Journal of the Korean Data & Information Science Society
2014, **25**(1), 263–269

http://dx.doi.org/10.7465/jkdi.2014.25.1.263
한국데이터정보과학회지

Estimation using response probability when missing data happen on the second occasion[†]

Hyeonah Park[1] · Seongryong Na[2]

[1]Department of Statistics, Seoul National University
[2]Department of Information and Statistics, Yonsei University

Received 4 December 2013, revised 7 January 2014, accepted 15 January 2014

Abstract

When the loss of samples appears under repeated surveys, new samples can often replace missing values. Estimators using response probability can be considered under repeated surveys on two occasions where new samples are selected instead of missing data on the second occasion. We propose a new estimator that uses both respondents and new samples on the second occasion. It is considered for the simulation setting that missing values can happen at the second occasion and are replaced by new samples. We can see that the proposed estimator is more efficient than that using a weighting adjustment method for respondents at the second occasion.

Keywords: Repeated survey, response probability, two occasions.

1. Introduction

In many repeated surveys, the samples which are selected from a sampling design are continuously surveyed for a given period. The loss of samples, however, often happens under repeated surveys during a given term because of moving, nonresponse, refusal and so on. We can replace the missing values with the values of new samples. In general, the estimation of a population parameter requires the revision of design weight by the reciprocal of response probability under a given period (Lohr, 1999). Thus, it is necessary to consider response probability in repeated surveys.

In this paper, we consider repeated surveys where the same samples are to be surveyed on two occasions. It is assumed that all sample units respond at the first occasion but some units can be lost at the second occasion. Unit nonresponses at the second occasion are assumed to be replaced by new samples. A new estimator combining an estimator based on samples which response on the two occasions and an estimator based on new samples instead of missing data on the second occasion is proposed to estimate the population mean. In the new estimator, the response probability differs at each observation. Different response

[†] 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 (NRF-2012R1A1A3003761).
[1] Corresponding author: Postdoc, Department of Statistics, Seoul National University, Seoul 151-742, Korea. E-mail: naparkna@daum.net
[2] Professor, Department of Information and Statistics, Yonsei University, Wonju 220-710, Korea.

probability has been considered in many cases of sample survey. For example, under the assumption that the response probability is constant in the same cell but different among cells, Little (1986) provided several methods to divide the cells. Eltinge and Yansaneh (1997) discussed diagnostics for the formation of the cells. Rosenbaum (1987), Robins *et al.* (1994) and Lipsitz *et al.* (1999) have used the logistic model to represent the different response probability in the non-imputation context. Park and Park (2013) researched imputation method using response probability.

We first prove the unbiasedness and the efficiency of the proposed estimator under the assumption of known response probability. We also consider the estimation of response probability because we cannot know response probability previously. We compare the proposed estimator using estimated response probability with a conventional estimator and show that the former has smaller variance in simulation study. This paper is organized as follows. In Section 2, the estimator using the response probability is studied. In Section 3, results from a limited simulation study are presented.

2. An estimator using response probability

We assume that the finite population of size N is indexed from 1 to N at each occasion. Let the population parameter be the mean $\bar{Y}_t = N^{-1}\sum_{i=1}^{N} y_{ti}$, where y_{ti} is the study variable of unit i on the t-th occasion for $t = 1, 2$. We assume that samples are selected once using simple random sampling and the same samples are used on both occasions. It is assumed that when missing values appear at the second occasion because of nonresponse, moving, refusal and so on, they are replaced by new samples.

In this section we consider the case that the response probability of each observation at the second occasion under repeated surveys is not the same. First, define the response indicator function at the second occasion as

$$R_i = \begin{cases} 1, & \text{if unit } i \text{ responds} \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and n is the sample size. Let $\pi_i = P\ (R_i = 1)$ be the response probability of unit i surveyed at the second occasion. We assume that the response is ignorable such that π_i depends on an auxiliary variable but not on y_{2i}.

