Efficient Small Area Estimation in the Presence of Measurement Error in Covariates

Dr. Trijya Singh
singht@lemoyne.edu

Department of Mathematics and Statistics
Le Moyne College, Syracuse, New York

Chulalongkorn University, Bangkok
September 2, 2013

Outline

1. What is small area estimation?
2. The Fay-Herriot Model
3. Bias Correction Using the Simulation-Extrapolation Method
4. Bias Correction Using Corrected Scores
5. Simulation Study
6. Data Example

What is a small area?

- Finite population $\mathcal{U} = \{1, ..., k, ..., N\}$. k's are labels of units. Population may be a nation, a state or any other geographical area or a large demographic group.
- A large scale survey carried out in $\mathcal{U}$, to estimate parameters like total, mean, variance, quartiles, proportions. For eg., average income or proportion of smokers.
- Later, policy makers may become interested in estimating these parameters from large scale survey data for subpopulations or domains called "small areas". These areas may be districts or counties.
- Survey was not planned for these areas. Number of units in large scale sample falling in these areas may be very small or may be even zero. So it's impossible to produce reliable estimates for small areas.

More Examples: Drug Use Survey in Nebraska

- A large scale survey of $n = 4300$ individuals for estimating percentage of drug users in Nebraska.
- Later, it was decided to produce estimates of counties of Nebraska.
- It was found that out of 4300 only 14 persons were from Boone county and only one Caucasian woman in the age group 25-44.
- No reliable estimate of percentage of drug users in Boone or Caucasian people in age group 25-44.

Estimation Approach

- We use sample information for the areas of interest and the auxiliary information from the census or administrative registers to build estimates for small areas.
- We borrow strength from other area either through regression or through a model.
- **Composite Estimators:** A convex combination (weighted average) of two estimators (eg. direct and indirect estimators)
    - Weights chosen by minimizing MSE of composite estimator.
    - Weights control shrinkage of the two estimators.
    - Larger weights for direct estimator if sample size is large, otherwise larger weights for indirect estimator.

- $m =$ No. of small areas of interest, $Y_i =$ population characteristic of interest in area i.
- $y_i =$ direct design-based estimator of Y_i using data from large scale survey for area i.
- Assume $E(y_i) = Y_i$, auxiliary information X_i (p-vector of population characteristics) from the ith small area known exactly.
- Fay-Herriot model:

$$y_i = X_i^T \beta + v_i + e_i,$$

v_i and e_j independent r.v.'s with mean 0 for all i and j.
$v_i's \sim N(0, \sigma_v^2)$, $e_i \sim N(0, \psi_i)$.

Fay-Herriot Model with Measurement Error

- But what if X_i's, considered to be fixed constants, are unknown & are themselves measured with error? Causes bias in parameter estimation & loss of power in detecting relationships among variables.
- Lohr & Ybarra assumed W_i, estimator of X_i provided by auxiliary information, exists for each area i. Consider $W_i = X_i + U_i$, where $U_i =$ measurement error for the auxiliary information in the i^{th} small area and $U_i \sim N(0, C_i)$.
- They expressed the Fay-Herriot model as:

$$y_i = W_i^T \beta + r_i(W_i, X_i) + e_i,$$

where $r_i(W_i, X_i) = v_i + (X_i - W_i)^T \beta$.

- Assume v_i independent of both W_i and e_i, random variables in different small areas are independent, W_i and y_i independent for each area i.
- Lohr-Ybarra estimator:

$$\widehat{Y}_{iME} = \widehat{\gamma}_i y_i + (1 - \widehat{\gamma}_i) W_i^T \widehat{\beta},$$

where $\widehat{\gamma}_i = \frac{\widehat{\sigma}_v^2 + \widehat{\beta}^T C_i \widehat{\beta}}{\widehat{\sigma}_v^2 + \widehat{\beta}^T C_i \widehat{\beta} + \psi_i} = \frac{\widehat{MSE(r_i)}}{\widehat{MSE(r_i)} + \psi_i}$

- On intuitive grounds they advocate larger weights to direct estimator if X_i is measured with error, larger weights to regression predictor otherwise.
- Takes care of measurement error to some extent, but estimator is still biased and improvement in efficiency not much.
- We use indirect estimates corrected for the bias in $\widehat{\beta}$ induced by measurement error.

