1}$$

where $r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) = v_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)\beta$ and MSE(r_i) = $\sigma_v^2 + \beta^T \mathbf{C}_i \beta$. As before, $V(e_i) = \psi_i$ is the sampling variance of y_i. We assume here that y_i, v_i and $\hat{\mathbf{X}}_i$ are mutually independent. Now let

$$\tilde{Y}_i = \gamma_i y_i + (1-\gamma_i)\{\mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta\}, \tag{4.2}$$

with $\gamma_i = (\sigma_v^2 + \beta^T \mathbf{C}_i \beta)/(\sigma_v^2 + \beta^T \mathbf{C}_i \beta + \psi_i)$. If y_i is measured in area *i* then MSE $(\tilde{Y}_i) = \gamma_i \psi_i$, which is at most as large as the variance ψ_i of the direct estimator, y_i. If y_i is not measured in area *i* then MSE $(\tilde{Y}_i) = \sigma_v^2 + \beta^T \mathbf{C}_i \beta$.

Note that MSE $(\tilde{Y}_i) \le$ MSE $(\tilde{Y}_i^*)$, where $\tilde{Y}_i^*$ is the substitution estimator from Section 2: Equality is attained if $\beta^T \mathbf{C}_i \beta = 0$. The MSE is also smaller than the MSE that would result from using the model $y_i = \mathbf{A}_i^T \alpha + t_i + e_i$ instead of (4.1). If the empirical Bayes estimator from Section 2 is used for $\hat{\mathbf{X}}_i$, then it can be shown that the estimator in (4.2) is equivalent to the multivariate Fay-Herriot estimator.

4.2 Estimating the regression parameters and variances

In practice, the quantities σ_v^2, α and β are unknown and must be estimated from the data. Although most research using area-level models has assumed that ψ_i is known, ψ_i and $\mathbf{C}_i$ may also need to be estimated.

Lindley (1947, p. 243) suggested using weighted least squares to estimate the regression parameters. For our model, the MSE of the error terms is MSE $(r_i + e_i) = \sigma_v^2 + \psi_i + \beta^T \mathbf{C}_i \beta$. Thus, one can solve for the unknown parameters by minimizing

$$Q_1(\alpha, \beta) = \sum_i \frac{(y_i - \mathbf{A}_i^T \alpha - \hat{\mathbf{X}}_i^T \beta)^2}{\psi_i + \sigma_v^2 + \beta^T \mathbf{C}_i \beta} \tag{4.3}$$

where the sum is over areas *i* where *y* is measured. Gleser (1981) gives large sample properties of the resulting estimates of the regression parameters.

Use of (4.3), however, requires that σ_v^2 be known. If it is unknown, we can use modified least squares to estimate the parameters (Cheng and Van Ness, 1999, pp. 85 and 146). In this case an unbiased estimator of σ_v^2 is

$$Q_2(\alpha,\beta) = \frac{1}{m}\sum_i [(y_i - \mathbf{A}_i^T\alpha - \hat{\mathbf{X}}_i^T\beta)^2 - \psi_i - \beta^T\mathbf{C}_i\beta] \tag{4.4}$$

Minimizing Q_2 with respect to α and β gives estimates of the regression parameters. Note, though, that terms in (4.4) may be negative and it is possible that minimization will occur on the boundaries of the parameter space.

5. DISCUSSION

In this paper, we have discussed and compared the first order properties of three methods—empirical Bayes, multivariate Fay-Herriot, and measurement error models—for incorporating auxiliary information into small area estimation, when the auxiliary information is subject to error. All three methods account for the extra variability in not knowing the auxiliary information exactly. The multivariate Fay-Herriot and measurement error models also change the relative weightings of the direct estimator and the predicted value from the regression equation, so that more reliance is placed on the direct estimator if $\mathbf{X}_i$ is measured inaccurately. The measurement error model is the most flexible in terms of choice of estimator of $\mathbf{X}_i$; it gives the same results as the multivariate Fay-Herriot estimator in the normal case when there are no administrative covariates and the shrinkage estimator from Section 2 is used to estimate $\mathbf{X}_i$.

