Sample Design in 3-Stage Household Surveys Supplemented by Commercial Lists

Alena Maze
PhD Candidate JPSM
Household Surveys with Over- or Under-Sampling

- Household surveys often target particular demographic subgroups
  - Blacks
  - Hispanics
  - Age groups

- Different ways to obtain target sample sizes
  - Select equal probability sample of HUs, screen persons for eligibility, retain at rates to obtain sample size
  - Stratify SSUs by census or ACS data related to target groups; sample SSU strata at different rates
  - Use commercial lists with demographic info on HUs
Pros and Cons

(1) Equal probability with screening
   - Expensive if oversampling rates differ by group
   - Many HUs may be screened out and dropped

(2) SSU stratification
   - More efficient than (1) if strata directly related to target groups
   - Info is at block group level not HU

(3) Commercial lists
   - Info is at HU level
   - Only ~60% of HUs have demographic info & may be wrong
**Goals of Dissertation**

- Estimate accuracy of commercial lists for identifying households with certain characteristics (e.g. Hispanics, non-hispanic blacks, teens (15-19), females, etc.)

- Determine how to allocate two and three stage samples supplemented with commercial lists accounting for:
  - Inaccuracy of listings
  - Costs at each stage of sampling
  - Target sample sizes and CVs for estimates of subgroups
  - Stratification of SSUs by area characteristics (e.g. density of blacks, hispanics, others)
  - Stratification of HU’s by list characteristics (e.g. Race/ethnicity, ages of persons in HU, etc.)
  - Characteristics of different variables of interest
Goals of Dissertation (continued)

- Study alternative variance component estimators
  - Design-based (ANOVA)
  - Anticipated variances
  - Bayes
Previous Literature


Example from Health & Retirement Study

Example based on HRS; LBB = Late Baby Boomers

Data from screening results in National Survey of Family Growth compared to commercial list records

<table>
<thead>
<tr>
<th>Commercial list stratum</th>
<th>LBB; B</th>
<th>LBB; H</th>
<th>LBB; Other</th>
<th>Not LBB</th>
<th>Unoccupied</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LBB; no race-eth</td>
<td>0.0000</td>
<td>0.0125</td>
<td><strong>0.5322</strong></td>
<td>0.4065</td>
<td>0.0487</td>
<td>1</td>
</tr>
<tr>
<td>2 LBB; B</td>
<td>0.2213</td>
<td>0.0163</td>
<td>0.5384</td>
<td>0.1586</td>
<td>0.0654</td>
<td>1</td>
</tr>
<tr>
<td>3 LBB; H</td>
<td>0.0081</td>
<td><strong>0.2730</strong></td>
<td>0.1400</td>
<td>0.5336</td>
<td>0.0453</td>
<td>1</td>
</tr>
<tr>
<td>4 LBB; Other</td>
<td>0.0238</td>
<td>0.0101</td>
<td><strong>0.4493</strong></td>
<td>0.1400</td>
<td>0.0654</td>
<td>1</td>
</tr>
<tr>
<td>5 Has record; Not LBB</td>
<td>0.0139</td>
<td>0.0101</td>
<td>0.8691</td>
<td>0.0566</td>
<td>0.0503</td>
<td>1</td>
</tr>
<tr>
<td>6 Has record; No age info</td>
<td>0.0159</td>
<td>0.0198</td>
<td><strong>0.7995</strong></td>
<td>0.0496</td>
<td><strong>0.1152</strong></td>
<td>1</td>
</tr>
<tr>
<td>7 No record</td>
<td>0.0121</td>
<td>0.0136</td>
<td><strong>0.6933</strong></td>
<td>0.0635</td>
<td><strong>0.2175</strong></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>0.0163</td>
<td>0.0159</td>
<td>0.0883</td>
<td>0.7553</td>
<td>0.1241</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- Commercial list info accurate enough to be useful but far from perfect
- MP allocation accounts for inaccuracies in finding sampling rates
Variance of an Estimator of Total

- 3-stage sample
  
  o $m$ PSUs selected with pps with replacement
  o $\bar{n}_a$ SSUs stratified and selected ppswr within stratum $a$
  o $q_{ab}$ HUs selected by stsr with SSU stratum $a$, list stratum $b$

\[
V(\hat{t}_{pwr}) = \frac{1}{m} \frac{S^{2}_{U1(pwr)}}{t_{U}^{2}} + \frac{1}{mt_{U}^{2}} \left\{ \sum_{i \in U} \frac{1}{p_i} \sum_{a=1}^{D} \frac{S^{2}_{U2ia(pwr)}}{\bar{n}_a} + \right. \\
\sum_{i \in U} \frac{1}{p_i} \sum_{a=1}^{D} \frac{1}{\bar{n}_a} \sum_{j \in U_{ia}} \frac{1}{p_{j|ia}} \sum_{b=1}^{B} \frac{Q^{2}_{iajb}}{q_{ab}} S^{2}_{3iaj} \left. \right\} \\
\equiv \frac{B^{2}}{m} + \sum_{a=1}^{D} \frac{W^{2}_{2a}}{m\bar{n}_a} + \sum_{a=1}^{D} \sum_{b=1}^{B} \frac{W^{2}_{3ab}}{m\bar{n}_a q_{ab}}
\]
- $B^2$, $W_{2a}^2$, and $W_{3ab}^2$ are relvariance components to be estimated

- Random effects model

$$y_{iajbk} = \mu_{iajbk} + \alpha_i + \gamma_{iaj} + \varepsilon_{iajbk}$$

$$\mu_{iajbk} = x_{iajbk}^T \beta$$

$$\alpha_i \sim (0, \sigma_{\alpha}^2), \quad \gamma_{iaj} \sim (0, \sigma_{\gamma a}^2), \quad \varepsilon_{iajbk} \sim (0, \sigma_{\varepsilon}^2)$$

- Anticipated variance

Compute $EMV_{\pi}(\hat{t}_{pwr})$;

Estimate model variance components via ML, REML, Bayes
Sample Allocation is a Math Programming Problem

- Allocation problem

\[
\text{Find } \{m, n_a, q_{ab}\} \text{ to } \min \left[ V\left( \hat{t}_{pwr} \right) \right]
\]

subject to

- minimum values of \( m, n_a, q_{ab} \)

- CV constraints on subgroup estimates (e.g., Blacks, Hispanics, Others)

- \( \text{deff} (w) \leq d_{\text{max}} \) for different subgroups

Or, could minimize cost s.t. constraints on sample sizes, CVs, etc.
Data & Analysis

- NSFG or HRS screening/interview results matched to commercial list information

- Estimate
  - List accuracy
  - Variance components

- Evaluate cost of MP allocations vs.
  - Equal probability allocation + screening
  - SSU stratification only + MP allocation
  - MP allocation to list strata and no SSU stratification