Efficient Small Area Estimation When Covariates Are Measured With Error Using Simulation Extrapolation

Trijya Singh*
Le Moyne College, Syracuse, U.S.A. - singht@lemoyne.edu

Suojin Wang
Texas A & M University, College Station, U.S.A. - sjwang@stat.tamu.edu
Raymond J. Carroll
Texas A & M University, College Station, U.S.A. - carroll@stat.tamu.edu

Abstract

Small area estimation methods typically combine direct estimates from a survey with predictions from a model in order to obtain estimates of population quantities such that the mean squared error of the predictor is minimized. The Fay-Herriot model, which uses area level auxiliary information to develop such an predictor is a popular choice among the practitioners. But covariates may be measured with error and in such situations predictors will have enhanced mean square error and may be even worse than the direct domain predictor. Recently, improved predictors have been proposed for small areas incorporating the mean squared error of the auxiliary variables measured with error. We propose an alternative strategy to improve the predictor of the parameter of interest in the small area and increase their efficiency by considering an additive measurement error model and applying the well-known bias correction technique called simulation-extrapolation (SIMEX). The performance of this corrected predictor is compared empirically with that of other predictors and our proposed predictor is shown to perform better. We also empirically illustrate the performance of the predictors under mild departures from normality of measurement error.

Keywords:Fay-Herriot model; Mean squared error; Jackknife estimator.

1. Introduction

Sample surveys are normally designed and conducted to produce reliable estimates of population parameters related to a characteristic of interest at a higher (national, state or district) level. Let $P = (U_1, \ldots, U_N)$ be a finite population and $Y_1, \ldots, Y_N$ be corresponding values of characteristic Y measured on the units. We are interested in estimating population parameters like mean, total, median or quartiles of Y or the proportion of units of the population which belong to a specified group in case the characteristic is categorical. To this end, a large scale survey is typically conducted using an appropriate sampling design to collect information on n units in the sample. A predictor having good design properties such as unbiasedness and minimum variance is used to estimate the parameter θ of interest. However, at this stage, planners may also need the estimates for some part or domain of the surveyed population.These domains are referred to as 'small areas'. Since there may be a very small number of units in the sample belonging to this small area, estimates based on them will be highly unstable and unreliable. The estimates may be unbiased but will have large variance. To tackle this problem, small area estimates have been developed which use auxiliary information available from census or previously conducted surveys. These estimates "borrow strength" from other small areas of the population surveyed. For an excellent account of the topic, one may refer to Rao (2003).

We consider the population as composed of m non-overlapping small areas $A_1, \ldots, A_m$, one of which is the small area of our interest and sampled units are also divided according to these areas. Let $\theta^T = (\theta_1, \ldots, \theta_m)$ be the vector of parameters of interest of m small areas and we are interested in estimating θ_i the parameter corresponding to the i^{th} small area. In addition to the area level sample information on Y for the area A_i, we also have area levelinformation on p covariates $x_1, \ldots, x_p$ from census, registers or past surveys. That is, area level information $x_i^T = (x_{i1}, \ldots, x_{ip})$ is known for each $i = 1, \ldots, m$. We use this information from other areas which is relevant in estimating θ_i. Thus, direct estimate of θ_i based on Y - values is augmented by indirect estimate of θ_i based on information from all other small areas. Such a predictor is, in fact, a

composite predictor- a linear combination of 'direct' and 'indirect' predictors. Fay & Herriot (1979) developed one such predictor which is Empirical Best Linear Unbiased Predictor (EBLUP) for θ_i and uses area level sample and auxiliary information.

