Discussion of Professor Rao's Paper

Partha Lahiri

JPSM, Univ. of Maryland, College Park

September 30, 2011

Maps

- A convenient way to display spatial variations of different socio-economic and health related estimates
  - Disease mapping
  - Poverty Mapping
- Data: survey/administrative/census data
- Such maps are useful to public policymakers in planning intervention and allocation of government resources.

Design-based Direct Estimation (Cochran 1977; Rao 2003)

$$F_{\alpha i}(\mathbf{y}_i) = N_i^{-1} \sum_{k \in U_i} u_k,$$

where

$$u_k = \left(\frac{z - y_k}{z} \right)^{\alpha} I(y_k < z).$$

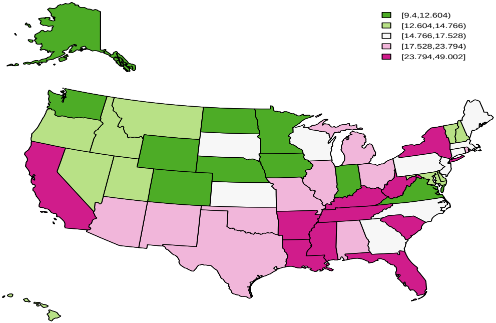
Define

- 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 $k \in s$

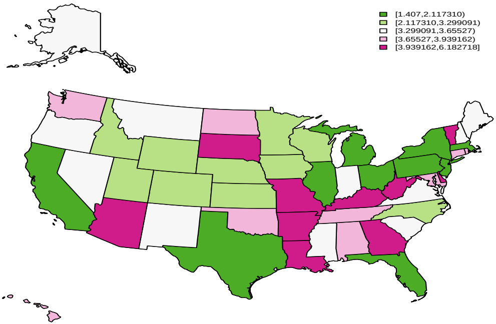
$$\hat{F}_{\alpha i}^{Dir} = \sum_{k \in s_i} w_k u_k / \sum_{k \in s_i} w_k$$

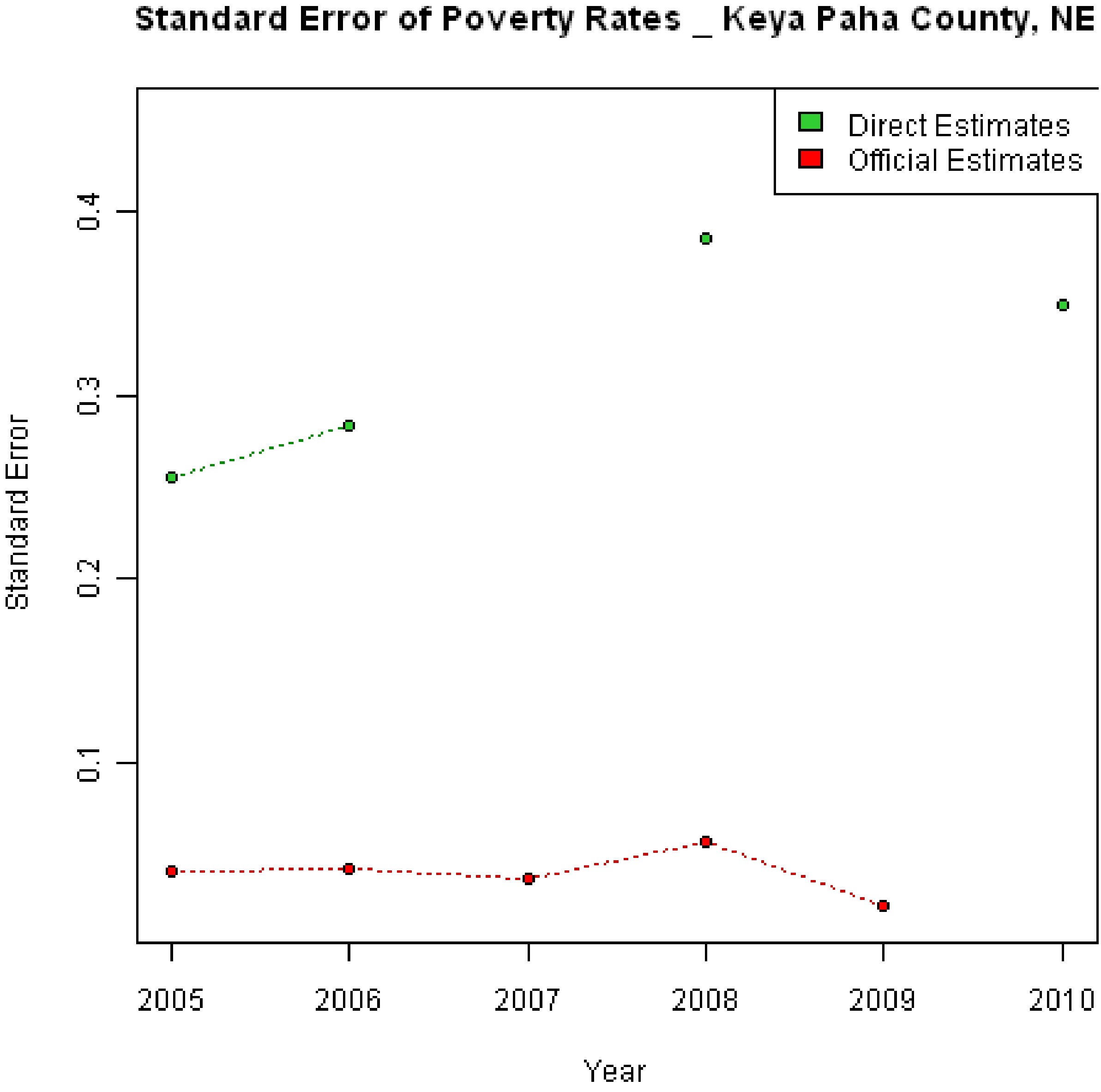
Note: The direct estimators are highly unreliable due to small sample sizes in the areas.

