Survey Estimating Equations Under Nonstandard MAR Models

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

M.Ghosh Conference

May 2014

Outline

- 1. Standard Household Survey Data Structure
- 2. Nonresponse Adjusted Calibrated Estimating Equation
- 3a. Propensity Covariates without National Totals
  b. Propensity Covariates observed at Interview
- 4. Modified Estimating Equation for these Cases
- 5. Directions for Further Research

Survey Sampling Motivation

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

$S \subset U$ is a probability sample drawn from frame U with known inclusion prob's π_i (may depend on $X_i^{(1)}$)

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

Y_i is unit level attribute of interest with desired population total t_Y, while totals of X_i vectors are known

R_i is a unit-response indicator, and $R_i \perp\!\!\!\perp Y_i \,|\, X_i$ (**MAR**)

For categorical X: Y, R uncorrelated within frame X-cells

Stages of Observability

Household Surveys

- $X_i^{(1)}$ geographic, neighborhood, housing-type info
- $X_i^{(2)}$ demographic, maybe economic background
- Y_i survey attribute (e.g., income, poverty, govt. program status, or whatever)

Contrast with 2-phase Sampling (in biostat)

- Y_i, $X_i^{(1)}$ disease outcome & 'cheap' measurement
- $X_i^{(2)}$ expensive accurate measurement

Standard Double-Robust Estimating Eq'n

$$\sum_i \frac{R_i}{\rho(X_i^{(1)}, \widehat{\eta})} a(X_i) \{Y_i - \mu(X_i, \beta)\} = 0$$

• X_i may contain components of both $X_i^{(1)}$, $X_i^{(2)}$

• Usual outcome model $\mu(X_i, \beta) = X_i' \beta$, $a(X_i) = X_i$.

• Propensity model $\rho(x) = P(R_i = 1 \,|\, X_i^{(1)} = x)$ fitted (e.g., by logistic regression) on same study data.

• Semiparametric theory shows this form of equation is optimal when residuals $Y_i - \mu(X_i, \beta) \perp\!\!\!\perp (X_i^{(1)}, X_i^{(2)})$.

Extensions for Household Surveys

(1) $X_i^{(1)}$ may contain additional components [e.g., from paradata, on modes of interim refusal in multistage attempts at contact] without known national totals.

(2) Regression may not include enough terms to make residuals independent of propensity predictors $X_i^{(1)}$.

(3) MAR assumption may hold with conditioning on X_i but not on $X_i^{(1)}$.

Two New Elements

(A) "Augmented" terms in survey estimating equations can improve precision when $E(Y - X_i'\beta \,|\, X_i^{(1)}) \neq 0$, e.g., when $X^{(1)}$ cannot be incorporated in regression.

(B) Estimating equation may be valid only with propensity depending on $X_i^{(2)}$: then when $p(X_i^{(1)}, X_i^{(2)})$ is known, estimate extended propensity

$$P(R = 1 \,|\, X^{(1)}, X^{(2)}) \;=\; P(R = 1 \,|\, X^{(1)}) \,\frac{P(X^{(2)} \,|\, X^{(1)}, R)}{p(X^{(2)} \,|\, X^{(1)})}$$

Augmented Estimating Equations

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

Robins, Rotnitzky, Zhao (1994) and Tsiatis (2006) advance "augmented" estimating equations in MAR cases:

$$\sum_i \frac{R_i}{\rho(X_i^{(1)}, \widehat{\eta})} a(X_i) \{Y_i - \mu(X_i, \beta)\} - \sum_i \frac{R_i - \rho(X_i^{(1)}, \widehat{\eta})}{\rho(X_i^{(1)}, \widehat{\eta})} L(X_i)$$

including outcome $(E(Y|X) = \mu(X, \beta))$ and response propensity $(P(R = 1|X^{(1)}) = \rho(X^{(1)}, \eta))$ models, via influence functions for Regular Asymptotically Linear estimators.

Augmented Estimating Eq'ns, Continued

$$\sum_i \frac{R_i}{\rho(X_i^{(1)}, \widehat{\eta})} a(X_i) \{Y_i - \mu(X_i, \beta)\} \; - \; \sum_i \frac{R_i - \rho(X_i^{(1)}, \widehat{\eta})}{\rho(X_i^{(1)}, \widehat{\eta})} L(X_i^{(1)})$$

Augmented (incomplete-case) terms help only if

$$E(a(X_i)(Y_i - \mu(X_i)) \,|\, X_i^{(1)}) \neq 0.$$

In that case, the optimal L is $E(a(X_i)(Y_i - \mu(X_i)) \,|\, X_i^{(1)})$.

Can estimate conditional expectations if $X^{(1)}$, X discrete.

Joint Distributional Calculations

We saw that estimating equations involving extended MAR conditions or augmentation terms arise in realistic survey settings. To calculate the necessary conditional probabilities, must fit models jointly for X_i variables (some $X_i^{(1)}$ and some $X_i^{(2)}$).

Natural in surveys to model R_i given $X_i^{(1)}$ and within responder-set $X_i^{(2)}$ given $(X_i^{(1)}, R_i = 1)$.

If convenient but unlikely assumption $X_i^{(2)} \perp\!\!\!\perp R_i \mid X_i^{(1)}$ holds, then propensity depends only on $X_i^{(1)}$.

With known cross-classified totals for X_i

To model many categorical variables jointly, with or without survey weights, try loglinear models with some suppressed interactions.

Such 'small-domain' models for conditionals $X_i^{(2)}$ given $X_i^{(1)}$ (within full population or responder-set) will yield extended propensity models in terms of X_i, beyond $\rho(X_i^{(1)})$.

This is a promising future direction for household-survey research.

References

Deville and Särndal (1992 JASA), *classic calibration paper*

Fuller, W. (2009) **Sampling Statistics**

Tsiatis (2006) **Semiparametric Theory and Missing Data**

Robins, Rotnitzky & Zhao (1994 JASA), *introduced augmented estimating equations for propensity & outcome models*