Small Area Estimation: An Overview

Stas Kolenikov

Abt SRBI

AAPOR 2014
Motivation

In many social, behavioral or health studies, there may be interest in obtaining estimates for small subgroups of population

- National study → estimates for
  - states
  - counties
  - school districts
  - metro areas (NHIS phone use)
  - health service areas

- Statewide study → estimates for counties or cities

- City-wide study → estimates for neighborhoods

- Detailed industry by size by region classifications (some BLS work)
US official statistics

- Census Bureau: Small Area Income and Poverty Estimates
  - States, Counties, School districts
  - [http://www.census.gov/did/www/saipe/](http://www.census.gov/did/www/saipe/)

- National Cancer Institute: Small Area Estimates for Cancer Risk Factors and Screening Behaviors
  - States, Counties
  - Combines BRFSS and NHIS (Raghunathan et al. 2007)

- National Center for Health Statistics: NHIS phone use data (Blumberg et al. 2013, Stephen Blumberg’s presentation)
Small Area Income and Poverty Estimates

Map Legend
- Percent in Poverty, Ages 5-17 in Families
- 46.1 to 100.0
- 32.0 to 45.0
- 21.1 to 32.0
- 17.3 to 21.0
- 10.5 to 17.2
- 0.0 to 10.4
- No Data

Filter Data
- 0%
- 100%
- Reset

Map transparency
0% to 100%

About this application

<table>
<thead>
<tr>
<th>State</th>
<th>ID</th>
<th>Name</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>060251D</td>
<td>Anaheim Elementary School District</td>
<td>31.3</td>
</tr>
<tr>
<td>CA</td>
<td>061044D</td>
<td>Cypress Elementary School District</td>
<td>10.9</td>
</tr>
<tr>
<td>CA</td>
<td>061473D</td>
<td>Fullerton Elementary School District</td>
<td>19.4</td>
</tr>
<tr>
<td>CA</td>
<td>062814D</td>
<td>Ocean View Elementary School District</td>
<td>14.9</td>
</tr>
<tr>
<td>CA</td>
<td>064215D</td>
<td>Westminster Elementary School District</td>
<td>26.1</td>
</tr>
</tbody>
</table>

Source: U.S. Census Bureau, Small Area Income and Poverty Estimates
Some work at Abt SRBI

- 2013 National Survey of American Jews for the Pew Research Center
  - County-level estimates of Jewish population
  - Coverage decisions (dial 90% of US population, 99% Jewish population)
  - Stratification decisions
  - See also Abt SRBI news release, Pew Research Center Methods report, AAPOR 2013 presentations by Ben Phillips and Stas Kolenikov

- Small area phone usage figures
  - ACS PUMA level estimates
  - Sample design for current and future surveys
  - Weighting targets
  - See AAPOR 2013 presentation by Kolenikov and ZuWallack

- Ongoing work in health surveys conducted by Abt SRBI
  - Small area = community within a city, county within a state
Decent estimates are needed!

Given the low sample sizes for small areas (sometimes \( n = \) double digits, sometimes \( n = \) single digits, sometimes \( n = 0 \)), can any reasonable estimates be obtained? If yes, can any reasonable measures of precision be attached to these estimates?

Since survey means/proportions (direct estimates in SAE jargon) may not be available, or are of insufficient accuracy, statistical models have to be used, and weaved into complex survey estimation.
Small area estimation approaches

Statistical approaches growing from traditional survey statistics:

- Apply the ratios/proportions
- Treat as mixed/multilevel models
- Filter/reweight a “big” data set to look like a small area
- Fit models on one data set, apply on another


Data requirements

- A large data set that would allow precise estimation of the SAE model for an outcome $y$ using variables $x$

- Identifiers of the small areas (if a part of the large data set)

- The same variables $x$ for the small area as those used in the SAE model in the large data set
Groupwise ratio estimator

Applies means or proportions of the outcome in population group

1. Split large sample into groups (e.g., age-gender-education) \( g = 1, \ldots, G \)
2. Estimate the nationwide mean outcome in each group \( \bar{y}_g \)
3. Apply the small area proportions \( \gamma_g \) of the groups as weights to obtain the SAE estimate \( \sum_g \gamma_g \bar{y}_g \)