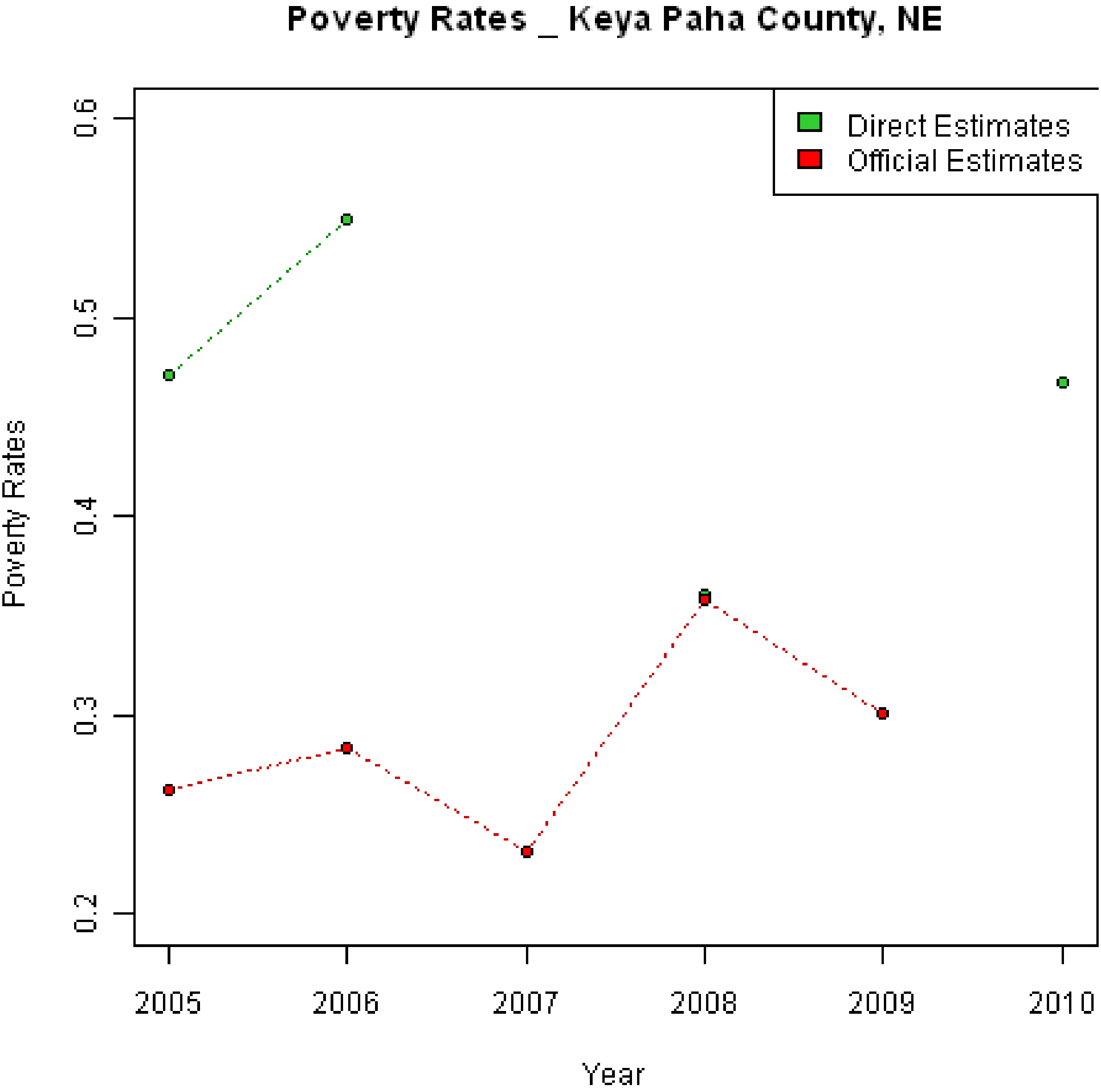
SAIPE 93’ Direct Estimate of Poverty

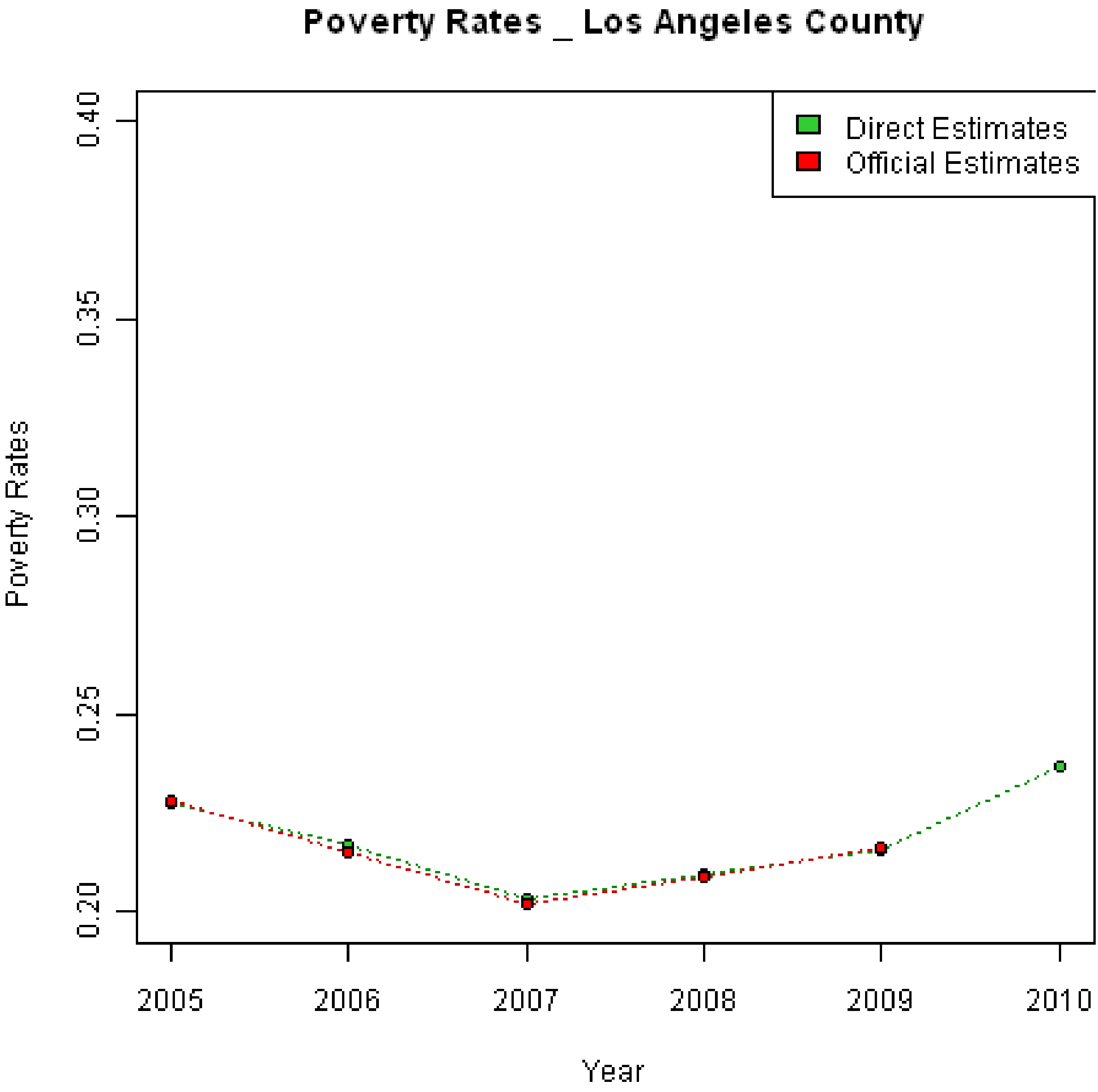


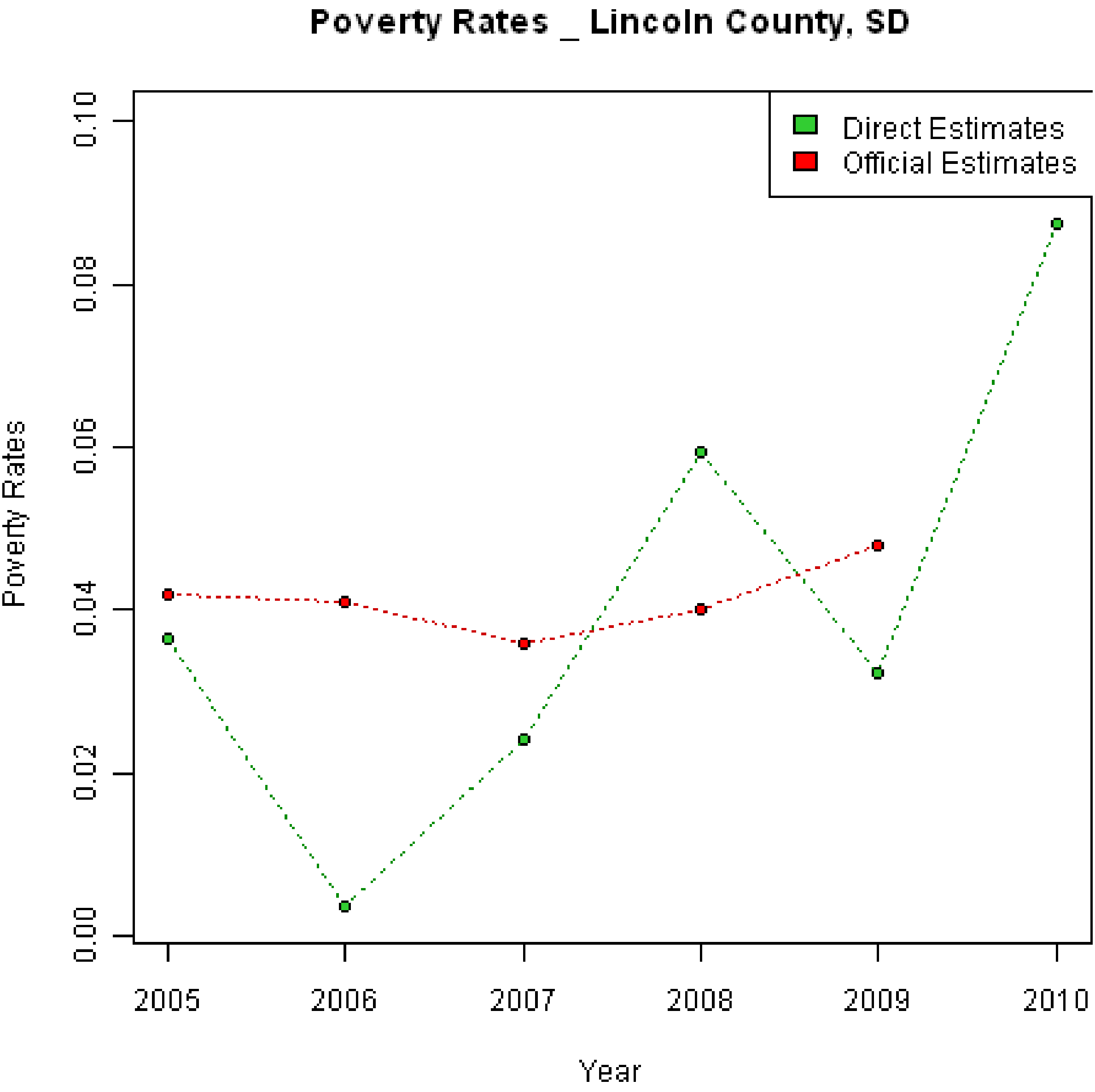
SAIPE 93' sqrt(Di) of Poverty

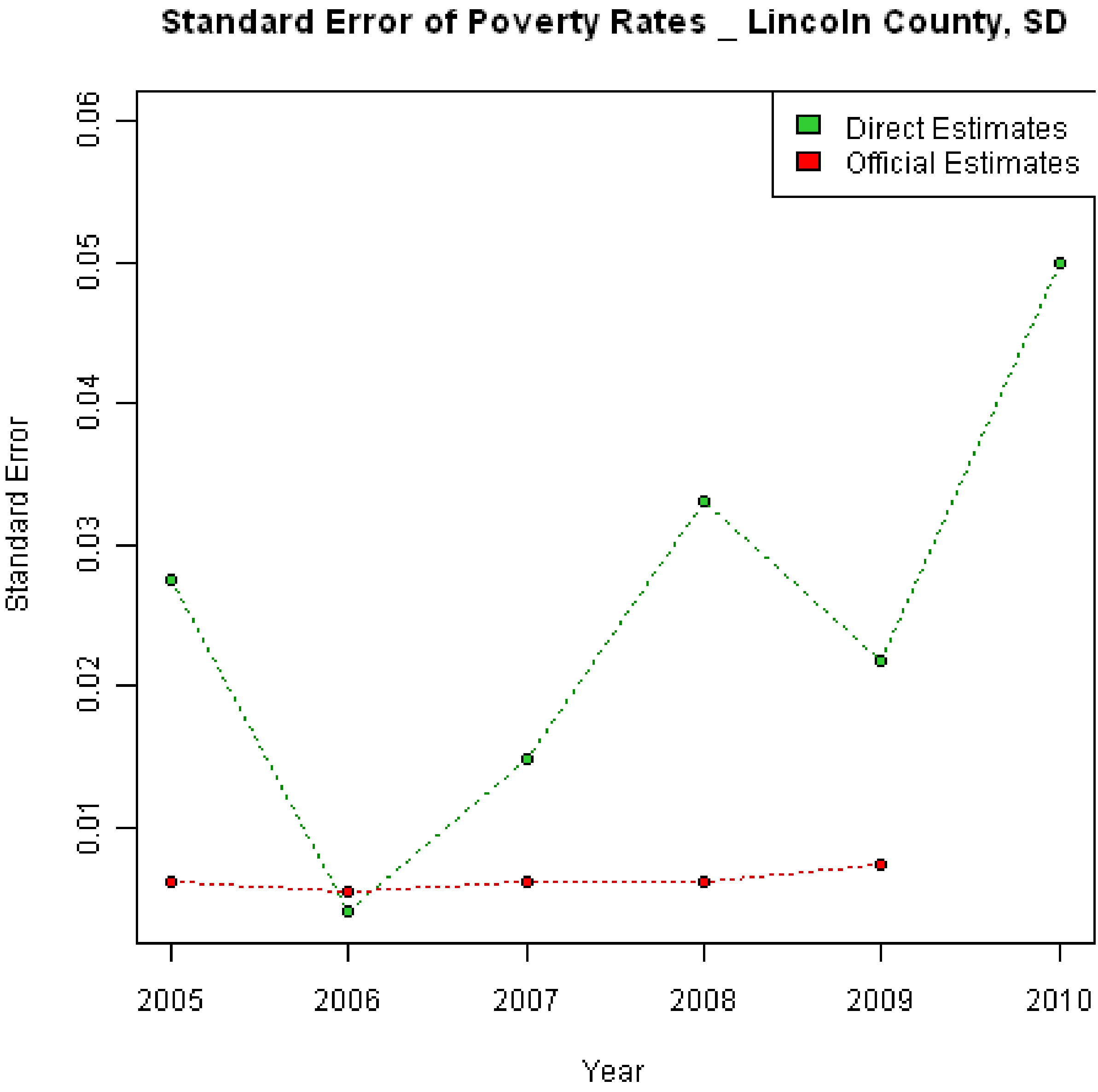



Standard Error of Poverty Rates _ Keya Paha County, NE
🟩 Direct Estimates
🟥 Official Estimates
Standard Error
0.1
0.2
0.3
0.4
2005
2006
2007
2008
2009
2010
Year


Poverty Rates _ Keya Paha County, NE
🟩 Direct Estimates
🟥 Official Estimates
Poverty Rates
0.2
0.3
0.4
0.5
0.6
2005
2006
2007
2008
2009
2010
Year


Poverty Rates _ Los Angeles County
🟩 Direct Estimates
🟥 Official Estimates
Poverty Rates
0.40
0.35
0.30
0.25
0.20
2005
2006
2007
2008
2009
2010
Year


Poverty Rates _ Lincoln County, SD
🟩 Direct Estimates
🟥 Official Estimates
Poverty Rates
0.00
0.02
0.04
0.06
0.08
0.10
Year