Let us introduce some notations for the estimators at two occasions. First, let

$$\bar{y}_1 = \sum_{i=1}^{n} w_i y_{1i}$$

be the estimator at the first occasion with $w_i = n^{-1}$. Note that w_i can be different from n^{-1} for another setting. Generally, if new samples aren't selected, the weight for a respondent at the second occasion becomes $w_i\pi_i^{-1}$ and the estimator at the second occasion is

$$\bar{y}_{2m} = \sum_{i=1}^{n} w_i \pi_i^{-1} R_i y_{2i},$$

which satisfies unbiasedness (Lohr, 1999). Define

$$\bar{y}_{2u} = \sum_{i=1}^{n} w_i (1-\pi_i)^{-1}(1-R_i) y_{2i}^{*}$$

as an estimator based on the new samples y_{2i}^* at the second occasion and

$$\bar{y}_{1m} = \sum_{i=1}^{n} w_i \pi_i^{-1} R_i y_{1i}$$

as an estimator based on the samples at the first occasion which respond at the second occasion.

Considering new samples instead of missing data on the second occasion, we propose a new estimator using the response probability

$$\bar{y}_2' = \phi_2 \bar{y}_{2u} + (1 - \phi_2)\bar{y}_{2m}', \tag{2.1}$$

where $\bar{y}_{2m}' = \hat{r}\bar{y}_1$ and $\hat{r} = \bar{y}_{1m}^{-1}\bar{y}_{2m}$. The constant ϕ_2 is selected to minimize $V(\bar{y}_2')$, such that

$$\phi_2 = [V(\bar{y}_{2u}) + V(\bar{y}_{2m}') - 2Cov(\bar{y}_{2u}, \bar{y}_{2m}')]^{-1}[V(\bar{y}_{2m}') - Cov(\bar{y}_{2u}, \bar{y}_{2m}')].$$

In the following theorem, we derive the asymptotic properties of the estimator $\bar{y}_2'$.

Theorem 2.1 Let us assume a sequence of finite populations with finite second moment of y_{ki} as defined in Isaki and Fuller (1982). Assume also that the response mechanism satisfies the condition that

$$\pi_i > K \tag{2.2}$$

for some nonnegative constant K and

$$P(R_i = 1, R_j = 1) = P(R_i = 1)P(R_j = 1) \tag{2.3}$$

for all i and j with $i \neq j$. Then,

$$E(\bar{y}_2') = \bar{Y}_2 + o(n^{-1/2}) \tag{2.4}$$

and

$$V(\bar{y}_2') = \left[\frac{2S_2^2}{n} + E_1 + E_2 + 2E_3\right]^{-1} \left[\left(\frac{S_2^2}{n}\right)^2 + \frac{S_2^2}{n}(E_1 + E_2) + E_1E_2 - E_3^2\right] + o(n^{-1}), \tag{2.5}$$

where $r = \bar{Y}_1^{-1}\bar{Y}_2$, $E_1 = E[\sum_{i=1}^n w_i^2(\pi_i^{-1} - 1)^{-1}y_{2i}^{*2})]$, $E_2 = E[\sum_{i=1}^n w_i^2(\pi_i^{-1} - 1)(y_{2i} - ry_{1i})^2]$, $E_3 = E[\sum_{i=1}^n w_i^2 y_{2i}^*(y_{2i} - ry_{1i})]$ and $S_2^2 = (N-1)^{-1}\sum_{i=1}^N (y_{2i} - \bar{Y}_2)^2$.

