Small Area Estimation: Part II

Partha Lahiri

JPSM, Univ. of Maryland, College Park, USA

May 18, 2011

Notation

- U: **set of units [e.g., households (HH)] in the finite population (e.g., country) of interest, which contains m small areas (e.g., states) U_i**
- N_i: **size of U_i, $(N = \sum_{i=1}^{m} N_i)$**

Notation

y_k: a welfare variable (income, expenditure, etc.) of interest for the kth unit in U

- Brazil: per-capita household expenditure
- U.S.A.: household income in the Small Area Income and Poverty Estimates (SAIPE) program.

Notation

z: threshold under(s) which a unit is under poverty

- **In Brazil, the IBGE used 20 different thresholds, varying by geographic region and rural/urban areas.**
- **In the U.S. SAIPE program, different thresholds are used depending on the household composition.**

FGT Poverty Indices (Foster, Greer and Thornbecke, 1984)

$F_{\alpha i}(\mathbf{y}_i) = \frac{1}{N_i} \sum_{k \in U_i} \left(\frac{z - y_k}{z}\right)^{\alpha} I(y_k < z)$, **where**

$$I(y_k < z) = \begin{cases} 1 & \textbf{if } y_k < z \textbf{ ,} \\ 0 & \textbf{otherwise,} \end{cases}$$

where α is a measure of the sensitivity of the index to poverty and $\mathbf{y}_i = (y_k,\ k \in U_i)$.

Poverty Incidence

Poverty Incidence $(\alpha = 0)$**:**

$$F_{\alpha i}(\mathbf{y}_i) = \frac{1}{N_i} \sum_{k \in U_i} I(y_k < z)$$

- **proportion of units in that area living below the poverty line**
- **does not measure the intensity of poor**

Poverty Gap

Poverty Gap ($\alpha = 1$)**:**

$$F_{\alpha i}(\mathbf{y}_i) = \frac{1}{N_i} \sum_{k \in U_i} \left(\frac{z - y_k}{z} \right) I(y_k < z)$$

- **measure poverty intensity**
- **can be interpreted as cost of eliminating poverty**

Poverty Severity

Poverty Severity ($\alpha = 2$):

$$F_{\alpha i}(\mathbf{y}_i) = \frac{1}{N_i} \sum_{k \in U_i} \left(\frac{z - y_k}{z} \right)^2 I(y_k < z)$$

- **gives more emphasis to the very poor.**

Design-based Direct Estimation

Define

- s: **set of units in the sample (size** n**)**
- s_i: **set of units in** s **that belong to area** i **(size** n_i**),** $\sum_{i=1}^{m} n_i = n$
- w_k: **survey weight associated with unit** $k \in s$
- $u_k = \left(\frac{z-y_k}{z}\right)^{\alpha} I(y_k < z)$.

$$\hat{F}_{\alpha i}^{Dir} = \sum_{k \in s_i} w_k u_k / \sum_{k \in s_i} w_k$$

The World Bank or the ELL Method (Elbers, Lanjouw and Lanjouw, 2003)

Basic data requirements

- **Micro level census data**
- **Micro level survey data containing the welfare variable of interest**
- **Common auxiliary variables between the survey and the census**

The World Bank or the ELL Method

Issues to think about

- **Time gap between the census and the survey**
- **Incomparability of the auxiliary variables between the survey and the census**

The ELL Method Contd.

- **Assume a linear mixed model on the log-transformed welfare variable of interest.**
- **Obtain L synthetic "census" files**

  $\tilde{\mathbf{y}}^{*}_{i;l},\ (l = 1, \ldots, L)$**.**

The ELL Method Contd.

- **The ELL estimate of $F^*_{\alpha i}(\mathbf{y}_i)$ is then obtained as $\bar{F}_{\alpha i} = L^{-1} \sum_{l=1}^{L} F_{\alpha i}(\tilde{\mathbf{y}}^*_{i;l})$.**
- **The measure of uncertainty of the ELL estimate is given by**

$$\frac{1}{L-1} \sum_{l=1}^{L} \left(F_{\alpha i}(\tilde{\mathbf{y}}^*_{i;l}) - \bar{F}^*_{\alpha i} \right)^2 .$$

A correction $1 + 1/L$ is often applied to capture variation due to imputation.