2005
2006
2007
2008
2009
2010


Standard Error of Poverty Rates _ Lincoln County, SD
🟩 Direct Estimates
🟥 Official Estimates
Standard Error
0.01
0.02
0.03
0.04
0.05
0.06
2005
2006
2007
2008
2009
2010
Year

Borrowing Strength

"..the client will always require more than specified at the design stage" (Fuller, 1999)

- Relevant Source of information
  - Census data
  - Administrative information
  - Related surveys
- Method of Combining Information
  - Choices of good small area models
  - Use of a good statistical methodology

Micro (unit) Unit Level Methods (Elbers, Lanjouw and Lanjouw, 2003; Molina and Rao, 2010)

Basic data requirements

- Micro level (e.g. person level) survey data containing the study variables (e.g. welfare variable of interest), auxiliary variables (e.g. demographic characteristics), and survey design variables (e,g, weights, stratification and cluster identifiers) are available.
- Auxiliary variables that are potentially related to the the welfare variable of interest are available for **each unit** in the population.

Issues to think about

- Frame imperfection could be a serious issue.
- Time gap between the census and the survey data
- The definition of the auxiliary variables between the survey and the census may be different.
- What to do if the selected auxiliary variables are missing for some units?
- "All models are wrong but some are useful" Box (1979)
- How easy is it to find a good unit level working model for a complex finite population?
- Does the sample follow the same population model? Or, should it be adjusted if the sample design is informative?

Statistical Inference

ELL Method

- The method is essentially synthetic, that is, the welfare variable is not used *directly* in the estimation method.
- 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.
- In some public policymaking, unlike the EB/HB, the ELL method may be viewed as a "fair" (not necessarily optimal) method to the public as the suggests the same method for all areas. However, it may be inferior to the EB/HB when the goal is to make best possible prediction.

ELL

- The ELL mixed model attempts to capture different features of the survey design except possibly for the survey weights. ELL does have a cluster specific random effects, but clusters may not be identical with the small areas.
- The method may not be design-consistent in general.
- ELL does not emphasize the need to incorporate area-specific auxiliary variables or area specific random effects. Even if the model incorporates all these, ELL method will be different from EB/HB. This is due to fact the data generation is done using the marginal model rather than the conditional distribution given the data.

The Molina-Rao Method

- The method and the various proposed extensions to the current methodology are promising for poverty mapping situation where the areas have some sample from the sample survey. For areas with no sample from the survey, EB/HB will be similar to the ELL method.
- Is the method design consistent?
- It may be important to incorporate survey weights and address robust methods to protect against potential outliers.
- Does the method has an in-built benchmarking property?
- Can proposed bootstrap method be used to improve on the ELL standard errors?

