Communications for Statistical Applications and Methods
2015, Vol. 22, No. 6, 655–664

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.6.655
Print ISSN 2287-7843 / Online ISSN 2383-4757

Nonresponse Adjusted Raking Ratio Estimation

Mingue Park[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

A nonresponse adjusted raking ratio estimator that consists of weighting adjustment using estimated response probability and raking procedure is often used to reduce the nonresponse bias and keep the calibration property of the estimator. We investigated asymptotic properties of nonresponse adjusted raking ratio estimator and proposed a variance estimator. A simulation study is used to examine the performance of suggested estimators.

Keywords: raking ratio estimator, logistic regression, propensity score, nonresponse, regression estimator

1. Introduction

Most surveys of human respondents entail a certain degree of nonresponse. Many studies on reducing or removing nonresponse bias have been done in survey statistics. A common way to handle unit nonresponse is weight adjustment in which the sampling weights for the respondents are adjusted so that the estimator based on the respondents only is (approximately) unbiased to the parameter of interest.

Using the framework of two-phase sampling design, weight adjustment is performed by deriving the response probability (often called propensity score). Nonresponse adjusted weights are then obtained by multiplying the inverse of estimated propensity score to the original sampling weight. Statistical models are often employed for the derivation of estimated response probability. One of the commonly used models, known as cell response model, assumes independent and identical response distribution for every elements in the same cell. The other popular model is logistic regression model that relates the binary response variable to the set of explanatory variables. The cell response model can be viewed as a logistic regression model with categorical explanatory variables that define cells. Results on the weight adjustment under the two-phase sampling framework were reviewed by Särndal and Lundström (2005) and Särndal *et al.* (1992). Kim and Kim (2007) gave the asymptotic results of weight adjusted estimator and proposed a possible variance estimator. Ekholm and Laaksonen (1991) is a practical example of this method.

Calibration estimators introduced by Deville and Särndal (1992) are often used to improve the efficiency of the estimator or to handle nonresponse when auxiliary information on the population (such as population total of auxiliary variables) is available. One typical example of such estimators is the raking ratio method in which the marginal distributions of several auxiliary categorical variables were used as auxiliary information. Raking ratio method, originally introduced by Deming and Stephan

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2006363).

[1] Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: mpark2@korea.ac.kr

Published 30 November 2015 / journal homepage: http://csam.or.kr

(1940) is often used to ensure that the estimator of the marginal distribution of auxiliary variables is equivalent to the known population distribution. The asymptotic results and small sample properties of the raking ratio estimator were investigated by Deville and Särndal (1992) and Deville *et al.* (1993).

In practice, a set of adjusted weights is often defined through two steps. In the first step, response probability, is estimated from a logistic regression model using appropriate set of explanatory variables. Then the final set of adjusted weights is obtained through raking ratio method or post stratification. Nonresponse adjusted raking ratio weights for both household and individuals were derived for the analysis of California Health Interviewing Survey (2011), that is a combined landline and cell telephone survey. Other examples in which the final weights were derived through two steps, nonresponse adjustment followed by raking, are National Health and Nutrition Examination Survey (NHANES) of USA (1996) and National Health and Nutrition Survey in South Korea (2010).

In our study, we derive the asymptotic properties of the nonresponse adjusted raking ratio estimator and suggest an explicit form of asymptotically unbiased variance estimator using the result of Deville and Särndal (1992) and Kim and Kim (2007). We also performed a small simulation study to investigate the properties of the nonresponse adjusted raking ratio estimator and variance estimator.

2. Nonresponse Adjusted Raking Ratio Estimator

Consider the finite population $U = \{1, 2, \ldots, N\}$ where the population size N is known. The parameter of interest is the population mean of the variable of interest, y, denoted by $\bar{y}_N = N^{-1} \sum_{i \in U} y_i$. If all sampled elements were observed, we consider the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of the form, $\bar{y}_{HT} = N^{-1} \sum_{i \in A} \pi_i^{-1} y_i$, as a design unbiased estimator of the population mean where $A \in U$ is the set of indices in the sample and $\pi_i = \Pr\{i \in A\}$.

