Design of Sample Surveys which Complement Observational Data to Achieve Population Coverage

Eric V. Slud, U.S. Census Bureau, CSRM
Univ. of Maryland, Math. Dept.

National Center for Health Statistics, Oct. 2016

Based on joint work with Robert Ashmead and Anindya Roy

United States Census Bureau

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are the author's and not necessarily the Census Bureau's.

Observational Versus Survey Data

$$\text{In future: observational} \left\{ \begin{array}{c} \text{administrative} \\ \text{web-search} \\ \text{web opt-in} \end{array} \right\} \text{data collection}$$

will play a large role in statistical agencies' operations

Question: if agencies continue to publish data with assessments of variability, quality and coverage, then what statistical methodology can support probability sampling and combined analysis of survey and observational data ?

Key is joint modeling of inclusion/response indicators for observational list and sample survey

Joint Inclusion Modeling not the Most Common Approach

Recent paper of Lohr & Raghunathan (2016, Statist. Sci.) surveys approaches to merging data across different sources – linkage, imputation, multiple frame methods, empirical Bayes & hierarchical small-area models

Modelling for joint inclusion not mentioned at all !

despite interest in supplementary data-collection from ‘non-traditional’ sources

Notes on Data Definitions

(a) 'Inclusion' for admin-rec list requires record-linkage; models needed for linkage errors in terms of covariates not used in linkage

(b) Frames (e.g., Census Master Address File) not error-free; Models needed for unit frame errors in terms of covariates X_i

(c) Sample, linkage indicators for units (persons or Households)

frame	**admin. list**	**sample**	**respondent**
$I_{[i \in \mathcal{U}]}$	A_i	$I_{[i \in \mathcal{S}]}$	R_i

R_i pseudorandomization defined for $i \in \mathcal{U}$, observed for $i \in \mathcal{S}$

(d) *Assume same values for covariates & outcomes observed both in admin list & survey*

Agreement between Covariates Measured in Survey or Census and Administrative List

Covariates nominally the same, different in two measuring instruments

Measurement Error modeling problem has been considered in the context of Survey & Register data by European statisticians

latent class model in book chapter by D. Oberski (2013)

Multi-trait Multi-method (latent class) models have been generalized by Oberski, Kirchner, Eckman and Kreuter (2015)to allow other features such as censoring (top-coding)

Latent class idea is generally to assume conditional independence given an unobserved discrete factor

Application of Joint Inclusion Model to Survey Design

How to design supplemental surveys if we have a strong model
$p(R_i \,|\, A_i = a,\, X_i, Y_i),\;\; a = 0, 1$?

If $\mathcal{A} \subset \mathcal{U}$, and $\mathcal{A}^c$ accessible, sample on $\mathcal{A}^c$ with inclusion probabilities $\pi_i = \pi_i(X_i)$ and estimate Y-totals by

$$\sum_{i \in \mathcal{A}} Y_i \;+\; \sum_{i \in \mathcal{A}^c \cap \mathcal{S}} R_i\, Y_i \,/\, \{\pi_i\, p(R_i \,|\, A_i = 0,\, X_i, Y_i)\}$$

or by GREG variants. Weights $w_i = 1/\pi_i(X_i)$ freely chosen: to minimize variability of $w_i/p(R_i \,|\, A_i = 0,\, X_i, Y_i)$.

Idealized Data Structure

Generally: geographic covariates X_i observable for all $i \in \mathcal{U}$

Other covariates V_i and outcomes Y_i generally observable only for Admin Rec list $\mathcal{A}$ units **or** survey/census respondents

Assume $\mathcal{U}$ covers all residential addresses, $\mathcal{A} \subset \mathcal{U}$

$$\mathcal{D} = \Big\{A_i,\, X_i,\, I_{[i\in\mathcal{S}]}\cdot(1,\, R_i),\, (A_i + (1-A_i)\, R_i \cdot I_{[i\in\mathcal{S}]})\cdot(V_i,\, Y_i)\Big\}_{i\in\mathcal{U}}$$

$A_i,\ R_i$ *dependent* given $\mathcal{S}$, and Y_i dependent on both

Initially assume no V_i is present

Joint Models for Indicators & Outcomes

(1) Missing-at-Random (MAR): R_i, Y_i (maybe also A_i) indep. given X_i as in capture-recapture (Alho 1990)

(2) NMAR variants, e.g. logistic regression for R_i on X_i, Y_i and A_i terms (as in Robins & Rotnitzky 1994)

(3) Log-linear models for categorical R_i, Y_i, A_i, X_i ; suppressed interactions as in Darroch et al. (1993) *triple system*

(4) ANOVA for Y_i in terms of A_i, R_i factors, linear in X_i [idea of Prentice et al. (2006) for outcome log-hazards]

(5) mixture model for R_i, A_i given X_i, as below

Previous multi-list Inclusion Models: Capture-Recapture & Census Coverage

Simplest case: A_i, R_i independent condtionally within poststrata

Loglinear approach: “triple system” with suppressed highest-order interactions

Dual-system logistic regression, as in 2010 Census coverage:
Alho, Mulry, Wurdeman & Kim (JASA 1993)
conditional independence given covariates
Model identifiable from captured data, but data issues compel census application to be done marginally

Specific Simulation Model

Consider scalar (continuous) X_i, 12-dim model for illustration:

(A.1)(*SRS or Poisson sampling*)