Remarks

- **In the ELL model, area specific auxiliary variables from different administrative records can be incorporated.**
- **The ELL mixed model attempts to capture different features of the survey design, but not any small area specific effect.**

Remarks

- **Even when small area random effects are introduced, the ELL method is still synthetic and is inferior to the empirical Bayes (Molina and Rao, 2010).**
- **Just like any other synthetic small area methods, the ELL method is capable of producing poverty estimates even when there is no survey data from the area.**

Mean Squared Error of A Synthetic Estimator

$$\mathrm{MSE}(\hat{\bar{Y}}_i) \equiv M_i = E(\hat{\bar{Y}}_i - \bar{Y}_i)^2 = V_i + B_i^2,$$

where

- $V_i = V(\hat{\bar{Y}}_i)$: **variance of** $\hat{\bar{Y}}_i$
- $B_i = E(\hat{\bar{Y}}_i) - \bar{Y}_i$: **bias of** $\hat{\bar{Y}}_i$

The expectations and variances are with respect to the sample design ($i = 1, \cdots, m$).

Remark

- **The variances V_i are generally small**
- **B_i^2 does not depend on the sample size. Its magnitude depends on the synthetic assumption that generates the synthetic estimators**

Average MSE

$$\text{AvMSE} \equiv M = \bar{V} + \eta,$$

- $\bar{V} = m^{-1} \sum_{i=1}^{m} V_i$
- $\eta = m^{-1} \sum_{i=1}^{m} B_i^2$

Approximately design-unbiased AMSE estimator was proposed by Gonzales and Waksberg (1973; GW in the graphs).

Design-based Simulation

- **Simulation set-up: Molina and Rao (2010)**
- **Finite populations are generated from the following nested error regression:**

$$\log(y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + v_i + e_k,\ k \in U_i,$$

where $\{v_i\}$ **and** $\{e_k\}$ **are independent with** $v_i \sim N(0, \sigma_v^2)$ **and** $e_k \sim N(0, \sigma_e^2)$

Design-based Simulation

- $m = 40,\ N_i = 250, n_i = 3, L = 50, \beta = (3, .03, -.04)', \sigma_e^2 = 0.25,\ \sigma_v^2 = 1, R = 1000.$
- **Case (i) ELL uncertainty measure is based on the correct model**
- **Case (ii) ELL uncertainty measure is based on the an incorrect model (covariate x_1 not included)**

DESIGN-BASED EXPERIMENT

✓ Generate only ONE population from the nested error model;

✓ Draw $I = 1000$ stratified samples with SRSWOR from each prov.