We define the response indicator variable of the unit i by

$$R_i = \begin{cases} 1, & \text{if unit } i \text{ responds,} \\ 0, & \text{if unit } i \text{ does not respond,} \end{cases} \tag{2.1}$$

for $i \in A$ and define $0 < p_{i|A} = \Pr\{R_i = 1 | i \in A\}$ as the response probability of sampled unit i. We assume R_i's are independent with $\mathrm{Var}(R_i|A) = p_{i|A}(1 - p_{i|A})$. By using the estimated response probability $\hat{p}_{i|A}$ based on the logistic regression model,

$$\mathrm{logit}\,(p_{i|A}|A) = \mathbf{z}_i' \boldsymbol{\alpha}_A, \tag{2.2}$$

we define the nonresponse adjusted estimator

$$\bar{y}_{NA} = N^{-1} \sum_{i \in A} \pi_i^{-1}\ \hat{p}_{i|A}^{-1}\ R_i\ y_i. \tag{2.3}$$

The estimator of (2.3) is also called nonresponse weights adjustment (NWA) estimator, see Rosenbaum (1987).

Assume there exist P-variables of which population marginal distributions are known and they are used for defining raking ratio estimator. Let $\mathbf{x}_i = (\mathbf{x}_{1i}, \ldots, \mathbf{x}_{pi}, \ldots, \mathbf{x}_{Pi})$ and $\mathbf{x}_{pi} = (x_{1,pi}, \ldots, x_{d_p,pi})$ be a set of indicators such that $x_{j,pi} = 1$ if element i is in j^{th} category of p^{th} categorical variable and $x_{j,pi} = 0$ otherwise, where $d_p + 1$ is the number of categories defined by $\mathbf{x}_p$. Define the nonresponse adjusted raking ratio estimator by

$$\bar{y}_{NARR} = N^{-1} \sum_{i \in A} \pi_i^{-1}\ \hat{p}_{i|A}^{-1}\ \exp\left(\mathbf{x}_i' \hat{\boldsymbol{\lambda}}\right)\ R_i\ y_i, \tag{2.4}$$

where $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\alpha}}$ are the solution to

$$N^{-1}\sum_{i\in A}\pi_i^{-1}\,\hat{p}_{i|A}^{-1}\,\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\lambda}}\right)\,R_i\,\mathbf{x}_i=\bar{\mathbf{x}}_N, \tag{2.5}$$

and

$$\sum_{i\in A}\pi_i^{-1}\left\{R_i-\left[1+\exp\left(-\mathbf{z}_i'\hat{\boldsymbol{\alpha}}_A\right)\right]\right\}\mathbf{z}_i=\mathbf{0}, \tag{2.6}$$

respectively, and $\hat{p}_{i|A}=[1+\exp(-\mathbf{z}_i'\hat{\boldsymbol{\alpha}}_A)]^{-1}$. We assume that there exists a unique solution to (2.5) and (2.6) and the solution $\hat{\boldsymbol{\alpha}}_A$ of (2.6) is a consistent estimator with respect to the response mechanism. Note that same variables could be used for both $\mathbf{x}_i$ and $\mathbf{z}_i$. Equation (2.5) can be understood as a calibration equation so that the nonresponse adjusted raking ratio estimators of x-variables are equivalent to known population values. Equation (2.6) provides the (weighted) maximum likelihood estimator of $\boldsymbol{\alpha}_A$.

To investigate the asymptotic properties of the $\bar{y}_{NARR}$, we consider a sequence of populations, samples and sampling designs assumed by Isaki and Fuller (1982). Assume population size, $N_n(>n)$ increases as n increases. Assume $\lim_{n\to\infty}\bar{\boldsymbol{\omega}}_N$ exists, where $\bar{\boldsymbol{\omega}}_N=N^{-1}\sum_{i=1}^N\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_i=(1,\mathbf{x}_i',y_i)'$. Also assume $\boldsymbol{\omega}_i$ has the finite fourth moments and its sample moments converge to the population moments such that

$$\left(N^{-1}\sum_{i\in A}\pi_i^{-1}\boldsymbol{\omega}_i\boldsymbol{\omega}_i'-N^{-1}\sum_{i=1}^N\boldsymbol{\omega}_i\boldsymbol{\omega}_i'\right)\Big|\mathcal{F}_N=O_p\left(n^{-\frac{1}{2}}\right), \tag{2.7}$$

where $\mathcal{F}_N=(\boldsymbol{\omega}_1',\ldots,\boldsymbol{\omega}_N')'$ is the finite population. The assumption (2.7) means that mean and variance of Horvitz-Thompson estimator are well defined and converge to the corresponding population parameter with the order $n^{-1/2}$ in probability. In most of sampling designs, Horvitz-Thompson estimator of the mean is unbiased and also has the finite variance of order n^{-1} and thus assumption, (2.7) is usually satisfied.

