Propensity score adjustment under non-ignorable non-response using information from paradata

Jongho Im

Center for Survey Statistics and Methodology
Department of Statistics
Iowa State University

June 3, 2013

Joint work with Jae-Kwang Kim

Outline

1. Introduction
   - Motivation: Korean Labor Force Survey (KLFS)
   - Paradata and contact data
2. Methods
   - Existing methods: Drew and Fuller (1980), Alho (1990)
   - Proposed method: Calibration weighting method (or GMM method)
3. Simulation and Application
4. Discussion

Introduction: Korean Labor Force Survey (KLFS)

- Household survey (157,205 sample households)
  Universe=employed+unemployed+not in LF
- Face-to-face survey with 3 followups
- We still have nonresponse (9.8%) after 3 followups.

Table: 2009 KLFS data

Status	First	Second	Third	Fourth	No response
Employment	81,685	46,926	28,124	15,992	
Unemployment	1,509	948	597	352	32,350
Not in LF	57,882	32,308	19,086	10,790	
Unemp. rate	1.81%	1.98%	2.08%	2.15%	

Introduction: KLFS (Continued)

From KLFS,

- Followups are helpful to reduce nonresponse rate $(57.0\% \rightarrow 9.8\%)$.
- But we still have some nonresponses after the final contact attempt.
- Unempolyment rate increases as the contact attempt increase, which suggests for nonignorable nonresponse.
- Nonignorable nonresponse can be adjusted by using the contact information data as paradata.

Introduction: Paradata and Contact data

-Paradata

- Data about the survey process as by-product.
- Originally conceptualized as the data automatically generated as the by product of computer-assisted survey method (Couper, 1998).
- Applied to telephone surveys (call records) and mail surveys (stamps).
- Expanded to include data collection process and response process.
- Couper and Lyberg (2005), Scheuren (2005), O'Reilly (2009) and Kreuter (2010).

-Contact data

- Call records data: the time of contact (day and time), the number of contacts, callbacks or followups
- Interviewer observations data: attitude for interview

Introduction: Contact data and nonreponse error

- Adjusting of nonignorable nonresponse error with contact information
  - Later respondents are more similar to nonrespondents than early respondents.
    Ex) Drew and Fuller (1980), Alho (1990), Potthoff et al. (1993) and Biemer et al. (2012)
  - Classification of interviewees based on "attitude for interview" can be used to account response propensity.
    EX) Peress (2010)

Methods

- Our goal is to correct for nonignorable nonresponse bias with followups.
  - Two existing methods
    - Drew and Fuller (1980): works for categorical data.
    - Alho (1990): Conditional likelihood approach
  - New approach
    - Calibration weighting method with the same conditional response probability of Alho (1990).

Methods: Drew and Fuller (1980)'s method

- Multinomial Likelihood with categorical data
- $T \times K$ contigency table + one nonresponse category
- Hardcore nonresponse

$$l(\pi_{11}, \cdots, \pi_{TK}, \pi_0) = \sum_{t=1}^{T} \sum_{k=1}^{K} n_{tk} \log \pi_{tk} + n_0 \log \pi_0$$

where

- $\pi_{tk} = \gamma(1-q_k)^{t-1} q_k f_k$ for $t = 1, \cdots, T$ and $k = 1, \cdots, K$.
- $\pi_0 = (1-\gamma) + \gamma \sum_{k=1}^{K} (1-q_k)^T f_k$
- f_k: population proportion for category k $(0 < f_k < 1)$
- q_k: the condtional response probability that the unit in category k responds when the unit is sampled
- $1-\gamma$: fraction of hardcore nonresponse

Proposed method

Drew and Fuller(1980)'s method (Continued)

Contigency table with T=4 and K=3

status	First	Second	Third	Fourth	No response
k=1	$n_{11}(\pi_{11})$	$n_{21}(\pi_{21})$	$n_{31}(\pi_{31})$	$n_{41}(\pi_{41})$	
k=2	$n_{12}(\pi_{12})$	$n_{22}(\pi_{22})$	$n_{32}(\pi_{32})$	$n_{42}(\pi_{42})$	$n_0(\pi_0)$
k=3	$n_{13}(\pi_{13})$	$n_{23}(\pi_{23})$	$n_{33}(\pi_{33})$	$n_{43}(\pi_{43})$	

Decomposed No responce cell

	No response	
No response	non hardcore	hardcore
n_0	n_{01}	n_{02}

Note that $n_{02} = n \times (1 - \gamma)$.

Methods: Basic setup

- A: original sample set of size n
- T: total number of trials (First trial and T-1 follow-ups)
- A_1: set of initial respondents
- A_t: set of respondents at (t-1) followups $t = 2, \cdots, T$

$$A_1 \subset \cdots \subset A_T \subset A$$

- Response indicator function δ_{it}

$$\delta_{it} = \begin{cases} 1 & \text{if } i \in A_t; \\ 0 & \text{o/w.} \end{cases}$$

$$0 \leq \delta_{i1} \leq \cdots \leq \delta_{iT} \leq 1$$

  - δ_{iT} is a indicator of reponse/nonresponse in survey.