This produces a synthetic estimate, i.e., the one that is based only on aggregate data subset to the small area values.
### Group-wise ratio estimator

<table>
<thead>
<tr>
<th>Education level</th>
<th>% smoking</th>
<th>Nationwide</th>
<th>Columbia, MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than college</td>
<td>22.6%</td>
<td>69.2%</td>
<td>44.1%</td>
</tr>
<tr>
<td>Bachelor+</td>
<td>7.9%</td>
<td>30.8%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Overall</td>
<td>18.0%</td>
<td>14.4%</td>
<td></td>
</tr>
</tbody>
</table>

**Source**
- NHIS 2012
- NHIS 2012
- ACS 2012 (via FactFinder)

Note: age 25+.
## Group-wise ratio estimator

<table>
<thead>
<tr>
<th>Education level</th>
<th>% smoking</th>
<th>Nationwide</th>
<th>Columbia, MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than college</td>
<td>22.6%</td>
<td>69.2%</td>
<td>44.1%</td>
</tr>
<tr>
<td>Bachelor+</td>
<td>7.9%</td>
<td>30.8%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Overall</td>
<td>18.0%</td>
<td>14.4%</td>
<td>Direct</td>
</tr>
<tr>
<td>Source</td>
<td>NHIS 2012</td>
<td>NHIS 2012</td>
<td>ACS 2012 (via FactFinder)</td>
</tr>
</tbody>
</table>

Note: age 25+. 

**Source**: NHIS 2012, NHIS 2012, ACS 2012 (via FactFinder)
### Group-wise ratio estimator

<table>
<thead>
<tr>
<th>Education level</th>
<th>% smoking</th>
<th>Nationwide</th>
<th>Columbia, MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than college</td>
<td>22.6%</td>
<td>69.2%</td>
<td>44.1%</td>
</tr>
<tr>
<td>Bachelor+</td>
<td>7.9%</td>
<td>30.8%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Overall</td>
<td>18.0%</td>
<td></td>
<td>14.4%</td>
</tr>
</tbody>
</table>

Source: NHIS 2012  
NHIS 2012  
ACS 2012 (via FactFinder)

Note: age 25+. 

Stas Kolenikov (Abt SRBI)
Issues

- Simplistic: assumes group structure explains all variations in outcome
- Synthetic estimate only: even if there are data available for Columbia, MO, they are ignored
- No telling how accurate the (implicit) model of ratios $y \propto x$ is
Area model: SAIPE history

Fay & Herriot (1979) analyzed per capita incomes for small places in the US with population less than 1,000:

\[
\begin{align*}
\theta_i &= x_i' \beta + v_i \\
\hat{\theta}_i &= \theta_i + e_i = x_i' \beta + v_i + e_i
\end{align*}
\]  

(1)

where

- \(\theta_i = \log \bar{Y}_i\) is the log of the mean per capita income
- \(X_i\) are demographic explanatory variables
- \(v_i\) is the model error
- \(\hat{\theta}_i\) is the *direct* estimator based only on sample data on \(y_i\) (if \(n_i > 0\))
- \(e_i\) is sampling error
Area model: composite estimate

Fay-Herriot model (1) → composite estimate

\[ \tilde{\theta}_i = \gamma_i \hat{\theta}_i + (1 - \gamma_i) x'_i \tilde{\beta} \]  

(2)

\[ \gamma_i = \frac{\sigma^2_v}{\sigma^2_v + \psi_i} = \text{contribution of the direct estimate,} \]  

(3)

\[ \psi_i = \text{sampling variance of } \hat{\theta}_i, \]

\[ \tilde{\beta} = \text{estimate of } \beta \text{ from (1)} \]

That is,

composite estimate = precision weight \times direct estimate + precision weight \times synthetic estimate

Stas Kolenikov  (Abt SRBI)  Small Area Estimation: An Overview  AAPOR 2014  13 / 28
Unit models