Result 1. *Under the assumptions on response variable R_i of (2.1), logistic regression model of (2.2) and the assumption (2.7) on the sequence of populations and samples,*

$$\bar{y}_{NARR}-\bar{y}_N=N^{-1}\sum_{i\in A}\pi_i^{-1}\left\{p_{i|A}\mathbf{z}_i'\boldsymbol{\delta}_N+p_{i|A}^{-1}R_i\left(e_i-p_{i|A}\mathbf{z}_i'\boldsymbol{\delta}_N\right)\right\}+O_p\left(n^{-1}\right), \tag{2.8}$$

where $e_i=y_i-\mathbf{x}_i'\boldsymbol{\beta}_N$,

$$\boldsymbol{\delta}_N=\left[\sum_{i\in U}p_{i|A}\left(1-p_{i|A}\right)\mathbf{z}_i\mathbf{z}_i'\right]^{-1}\sum_{i\in U}\left(1-p_{i|A}\right)\mathbf{z}_ie_i,$$

and

$$\boldsymbol{\beta}_N=\left(\sum_{i=1}^N\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^N\mathbf{x}_iy_i.$$

The brief proof of the result is given in Appendix. Using the result of (2.8), we could derive an asymptotic variance of the nonresponse adjusted raking ratio estimator as below.

$$V\left\{\bar{y}_{NARR}-\bar{y}_N|\mathcal{F}_N\right\}=E\left[V\left\{\bar{y}_{NARR}-\bar{y}_N|A_N,\mathcal{F}_N\right\}|\mathcal{F}_N\right]+V\left[E\left\{\bar{y}_{NARR}-\bar{y}_N|A_N,\mathcal{F}_N\right\}|\mathcal{F}_N\right], \tag{2.9}$$

where

$$E\left[V\left\{\bar{y}_{NARR}-\bar{y}_N|A_N,\mathcal{F}_N\right\}|\mathcal{F}_N\right]\approx E\left[N^{-2}\sum_{i\in A}\pi_i^{-2}p_{i|A}^{-1}\left(1-p_{i|A}\right)\left(e_i-p_{i|A}\mathbf{z}_i'\boldsymbol{\delta}_N\right)^2\Big|\mathcal{F}_N\right],$$

$$V\left[E\left\{\bar{y}_{NARR}-\bar{y}_N|A_N,\mathcal{F}_N\right\}|\mathcal{F}_N\right]\approx V\left(N^{-1}\sum_{i\in A}\pi_i^{-1}e_i\Big|\mathcal{F}_N\right),$$

A_N is the set of indices in the sample selected from the population U_N, and U_N is the N^{th} population of size N in the sequence.

The first term of (2.9) is the design expectation of conditional variance of the estimator conditioning on population and sample. The second term is the design variance of the Horvitz-Thompson estimator of the population mean of residuals.

For the definition of an asymptotically unbiased variance estimator of $\bar{y}_{NARR}$, assume that there exists a unbiased variance estimator under the full responses as a form of

$$\hat{V}\left(\bar{y}_{HT}\right)=\sum_{i\in A}\sum_{j\in A}\Omega_{ij}y_iy_j. \tag{2.10}$$

As an example of (2.10), we could consider Horvitz-Thompson variance estimator or Sen-Yates-Grundy variance estimator. For the definition of Horvitz-Thompson variance estimator and Sen-Yates-Grundy variance estimator, see Särndal *et al.* (1992).