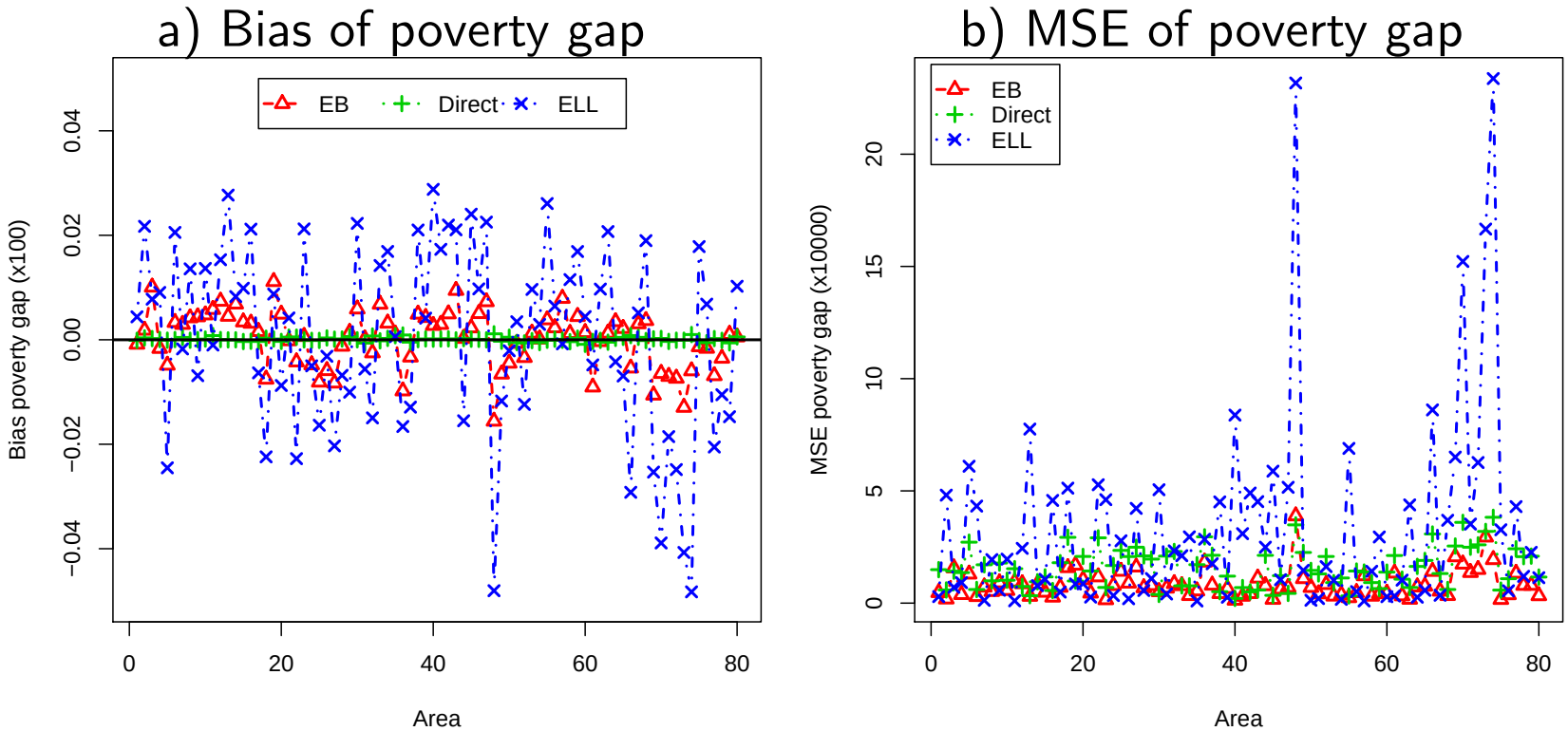
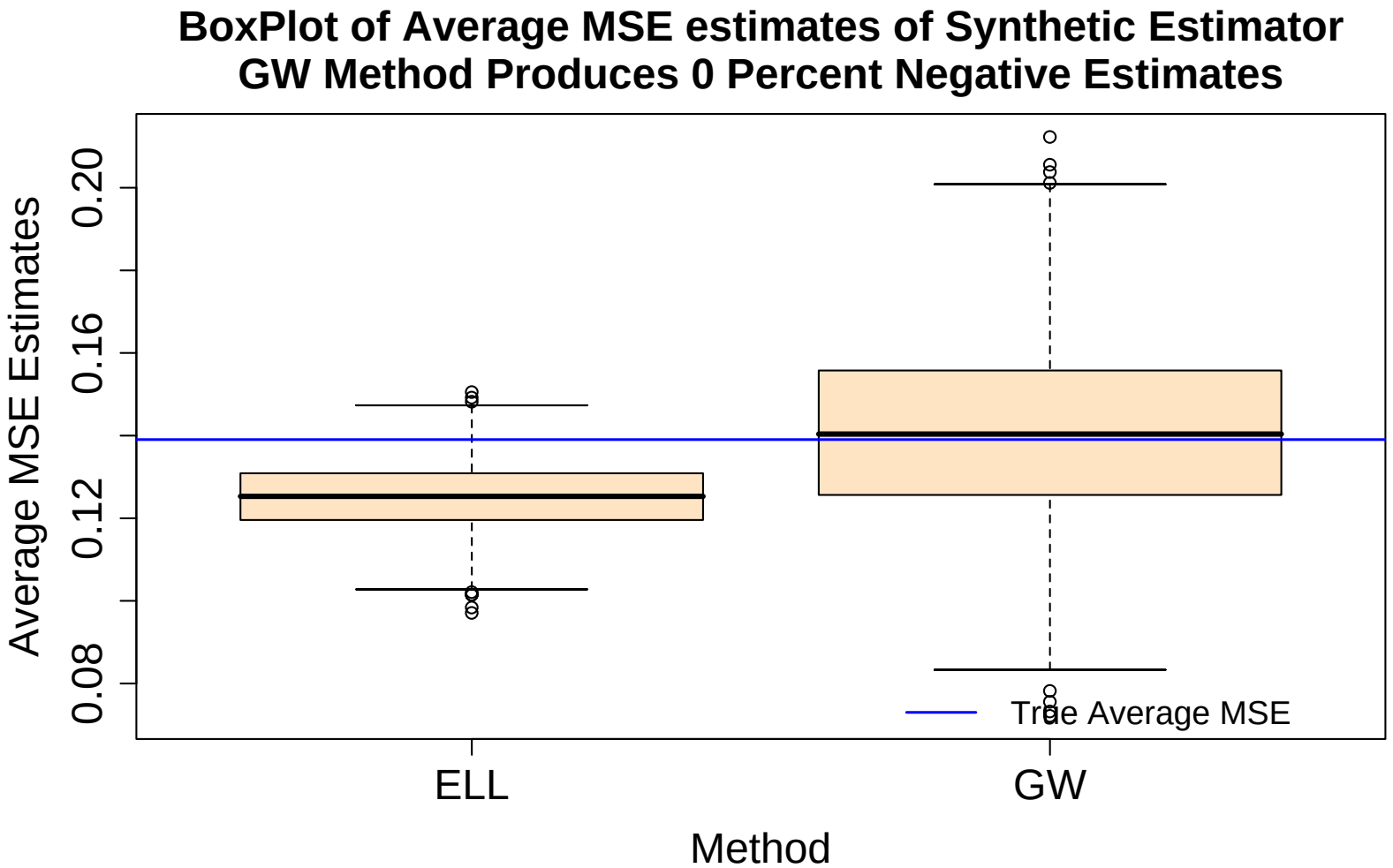