Methods: Basic setup (Cont'd)

- Assume logistic conditional response probability

$$p_{it} \equiv P(\delta_{it} = 1 \mid \delta_{i,t-1} = 0, y_i) = \frac{\exp(\alpha_t + \phi y_i)}{1 + \exp(\alpha_t + \phi y_i)} \tag{1}$$

  for $t = 1, \cdots, T$ with $\delta_{i0} = 0$.

- Alho (1990) considered maximizing conditional likelihood:

$$\begin{aligned} L_c(\alpha, \phi) &= \prod_{\delta_{iT}=1} \prod_{t=1}^{T} P(\delta_{it} = 1 | y_i, \delta_{i,t-1} = 0, \delta_{iT} = 1)^{\delta_{it} - \delta_{i,t-1}} \\ &= \prod_{\delta_{iT}=1} \prod_{t=1}^{T} \left(\frac{\pi_{it}}{\sum_{t=1}^{T} \pi_{it}} \right)^{\delta_{it} - \delta_{i,t-1}} \end{aligned}$$

  where $\pi_{it} = p_{it} \prod_{k=1}^{t-1} (1 - p_{ik})$.

Methods: Calibration weighting method

- Our goal is to estimate $(\hat{\alpha}, \hat{\phi})$ of the conditional probability in (1).
- We will use some calibration equations to estimate parameters rather than maximizing conditional likelihood.

First consider the case of T=2. From the set of respondents A_1 and A_2, we have

$$E\left\{\sum_{i\in A} d_i \frac{\delta_{i1}}{p_{i1}}(1, y_i)\right\} = (N, Y) \tag{2}$$

$$E\left\{\sum_{i\in A} d_i \delta_{i1}(1, y_i) + \sum_{i\in A} d_i \frac{(1-\delta_{i1})\delta_{i2}}{p_{i2}}(1, y_i)\right\} = (N, Y). \tag{3}$$

Note that $E\left[\delta_{it} \mid \delta_{i,t-1}, y_i\right] = p_{it}$ by the construction of p_{it}.

Methods

Calibration weighting method (Continued)

Combining (2) and (3),

$$\sum_{i\in A} d_i \frac{\delta_{i1}}{p_{i1}}(1, y_i) = \sum_{i\in A} d_i \delta_{i1}(1, y_i) + \sum_{i\in A} d_i \frac{(1-\delta_{i1})\delta_{i2}}{p_{i2}}(1, y_i)$$

Writing again with conditional response model in (1),

$$\sum_{i\in A} d_i \delta_{i1} \left\{1 + \exp(-\alpha_1 - \phi y_i)\right\}(1, y_i)$$

$$= \sum_{i\in A} d_i \delta_{i1}(1, y_i) + \sum_{i\in A} d_i (1-\delta_{i1})\delta_{i2} \left\{1 + \exp(-\alpha_2 - \phi y_i)\right\}(1, y_i)$$

Also, add

$$\sum_{i\in A} d_i \delta_{i1} \left\{1 + \exp(-\alpha_1 - \phi y_i)\right\} = \sum_{i\in A} d_i.$$

We have 3 equations with 3 parameters. Uniquely determine $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\phi})$.

Methods: Calibration weighting method

Now consider the case of $T \geq 2$. Given sampling weight d_i, we have

$$\sum_{i \in A} d_i \delta_{i,t-1}(1, y_i) + \sum_{i \in A} d_i (1 - \delta_{i,t-1}) \frac{\delta_{it}}{p_{it}} (1, y_i) = (N, Y) \tag{4}$$

where $t = 1, \cdots, T$ and

$$\sum_{i \in A} d_i = N \tag{5}$$

where N is the population size and $Y = \sum_{i=1}^{N} y_i$.

- We have $T + 2p + 1$ parameters with $(p + 1) \times T + 1$ equations, where $p = dim(y)$.
- May use GMM (Generalized method of moment) idea for parameter estimation.

Methods

Calibration weighting method (Continued)

Writing $\eta = (\alpha_1, \cdots, \alpha_T, \phi, Y)$, the GMM estimates $\hat{\eta}$ can be obtained by minimizing

$$Q = \hat{U}^T(\eta) \left[\hat{V} \left\{ \hat{U}(\eta) \right\} \right]^{-1} \hat{U}(\eta)$$

where $\hat{U}(\eta)$ is the system of estimating equations in (4) and (5) and $\hat{V} \left\{ \hat{U}(\eta) \right\}$ is a design-consistent variance estimator of $\hat{U}(\eta)$ for fixed value of η.

Also the GMM estimator of η has asymptotic variance estimated by

$$\hat{V}(\hat{\eta}) = \left\{ \hat{\tau} \left[\hat{V} \left\{ \hat{U}(\hat{\eta}) \right\} \right]^{-1} \hat{\tau}^T \right\}^{-1}$$

where $\hat{\tau} = \partial \hat{U}(\hat{\eta}) / \partial \eta^T$.