One possible variance estimator of $\bar{y}_{NARR}$ is

$$\hat{V}=\hat{V}_{res}+\hat{V}_{sam}, \tag{2.11}$$

where

$$\begin{aligned}
\hat{V}_{sam}&=\sum_{i\in A_R}\Omega_{ii}\hat{p}_i^{-1}\left(g_i\hat{e}_i\right)^2+\sum_{i\neq j}\sum_{i,j\in A_R}\Omega_{ij}\hat{p}_i^{-1}\hat{p}_j^{-1}\left(g_i\hat{e}_i\right)\left(g_j\hat{e}_j\right),\\
\hat{V}_{res}&=N^{-2}\sum_{i\in A_R}\pi_i^{-2}\hat{p}_i^{-2}\left(1-\hat{p}_i\right)\left(\hat{e}_i-\hat{p}_i\mathbf{z}_i'\hat{\boldsymbol{\gamma}}\right)^2,\\
\hat{e}_i&=y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{NA},\\
g_i&=\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\lambda}}\right),\\
\hat{\boldsymbol{\gamma}}&=\left(\sum_{A_R}\pi_i^{-1}\hat{p}_i^{-1}\left(1-\hat{p}_i\right)\mathbf{z}_i\mathbf{z}_i'\right)^{-1}\left(\sum_{A_R}\pi_i^{-1}\left(1-\hat{p}_i\right)\hat{e}_i'\right),
\end{aligned}$$

where Ω_{ij} is the term that are used to define an unbiased variance estimator of $\bar{y}_{HT}$ in (2.10). The variance estimator is obtained by estimating each term of (2.9) unbiasedly. For the definition of the variance estimator, we considered unbiased estimator with respect to design and response mechanism. We also use g-weight suggested by Särndal, Swensson and Wretman (1989) for the variance estimator of (2.11).

3. Simulation Study

A simulation study was performed to investigate the properties of the nonresponse adjusted raking ratio estimator and its variance estimator investigated at Section 2. For the simulation study, 18 different stratified populations of size 10,000 were generated from multivariate normal distribution

$$\begin{pmatrix} x_{hi} \\ z_{hi} \\ y_{hi} \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ & 1 & \rho_3 \\ & & 1 \end{pmatrix}\right], \quad h = 1,2,3,4,\ i = 1,\ldots,N_h, \tag{3.1}$$

where N_h = 1,000, 2,000, 3,000, 4,000 for $h = 1,2,3,4$ respectively. For ρ_1, we considered two levels 0.0 and 0.3, and three levels of (0.0, 0.3, 0.6) are considered for ρ_2 and ρ_3. We also considered the variable w_{hi} generated from χ^2 distribution with 2 degrees of freedom.

The variable z generated the response probability using the logistic regression model,

$$R_i|A \sim \text{Bernoulli}(p_{i|A}), \qquad \text{logit}(p_{i|A}) = -1 + z_i. \tag{3.2}$$

Variable x and w were used to define membership variables that were used to calculate raking ratio estimator. The vector of indicator variables defined is

$$\mathbf{I}_i = \left(\mathbf{I}'_{i1}, \mathbf{I}'_{i2}\right)', \tag{3.3}$$

where

$$\mathbf{I}'_{i1} = \begin{cases} (1,0,0,0), & \text{if } x_i < q^N_{0.25}, \\ (0,1,0,0), & \text{if } q^N_{0.25} \le x_i < q^N_{0.50}, \\ (0,0,1,0), & \text{if } q^N_{0.50} \le x_i < q^N_{0.75}, \\ (0,0,0,1), & \text{if } q^N_{0.75} \le x_i, \end{cases}, \qquad \mathbf{I}'_{i2} = \begin{cases} (1,0,0), & \text{if } w_i < q^{\chi^2}_{0.25}, \\ (0,1,0), & \text{if } q^{\chi^2}_{0.25} \le w_i < q^{\chi^2}_{0.50}, \\ (0,0,1), & \text{if } q^{\chi^2}_{0.50} \le w_i < q^{\chi^2}_{0.75}, \end{cases}$$

and q^N_p and $q^{\chi^2}_p$ denote the p^{th} percentiles of the normal distribution with mean 2 and variance 1 and χ^2 distribution with 2 degree of freedom, respectively.

From each generated finite population, two sets of stratified random samples of size $n = 400$ and $n = 800$ were selected and the stratum sample size were equal as $n_h = n/4$ for $h = 1,2,3,4$. We assumed the value of z_i is known for every element in the sample but the vector of membership indicators and variables of interest are known only for the responding element, that is, $R_i = 1$. We also assume that population total (or mean) of $\mathbf{I}_i$ is known so that we could calculate the raking ratio estimator. The Monte Carlo sample sizes are all 20,000 and average response rate is about 0.70. As an estimator for the population mean, we considered $\bar{y}_{NA}$ of (2.3) and $\bar{y}_{NARR}$ of (2.4) with $\mathbf{x}_i = \mathbf{I}_i$ of (3.3).