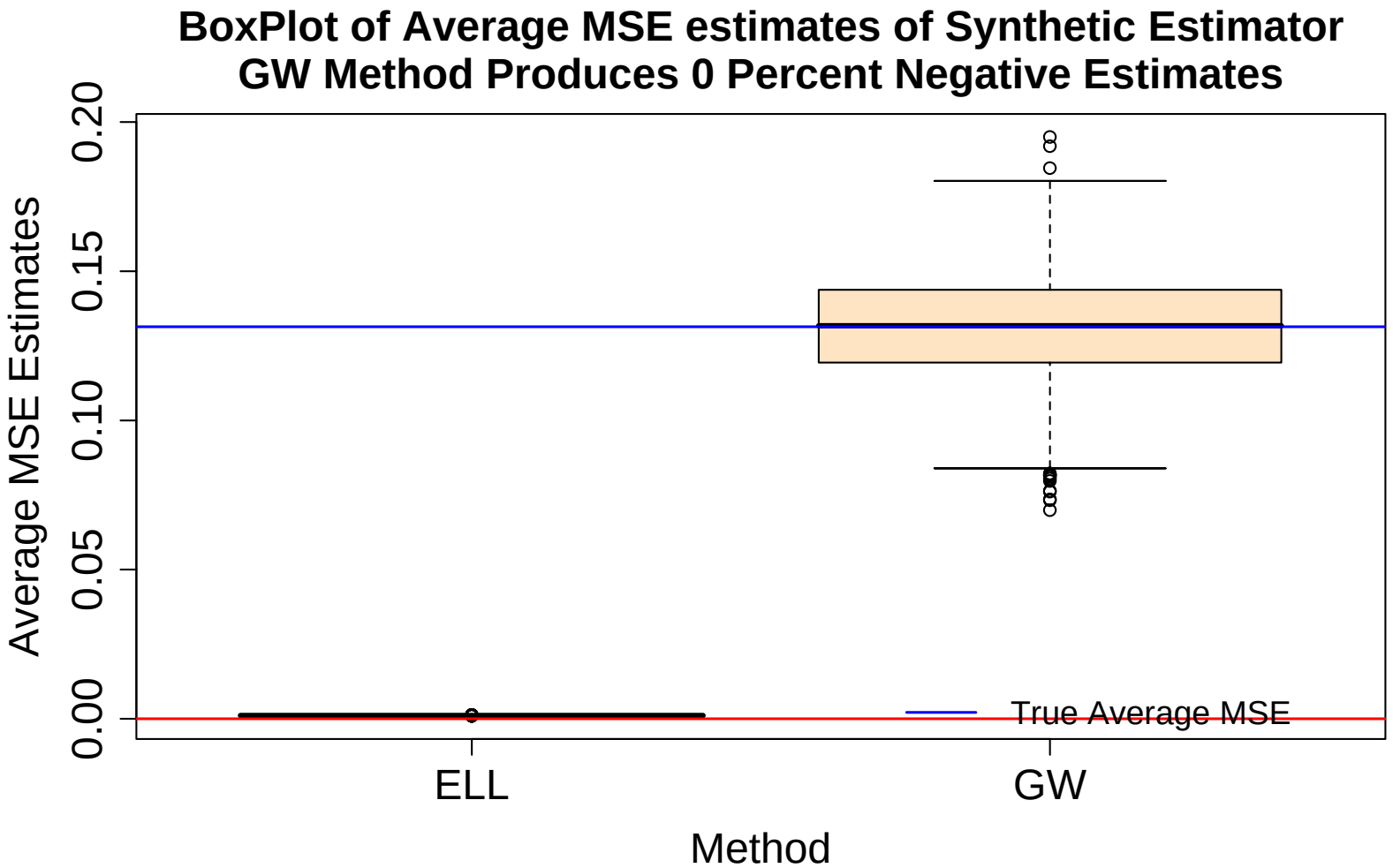