SIMEX Steps:

- Simulation of *pseudo-errors* with variance ζC_i.
- A *re-measurement* of the auxiliary data W_i. New *pseudo-variable* $\tilde{W}_i$ for the b^{th} iteration ($b = 1, ..., B$):

$$\tilde{W}_{b,i} = W_i + \sqrt{\zeta} U_{b,i}.$$

- Estimates obtained from each of the generated, contaminated data sets in each area i.
- Above steps repeated large number of times. Average value of estimate for each level of contamination (different values of ζ) calculated. Averages plotted against ζ values (an extrapolant function fitted to averaged, error-contaminated estimates).
- Extrapolation to the ideal case of no pseudo-measurement error ($\zeta = -1$) yields the SIMEX estimate.

What are Corrected Scores?

- For the i^{th} sample observation, estimating function $\Psi_i(\beta; Y_i, X_i, v_i)$ (based on least squares, likelihood, etc.) is unbiased if:

  $$E\{\Psi_i(\beta; Y_i, X_i, v_i)\} = 0,$$

  for $i = 1, 2, ..., m$.
- Solution of $\sum_{i=1}^{n} \Psi_i(\beta; Y_i, X_i, v_i) = 0$ gives consistent estimator for β (Nakamura, 1990).
- Let $W_i = X_i + U_i$ be observed where U_i is the measurement error.

- Principle behind corrected scores: Construct unbiased $\Psi_i^*(\beta; Y_i, W_i, v_i)$ such that,

$$E^*_{W/Y,X,v}\{\Psi_i^*(\beta; Y_i, W_i, v_i)\} = \Psi_i(\beta; Y_i, X_i, v_i).$$

- $\Psi_i^*(\cdot)$ will be unbiased if $\Psi_i(\cdot)$, in the absence of measurement error, was unbiased to begin with.
- $\sum_{i=1}^n \Psi_i^*(\beta; Y_i, W_i, v_i) = 0$ yields consistent corrected score estimator of β.

- The Fay-Herriot model with measurement error:

$$\underline{y}_{m\times 1} = \underline{X}\beta + \underline{v} + \underline{e},$$

  $\underline{v}$ and $\underline{e}$ are distributed as $N_m(0, \sigma_v^2 I)$ and $Normal_m(0, \Sigma)$ respectively, where $\Sigma = Diag(\psi_1, \psi_2, ..., \psi_m)$.
- But we observe $W_i = X_i + U_i$, $U_i \sim N(O, \Lambda)$.

The corrected score estimators (using corrected log-likelihoods) for Fay-Herriot model:

$$\widehat{v}_{iFHCS} = \frac{\sigma_v^2}{\sigma_v^2 + \psi_i}(y_i - W_i^t \widehat{\beta}_{FHCS}),$$

and

$$\widehat{\beta}_{iFHCS} = \left\{ \sum_{i=1}^{m} \frac{W_i W_i^t}{\sigma_v^2 + \psi_i} - tr(\mathbb{P})\Lambda \right\}^{-1} \sum_{i=1}^{m} W_i y_i.$$

where $\mathbb{P} = Diag\left\{ \frac{1}{(\sigma_v^2+\psi_1)}, \frac{1}{(\sigma_v^2+\psi_2)}, \ldots.., \frac{1}{(\sigma_v^2+\psi_m)} \right\}$.

Estimation of Variance Components for CS Estimators

- Corrected score estimating equations:

$$\begin{pmatrix} W^t\Sigma^{-1}W - tr(\mathbb{P}).\Lambda & W^t\Sigma^{-1} \\ \Sigma^{-1}W & \Sigma^{-1} + \frac{1}{\sigma_v^2} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{v} \end{pmatrix} = \begin{pmatrix} W^t\Sigma^{-1}y \\ \Sigma^{-1}y \end{pmatrix}$$

- Equating the partial derivative of corrected log-likelihood with respect to σ_v^2 we obtain,

$$\widehat{\sigma}_v^2 = \frac{\widehat{v}^t\widehat{v}}{m} - \frac{1}{m}\sum_{i=1}^{m} \frac{\widehat{\beta}^t\Lambda\widehat{\beta}}{\left\{1 + \frac{\psi_i}{\widehat{\sigma}_v^2}\right\}^2}.$$

- Λ estimated using method of moments.