The first order MSE's presented in this paper are calculated assuming that the regression parameters and covariances are known. If these are unknown, and must be estimated from the data, the MSE's of the estimators will be larger than stated here. The second order properties depend on the relative magnitudes of m, the number of areas, and $\mathbf{C}_i$, the error in estimating $\mathbf{X}_i$. Ybarra (2003) compares estimators of the unknown quantities and derives the second order terms of the MSE's of the estimators.

Although in some situations the measurement error model and multivariate Fay-Herriot method give the same results, we prefer the measurement error model for many practical situations. It is more flexible for choice of estimator of $\mathbf{X}_i$. In addition, robust methods may be used for estimating the regression parameters and variance terms, so that the measurement error model is adaptable for situations in which some of the $\mathbf{x}_i$'s are outliers due to variable quality of the data sources.

ACKNOWLEDGEMENTS

This research was partially supported by the U.S. National Science Foundation under Grant No. 0105852.

REFERENCES

Botman, S. L., T. F. Moore, C. L. Moriarity, and V. L. and Parsons (2000), "Design and Estimation for the National Health Interview Survey, 1995-2004", National Center for Health Statistics. *Vital Health Statistics* 2(130).

Campbell, J., R. Fisher and D. Waddington (2002), "Preliminary Research for Small Area Health Insurance Estimates", paper presented at the October meeting of the Census Advisory Committee of Professional Associations, Arlington, Virginia.

Cheng, C.-L. and J. W. Van Ness (1999), *Statistical Regression with Measurement Error,* London: Arnold.

Citro, C. F. and G. Kalton (eds.) (1999), *Small-area Estimates of School-age Children in Poverty: Interim Report 3*, Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

Datta, G. S., R. E. Fay and M. Ghosh (1991), “Hierarchical and Empirical Multivariate Analysis in Small Area Estimation”, *Proceedings of the Bureau of the Census Annual Research Conference*, Washington: Bureau of the Census, pp. 63-79.

Datta, G. S., J. N. K. Rao and Smith, D. D. (2001), “On Measures of Uncertainty of Small Area Estimators in the Fay-Herriot Model”, Technical Report, University of Georgia, Athens, Georgia.

Fay, R.E. (1987), “Application of Multivariate Regression to Small Domain Estimation”, in R. Platek et al. (eds.), *Small Area Statistics*, New York: Wiley, pp. 91-102.

Fay, R. E. and R. A. Herriot (1979), “Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data”, *Journal of the American Statistical Association*, 78, pp. 269-277.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Fuller, W. A. (1990), “Prediction of True Values for the Measurement Error Model”, in P. J. Brown and W. A. Fuller (eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics Vol. 112,* Providence, RI: AMS, pp. 41-57.

Gleser, L. J. (1981), “Estimation in a Multivariate ‘Errors-in-Variables’ Regression Model: Large Sample Results”, *Annals of Statistics*, 9, pp. 24-44.

Lindley, D.V. (1947), “Regression Lines and the Linear Functional Relationship”. *Journal of the Royal Statistical Society Supp*., 9, pp. 218-244.

Lohr, S. and N. G. N. Prasad (2001), “Small Area Estimation with Auxiliary Survey Data”, Technical Report 01.08, Statistics Centre, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton.

Prasad, N.G.N. and Rao, J.N.K. (1990), “The estimation of the mean squared error of small-area estimators”, *Journal of the American Statistical Association*, 85, pp. 163-171.

Rao, J. N. K. (2003), *Small Area Estimation*, New York: Wiley, in press.

Schaible, W. L. (ed.) (1996), *Indirect Estimators in U.S. Federal Programs*. Lecture Notes in Statistics, 108, New York: Springer-Verlag.

Ybarra, L. M. R. (2003), *Small Area Estimation using Data from Multiple Surveys*, unpublished Ph.D. dissertation, Arizona State University.