Proof. Note that it follows from (2.2) and (2.3) that

$$E[(\bar{y}_{1m} - \bar{Y}_1)^2] = V\left(\sum_{i=1}^n w_i y_{1i}\right) + E\left(\sum_{i=1}^n w_i^2(\pi_i^{-1} - 1)y_{1i}^2\right) = O(n^{-1})$$

and

$$E[(\bar{y}_{2m} - \bar{Y}_2)^2] = V\left(\sum_{i=1}^n w_i y_{2i}\right) + E\left(\sum_{i=1}^n w_i^2(\pi_i^{-1} - 1)y_{2i}^2\right) = O(n^{-1}).$$

Then, using Corollary 5.1.1.1 of Fuller (1996), we obtain that

$$\bar{y}_{1m} - \bar{Y}_1 = O_P(n^{-1/2})$$

and

$$\bar{y}_{2m} - \bar{Y}_2 = O_P(n^{-1/2}).$$

Then, by Taylor expansion, we obtain that

$$\hat{r} = r + \bar{Y}_1^{-1}[(\bar{y}_{2m} - \bar{Y}_2) - r(\bar{y}_{1m} - \bar{Y}_1)] + o_p(n^{-1/2}).$$

Write $\bar{y}'_{2m}$ as

$$\bar{y}'_{2m} = (\hat{r} - r)(\bar{y}_1 - \bar{Y}_1) + \hat{r}\bar{Y}_1 + r(\bar{y}_1 - \bar{Y}_1) = \bar{y}_{2m} - r(\bar{y}_{1m} - \bar{y}_1) + o_p(n^{-1/2}),$$

and we obtain (2.4) by unbiasedness of $\bar{y}_{2u}$.

Observe that

$$V(\bar{y}'_2) = [V(\bar{y}_{2u}) + V(\bar{y}'_{2m}) - 2Cov(\bar{y}_{2u}, \bar{y}'_{2m})]^{-1}[V(\bar{y}_{2u})V(\bar{y}'_{2m}) - Cov(\bar{y}_{2u}, \bar{y}'_{2m})^2],$$

which follows the definition of ϕ_2. First, from (2.3), we have that

$$V(\bar{y}_{2u}) = n^{-1}S_2^2 + E_1. \tag{2.6}$$

In the second place, by (2.2) and (2.3),

$$V(\bar{y}'_{2m}) = n^{-1}S_2^2 + E_2 + o(n^{-1}). \tag{2.7}$$

Again by (2.2) and (2.3),

$$Cov(\bar{y}_{2u}, \bar{y}'_{2m}) = -E_3 + o(n^{-1}). \tag{2.8}$$

Finally, from (2.6), (2.7) and (2.8), we obtain (2.5). □

So far, we have assumed that π_i and ϕ_2 are all known. Because we cannot know them in realistic case, we must estimate π_i and ϕ_2. For the case of different response probability, we use the logistic regression model

$$\pi_i = exp(\alpha_0 + \alpha_1 x_i)/[1 + exp(\alpha_0 + \alpha_1 x_i)], \tag{2.9}$$

where x_i is the value of the auxiliary variable of unit i. The maximum likelihood parameter estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$of the logistic regression model are computed iteratively using the Newton-Raphson method. The estimated response probability is

$$\hat{\pi}_i = exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_i)/[1 + exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_i)].$$

We also consider variance and covariance estimators for $\bar{y}_{2u}$ and $\bar{y}'_{2m}$ to estimate ϕ_2. The jackknife variance estimator of $V(\bar{y}_{2u})$ is

$$\hat{V}(\bar{y}_{2u}) = \sum_{k=1}^{n} n^{-1}(n-1)(\bar{y}_{2u}^{(k)} - \bar{y}_{2u})^2,$$

where $\bar{y}_{2u}^{(k)} = \sum_{i \neq k}^{n}(n-1)^{-1}nw_i(1 - \hat{\pi}_i^{(k)})^{-1}(1 - R_i)y_{2i}^*$ and $\hat{\pi}_i^{(k)}$ is the estimated $\hat{\pi}_i$ after the deletion of the k-th observation. The jackknife variance estimator of $V(\bar{y}'_{2m})$ is

$$\hat{V}(\bar{y}'_{2m}) = \sum_{k=1}^{n} n^{-1}(n-1)(\bar{y}_{2m}^{\prime(k)} - \bar{y}'_{2m})^2,$$