Monte Carlo Corrected Scores

- *What if corrected estimating equations cannot be solved analytically?*
- For $b = 1, \ldots., B$, generate random variables $\mathbf{Q}_{b,i}$, independent normal random vectors with mean zero and covariance matrix Σ_{uu}.
- Consider complex-valued random variate $\widetilde{\mathbf{W}}_{b,i} = \mathbf{W}_i + \mathbf{i}\mathbf{Q}_{b,i}$, where $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$.
- Replace $\mathbf{X}_i$ with $\widetilde{\mathbf{W}}_{b,i}$ in Ψ_{True}. (Ψ_{True}= estimating equation unbiased in absence of measurement error)
- Define the Monte Carlo Corrected Scores as:

$$\Psi_{MCCS,B}(\mathbf{Y}_i, \mathbf{W}_i, \Theta) = B^{-1} \sum_{b=1}^{B} Re\{\Psi_{True}(\mathbf{Y}_i, \mathbf{W}_{b,i}, \Theta)\}.$$

Steps (Contd.):

- Average over multiple sets of pseudorandom vectors, $b = 1, ..., B$.
- Solve the estimating equations:

$$\sum_{i=1}^{m} \widetilde{\Psi}_{MCCS,B}(\mathbf{Y}_i, \mathbf{W}_i, \Theta) = \mathbf{0}$$

for estimates of Θ, the vector the parameters in the model.

- It has been shown that:

$$E[Re\{\Psi_{True}(Y_i, \widetilde{W}_{b,i}, \Theta)\}|Y_i, X_i] = \Psi_{True}(\mathbf{Y}_i, \mathbf{X}_i, \Theta).$$

That is, $Re\{\Psi_{True}(Y_i, \widetilde{W}_{b,i}, \Theta)\}$ is a corrected score.

- Generation of $X_i \sim N(4, 9)$, $\psi_i \sim$ Gamma(5, 2).
- For each iteration we generated $Y_i = 1 + 4x_i + v_i$, $y_i = Y_i + e_i$ 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_v^2, ψ_i and c_i respectively.
- Consider 3 factors (Lohr and Ybarra, 2008) : Factor 1: $\sigma_v^2 = 2$ or 4. Factor 2: $c_i \in \{0, d\}$ for $d=$ $\{$2,3, or 4$\}$; Factor 3: $m=$ 20, 50 or 100. No. of iterations for each combination $=$ 10000.

Simulation study(contd.)

- 3 different scenarios w.r.t. X_i, i.e., ALL of them being measured with error (k=100), some (specified percentage k) measured with error and NONE ($k = 0$) of them measured with error.
- SIMEX estimates obtained after generating pseudo-variables.
- Find empirical MSE's, for each area i, for the direct, Fay-Herriot (ignoring measurement error), Lohr-Ybarra and SIMEX estimators, $\Sigma_{l=1}^{10000}(\widehat{Y}_{i(l)} - Y_{i(l)})^2/10000$ where $Y_{i(l)}$ and $\widehat{Y}_{i(l)}$ are the true and predicted values of $X_i^T\beta + v_i$ in l^{th} iteration.

Table 1 : Empirical MSE's for estimators, y_i (direct), $\widetilde{Y}_{iS}$ (Fay-Herriot estimator ignoring measurement error), $\widehat{Y}_{iME}$ (Lohr-Ybarra), $\widehat{Y}_{iSIMEX}$ (SIMEX), $\widehat{Y}_{iFHCS}$ (ordinary corrected scores), $\widehat{Y}_{iMCCS}$ (Monte Carlo Corrected Scores) when the number of small areas is 100, measurement error variance $C_i = 4$ and $\sigma_v^2 = 4$. k is the percentage of areas having auxiliary information measured with error.

k	C_i	y_i	$\widetilde{Y}_{iS}$	$\widehat{Y}_{iME}$	$\widehat{Y}_{iSIMEX}$	$\widehat{Y}_{iFHCS}$	$\widehat{Y}_{iMCCS}$
0	0	8.1	3.8	3.7	3.8	3.7	3.7
20	4	9.2	7.3	6.4	3.9	4.0	4.1
50	4	9.3	6.5	6.5	4.2	4.3	4.3
80	4	10.8	6.7	7.4	5.7	5.6	5.5
100	4	10.9	7.5	7.3	5.5	5.4	5.3

