Accounting for Group Classification Error in Variance Estimates Using the American Community Survey

Matthew W. Brault

For the American Community Survey Data Users Conference
May 29, 2014

Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily that of the U.S. Census Bureau.
Motivation

- Wanted to classify occupations as “high” or “low” wage jobs.
  - Great! ACS can do that!
  - Calculate median earnings for individual occupations. Set a criteria. Define occupations.
  - But how certain am I about that classification?
  - How do I account for that uncertainty in my final estimate?
Outline

- General case
  - Naïve method
  - Sophisticated method
- Example using areas of concentrated poverty
- Simulation
General Case

- Individuals are arranged in $K$ groups.
- For each group $K = k$, a statistic $\theta_k$ is calculated with standard error $\sigma_k$.
- $\theta_k$ is compared against some critical value $\tau$ and an indicator $y_k$ is set to 1 or 0.

\[
y_k \begin{cases} 
  0, & \text{if } \theta_k \leq \tau \\ 
  1, & \text{if } \theta_k > \tau 
\end{cases}
\]
General Case
Naïve Variance

- Final estimate is \[ Y = \frac{\sum (y_k W_i)}{\sum W_i} \]
- Naïve variance

\[ \tilde{Y}_r = \frac{\sum (y_k W_{i,r})}{\sum W_{i,r}} \]

\[ \sigma^2_Y = \text{var}(Y) = \frac{\sum (\tilde{Y}_0 - \tilde{Y}_r)^2}{R(1 - \varepsilon)^2} \]
General Case Measurement Error

- The assignment of $y_k$ is a measurement error problem

\[ y = \hat{\theta} + \eta, \quad \eta \sim N(0, \sigma_\eta) \]

\[ \sigma_y^2 = \sigma_{\hat{\theta}}^2 + \sigma_\eta^2 + 2\sigma_{\hat{\theta}, \eta} \]

- $\sigma_\eta^2$ is related to the variances of $\theta_k$
General Case
Sophisticated Variance

\[ y_{k,r} \begin{cases} = 0, & \text{if } \theta_{k,r} \leq \tau \\ = 1, & \text{if } \theta_{k,r} > \tau \end{cases} \]

\[ Y_r = \frac{\sum (y_{k,r} W_{i,r})}{\sum W_{i,r}} \]

\[ \sigma_y^2 = \text{var}(Y) = \frac{\sum(Y_0 - Y_r)^2}{R(1 - \varepsilon)^2} \]
General Case
Measurement Error (Cont.)

- Variance attributed to indicator alone

\[ \delta_{k,r} = y_{k,0} - y_{k,r} \]

\[ \Delta_r = \frac{\sum (\delta_{k,r} W_{i,r})}{\sum W_{i,r}} = \bar{Y}_r - Y_r \]

\[ \sigma^2_{\eta} = \text{var}(\Delta) = \frac{\sum (\Delta_0 - \Delta_r)^2}{R(1 - \varepsilon)^2} \]
How they relate

\[ \sigma_Y^2 = \text{var}(Y) = \frac{\sum_R (Y_0 - Y_r)^2}{R(1 - \varepsilon)^2} = \frac{\sum_R (Y_0 - \bar{Y}_r + \bar{Y}_r - Y_r)^2}{R(1 - \varepsilon)^2} \]

\[ = \frac{\sum_R \left[ (Y_0 - \bar{Y}_r)^2 + (\bar{Y}_r - Y_r)^2 + 2(Y_0 - \bar{Y}_r)(\bar{Y}_r - Y_r) \right]}{R(1 - \varepsilon)^2} \]

\[ = \frac{\sum_R (Y_0 - \bar{Y}_r)^2}{R(1 - \varepsilon)^2} + \frac{\sum_R (\bar{Y}_r - Y_r)^2}{R(1 - \varepsilon)^2} + 2 \frac{\sum_R (Y_0 - \bar{Y}_r)(\bar{Y}_r - Y_r)}{R(1 - \varepsilon)^2} \]

\[ = \sigma_Y^2 + \sigma_{\eta}^2 + 2\sigma_{Y,\eta} \]
Applications

- Industries as generous providers of health insurance
- Foreign born groups (by country of birth) as “new/emerging” immigrants
- Neighborhoods as impoverished
  - Bishaw, 2011
- Etc...
Poverty Areas Example

- **Areas With Concentrated Poverty: 2006-2010**
  - ACS Brief that examines census tracts by poverty rate:
    - Category I (0-13.7%)
    - Category II (13.8%-19.9%)
    - Category III (20.0%-39.9%)
    - Category IV (40.0%-100.0%)
Methods