In practice, the auxiliary information on some or all covariates may contain measurement error. Instead of measuring x_i, we may obtain w_i. In such situations, not only the bias but also the mean squared error of the predictor may be enhanced to unacceptable levels. In certain cases, Fay-Herriot's Best Linear Unbiased Predictor (BLUP) may be worse than even the direct estimate of the small area, as shown by Lohr & Ybarra (2008). For functional model where x_i's are treated as fixed and classical additive measurement error is considered, that is, $w_i = x_i + u_i$, $E(w_i|x_i) = x_i$ and $MSE(w_i|x_i) = C_i$, Lohr & Ybarra (2008) proposed a predictor of θ_i, incorporating the C_i's (assumed to be known) in the predictor itself and obtained the mean squared error of the predictor. They showed that their predictor is approximately unbiased. Since the enhanced inconsistency in the predictor and the enhanced mean squared error is due to the enhanced inconsistency of the estimate of the regression coefficient β (caused by measurement error) used in the model, we first correct the estimator of β for bias using the simulation-extrapolation (SIMEX) method developed by Cook & Stefanski (1994) and then use it in the Fay-Herriot predictor. We investigate the properties of the resulting predictor empirically.

The remaining part of the paper is organized as follows. In Section 2, we briefly describe the Fay-Herriot model for small areas when the area level information is available, discuss the consequences of the measurement error in the covariates and describe the Lohr-Ybarra model for measurement error and their predictor. In Section 3, we describe the SIMEX method for bias correction and obtain a modified multivariate version of SIMEX method for bias correction suitable for the Fay-Herriot model. We then use the bias corrected estimator of regression coefficient to develop a new predictor for estimating the small area parameter.Section 4 is devoted to simulation studies for investigating the properties of SIMEX based predictor and its comparison with the Fay-Herriot and Lohr-Ybarra predictors. The simulation studies were carried out under different error structures. Section 5 concludes the paper with a discussion.

2. Measurement Error in the Fay-Herriot Model

The Fay-Herriot model is described as follows. Assume that the auxiliary information $x_i^T = (x_{i1}, \ldots, x_{ip})$ on p covariates is available for each area i. The Fay-Herriot model relates the small area direct design based unbiased predictor θ_i to the area level auxiliary data x_i, $i = 1, \ldots, m$ through a linking model. Here, we shall take $\theta_i = y_i$, that is, y_i is the direct sample survey estimate of θ_i based on units of large scale survey belonging to small area A_i. The Fay-Herriot model is given by

$$y_i = x_i^T \beta + v_i + e_i \tag{1}$$

for $i = 1, \ldots, m$, where y_i is the design based predictor of θ_i. It is assumed that the random effect v_i $N(0, \sigma_v^2)$ and sampling error e_i $N(0, \psi_i)$ are independently distributed. $\beta^T = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients also known as fixed effects. The elements v_i's are the unobservable random effects (also called model errors) which capture the additional "unstructured" area specific effects which could not be accounted for by the covariates x_i's.

If instead of measuring the true x_i's, we obtain contaminated observations with additive error: $w_i = x_i + u_i$, where u_i $N(0, C_i)$, then the naive generalized least squares estimator of β is given by $\beta_{GLS(naive)} = (W^T V^{-1} W)^{-1} W^T V^{-1} y$, where $W^T = (w_1, \ldots, w_m)$, $y^T = (y_1, \ldots, y_m)$ and $V = Diag(\psi_1 + \sigma_v^2, \ldots, \psi_m + \sigma_v^2)$. It can be shown that the estimator $\beta_{GLS(naive)}$ is an inconsistent, biased estimator of β and is attenuated towards the null vector. Since this estimator is used in obtaining the best linear unbiased predictor of θ, the resulting predictor will no longer remain unbiased. Also, measurement error inflates the mean squared error of the best predictor that uses the Fay-Herriot model to the extent that in the case of high heterogeneity of auxiliary information in small area i, Best Predictor (BP) could be worse than even direct predictor y_i for the i^{th} small area. Using $\beta_{GLS(naive)}$ to get the BLUP or σ_v^2 and $\beta_{GLS(naive)}$ to get EBLUP will worsen the situation manifold. In fact, in the presence of measurement error, the Fay-Herriot predictor may borrow bias from other areas rather than strength.