(A.2)(*Mixture propensities*) (idea from education statistics)
(A_i, R_i) dist'n mixture of indep. & degenerate $A = R = 1$:

$$P(A_i = j,\ R_i = k \,|\, X_i) \ = \ \gamma(X_i)\, I_{[j=k=1]} +$$

$$(1-\gamma(X_i))\, a(X_i)^j\, (1-a(X_i))^{1-j}\, r(X_i)^k\, (1-r(X_i))^{1-k}$$

(A.3)(*ANOVA outcome, factors* A_i, R_i)

$$Y_i \ = \ \alpha_0 + \alpha_1 X_i + (\alpha_2 + \alpha_3 X_i) \cdot R_i + (\alpha_4 + \alpha_5 X_i) \cdot A_i + \epsilon_i$$

$$\gamma(x) \equiv \gamma, \ \ a(x) = \texttt{plogis}(\theta_1 + \theta_2 x), \ \ r(x) = \texttt{plogis}(\beta_1 + \beta_2 x)$$

Simulation Objectives

- illustrate feasibility of estimation
- illustrate information due to joint model for indicators
- illustrate estimation errors based on MAR model

Begin with $N = 10^4$, $n = 500$:

Logistic regression coeff's $\underline{\theta}, \underline{\beta}$ & mixture γ:

ML scores preferred to EM which is too slow

Parameters $\underline{\alpha}$ can be estimated here by (weighted) least-squares

$$Y \sim \begin{cases} (1, X, R, RX, A, AX) & \text{on} \quad \mathcal{S} \\ (1, X, E(R|X, A=1), X\,E(R|X, A=1)) & \text{on} \quad \mathcal{A} \cap \mathcal{S}^c \end{cases}$$

Estimating Equations for Outcome Regression

Given X_i, $A_i = 1$:

$$P(R_i = 1 \,|\, A_i = 1\, X_i) = r^*(X_i) \;=\; \frac{\gamma + (1-\gamma) r(X_i),\, a(X_i)}{\gamma + (1-\gamma) a(X_i)}$$

$$Y_i \;=\; (\alpha_0 + \alpha_4) + (\alpha_1 + \alpha_5) X_i + (b_2 + b_3 X_i) r^*(Xi)$$

$$+ (b_2 + b_3 X_i) \cdot \{R_i - r^*(X_i)\} \;+\; \epsilon_i$$

Compute variance for weighted least squares using model parameters θ, γ, β fitted in the list-inclusion joint model.

Estimate outcome-coefficients combining least-squares for $\mathcal{S}$ and for $\mathcal{S}^c \cap \mathcal{A}$ data

Simulation Results

(I) Contrast between estimation accuracy based on (A_i, X_i) versus $\mathcal{D} = \{(A_i, X_i, R_i)\}$ data with n=500

Data	N	Param	θ_1	θ_2	γ	β_1	β_2
A,X	1e4	True	-0.400	2.500	0.300	0.600	1.600
		Avg	-0.982	2.985	0.405	*	*
		SD	0.448	0.356	0.091	*	*
A,R,X		Avg	-0.586	2.709	0.302	0.601	1.620
		SD	0.226	0.149	0.068	0.272	0.486
A,X	2e5	True	-0.200	1.700	0.200	0.800	1.200
		Avg	0.202	1.556	0.000	*	*
		SD	0.279	0.105	0.160	*	*
A,R,X		Avg	-0.077	1.671	0.139	0.930	1.166
		SD	0.104	0.045	0.233	0.393	0.048

Additional Results

(II) Contrast estimates & precision based on mixture model vs. MAR (conditional independence) model with same form of $p(A|X)$, $p(R|X)$, N=10,000 and n=500, with $\mathcal{D}$ data

Model	Stat	θ_1	θ_2	γ	β_1	β_2
True		-0.400	2.500	0.300	0.600	1.600
Correct	Avg	-0.586	2.709	0.302	0.601	1.620
	SD	0.226	0.149	0.068	0.272	0.486
Misspec.	Avg	-0.978	2.982	0.404	0.357	1.703
	SD	0.020	0.015	0.004	0.263	0.516

Summary & Further Research

- Advocated joint modeling of list & response indicators for admin-records/survey research

- Models similar to capture-recapture coverage estimation, but data are different than those in Census coverage estimation

- Need extensions of models & estimates to realistic data including covariates V_i observed only within samples or admin recs

Other related research problems:

(i) record-linkage strengths and accuracy in terms of covariates

(ii) research on frame accuracy in terms of covariates

References

Alho, J. (1990), *Biometrics* **46**, 623-635.

Alho, J., Mulry, M., Wurdeman, K. and Kim, J. (1993, JASA)

Darroch, J., Fienberg, S., Glonek, G. and Junker, B. (1993), *Jour. Amer. Statist. Assoc.* **88**, 1137-1148.

Lohr, S. & Raghunathan, T. (2016, Statist. Sci.)

Oberski, D. (2013) book-chapter on Latent Class Models (register/survey)

Prentice, R., Langer, R., Women's Health Initiative investigators, et al. (2006), *Amer. Jour. Epidemiol.* **163**, 589-599.

Robins, J., Rotnitzky, A. and Zhao, L-P. (1995) *Jour. Amer. Statist. Assoc.* **90**, 106-121.

Thank you !

Eric.V.Slud@census.gov , evs@math.umd.edu