Table 1 shows the Monte Carlo relative bias in percent and variance of the two estimators. Monte Carlo relative bias of $\hat{\theta}$, that is an estimator of θ, is defined as $\theta^{-1}[E_{MC}(\hat{\theta}) - \theta]$, where $E_{MC}(\hat{\theta})$ is the Monte Carlo mean of $\hat{\theta}$ and θ is the parameter. With a sample of size 400 and 800, both estimators have negligible bias with less than 1% in absolute. Variance of $\bar{y}_{NA}$ gets smaller as the correlation between z and y gets stronger as expected. The variance of $\bar{y}_{NARR}$ decreases as the correlation between x and y increases because the raking ratio estimator is asymptotically equivalent to the regression estimator with the regression estimator efficient when the auxiliary variables are correlated with the variable of interest. Across all samples, the average reduction of the variance obtained by using $\bar{y}_{NARR}$ is about 13% with maximum 36%.

Table 1: Monte Carlo relative bias and variance of estimators

Sample size	ρ_1	ρ_2	ρ_3	Relative Bias in %		Variance $\times 10^3$	
				$\bar{y}_{NA}$	$\bar{y}_{NARR}$	$\bar{y}_{NA}$	$\bar{y}_{NARR}$
$n = 400$	0.0	0.0	0.0	−0.293	−0.284	4.940	4.738
			0.3	0.058	0.407	4.616	4.665
			0.6	−0.522	0.077	4.263	4.512
		0.3	0.0	0.502	0.435	5.058	4.268
			0.3	0.121	0.010	4.538	4.236
			0.6	0.591	0.826	3.922	3.847
		0.6	0.0	0.336	0.498	5.363	3.272
			0.3	−0.371	−0.031	4.610	3.249
			0.6	−0.190	0.154	4.096	2.804
	0.3	0.0	0.0	0.784	0.799	4.822	4.565
			0.3	−0.419	−0.381	4.429	4.409
			0.6	0.516	0.648	4.161	4.213
		0.3	0.0	−0.534	−0.589	4.911	4.266
			0.3	0.085	0.226	4.486	4.211
			0.6	−0.581	−0.115	4.137	4.097
		0.6	0.0	0.107	0.192	5.021	3.227
			0.3	−0.168	0.041	4.493	3.187
			0.6	−0.162	0.257	4.131	3.154
$n = 800$	0.0	0.0	0.0	−0.352	−0.341	2.323	2.256
			0.3	0.691	0.592	2.252	2.261
			0.6	0.136	0.330	1.966	2.064
		0.3	0.0	0.521	0.434	2.419	2.050
			0.3	0.399	0.559	2.039	1.878
			0.6	0.247	0.261	1.965	1.979
		0.6	0.0	0.361	0.526	2.537	1.588
			0.3	−0.048	0.035	2.193	1.483
			0.6	0.268	0.402	1.978	1.409
	0.3	0.0	0.0	0.784	0.802	2.323	2.197
			0.3	−0.436	−0.592	2.221	2.214
			0.6	0.356	0.400	1.884	1.990
		0.3	0.0	−0.493	−0.541	2.300	2.006
			0.3	0.272	0.159	2.300	2.065
			0.6	0.514	0.537	1.830	1.814
		0.6	0.0	0.130	0.244	2.315	1.515
			0.3	−0.087	−0.094	2.239	1.543
			0.6	0.449	0.581	1.965	1.532

We also calculate the variance estimator of $\bar{y}_{NARR}$ given in (2.11). Table 2 shows the Monte Carlo properties of $\hat{V}_{sam}$, $\hat{V}_{res}$ and $\hat{V}$. The relative bias of the variance estimator in percent is

$$\left[V_{MC}(\bar{y}_{NARR})\right]^{-1}\left[E_{MC}\left(\hat{V}\right) - V_{MC}\left(\bar{y}_{NARR}\right)\right].$$

For the samples of size 400, suggested variance estimator underestimates the true variance in all population. Relative bias of the variance estimator is about −4% to −7%. Underestimation of the variance estimator with sample size 400 is mainly due to the Taylor approximation in which the second and higher order term of the approximation are not considered for the variance estimation. The absolute relative bias of the suggested variance estimator with a sample of size 800 is less than 2.2%. For both sample sizes, the component of the variance estimator, $\hat{V}_{res}$ makes up about 34% to 36% of the $\hat{V}$. Thus if only $\hat{V}_{sam}$ is used to estimate the variance of $\bar{y}_{NARR}$, which is common in practice, it is expected to underestimate the true variance severely.