Treating x_i's as fixed, Lohr & Ybarra (2008) considered classical additive error and expressed the Fay-Herriot model as $y_i = w_i^T \beta + r_i(w_i, x_i) + e_i$, where $r_i(w_i, x_i) = v_i - \beta^T u_i$. We can easily check that $Var(r_i) = \sigma_v^2 + \beta^T C_i \beta$.

Lohr & Ybarra (2008) proposed the empirical predictor as

$$\theta_{iME} = \gamma_i y_i + (1 - \gamma_i) w_i \beta_w, \tag{2}$$

where

$$\gamma_i = \frac{\sigma_v^2 + \beta_w^T C_i \beta_w}{\sigma_v^2 + \beta_w^T C_i \beta_w + \psi_i}, \tag{3}$$

is the empirical estimate of the weight γ_i. The authors used the modified least squares method of Cheng & Van Ness (1999) and applied a recursive algorithm to obtain

$$\beta_w = \left(\sum_{i=1}^{m} d_i (w_i w_i^T - C_i) \right)^{-1} \sum_{i=1}^{m} d_i w_i y_i, \tag{4}$$

provided the inverse exists. Here, d_i's are suitable finite weights bounded away from zero. Their estimate of σ_v^2 is $\sigma^2_{v(w)} = (m-p)^{-1} \sum_{i=1}^{m} \{(y_i - w_i^T \beta_w)^2 - \psi_i - \beta_w^T C_i \beta_w\}$. Their method involves the choice of initial weights d_i's in the normal equation, the existence and convergence of the inverse in order to find the estimate β_w given in Equation (4).

3. A SIMEX Predictor

Simulation-extrapolation (SIMEX) is a simulation based bias correction method for functional models when the bias is induced into the regression coefficient estimator due to measurement error. This method was first developed by Cook & Stefanski (1994). It is designed to measure the effect of different levels of measurement error through simulation in a resampling-like situation and to establish a trend in induced bias for different measurement levels. From this trend, we extrapolate the estimate for the situation when there is no measurement error.

We propose to use the SIMEX estimator of β in Equations (2) and (3) in order to get the SIMEX predictor for Fay-Herriot model. We shall now develop the SIMEX estimator for β under the Fay-Herriot model from Equation (1). The SIMEX algorithm consists of two steps, which are applied in the case of the Fay-Herriot model containing measurement error as follows.

Simulation Step:

For a given data set $(y_1, w_1), \ldots, (y_m, w_m)$, we generate B (large) additional data sets by simulating multivariate pseudo-errors u_{bi} from multivariate normal distribution $N(0, C_i)$ as follows: for each area $i = 1, \ldots, m$, obtain

$$w_{bi}(\lambda_m) = w_i + \sqrt{\lambda_m} u_{bi}, \tag{5}$$

where $u_{bi} \overset{iid}{\sim} MVN(0, C_i)$ and $b = 1, \ldots, B$. From Equation (5), it can be seen that $Var(w_{bi}(\lambda_m)) = (1+\lambda_m) C_i$. For the b^{th} data set, we obtain, $\beta_b(\lambda_m) = (W_b^T V^{-1} W_b)^{-1} W_b^T V^{-1} y$, where $W_b^T = (w_{b1}, \ldots, w_{bm})$. From the B estimates $\beta_b(\lambda_m)$, $b = 1, \ldots, B$, for a given λ_m, we define the SIMEX estimator as the average of all the values of $\beta_b(\lambda_m)$, that is, $\beta_{SIMEX}(\lambda_m) = B^{-1} \sum_{b=1}^{B} \beta_b(\lambda_m)$.

Extrapolation Step:

We obtain a set $(\lambda_m, \beta_{SIMEX}(\lambda_m))$, $m = 0, \ldots, M$ of estimates for $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_M$. Normally, we use 0.0, 0.5, 1.0, 1.5, 2.0 as values of λ_m. For each co-ordinate β_j, $j = 1, \ldots, p$ of β, we fit a suitable model $g_j(\lambda_m, \Gamma)$ for $\beta_j(\lambda_m)$ as a function of λ_m, where Γ is a vector of parameters of the function $g(\cdot, \cdot)$. The SIMEX estimator of β_j is extrapolated from this function.

