Weighted Estimating Equations with Response Propensities in Terms of Covariates Observed only for Responders

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

NISS Missing Data Workshop,
November 2015

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.

Outline

1. Standard Household Survey Data Structure
   - Propensity Covariates observed at Interview
   - MAR & Conditional Independence of Y, R given X

2. Modified Estimating Equations – Alternative Forms

3. Consequences for Nonresponse Adjustment
   in Complex Surveys

Survey- or Biased- Sampling Motivation

Data $\{X_i^{(1)}, R_i, R_i \cdot (X_i^{(2)}, Y_i) \; : \; i \in S\}$

$S \subset U$ probability sample from frame U

Inclusion prob's π_i , R_i response indicator
(likely depend on both $X_i^{(1)}$, and $X_i^{(2)}$)

$X_i^{(1)}, X_i^{(2)}$ predictive (unit-level) covariates

Y_i attribute of interest with desired population mean μ_Y

Predictive covariates $\quad X_i^{(1)} \equiv \begin{pmatrix} X_i^{(11)} \\ X_i^{(12)} \end{pmatrix}, \quad X_i^{(2)} \equiv \begin{pmatrix} X_i^{(21)} \\ X_i^{(22)} \end{pmatrix}$

Known totals $\mu_{X(11)}$, $\mu_{X(21)}$ of $X_i^{(11)}$ and $X_i^{(21)}$
$X_i^{(11)}$ includes 1 (intercept)

R_i response indicator conditionally indep. of Y_i given X_i

(1) $X_i^{(1)}$ components, e.g., from paradata on modes of interim refusal in multiple contact attempts, without known means.

(2) Regression on $X_i^{(a)} = (X_i^{(11)}, X_i^{(21)})$ may leave residuals dependent on propensity predictors X_i.

(3) Cond. indep. R_i, Y_i may hold given X_i but not given $X_i^{(1)}$.
(therefore informative)

Problem Setting

Working linear outcome model $E(Y \,|\, X^{(a)}) \;=\; \beta' \, X^{(a)}$
in terms of $X^{(a)} \;=\; (X^{(11)}, \, X^{(21)})$

$$E(X^{(a)}) \;=\; \mu_a \quad \text{known}$$

Estimate mean of Y as $\widehat{\beta}' \, \mu_a$

Nonresponse adjustment via **Inverse Probability Weighting** with respect to '*propensity*'

$$p_0(X, \gamma) \;=\; P(R = 1 \,|\, X)$$

Survey analysts do not use estimating equations with such propensities; instead, do post-stratified ratio adjustment for nonresponse, followed by regression estimation.

American Community Survey Variables

Covariates:

$X^{(1)} =$ (Multi-unit, Base-Wt, URBAN, CTY, Nghbd*),
$X^{(11)} =$ (Geography down to block-group, Multi-Unit)
$X^{(2)} =$ (BLD-type, OWNER, AGE, SEX, HISP, RACE)
$X^{(21)} =$ (AGE, SEX, HISP, RACE)

* summary in *planning data base* (PDB) at block-gp level

Housing-type covariates not available in ACS before interview

- Individual ACS covariates may be missing and imputed
- unit-level covariates displace PDB covariates
- Imputations do not much affect block-group ACS covariates

Notes from Semiparametric Theory, I

I.I.D. Data $\quad (R, X^{(1)}, R \cdot (X^{(2)}, Y))$ observable

$X = (X^{(1)}, X^{(2)}), \;\; X^{(j)} = (X^{(j1)}, X^{(j2)}), \quad X^{(a)} = (X^{(11)}, X^{(21)})$

Ignore survey (biased-sampling) aspect and restrict (X, Y, R) only by joint densities satisfying

