Statistical Modeling of NSCG and ACS variables

Michael D. Larsen
The George Washington University
Monday, November 4, 2013
FCSM, Washington, D.C.
Outline

A. National Survey of College Graduates and the American Community Survey
B. Overview of the project
C. Small area estimation
D. Models and Data
E. Preliminary results
Acknowledgments and Disclaimer

This work was supported by contract number YA-1323-SE-0066 with the U.S. Census Bureau. The work and ideas reported here are the responsibility of the author only and not necessarily the opinions of the U.S. Census Bureau, the NSCG, the ACS, the NSF, or the GWU.

Thanks to Benjamin Reist (Census), Cassandra Logan (Census), Michael White (Census), Stephanie Coffeey (Census), Yang Cheng (Census), John Finnamore (NSF), Stephen Cohen (NSF), and others for support and input.
National Survey of College Graduates

- Conducted by Census Bureau for NSF since 1960s
- Detailed statistics on S&E labor force
- Longitudinal survey; two-phase sampling
- Data on number and characteristics of individuals with education/employment in S&E fields
- NSCG+NSRCG+SDR = SESTAT
- DSMD of Census Bureau provides NSF statistical support
- NSCG is undergoing design/frame changes
### NSCG old design – Decennial frame

<table>
<thead>
<tr>
<th>Survey year</th>
<th>Frames</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 NSCG</td>
<td>2000 decennial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003 NSCG</td>
<td>2001 NSCG</td>
<td>2000</td>
<td>decennial</td>
<td></td>
</tr>
<tr>
<td>2006 NSCG</td>
<td>2003 NSCG</td>
<td>2001 NSCG</td>
<td>2000 decennial</td>
<td></td>
</tr>
</tbody>
</table>
NSCG new design

- Eventually, NSCG subsample from 4 (odd) ACS years (1/4 sample each)

<table>
<thead>
<tr>
<th>NSCG Survey Year</th>
<th>NSCG interview round</th>
<th>ACS source for subsample</th>
<th>ACS years unused in NSCG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
NSCG 2010 new cohort

2009 ACS was sub-sampled to add to the 2010 NSCG: *New Cohort*

- \( n = 65,195 \)
- Non institutionalized, less than 76, at least a bachelor’s degree in ACS
- NSCG 2010 new cohort has *both ACS and NSG variables*
Project overview

1. Gather documentation on NSCG and ACS design and estimation
2. Learn about the formation/use of survey weights, estimation, and variance estimation in NSCG (and ACS)
3. Investigate models for data in and between the NSCG and ACS
4. Conduct analysis on focal questions
NSCG 2010 Estimation

- **Estimation**: Use weights in estimation of totals, means, and proportions
- **Variance estimation**: 80 replicates; successive difference replication variance estimation (ACS documentation; Fay and Train 1995)
- Issues studied by White and Opsomer (2011, 2012, SRMS proceedings)
Model overview

- Statistical models relating variables to one another within and across surveys
  - ACS in year $t$, ACS in year $t+1$, ACS in year $t+2$ (aggregates, not longitudinal)
  - NSCG in year $t+1$ and NSCG in year $t+3$ (aggregate and longitudinal)
  - ACS in year $t$ and NSCG in year $t+1$ (aggregate, subsample)

<table>
<thead>
<tr>
<th>Year $t$</th>
<th>Year $t+1$</th>
<th>Year $t+2$</th>
<th>Year $t+3$</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACS</td>
<td>ACS</td>
<td>ACS</td>
<td>ACS</td>
<td>....</td>
</tr>
<tr>
<td>NSCG</td>
<td></td>
<td></td>
<td>NSCG</td>
<td></td>
</tr>
</tbody>
</table>
Types of variables

- Discrete and continuous variables
- Suggest some relationships ...
Five Analysis Topics

1. Estimation for small domains (small area estimation)
2. Updating NSCG survey weights for intermediate year ACS – does this improve estimation?
3. Estimation for NSCG variables in intermediate years when an ACS is collected but not a NSCG sample – can this provide adequate estimates between survey years?
4. Question block rotation strategies – reduce respondent burden and survey cost over time by rotating blocks of questions across time?
5. Aggregate data to form periodic estimates (as in ACS). This strategy implies a reduction in sample size and estimates every other survey year.
Topic of the present study: SAE

- The NSCG is designed to give sufficient accuracy at the national level and at the level of large regions of the country.
- There is an interest in estimation in small areas (e.g., states) and small domains (e.g., subgroups by demographics, including female/male, race/ethnicity, age, and other factors).
- Estimation methods that “borrow strength” across areas/domains could produce reductions in mean square error (MSE).
- Estimation methods that utilize information from multiple surveys (NSCG, ACS) could also produce gains in MSE.
Small “areas” of interest to NSF
Sizes in NSCG 2010 – public data

- USCAB Hispanic by Broad Occupation (12 levels; part of Primary Analysis Domains 1)
  - n=7533 (9.8% of sample Hispanic)

- USCAB AIAN/NHPI by Broad Occupation (12 levels; part of PAD 1)
  - n=317/307 (0.4% of sample each AIAN and NHPI)