- Data set from the 2003-2004 U.S. National Health and Nutrition Examination Survey (NHANES), 2004 U.S. National Health Interview Survey as auxiliary information.
- Small areas = 30 demographic subgroups cross-classified by race and ethnicity (Mexican American, Non-Hispanic Black and Non-Hispanic White), by age group (20-39, 40-59, 60 years and above) and by gender.
- Height and weight for each respondent are measured National Health and Nutrition Examination Survey medical examination by government interviewers.
- The body mass index (BMI) is calculated as $height/weight^2$.
- In the NHIS, BMI calculated using responses reported by the subjects themselves through a questionnaire, hence the presence of measurement error.

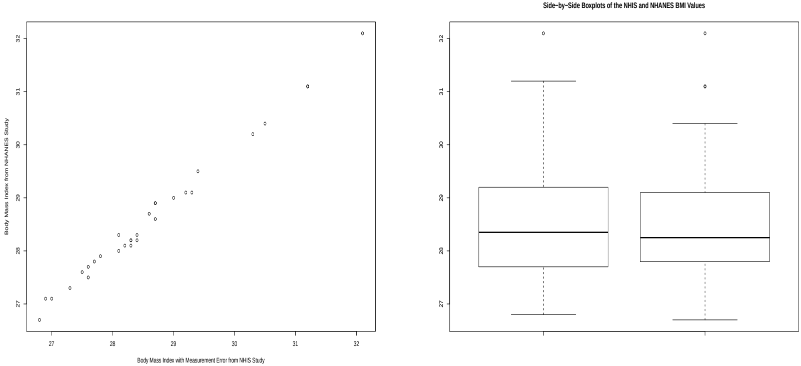

Figure 1 : Left: BMI's values from NHANES v/s BMI's from NHIS. Right: Box-plots for BMI's from NHIS and NHANES

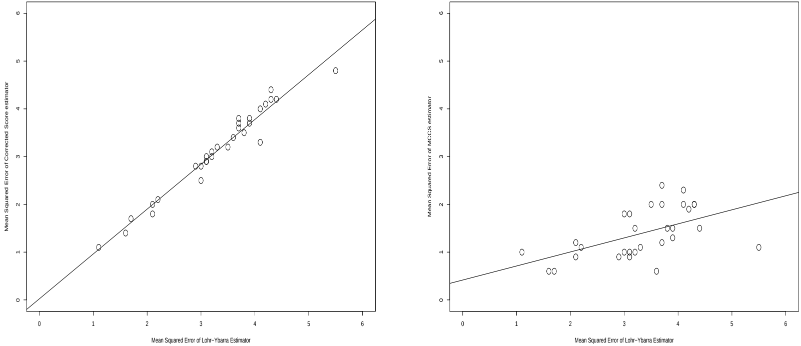

Figure 2 : Left: Jackknife MSE's of corrected score estimates v/s Lohr-Ybarra estimates. Right: MCCS v/s Lohr-Ybarra estimates

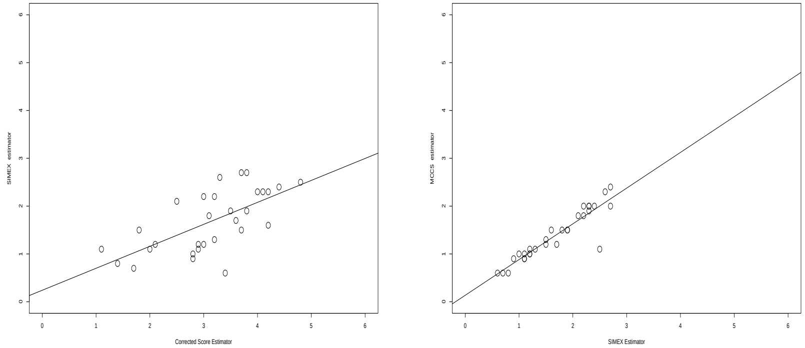

Figure 3 : Left: Jackknife MSE's of SIMEX estimates v/s corrected score estimates. Right: MCCS v/s SIMEX estimates

THANK YOU !!