(i°) $\quad Y, R$ conditionally independent given X

(ii°) $\quad \mu_a \;=\; E(X^{(a)})$ known

(iii°) $\quad E(Y \,|\, X^{(a)}) \;=\; \beta' X^{(a)}$

(iv°) $\quad p_2(x_2|x_1) \;\equiv\; P(X^{(2)} = x_2 \,|\, X^{(1)} = x_1)$ known

Semiparametric theory (Tsiatis 2006) $\Rightarrow$ Regular Asympt. Linear Estimators of β satisfy estimating equation

$$\sum_{i=1}^{n} \frac{R_i}{p_0(X_i)}\, g(X_i^{(a)})\, (Y_i \;-\; \beta'\, X_i^{(a)}) \;=\; 0 \qquad (1)$$

In regression-type estimator

$$\hat{\mu}_Y \;=\; \hat{\beta}'\, \mu_a \;=\; n^{-1} \sum_{i=1}^{n} R_i\, Y_i/p_0(X_i) \;+\; \hat{\beta}'\, (\mu_a - \hat{\mu}^{\,\mathrm{IPW}}_{X^{(a)}})$$

is '*double-robust*' by def'n because *model-assisted design-based*.

Optimal Estimating Equation of form (1) has

$$g(X^{(a)}) = X^{(a)} \Big/ E\Big(\frac{(Y - X^{(a)\prime}\beta)^2}{p_0(X)} \Big|\, X^{(a)} \Big)$$

with

$$\text{a.var}\Big(\sqrt{n}\,(\hat{\beta}-\beta)\Big) \;=\; \Big\{\, E\Big[X^{(a)\,\otimes 2} \Big/ E\Big(\frac{(Y - X^{(a)\prime}\beta)^2}{p_0(X)} \Big|\, X^{(a)} \Big) \Big] \,\Big\}^{-1}$$

Semiparametric Theory, II

Idea of Pfeffermann and Sverchkov (1999, 2009):

to estimate μ_Y by $\hat{\beta}' \mu_a$ with $\hat{\beta}$ coefficients estimated from

$$\sum_{i \in S} \frac{w_i \, R_i}{\hat{E}_{RS}(w_i \,|\, X_i^{(a)})} X_i^{(a)} \, (Y_i - \beta' X_i^{(a)}) = 0$$

where $w_i / \hat{E}_{RS}(w_i \,|\, X_i^{(a)})$ is a *'smoothed weight'*, with cond. exp. given sample-inclusion and response.

w_i may depend on (Y_i, X_i); denominator uses (misspecified) in-sample parametric model, e.g. WLS regression of w_i on $X_i^{(a)}$.

If denom. converges in prob. to nonrandom function of $X_i^{(a)}$, at $1/\sqrt{n}$ rate, then β estimator is consistent in superpopulation if linear outcome model $E(Y_i \,|\, X_i^{(a)}) = \beta' X_i^{(a)}$ holds.

Alternative Estimating Equations for γ

For β use (1). Forms for γ include the following:

(I) (*with thanks to Z. Tan*) *When enough totals are known*

$$\dim(h(X_i^{(1)})) + \dim(X_i^{(21)}) = \dim(\gamma)$$

one general form is based on external calibration totals:

$$\sum_{i=1}^{n} h(X_i^{(1)}) \left(\frac{R_i}{p_0(X_i, \gamma)} - 1 \right) = 0 \tag{2}$$

$$\sum_{i=1}^{n} \left(X_i^{(21)} \frac{R_i}{p_0(X_i, \gamma)} - \mu_{X^{(21)}} \right) = 0 \tag{3}$$

(II) If $p_2(x_2|x_1)$ completely known, $p_0(\cdot,\gamma) = \frac{P(R=1, x_2\,|\,x_1,\gamma)}{p_2(x_2|x_1)}$.
For sufficient set of q's ,

$$\sum_{i=1}^{n} q(X_i^{(1)}) \Big(R_i \, I_{[X_i^{(2)}=x_2]} - P(R=1, x_2 \,|\, X_i^{(1)}, \gamma) \Big) = 0$$

More often, not all joint cell-values are known ('raking').

(III) Treat external calibration data as over-determining

a model $p_2(x_2|x_1,\alpha)$.