A few possible alternative approaches

- Synthetic method using area level model (SAIPE county level estimation using CPS)
- Traditional composite method (Haslett and Jones, 2009)
- Empirical best prediction (EBP) (hierarchical Bayes HB) method using area level model (Casas-Cordero, Herrera and others at UNDP)
- Empirical best prediction (EBP) (hierarchical Bayes HB) method using unit level model (Molina and Rao 2010)
- Nonparametric EB/HB using Dirichlet process prior (Ghosh, Lahiri, Tiwari 1989; Lahiri and Tiwari 1990)

Synthetic Estimation: Area level Model

Step 1: Fit a multiple regression model:

$$Y_i = x_i^T \beta + \epsilon_i,$$

- Y_i : direct estimator for area i;
- x_i: a vector of known auxiliary variables for area i;
- β: a vector of unknown regression coefficients;
- $\{\epsilon_i,\ i = 1, \cdots, m\}$ are uncorrelated errors with means 0 and known variances $\sigma^2 D_i$, where D_i are known, but σ^2 is unknown.

Step 2: A synthetic estimator of Y_i: $\hat{Y}_i^{Syn} = x_i^T \hat{\beta}$, where $\hat{\beta}$ is OLS or WLS of β.

Synthetic Assumption: The regression coefficients β are the same across areas.

A historical note

Estimate the median number of radio stations heard during the day for over 500 counties of the USA (small areas).

Ref: Hansen et al. (1953)
Two different survey data used:

- Mail Survey
  - large sample (1000 families/county) from an incomplete list frame
  - response rate was low (about 20%)
  - estimates x_i are biased due to non-response and incomplete coverage

- Personal Interview Survey: stratified multi-stage area frame
  - Nonresponse and coverage error properties were better than the mail survey
  - reliable estimates y_i for the 85 sampled counties were available, but no estimate can be produced for the remaining 415 counties
- Using (y_i, x_i) for the 85 sampled counties, the following fitted line (synthetic estimator) was obtained:

$$\hat{Y}_i^{Syn} = 0.52 + 0.74x_i$$

- Use y_i for the 85 sampled counties and $\hat{y}_i$ for the rest.

SAIPE

For county i

- Y_i: direct estimate of the proportion of 5-17 year old children in poverty
- x_{1i}: proportion of child exemptions reported by families in poverty on tax returns
- x_{2i}: proportion of people under age 65 not included in an income tax return
- x_{3i}: proportion of people receiving food stamps
- x_{4i}: census residual

An Area Level Model

Level 1: (Sampling Model) $\hat{\bar{Y}}_i \mid \bar{Y}_i \overset{ind}{\sim} N[\bar{Y}_i, D_i]$;
Level 2: (Linking Model) $\bar{Y}_i \overset{ind}{\sim} N[\mathbf{x}_i'\boldsymbol{\beta}, A]$.

- The hyper-parameters $\boldsymbol{\beta}$ and A are unknown,
- The sampling variances D_i are assumed to be known.
- In practice, smoothed estimates of D_i's are obtained (Bell, Otto, Hawala, Lahiri and others)
- EB: $\hat{\bar{Y}}_i^B = (1 - B_i)\hat{\bar{Y}}_i + B_i\mathbf{x}_i'\boldsymbol{\beta}$, where $B_i = D_i/(A + D_i)$.

Mean Squared Error of 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

An Example: Stratified SRS

A stratified SRS in which a SRS sample of size n is taken from each stratum (same as small area). Consider the estimation of $\bar{Y}_i$.

- Direct Estimator: $\hat{\bar{Y}}_i^D = \bar{y}_i$
- Synthetic Estimator: $\hat{\bar{Y}}_i = \bar{y}$

$$\begin{aligned} V_i^D &= \frac{S_i^2}{n} = O(n^{-1}), \;\; B_i = 0 \\ V_i &= \frac{\overline{S^2}}{nm} = O((mn)^{-1}), \;\; B_i = \bar{Y} - \bar{Y}_i, \end{aligned}$$

ignoring the fpc, where S_i^2 is the population variance for area i and $\overline{S^2} = m^{-1} \sum_{i=1}^{m} S_i^2$

Average MSE

Define

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

where

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

Naïve Estimator of Average MSE

$$\hat{M}^{\text{Naïve}} = m^{-1} \sum_{i=1}^{m} \hat{V}_i = \bar{\hat{V}}, \text{ say}$$

where $\hat{V}_i$ is computed using any standard design-based method.