Unit-level data on respondents in their specific areas is available → the unit model for responses:

\[ y_{ij} = x_{ij}' \beta + v_i + \epsilon_{ij} \]  

(4)

where

- \( i \) still enumerates the areas
- \( j \) enumerates individuals in areas
If sampling fractions within areas are small and the covariates are known for all units in the area, then the composite estimator is

\[ \tilde{\mu}_i = \gamma_i [\bar{y}_i + (\bar{X}_i - \bar{x}_i)\tilde{\beta}] + (1 - \gamma_i) \bar{X}_i'\tilde{\beta}, \]  

(5)

\[ \gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2/n_i} \]  

(6)
Variance estimation

In the area model framework,

\[ \mathbb{V}[\text{composite estimator } \tilde{\theta}_i] = \]

\[ + \left[ g_{2i}(\sigma_v^2, \psi_i, \mathbf{x}_i) \sim \text{sampling uncertainty due to the fact that } \beta \text{ needs to be estimated} \right] \]

\[ + \left[ g_{3i}(\sigma_v^2, \psi_i) \sim \text{sampling uncertainty due to the fact that } \sigma_v^2 \text{ needs to be estimated} \right] \]
What is being estimated is the mean squared error (MSE) rather than the variance of an estimate.

Contribution of $v_i$ is treated as a bias contribution rather than variance contribution.

Simply plugging in the parameter estimates underestimates MSE.

This is a large sample expression applicable when there are many small areas.
SAE Extensions

- Multivariate models (for several outcomes at a time; cancer SAE)
- Generalized linear models (e.g., logistic for binary outcome) (Jiang & Lahiri 2001)
- Models for quantities other than means and proportions (e.g., quantiles) (Chambers & Tzavidis 2006)
- Spatial covariance modeling for area effects
- Parametric bootstrap to account for uncertainty in variance component estimation (Lahiri 2003)
- Bayesian methods (Ghosh et al. 1998) (cancer SAE)
Simulation of non-sampled units

1. For each small area, prepare the set of calibration characteristics
2. Take a rich, large sample survey data set
3. Simulate the units in small area by either . . .
   ▶ . . . reweighting the observations in the large data set, or . . .
   ▶ . . . finding a subset of observations from the large data set via combinatorial optimization . . .

. . . so that the resulting sample matches the calibration characteristics of the small area of interest

Microdata analysis can be performed on the resulting microsample that is made to resemble the characteristics of the small area of interest.

See Williamson et al. (1998)
Two micro data sets

Blumberg et al. (2009) estimated phone usage at state level:

1. Phone usage data is available in NHIS
2. Fit the multinomial logistic model using NHIS outcome $y_i$ and NHIS demographics $x_i$
3. Microdata with both relevant demographic variables and state IDs are available through CPS
4. Take NHIS coefficients, apply the model with CPS demographics, get predicted probabilities for everybody
5. Estimate prevalences as average probabilities by state

Likewise, combine NHIS and ACS data (with PUMA geographic IDs) to estimate phone usage: Battaglia et al. (2010) for NYC neighborhoods and Kolenikov & ZuWallack (2013) for states.
What I covered today

1 Motivation
   - Federal statistics
   - Work at Abt SRBI

2 Methods
   - Ratio calculations
   - Area models
   - Unit models
   - Variance estimation
   - Extensions
   - Reweighting simulation
   - Two micro data sets

3 Discussion
SAE is the synthesis of...

Mixed models

GLM

Survey statistics

BLUP / MSE optimization

Bayesian methods

GIS

SMALL AREA ESTIMATION
Challenges

- Standard errors are always tough
- Need to match the outcomes data set with administrative data set
  - Definitions of explanatory variables may not match in different data sets
- Very detailed levels of geography may be protected due to confidentiality constraints
  - ACS only has PUMA $\approx 100,000$ people
- Methodological challenges specific to the components of SAE
  - Non-response as a survey methodology issue
  - Use of weights in multilevel models as multilevel modeling issue
  - Complicated custom code
THANK YOU!

Questions, comments, requests: kolenikovs@srbi.com


References III