Figure 4. Bias and MSE of EB, direct and ELL estimators of pov. gap.


BoxPlot of Average MSE estimates of Synthetic Estimator
GW Method Produces 0 Percent Negative Estimates
Average MSE Estimates
0.08
0.12
0.16
0.20
True Average MSE
ELL
GW
Method


BoxPlot of Average MSE estimates of Synthetic Estimator
GW Method Produces 0 Percent Negative Estimates
Average MSE Estimates
0.00
0.05
0.10
0.15
0.20
True Average MSE
ELL
GW
Method

Estimation of per-capita income of small places

Ref: Fay and Herriot (1979, JASA)

- **Estimation of 1969 per-capita income (PCI) for small places ($\approx$ 15,000 are for places with population $<$ 500 in 1970.)**
- **Income data was collected on the basis of about 20% sample in the 1970 census.**
- $\hat{\bar{Y}}_i$**= estimator**
- $\hat{N}_i = \sum_{j \in s_i} w_i$**=weighted sample count**

Estimation of per-capita income of small places

- **CV($\hat{\bar{Y}}_i$)$\approx \frac{3}{\sqrt{\hat{N}_i}}$**
- **CV: about 13% (population ≈ 500)**

  about 30% (population ≈ 100)

Standard deviation increases in direct proportion to the expected value.
Let $\hat{\theta}_i = \ln(\hat{\bar{Y}}_i)$ and $\hat{\psi}_i = 9/\hat{N}_i$