where $\bar{y}_{2m}^{\prime(k)} = (\bar{y}_{2m}^{(k)}/\bar{y}_{1m}^{(k)})\bar{y}_1^{(k)}$, $\bar{y}_{2m}^{(k)} = \sum_{i\neq k}^{n}(n-1)^{-1}nw_i(\hat{\pi}_i^{(k)})^{-1}R_iy_{2i}$, $\bar{y}_{1m}^{(k)} = \sum_{i\neq k}^{n}(n-1)^{-1}nw_i(\hat{\pi}_i^{(k)})^{-1}R_iy_{1i}$ and $\bar{y}_1^{(k)} = \sum_{i\neq k}^{n}(n-1)^{-1}nw_iy_{1i}$. The covariance of $\bar{y}_{2u}$ and $\bar{y}_{2m}^{\prime}$ is estimated by

$$\hat{C}(\bar{y}_{2u}, \bar{y}_{2m}^{\prime}) = \sum_{k=1}^{n} n^{-1}(n-1)(\bar{y}_{2u}^{(k)} - \bar{y}_{2u})(\bar{y}_{2m}^{\prime(k)} - \bar{y}_{2m}^{\prime}).$$

Then, the estimator of ϕ_2 is given by

$$\hat{\phi}_2 = [\hat{V}(\bar{y}_{2u}) + \hat{V}(\bar{y}_{2m}^{\prime}) - 2\hat{C}(\bar{y}_{2u}, \bar{y}_{2m}^{\prime})]^{-1}[\hat{V}(\bar{y}_{2m}^{\prime}) - \hat{C}(\bar{y}_{2u}, \bar{y}_{2m}^{\prime})].$$

Finally, we have the estimator (2.1) with $\hat{\pi}_i$ and $\hat{\phi}_2$ plugged in.

3. Simulation results

We provide the results of a limited simulation study performed to test the efficiency of our estimator. In the simulation study, $B = 1,000$ samples of size $n = 100$ were generated by

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & (1-\rho^2)^{1/2}\sigma_2 \end{pmatrix} \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

where $z_{ti} \sim$ iid $N(0,1)$ for $i = 1, \ldots, n$ and $t = 1, 2$ and $\sigma_1 = \sigma_2 = 15$, $\mu_1 = \mu_2 = 15$. The correlation coefficient ρ of y_{1i} and y_{2i} has the values of 1, 0.9, 0.8 and 0.7. The parameters of logistic regression model were assumed to be $(\alpha_0, \alpha_1) = (1.5, -1.0)$, $(-0.5, 1.0)$, $(-1.0, 0.0)$ and, thus, the overall response rates were respectively 0.73, 0.50, 0.27. The auxiliary variable was assumed to be distributed as $x_i \sim \text{Uniform}(0, 1)$. The maximum likelihood parameter estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$ of the logistic regression model are computed iteratively using the Newton-Raphson method. Note that the missing data at the second occasion were replaced by new samples.

Using B samples of $(y_{1i}, y_{2i}, R_i, x_i)$, $i = 1, ..., n$, we computed the empirical values of two types of variances $V(\bar{y}_{2m})$ and $V(\bar{y}_2^{\prime})$ and of two types of expectations $E(\bar{y}_{2m})$ and $E(\bar{y}_2^{\prime})$, where $\bar{y}_2^{\prime}$ is our proposed estimator using $\hat{\pi}_i$ and $\hat{\phi}_2$. Table 3.1 contains the information of the simulated values for expectation and variance for the case of different response probability. Two rows of each cell in Table 3.1 denote bias, MSE and SMSE of $\bar{y}_{2m}$ and $\bar{y}_2^{\prime}$, respectively, for different values of (α_0, α_1) and correlation coefficient ρ. Here, the standardized MSE (SMSE) is defined by $MSE(\bar{y}_2^{\prime})/MSE(\bar{y}_{2m})$.