In our simulation studies, we used the quadratic extrapolant function given by $g_j(\lambda, \Gamma) = c_1 + c_2\lambda + c_3\lambda^2$, where $\Gamma = (c_1, c_2, c_3)$. Hence, $\beta_{j(SIMEX)} = g_j(-1, \Gamma) = c_1 - c_2 + c_3$, where $\Gamma = (c_1, c_2, c_3)$ is a vector of fitted coefficients.

In Section 2, we saw that measurement error may affect the Fay-Herriot predictor through w_i and β. The adverse effect through β shall be more serious as it depends on the measurement errors of all the small areas.

The adverse effect through w_i is taken care of by considering Equation (2) with the choice of γ_i given by

$$\gamma_{i(SIMEX)} = \frac{\sigma^2_{v(SIMEX)} + \beta^T_{SIMEX} C_i \beta_{SIMEX}}{(\sigma^2_{v(SIMEX)} + \beta^T_{SIMEX} C_i \beta_{SIMEX} + \psi_i)},$$

where $\beta^T_{SIMEX} = (\beta_{1(SIMEX)}, \ldots, \beta_{p(SIMEX)})$ and $\sigma^2_{v(SIMEX)} = (m-p)^{-1} \sum_{i=1}^{m} \{(y_i - w_i^T \beta_{SIMEX})^2 - \psi_i - \beta^T_{SIMEX} C_i \beta_{SIMEX}\}$ is a consistent estimator. The effect of measurement error on the Fay-Herriot predictor would be reduced to a great extent when a good estimate of β is employed. Substituting the SIMEX estimates β_{SIMEX} and $\gamma_{i(SIMEX)}$ in Equations (2) and (3), for given ψ_i, we obtain the SIMEX predictor for i^{th} small area as

$$\theta_{i(SIMEX)} = \gamma_{i(SIMEX)} y_i + (1 - \gamma_{i(SIMEX)}) w_i \beta_{SIMEX} . \quad (6)$$

As in the Fay-Herriot model, for obtaining β_{SIMEX} also we need estimates of weights $(\psi_i + \sigma^2_v)^{-1}$, $i = 1, \ldots, m$. Fay & Herriot (1979) and Lohr & Ybarra (2008) both used iterative algorithms to arrive at final estimators of β and σ^2_v. In our case, we shall follow a similar approach. We shall start with an initial estimate of $\sigma^2_v = 0$, and obtain estimates β_{SIMEX} and $\sigma^2_{v(SIMEX)}$. Using these estimates, we iterate the process until convergence occurs.

4. Simulation Studies

In order to study and compare the performance of the proposed SIMEX predictor with other methods, we generated x_i from $N(5, 9)$ and ψ_i from a Gamma distribution with shape parameter 4.5 and scale parameter 2. Then, for each iteration, we generated $Y_i = 1 + 3x_i + v_i$, $y_i = Y_i + e_i$ and $w_i = x_i + u_i$, where v_i, e_i and u_i are independent normal variables with mean 0 and variance σ^2_v, ψ_i and c_i respectively. Thus, w_i^T used in Equation (6) will, in the case of this simulation, be a vector given by $(1, w_i)$ and C_i will be a matrix whose (2,2) entry is c_i with remaining entries being zero. We carried out simulations using three factors of choice of parameters: Factor 1: σ^2_v = 2, 3 or 4; Factor 2: c_i = 0, 2, 3, 4; Factor 3: m= 20, 50 or 100. For each