Area Specific Auxiliary Information

(1) **per-capita income for the county in which the place belongs**

(2) **value of housing for the place**

(3) **value of housing for the county**

(4) **Internal Revenue Service (IRS) adjusted gross income per exemption for the place**

(5) **IRS-adjusted gross income per exemption for the county**

Empirical Bayes Estimation

The Fay-Herriot Area Level Model:

For $i = 1, \cdots, m,$

$$\textbf{Level 1: } \hat{\theta}_i | \theta_i, \psi_i = \hat{\psi}_i \overset{ind}{\sim} N(\theta_i, \psi_i);$$

$$\textbf{Level 1: } \textit{Apriori, } \theta_i \overset{ind}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \tau^2),$$

- $\psi_i = \hat{\psi}_i$ **are known**
- $\mathbf{x}_i$ **is a** $p \times 1$ **column vector of known auxiliary variables**
- $\boldsymbol{\beta}$ **and** τ^2 **are unknown hyperparameters.**

The Bayes Estimator of θ_i

The Bayes estimator:

$$\hat{\theta}_i^B = \hat{\theta}_i^B(\boldsymbol{\phi}) = \mathbf{x}_i^T\boldsymbol{\beta} + \gamma_i(\hat{\theta}_i - \mathbf{x}_i^T\boldsymbol{\beta}),$$

where $\gamma_i = \frac{\tau^2}{\psi_i+\tau^2}$ **and** $\boldsymbol{\phi} = (\boldsymbol{\beta}, \tau^2)^T$.

Estimation of hyperparameters

The maximum likelihood estimator (MLE) of β when τ^2 is known

$$\begin{aligned}&\hat{\boldsymbol{\beta}}(\tau^2)\\ &= \left(\sum_{j=1}^{m} \frac{1}{\psi_j + \tau^2}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\left(\sum_{j=1}^{m} \frac{1}{\psi_j + \tau^2}\mathbf{x}_j\hat{\theta}_j\right).\end{aligned}$$

One can use MLE, REML, ANOVA or any other reasonable estimator of τ^2.

Empirical Bayes Estimator of θ_i

An empirical Bayes (EB) estimator of θ_i:

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^B(\hat{\tau}^2) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\hat{\tau}^2) + \hat{\gamma}_i[\hat{\theta}_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\hat{\tau}^2)],$$

where γ_i is by $\hat{\gamma}_i = \hat{\tau}^2/(1+\hat{\tau}^2)$.

Adjustments to EB to Achieve Robustness

(a) **Consider the following winsorized EB:**

$$\begin{aligned}\hat{\theta}_i^{*EB} &= \hat{\theta}_i^{EB} \text{ if } \hat{\theta}_i - c_i \leq \hat{\theta}_i^{EB} \leq \hat{\theta}_i + c_i \\ &= \hat{\theta}_i - c_i \text{ if } \hat{\theta}_i^{EB} < \hat{\theta}_i - c_i \\ &= \hat{\theta}_i + c_i \text{ if } \hat{\theta}_i^{EB} > \hat{\theta}_i + c_i,\end{aligned}$$

where $c_i = \sqrt{\psi_i}$.

(b) **A PCI estimator** $e^{\hat{\theta}_i^{*EB}}$ **is obtained using a simple back transformation.**

(c) **Apply a two-way raking.**

Evaluation

The Census Bureau conducted complete censuses of a random sample of places and townships in 1973 and collected income data for 1972 on a 100% basis.

$\#$ of places with population size $< 500 : 17.$

$\#$ of places with population size between 500 and 999: 7.

Evaluation

Estimates for 1972 were obtained by multiplying the estimates by updating factors f_i

Average Percent Difference

N	$\hat{\bar{Y}}_i$	$\hat{\bar{Y}}_i^*$	$\hat{\bar{Y}}_i^C$
$<$ 500	28.6	22.0	31.6
500-999	19.1	15.6	19.3

where $\hat{\bar{Y}}_i^C$= County estimate.