We have

$$\hat{M}^{\text{Naïve}} - M \approx -\eta$$

and so this underestimates the true MSE and the extent of underestimation depends on the accuracy of the synthetic assumption.

The Gonzales-Waksberg Estimator of Average MSE

Under the assumptions $\mathrm{Cov}(\hat{\bar{Y}}_i, \hat{\bar{Y}}_i^D) \approx 0$ and $\mathrm{E}(\hat{\bar{Y}}_i^D) \approx \bar{Y}_i$, one gets

$$\mathrm{MSE}(\hat{\bar{Y}}_i) \approx E(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - V_i^D,$$

which motivates the Gonzales-Waksberg (GW) AMSE estimator:

$$\hat{M}^{\mathsf{GW}} = m^{-1} \sum_{i=1}^{m} (\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^D)^2 - m^{-1} \sum_{i=1}^{m} V_i^D.$$

- The above estimator can produce negative estimates of average MSE.

Design-based Simulation

- We follow-up on the design-based simulation of Molina and Rao (2010), but considered evaluation of the ELL measure of uncertainty relative to the GW method.
- Finite populations are generated from the following nested error regression:

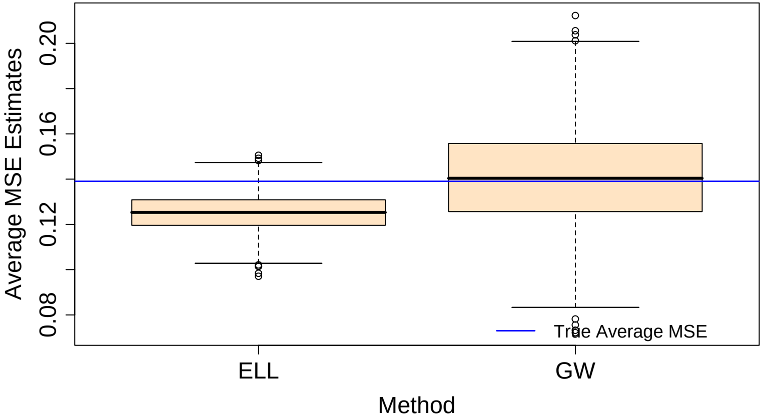
$$\log(y_k) = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + v_i + e_k,\ k \in U_i,$$

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

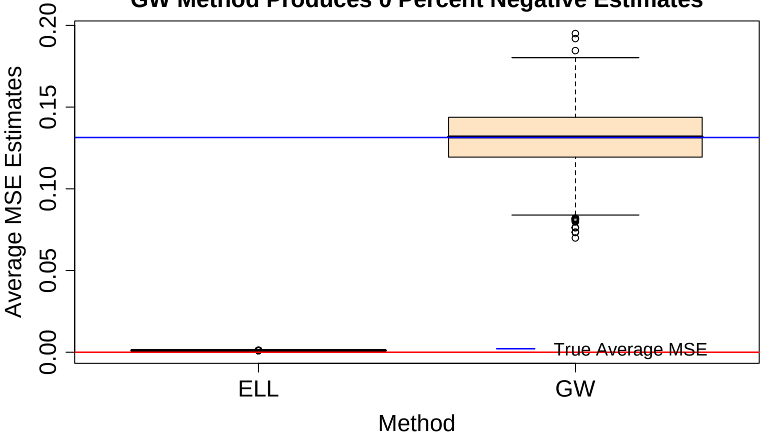
Design-based Simulation: Contd.

- $m = 40,\ N_i = 250, n_i = 3, L = 50, \beta = (3, .03, -.04)', \sigma_e^2 = 0.25,\ \sigma_v^2 = 1, R = 1000.$
- 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)

BoxPlot of Average MSE estimates of Synthetic Estimator
GW Method Produces 0 Percent Negative Estimates



BoxPlot of Average MSE estimates of Synthetic Estimator
GW Method Produces 0 Percent Negative Estimates


Average MSE Estimates
True Average MSE
ELL
GW
Method