Table 2: Monte Carlo properties of variance estimators

Sample size	ρ_1	ρ_2	ρ_3	Monte Carlo mean			Relative bias of $\hat{V}$ (%)	$V_{MC}(\hat{V})$ $\times 10^6$
				$\hat{V}_{sam}$	$\hat{V}_{res}$	$\hat{V}$		
$n = 400$	0.0	0.0	0.0	2.863	1.669	4.531	−4.358	0.624
			0.3	2.943	1.509	4.452	−4.570	0.438
			0.6	2.920	1.374	4.294	−4.835	0.700
		0.3	0.0	2.576	1.487	4.062	−4.821	0.398
			0.3	2.676	1.359	4.035	−4.747	0.341
			0.6	2.590	1.111	3.701	−3.788	0.433
		0.6	0.0	1.948	1.149	3.097	−5.343	0.328
			0.3	2.019	1.067	3.087	−4.999	0.216
			0.6	1.889	0.732	2.621	−6.532	0.308
	0.3	0.0	0.0	2.821	1.574	4.394	−3.748	0.520
			0.3	2.731	1.493	4.223	−4.217	0.448
			0.6	2.768	1.255	4.024	−4.480	0.476
		0.3	0.0	2.615	1.426	4.042	−5.253	0.322
			0.3	2.606	1.459	4.066	−3.444	0.345
			0.6	2.709	1.163	3.871	−5.510	0.423
		0.6	0.0	1.971	1.098	3.069	−4.897	0.219
			0.3	1.941	1.052	2.992	−6.108	0.286
			0.6	1.958	0.982	2.940	−6.764	0.606
$n = 800$	0.0	0.0	0.0	1.412	0.878	2.290	1.497	0.090
			0.3	1.402	0.890	2.293	1.428	0.111
			0.6	1.402	0.663	2.065	0.024	0.062
		0.3	0.0	1.265	0.782	2.047	−0.149	0.053
			0.3	1.225	0.677	1.902	1.250	0.036
			0.6	1.306	0.683	1.989	0.493	0.360
		0.6	0.0	0.962	0.613	1.574	−0.864	0.052
			0.3	0.966	0.527	1.494	0.706	0.029
			0.6	0.960	0.452	1.413	0.295	0.076
	0.3	0.0	0.0	1.388	0.824	2.211	0.646	0.076
			0.3	1.407	0.807	2.214	0.002	0.076
			0.6	1.403	0.629	2.032	2.125	0.084
		0.3	0.0	1.287	0.744	2.032	1.267	0.041
			0.3	1.278	0.772	2.049	−0.737	0.054
			0.6	1.229	0.595	1.824	0.540	0.039
		0.6	0.0	0.971	0.570	1.541	1.692	0.027
			0.3	0.982	0.588	1.570	1.740	0.031
			0.6	0.948	0.554	1.502	−1.985	0.073

Appendix A: Proof of result 1

Step 1. Show

$$N^{-1}\sum_{i\in A}\pi_i^{-1}\hat{p}_{i|A}^{-1}R_i\boldsymbol{\omega}_i\boldsymbol{\omega}_i' - N^{-1}\sum_{i=1}^{N}\boldsymbol{\omega}_i\boldsymbol{\omega}_i' = O_p\left(n^{-\frac{1}{2}}\right). \tag{A.1}$$