Simulation: Estimators

-Four estimators of μ_y,

- Full: full sample estimator assuming no nonresponse.
- Alho: Alho's estimator.
- CAL: Calibration estimator
- CK: Chang and Kott (2008)'s estimator obtained without use of followups information.

$$\hat{Y}_{CK} = \frac{1}{n} \frac{\sum_{i \in A} d_i \delta_{iT} y_i / \hat{\pi}_i}{\sum_{i \in A} d_i \delta_{iT} / \hat{\pi}_i}$$

where $\hat{\pi}$ is obtained by solving

$$\sum_{i \in A} d_i \left(\delta_{iT} / \pi_i - 1 \right) (1, x_i) = 0$$

wherex_i is an instrumental variable in $\pi_i = \{1 + \exp(-\alpha^* + \phi^* y_i)\}^{-1}$.

Simulation: Setup

Given two simulation data sets,

- Case 1: $y_i = 0.5x_i + e_i$ (Linear model)
- Case 2: $y_i = 0.5x_i^2 + e_i$ (Quadratic model)

where $x_i \sim N(1, 1)$ and $e_i \sim N(0, 1/2)$.

We assume two conditional response models with one followup (T=2),

- Model 1: Logistic model

$$p_{it} = \{1 + \exp(-\alpha_t + \phi y_i)\}^{-1}$$

with $(\alpha_1, \alpha_2, \phi) = (-1, .5, 1)$.

- Model 2: Beta model

$$p_{it} = \frac{\Gamma(\alpha_{t+1} + \phi)}{\Gamma(\alpha_{t+1})\Gamma(\phi)} z_i^{\alpha_{t+1}-1}(1 - z_i)^{\phi-1},$$

with $(\alpha_1, \alpha_2, \phi) = (1, .5, 3)$ and $z_i = y_i^2/(1 + y_i^2)$.

Simulation: Results

Table: Monte Carlo biases, variances, and mean squared errors (MSE) of the point estimates (under model 1)

Case	Estimator	Bias	Variance	MSE
Case 1 (Linear)	Full	0.0002	0.0020	0.0020
	Alho	0.0016	0.0033	0.0033
	CK	-0.0022	0.0034	0.0034
	CAL	0.0084	0.0029	0.0030
Case 2 (Quadratic)	Full	0.0018	0.0025	0.0025
	Alho	0.0025	0.0034	0.0034
	CK	0.2148	0.4701	0.5162
	CAL	-0.0003	0.0048	0.0048

- Calibration method estimates are comprable to those of Alho's method in both cases: linear and quadratic.
- Chang and Kott (2008) estimator does not work in quadratic case.

Simulation: Results (Continued)

Table: Monte Carlo biases, variances, and mean squared errors (MSE) of the point estimates under model 2 (Case 1)

Estimator	Bias	Variance	MSE
Full	0.0015	0.0020	0.0020
Alho	-0.1492	0.0030	0.0252
CK	-0.0008	0.0031	0.0031
CAL	0.0020	0.0023	0.0023

- We used the Beta response model as true model and the Logistic response model as working model.
- Alho's method is not robust because it is based on maximum likelihood approach.

Application: KLFS

Assume that the conditional response model is

$$p_{it} \equiv P(\delta_{it} = 1 \mid \delta_{i,t-1}, y_i) = \{1 + \exp(\alpha_t + \phi y_i)\}^{-1}$$

where y_i is the number of unemployment family member in ith household.

- θ_1 and θ_2 are the fraction of employment and unemploymet with respect to population total, respectively.

$$(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{328,549} \sum_{i=1}^{157,205} d_i \frac{\delta_{i1}}{p_{i1}} (x_i, y_i)$$

  where x_i is the number of employment family member in ith household.

- The unemployment rate is defined by $\theta_2/(\theta_1 + \theta_2)$.

Application: KLFS

KLFS estimates

Table: Estimated parameters for labor force in Korean LALF

Parameter	Method	Estimates	S.E($\times 10^{-4}$)
θ_1	Naive	0.5831	11.05
	Alho	0.5830	10.94
	Drew & Fuller	0.5847	10.90
	Calibration	0.5835	11.05
θ_2	Naive	0.0115	2.00
	Alho	0.0119	2.56
	Drew & Fuller	0.0119	2.46
	Calibration	0.0119	2.32

- Variance of Alho's estimator and Drew and Fuller's estimator are computed by the Jackknife method.

Discussion

- Motivated from a real survey problem.
- Contact information (ex.followups) as paradata can be used to reduce non-ignorable nonresponse error
- We proposed the calibration weighting method (or GMM) using the moment conditions obtained from conditional response model.
- In the simulation study, our proposed method shows robustness without losing the efficiency much.
- Variance estimation relatively easy (because it is a direct application of the GMM)

Thank you for your attention