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 probabillity 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.)7
    - Characteristics of different variables of interest

Goals of Dissertation (continued)

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

Previous Literature

Iannacchione, V., J. Staab, and D. Redden (2003), "Evaluating the Use of Residential Mailing Lists in a Metropolitan Household Survey," *Public Opinion Quarterly*, 67(2), 202–210.

Roth, S. B., J. Montaquila, and D. Han (2012), "The ABS Frame: Quality and Considerations," *Proceedings of the Section on Survey Research Methods*, 3779-3793.

Roth, S. B., D. Han, and J. Montaquila (2013), "The ABS Frame: Quality and Considerations," *Survey Practice*, 6(4), available at http://surveypractice.org/index.php/SurveyPractice/article/view/73/pdf.

Shook-Sa, B., D. Currivan, J. McMichael, and V. Iannacchione (2013), "Extending the Coverage of Address-Based Sampling Frames: Beyond the USPS Computerized Delivery Sequence File," *Public Opinion Quarterly*. doi:10.1093/poq/nft041.

Valliant, R., Hubbard, F., Lee, S. and Chang, C. (2014), "Efficient use of commercial lists in U.S. Household Sampling. *Journal of Survey Statistics and Methodology*," 2: 182- 209.

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

	Commercial list stratum	LBB; B	LBB; H	LBB; Other	Not LBB	Unoccupied	Total
1	LBB; no race-eth	0.0000	0.0125	**0.5322**	0.4065	0.0487	1
2	LBB; B	**0.2213**	0.0163	0.1586	0.5384	0.0654	1
3	LBB; H	0.0081	**0.2730**	0.1400	0.5336	0.0453	1
4	LBB; Other	0.0238	0.0101	**0.4493**	0.4657	0.0510	1
5	Has record; Not LBB	0.0139	0.0101	0.0566	**0.8691**	0.0503	1
6	Has record; No age info	0.0159	0.0198	0.0496	**0.7995**	**0.1152**	1
7	No record	0.0121	0.0136	0.0635	**0.6933**	**0.2175**	1
	Total	0.0163	0.0159	0.0883	0.7553	0.1241	1.000

- 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
  - m PSUs selected with *pps* with replacement
  - $\bar{n}_a$ SSUs stratified and selected *ppswr* within stratum a
  - $\bar{\bar{q}}_{ab}$ *HUs* selected by *stsrs* with SSU stratum a, list stratum b

$$V\left(\hat{t}_{pwr}\right) = \frac{1}{m}\frac{S^2_{U1(pwr)}}{t_U^2} + \frac{1}{mt_U^2}\left\{\sum_{i\in U}\frac{1}{p_i}\sum_{a=1}^{D}\frac{S^2_{U2ia(pwr)}}{\bar{n}_a} + \right.$$

$$\left. \sum_{i\in U}\frac{1}{p_i}\sum_{a=1}^{D}\frac{1}{\bar{n}_a}\sum_{j\in U_{ia}}\frac{1}{p_{j|ia}}\sum_{b=1}^{B}\frac{Q^2_{iajb}}{\bar{\bar{q}}_{ab}}S^2_{3iaj}\right\}$$

$$\equiv \frac{B^2}{m} + \sum_{a=1}^{D}\frac{W^2_{2a}}{m\bar{n}_a} + \sum_{a=1}^{D}\sum_{b=1}^{B}\frac{W^2_{3ab}}{m\bar{n}_a\bar{\bar{q}}_{ab}}$$

- B^2, W_{2a}^2, and W_{3ab}^2 are relvariance components to be estimated

- Random effects model

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

$$\mu_{iajbk} = \mathbf{x}_{iajbk}^T \boldsymbol{\beta}$$

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

- Anticipated variance

  Compute $E_M V_\pi \left(\hat{t}_{pwr}\right)$;

  Esimate model variance components via ML, REML, Bayes

Sample Allocation is a Math Programming Problem

- Allocation problem

  Find $\left\{m, \overline{n}_a, \overline{\overline{q}}_{ab}\right\}$ to $\min\left[V\left(\hat{t}_{pwr}\right)\right]$

  subject to

  - minimum values of $m, \overline{n}_a, \overline{\overline{q}}_{ab}$
  - CV constraints on subgroup estimates (e.g., Blacks, Hispanics, Others)
  - $deff\left(w\right) \leq d_{\max}$ 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