Proof: Let v_i be an any element of $\boldsymbol{\omega}_i\boldsymbol{\omega}_i'$. Then, by Kim and Kim (2007),

$$(\bar{v}_{NA} - \bar{v}_{NAL})\left|\mathcal{F}_N = O_p\left(n^{-1}\right),\right.$$

where

$$\bar{v}_{NAL} = N^{-1}\sum_{i\in A}\pi_i^{-1}\left\{p_{i|A}\mathbf{z}_i'\boldsymbol{\gamma}_n + p_{i|A}^{-1}R_i\left(v_i - p_{i|A}\mathbf{z}_i'\boldsymbol{\gamma}_n\right)\right\},$$

and

$$\boldsymbol{\gamma}_n = \left[\sum_{i\in A} \pi_i^{-1} p_{i|A}\left(1-p_{i|A}\right)\mathbf{z}_i\mathbf{z}_i'\right]^{-1} \sum_{i\in A} \pi_i^{-1}\left(1-p_{i|A}\right)\mathbf{z}_i\nu_i.$$

Note that

$$E\left(\bar{\nu}_{NAL}\middle|\mathcal{F}_N\right) = E\left\{E\left(\bar{\nu}_{NAL}\middle|A_N,\mathcal{F}_N\right)\middle|\mathcal{F}_N\right\} = \bar{\nu}_N$$

and

$$\begin{aligned} V\left(\bar{\nu}_{NAL}\middle|\mathcal{F}_N\right) &= V\left(N^{-1}\sum_{i\in A}\pi_i^{-1}\nu_i\middle|\mathcal{F}_N\right) + N^{-2}E\left\{\sum_{i\in A}\pi_i^{-2}\left[p_{i|A}^{-1}(1-p_{i|A})\left(\nu_i - p_{i|A}\mathbf{z}_i'\boldsymbol{\gamma}_n\right)^2\right]\middle|\mathcal{F}_N\right\} \\ &= O\left(n^{-1}\right). \end{aligned}$$

Thus, for all element ν_i of $\boldsymbol{\omega}_i\boldsymbol{\omega}_i'$,

$$\left(\bar{\nu}_{NA} - \bar{\nu}_N\right)\middle|\mathcal{F}_N = O_p\left(n^{-\frac{1}{2}}\right).$$

□

Step 2. Show

$$\bar{y}_{NARR} = \bar{y}_{NA} + \left(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{NA}\right)\hat{\boldsymbol{\beta}}_{NA} + O_p\left(n^{-1}\right),$$

where

$$\hat{\boldsymbol{\beta}}_{NA} = \left(\sum_A \pi_i^{-1}\hat{p}_{i|A}^{-1}R_i\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_A \pi_i^{-1}\hat{p}_{i|A}^{-1}R_i\mathbf{x}_i'y_i\right).$$

Proof: For any sample and a set of respondents, there exists a solution of $\hat{\boldsymbol{\lambda}}$ that satisfies the equation (2.5) as explained by Deville and Särndal (1992). Define the function $\phi(\cdot)$ as

$$\phi\left(\hat{\boldsymbol{\lambda}}\right) = N^{-1}\sum_{i\in A}\pi_i^{-1}\hat{p}_{i|A}^{-1}\left[\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\lambda}}\right) - 1\right]\mathbf{x}_i R_i. \tag{A.2}$$

Note that $\phi(\mathbf{0}) = 0$. By using the Taylor expansion of $\phi(\hat{\boldsymbol{\lambda}})$ at $\hat{\boldsymbol{\lambda}} = \mathbf{0}$, we obtain

$$\phi\left(\hat{\boldsymbol{\lambda}}\right) = \phi(\mathbf{0}) + \phi'(\boldsymbol{\xi})\left(\hat{\boldsymbol{\lambda}} - \mathbf{0}\right) = \phi'(\boldsymbol{\xi})\hat{\boldsymbol{\lambda}},$$

where $\boldsymbol{\xi}$ is a vector in the interval $\mathbf{0}$ and $\hat{\boldsymbol{\lambda}}$, called $\mathbf{B}$. Now let $\boldsymbol{\eta}$ be a vector such that

$$\boldsymbol{\eta} = \bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{NA}.$$