Table 3.1 Bias, MSE, standardized MSE of $\bar{y}_{2m}$ and $\bar{y}_2^{\prime}$

ρ		(α_0, α_1)								
		$(1.5, -1.0)$			$(-0.5, 1.0)$			$(-1.0, 0.0)$		
		Bias	MSE	SMSE	Bias	MSE	SMSE	Bias	MSE	SMSE
1.0	$\bar{y}_{2m}$	-0.074	2.987	1	-0.04	4.4	1	0.004	8.529	1
	$\bar{y}_2^{\prime}$	-0.059	1.768	0.592	-0.042	1.532	0.348	-0.031	1.294	0.152
0.9	$\bar{y}_{2m}$	-0.081	3.062	1	-0.05	4.708	1	-0.015	8.709	1
	$\bar{y}_2^{\prime}$	-0.07	1.981	0.647	-0.046	1.845	0.392	-0.012	1.777	0.204
0.8	$\bar{y}_{2m}$	-0.079	3.109	1	-0.052	4.861	1	-0.023	8.786	1
	$\bar{y}_2^{\prime}$	-0.069	2.144	0.689	-0.051	2.057	0.423	-0.027	2.053	0.234
0.7	$\bar{y}_{2m}$	-0.076	3.148	1	-0.051	4.969	1	-0.029	8.834	1
	$\bar{y}_2^{\prime}$	-0.068	2.287	0.726	-0.059	2.217	0.446	-0.047	2.24	0.254

As anticipated, we can see that biases of $\bar{y}_{2m}$ and $\bar{y}_2'$ are negligible. We can also know that the biases of estimators are decreased more and more when sample size is increased. Table 3.1 shows that $\bar{y}_2'$ is more efficient than $\bar{y}_{2m}$ for all response rates and correlation coefficients. See the values of SMSE. We can find that the value of MSE($\bar{y}_2'$) tends to decrease as the value of ρ increases. On the contrary, the simulation result shows that there is a tendency for MSE($\bar{y}_2'$) to decrease as the response probability decreases. We empirically conclude that the proposed estimator reveals more efficiency than the conventional estimator using the reciprocal of response probability especially when the correlation coefficient is not too small. Furthermore, it can be seen that our estimator becomes more efficient when the overall response probability is small. In repeated surveys when the correlation coefficient is properly large, the utilization of $\bar{y}_2'$ can be recommended under a proper response probability.

4. Concluding remarks

In this paper we proposed a new estimator of the population mean at the second occasion under repeated surveys when the loss of samples is possible at the second occasion. This estimator combines an estimator of the samples responding at both occasions and an estimator of the samples replacing missing data at the second occasion. The response probability is allowed to be different for each unit and assumed to be ignorable. First, we investigated the asymptotic mean and variance of the proposed estimator. It can be shown that the estimator is asymptotically unbiased under some mild conditions.

In most realistic situations, the response probability is unknown and must be estimated to be used in the new estimator. In this paper, through a simulation study, we investigated the asymptotic behaviors of several estimators including the proposed one with estimated response probabilities. It was empirically ascertained that our estimator has smaller variance than the estimator using the reciprocal of response probability while both estimators are asymptotically unbiased. This seems to be caused by the fact that the proposed estimator uses the information in new samples replacing missing data and the correlation between data of two occasions is not small. Thus, our new estimator will be an efficient candidate in the estimation of the population parameter at the second occasion when missing samples are replaced and the correlation is expected to be somewhat large. Theoretical evaluation of the proposed estimator under complex survey remains as a future work.

References

Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, **23**, 33-40.

Fuller, W. A. (1996). *Introduction to statistical time series*, Wiley, New York.

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

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, **94**, 1147-1160.

Little, R. J. A. (1986). Survey nonresponse adjustments. *International Statistical Review*, **54**, 139-157.

Lohr, S. L. (1999). *Sampling: Design and analysis*, Duxbury, California.

Park, H. and Park, W. (2013). Usage of auxiliary variable and neural network in doubly robust estimation. *Journal of the Korean Data & Information Science Society*, **24**, 659-667.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

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