Table 1: Average empirical mean squared error for the predictors, m = 100, c_i = 3, and σ_v^2 = 4

k	c_i	y_i	Y_{iS}	Y_{iME}	$Y_{i(SIMEX)}$	Y_{iFH}
0	0	8.9	3.3	3.2	3.4	3.2
20	3	10.4	6.9	7.8	3.2	3.3
50	3	9.2	6.3	7.4	3.2	3.2
80	3	9.8	6.7	7.0	5.6	3.2
100	3	11.8	5.4	5.6	5.0	3.2

Table 2: Average empirical MSE and average absolute bias for the Lohr-Ybarra predictor, Y_{iME} and the SIMEX predictor $Y_{i(SIMEX)}$ for non-normal measurement error distributions and different values of m and σ_v^2.

Average Empirical MSE			
Parameters	k	Y_{iME}	$Y_{i(SIMEX)}$
m = 50, σ_v^2 = 4, u_i t_5	0	5.87	5.86
	20	5.34	5.21
	50	6.68	5.20
	80	6.73	4.79
	100	6.74	4.72
Average Absolute Bias			
Parameters	k	Y_{iME}	$Y_{i(SIMEX)}$
m = 100, σ_v^2 = 4, u_i t_{10}	0	1.23	1.22
	20	1.31	1.06
	50	1.36	0.64
	80	1.57	0.62
	100	1.93	0.75

5. Conclusions

Measurement error has been referred to by Carroll et al.(2006) as the triple whammy since it has three main effects. It causes bias in parameter estimation for statistical models, leads to a loss of power for detecting relationship among variables, and it masks the features of the data, making graphical model analysis difficult. The bias caused in the slope estimate due to measurement error in the direction of zero is commonly referred to as attenuation or attenuation to the null. The effects of measurement error can range from simple attenuation to situations where real effects are hidden, observed data exhibit relationships that are not present in the error free data and even signs of estimated coefficients are reversed.

This paper has directly dealt with the triple whammy of measurement error described above and the havoc it can cause. There has been extensive research in the area of measurement error for decades now which has resulted in a wide variety of methods that answer the need for a solution to the measurement error problem. We make advantage of the flexibility and widespread applicability of these methods at our disposal and use them to tackle the problem with respect to small area estimation.

Since our model is a functional measurement error model, one of the most widely used methods in this scenario is simulation extrapolation (SIMEX) which we apply with success to the Fay-Herriot model with

measurement error in covariates. This proves to be an effective method of bias correction in this case and provides reduction in the estimated mean squared errors compared with predictors that are currently available. The beauty of the SIMEX method lies in its effectiveness even when there is departure from normality or additivity of the measurement error. This is evident from the simulation results provided in the paper when data were generated from heavy-tailed *t* and skew-normal distributions. One fact stands out very clearly from the simulation study, namely, when auxiliary information is available, that a survey statistician should always make use of it to obtain small area estimates and not merely use the direct design-based predictors. It is important to note that though SIMEX is a simulation based method, it is easy to implement. There is a SIMEX package available in R developed by Wolfgang and Lederer that is very convenient and flexible since it allows for different choices of extrapolant functions namely linear, quadratic or non linear.
Despite the usefulness of SIMEX for tackling the problem of measurement error in the Fay-Herriot model, there is much room for improvement and further research in this area. The availability of better estimators for variance components might further improve the performance of the SIMEX predictor and is the subject of ongoing research.Also, it might be interesting to study the performance of these bias correction methods in small area estimation models involving count data, where the covariates are measured with error.

References

Carroll, R. J., Ruppert, D., Stefanski, L. A. , & Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd edition), New York: Chapman & Hall.

Cook, J. R., & Stefanski, L. A. (1994), "Simulation Extrapolation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Fay, R., & Herriot. R. (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.

Lohr, S., & Ybarra, L. (2008), "Small Area Estimation When Auxiliary Information Is Measured With Error," *Biometrika*, 95, 919–931.

Rao, J. N. K. (2003), *Small Area Estimation*, (1st edition), Hoboken, New Jersey: John Wiley Sons Inc.