- 72,254 Census tracts in U.S.
- Used 2006-2010 ACS 5-year data to calculate poverty rates for tracts
  - Standard errors calculated using replicate weights.
  - 517 tracts had rates of 0 percent and 18 had rates of 100 percent
    - Standard errors calculated using ACS Production method (based on tract size and average weight in the state)
    - Replicate poverty rates simulated from SE
### Tract Poverty Rates

<table>
<thead>
<tr>
<th>Tract Group</th>
<th>Number of Tracts</th>
<th>Percentage of Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category I</td>
<td>42,383</td>
<td>61.4</td>
</tr>
<tr>
<td>Category II</td>
<td>11,574</td>
<td>16.0</td>
</tr>
<tr>
<td>Category III</td>
<td>14,823</td>
<td>19.1</td>
</tr>
<tr>
<td>Category IV</td>
<td>3,474</td>
<td>3.5</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>Category I</th>
<th>Category II</th>
<th>Category III</th>
<th>Category IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve variance ($\text{var}(Y)$)</td>
<td>0.000153</td>
<td>0.000079</td>
<td>0.000118</td>
<td>0.000025</td>
</tr>
<tr>
<td>Naïve standard error</td>
<td>0.012355</td>
<td>0.008887</td>
<td>0.010850</td>
<td>0.005017</td>
</tr>
<tr>
<td>Sophisticated variance ($\text{var}(Y)$)</td>
<td>0.058874</td>
<td>0.223126</td>
<td>0.023230</td>
<td>0.071863</td>
</tr>
<tr>
<td>Sophisticated standard error</td>
<td>0.242639</td>
<td>0.472362</td>
<td>0.152413</td>
<td>0.268073</td>
</tr>
<tr>
<td>Measurement error variance ($\text{var}(\Delta)$)</td>
<td>0.062034</td>
<td>0.225113</td>
<td>0.024111</td>
<td>0.071109</td>
</tr>
<tr>
<td>Covariance ($\text{cov}(Y, \Delta)$)</td>
<td>-0.001656</td>
<td>-0.001033</td>
<td>-0.000499</td>
<td>0.000364</td>
</tr>
<tr>
<td>Ratio of standard errors</td>
<td>19.64</td>
<td>53.15</td>
<td>14.05</td>
<td>53.43</td>
</tr>
</tbody>
</table>
Size of Standard errors

<table>
<thead>
<tr>
<th>Category</th>
<th>Naive Standard Error</th>
<th>Sophisticated Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.012</td>
<td>0.243</td>
</tr>
<tr>
<td>II</td>
<td>0.009</td>
<td>0.472</td>
</tr>
<tr>
<td>III</td>
<td>0.011</td>
<td>0.152</td>
</tr>
<tr>
<td>IV</td>
<td>0.005</td>
<td>0.268</td>
</tr>
</tbody>
</table>
## State Estimates

<table>
<thead>
<tr>
<th></th>
<th>Category I</th>
<th>Category II</th>
<th>Category III</th>
<th>Category IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median CV (Naïve)</td>
<td>0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0.014</td>
</tr>
<tr>
<td>Median CV (sophisticated)</td>
<td>0.029</td>
<td>0.118</td>
<td>0.085</td>
<td>0.190</td>
</tr>
<tr>
<td>Smallest Ratio of Standard Errors</td>
<td>12.2</td>
<td>18.7</td>
<td>12.1</td>
<td>7.1</td>
</tr>
<tr>
<td>Largest Ratio</td>
<td>37.5</td>
<td>55.0</td>
<td>34.6</td>
<td>30.2</td>
</tr>
<tr>
<td>Median Ratio</td>
<td>19.9</td>
<td>29.4</td>
<td>17.4</td>
<td>13.7</td>
</tr>
</tbody>
</table>
Simulation

- Attaching to other datasets
  - Different number of replicates
- Can’t get replicate estimates from public use data
- Use FactFinder Estimates/Standard Errors
- Simulate the Replicate Distribution
  
  - Normal distribution \( \sim N(\theta_k, \gamma \sigma_k^2) \), \( \gamma = \frac{R(1-\varepsilon)^2}{(R-1)} \)
Simulated and Replicate Based Standard Errors - States

\[ r = 0.983 \]
Conclusion

- Error can be quite large!!
- Provide greater utility to working with estimates for small domains as an aggregate
- Properly reflect the level of uncertainty associated with estimates

SAS code available in an appendix to the paper
Thank You!

Matthew W. Brault
Health and Disability Statistics Branch
Social Economic and Housing Statistics Division
matthew.w.brault@census.gov
301-763-9112