Compatibility conditions between external (α) and internal (γ) survey models: (for sufficiently large set of q, B)

$$\sum_{i=1}^{n} q(X_i^{(1)}) \Big(R_i \, I_{[X_i^{(2)} \in B]} - p_0(X_i,\gamma)\, p_2(X_i^{(2)} \in B \,|\, X_i^{(1)}, \alpha) \Big) = 0$$

External versus Current Data Model

α in $p_2(x_2 \,|\, x_1, \alpha)$ based on high-quality external data;

- *variability not always quantified*
- *estimation may also use current-survey data*

γ must use internal survey data relating X_i to R_i

"Control" information may be very highly detailed from sources such as US Pop. Estimates down to county-level demographically cross-classified (13 Age-Gp by 6 Race/Hisp by Sex), but many cells are too small to be 100% reliable, so can work with model $p(x, \alpha)$ suppressing highest-order interactions.

THEN eq'ns in **(II), (III)** can be used, to solve exactly or to minimize weighted sum of squares to estimate survey propensity parameters γ.

Survey Forms of Estimating Equations

1st step in transition: Poisson sampling, efficiency results

2nd step: "high-entropy" sampling (Hajek 1964, Tan 2014)
(includes SRS and other PPS rejective sampling)
still maintain efficiency results

General complex surveys: no likelihood-based optimality results, but apply same inverse-propensity-weighted estimating equations with survey weights.

w_i : (possibly adjusted, not yet calibrated) weights

$p_0(x, \gamma)$ d-dim logistic regression, $\dim(X^{(a)}) \ll d \ll \dim(X)$

Estimating equations

$$\sum_{i \in S} w_i \, h(X_i^{(1)}) \left(\frac{R_i}{p_0(X_i, \gamma)} - 1 \right) = 0$$

$$\sum_{i \in S} w_i \left(X_i^{(21)} \frac{R_i}{p_0(X_i, \gamma)} - \mu_{X^{(21)}} \right) = 0$$

$$\sum_{i \in S} w_i \, X_i^{(1)} \left(R_i \, I_{[X_i^{(2)} \in B_k]} - p_0(X_i, \gamma) \, p_2(X_i^{(2)} \in B_k \mid X_i^{(1)}, \alpha) \right) = 0$$

and

$$\sum_{i \in S} w_i \frac{R_i}{p_0(X_i, \gamma)} g(X_i^{(a)}) \, (Y_i - \beta' X_i^{(a)}) = 0$$

Discussion on Search for Covariates

Kreuter, Olson, Wagner et al. (2010), **Using Proxy Measures and other Correlates of Survey Outcomes to Adjust for Non-response**, **JRSSA** *highly cited*

• argue via correlations that variables highly dependent both on Survey Outcomes and Response indicator are hard to find.

• same assertion difficult to justify in large surveys if single variables can be replaced by blocks of interacting variables.

• $X^{(a)}$ outcome variables could simultaneously interact with a subset of X variables that strongly interact with block of key variables in propensity $p_0(X) = P(R=1|\, X)$

• Stronger possibility if propensity involves outcome variables.

Summary

(1) Propensities may involve covariates observed at interview; survey world does this only through poststratified regression.

(2) In IID/Poisson-sampling settings, weighted regression estimates from $\sum_{i \in S} w_i \frac{R_i \, h(X_i^{(a)})}{p_0(X_i, \hat{\gamma})} (Y_i - \beta' X_i^{(a)})$ are efficient.

(3) Weight-smoothing strategies may help but do not improve on (2) in noninformative-sampling settings.

(4) External control data can usually not supply fully cross-classified totals or stable calibrated survey weights. Must be incorporated through (α) models forced to be compatible with propensity (γ) parameter estimates.

This is a direction of further research.

References

ACS Design & Methodology (2014), Ch. 11, Weighting

Kreuter, F. et al. (2010), JRSSA

Pfeffermann, D. and Sverchkov (1999), Sankhya B

—— (2009) chapter in: **Handbook of Statistics: Survey Sampling** chapter, North-Holland

Tan, Z. (2014), *calibration ... high-entropy sampling*, Biometrika

Tsiatis, A. (2006) **Semiparametric Theory and Missing Data**, Springer

Thank you !

Eric.V.Slud@census.gov