Then, by (2.5),

$$\hat{\boldsymbol{\lambda}} = [\phi'(\boldsymbol{\xi})]^{-1}(\boldsymbol{\eta})$$

and thus, by the characteristics of exponential function,

$$\left\|\hat{\lambda}\right\| \leq K'_{\phi} \left\|\boldsymbol{\eta}\right\|,$$

where K'_{ϕ} is a positive constant such that $\|[\phi'(\boldsymbol{\xi})]^{-1}\| \leq K_{\phi}$ for all $\boldsymbol{\xi}$ in $\mathbf{B}$. By (2.8), $\|\boldsymbol{\eta}\| = O_p(n^{-1/2})$ and thus

$$\hat{\boldsymbol{\lambda}} = O_p\left(n^{-\frac{1}{2}}\right).$$

Now using $\exp(\mathbf{x}'_i\hat{\boldsymbol{\lambda}}) = 1 + \mathbf{x}'_i\hat{\boldsymbol{\lambda}} + \theta(\mathbf{x}'_i\boldsymbol{\xi})$ and (2.7), we have

$$\boldsymbol{\lambda} = \left[N^{-1}\sum_{i\in A}\pi_i^{-1}\hat{p}_{i|A}^{-1}R_i\mathbf{x}_i\mathbf{x}'_i\right]^{-1}(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{NA}) + O_p(n^{-1}),$$

because $\max_{\boldsymbol{\xi}}\theta(\mathbf{x}'_i\boldsymbol{\xi}) = O(n^{-1})$. Thus,

$$\begin{aligned}\bar{y}_{NARR} &= N^{-1}\sum_{i\in A}\pi_i^{-1}\hat{p}_{i|A}\exp\left(\mathbf{x}'_i\hat{\boldsymbol{\lambda}}\right)R_iy_i \\ &= \bar{y}_{NA} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{NA})\hat{\boldsymbol{\beta}}_{NA} + O_p\left(n^{-1}\right). \end{aligned} \tag{A.3}$$

□

Step 3. Show the final result.

Proof: By (A.1),

$$\hat{\boldsymbol{\beta}}_{NA} - \boldsymbol{\beta}_N = O_p\left(n^{-\frac{1}{2}}\right),$$

and thus, due to the result of Kim and Kim (2007),

$$\begin{aligned}\bar{y}_{NARR} &= \bar{y}_{NA} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{NA})\boldsymbol{\beta}_N + O_p\left(n^{-1}\right) \\ &= \bar{\mathbf{x}}_N\boldsymbol{\beta}_N + N^{-1}\sum_{A}\pi_i^{-1}\hat{p}_{i|A}^{-1}R_i\left(y_i - \mathbf{x}_i\boldsymbol{\beta}_N\right) + O_p\left(n^{-1}\right) \\ &= \bar{\mathbf{x}}_N\boldsymbol{\beta}_N + N^{-1}\sum_{i\in A}\pi_i^{-1}\left\{p_{i|A}\mathbf{z}'_i\boldsymbol{\delta}_N + p_{i|A}^{-1}R_i\left(e_i - p_{i|A}\mathbf{z}'_i\boldsymbol{\delta}_N\right)\right\} + O_p\left(n^{-1}\right). \end{aligned} \tag{A.4}$$

□

Appendix B: Concluding remark

In many survey practices, both weight adjustment using the estimated response probability and raking ratio procedure are employed to reduce nonresponse bias and obtain the known population auxiliary information if the estimator were applied to auxiliary variables. We proposed a consistent variance estimator using the asymptotic result of nonresponse adjusted raking ratio estimator. A simulation study indicated that a variance estimator that does not consider variability due to nonresponse and adjustment could severely underestimate true variance.

References

Center for Health Policy Research (2011). *Report 5: Weighting and Variance Estimation (CHIS 2009 Methodology Series)*, UCLA Center for Health Policy Research, Los Angeles, CA.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, **11**, 427–444.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

Deville, J. C., Sarndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88**, 1013–1020.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish household budget survey, *Journal of Official Statistics*, **7**, 325–337.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89–96.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability, *Canadian Journal of Statistics*, **35**, 501–514.

National Health and Nutrition Examination Survey (1996). *Analytic and Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANE III (1988–1994)*, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD.

National Health and Nutrition Survey (2010). *Analytic Guidelines: The fourth NHNS*, Ministry of Health and Welfare, Seoul, Korea.

Rosenbaum, P. R. (1987). Model-based direct adjustment, *Journal of the American Statistical Association*, **82**, 387–394.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, Wiley, New York.

Särndal, C. E., Swensson, B. and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika*, **76**, 527–537.

Särndal, C. E., Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Received September 3, 2015; Revised October 6, 2015; Accepted October 6, 2015