- USCAB is predicted to have U.S. bachelor’s degree
Asian and White categories have larger counts
Public use data on this variable has 9 levels

<table>
<thead>
<tr>
<th></th>
<th>American Indian/Alaska Native, non-Hispanic ONLY</th>
<th>Black, non-Hispanic ONLY</th>
<th>Hispanic, any race</th>
<th>B_JOB_OCC_GRP_MAJOR_NEW2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
</tr>
<tr>
<td>Computer and mathematical scientists</td>
<td>16</td>
<td>551</td>
<td>454</td>
<td>16</td>
</tr>
<tr>
<td>Biological, agricultural and other life scientists</td>
<td>10</td>
<td>133</td>
<td>249</td>
<td>10</td>
</tr>
<tr>
<td>Physical and related scientists</td>
<td>14</td>
<td>107</td>
<td>200</td>
<td>7</td>
</tr>
<tr>
<td>Social and related scientists</td>
<td>7</td>
<td>146</td>
<td>224</td>
<td>6</td>
</tr>
<tr>
<td>Engineers</td>
<td>23</td>
<td>521</td>
<td>692</td>
<td>32</td>
</tr>
<tr>
<td>S&amp;E related occupations</td>
<td>48</td>
<td>960</td>
<td>1,160</td>
<td>61</td>
</tr>
<tr>
<td>Non-S&amp;E Occupations</td>
<td>126</td>
<td>3,244</td>
<td>3,274</td>
<td>127</td>
</tr>
<tr>
<td>Logical Skip</td>
<td>73</td>
<td>1,418</td>
<td>1,280</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>317</td>
<td>7,080</td>
<td>7,533</td>
<td>307</td>
</tr>
<tr>
<td>Non-Hispanic Native Hawaiian/Other Pacific Islander ONLY</td>
<td>16</td>
<td>126</td>
<td>6,397</td>
<td>16</td>
</tr>
<tr>
<td>Multiple Race</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
</tr>
<tr>
<td>Total</td>
<td>317</td>
<td>7,080</td>
<td>7,533</td>
<td>307</td>
</tr>
</tbody>
</table>
Small “areas” in NSCG 2010 more generally

- ACS_RACETH has 6 levels
- ACS_SEX has 2 levels
- ACS_DEMGROUP includes two age groups
- ACS_SE has two levels (S&E versus not)
- ACS_HIDEG has 3 levels (BA/BS; MA/MS; PhD)
- Fully crossed, there are $12 \times 6 \times 2 \times 2 \times 2 \times 3 = 1,728$ cells.
- Other variables?

11/4/13 M. Larsen, GWU, FCSM
Small area models for cross classified nominal variables

- **Data** are multinomial with proportion parameters
- **Prior** distribution on proportions is Dirichlet
- **Posterior** distribution for proportions is Dirichlet: means, variances, simulated values are simple to produce
- **Predictive** distribution for unknown data: data are multinomial with sample size 1: simulated cell entries are possible based on observed cell information and draws of proportions from the posterior distribution
Small area models

- **Large model**: full cross classification produces a saturated log linear model
- **Reduced models**: a log linear model with some higher order interactions set to zero produces reduced models
Issues with the SAE models

- **How to select models?** Fully saturated, reduced, etc.
- **Use of design and other variables:** Additional variables (e.g., detailed occupations crossed with demographics) were used for sampling cells. Should models be made bigger to account for this? A unit level model could use additional variables for each person in the sample.
  - If the ACS frame includes all the unit level variables, then predictions can be formed for all ACS sample members.
Issues with the SAE models

- **Use of survey weights**: A population size by cell is implied by the sum of survey weights. Posterior mean value for proportions for unobserved cases could be used in estimation. Then the weighted posterior means could be used to produce a population-based estimate of small area size.

- **Replicate survey weights**: Replicate weights could be used in place of final survey weights in this procedure; this would enable use of successive difference replication variance estimation.
Possible model extension

- For each category (small area domain or cell), one could model the propensity of being in that category – this is multinomial (polytomous) logistic regression.
- Some variables (e.g., Highest degree, Sex, Age group) would then be used as predictors of cell membership in the logistic regression models. Models could have main effects and some interactions.
- Prior distributions would be placed on model regression parameters. This produces a hierarchical polytomous logistic regression model.
Work is ongoing

- I have access to Census (Sworn status) and access to NSCG data (think took awhile)
- The initial experiences and efforts have been important in setting up for continuing work.
- The shutdown was a setback to use of data for this conference and for establishing a new contract, but efforts are proceeding.
- Work is planned for rest of fiscal year.
Conclusion and future work

- **Small area estimation**: conditions seem right for trying small area estimation – many domains, large data set but small in some places/subgroups.

- **Models**: Bayesian log linear models can be one approach to try, others can be compared.

- **Plan for near future**: continue research on SAE for NSCG 2010 using the subset of NSCG 2010 drawn from the 2009 ACS.
Thanks!

mlarsen@bsc.gwu.edu

- Rao, 2003, *Small area estimation*