Estimation Using Response Probability Under Callbacks

Hyeonah Park*

Department of Statistics, Seoul National University, Seoul, 151-742, Korea

Abstract

Although the response model has been frequently applied to nonresponse weighting adjustment or imputation, the estimation under callbacks has been relatively underdeveloped in the response model. The estimation method using the response probability is developed under callbacks. A replication method for the estimation of the variance of the proposed estimation is also developed. Since the true response probability is usually unknown, we study the estimation of the response probability. Finally, we propose an estimator under callbacks using the ratio imputation as well as the response probability. The simulation study illustrates our techniques.

Key words : callbacks, response probability, ratio imputation, variance estimation.

1 Introduction

Generally, almost all surveys rely on callbacks to raise the response rate of persons who are not at home. The technique of callback and the estimation after callbacks have been considered numerous times by many survey researchers. A method of selecting subsamples from persons not at home and the estimation of double sampling were first considered by Hansen and Hurwitz (1946). Deming (1953) studied the estimation of the population mean when responses are collected until the i-th callback attempt. Groves (1989) provided an excellent summary of these approaches. Recently, Elliott, Little and Lewitzky (2000) considered the subsampling callback, where an efficient subsampling strategy considering variance and cost from the repeated callback attempts was established.

The estimators are relatively underdeveloped in the response model under callbacks though the response model has been applied to nonresponse weighting adjustment or imputation. There is vast literature on nonresponse weighting adjustment or imputation under uniform or non-uniform response model. See, for example, Rosenbaum (1987), Rao and Shao (1992), Robins, Rotnitzky and Zhao (1994), Rao and Sitter (1995), Lipsitz, Ibrahim and Zhao (1999) and Shao and Steel (1999).

In this Section, we propose an estimator using the response probability and an auxiliary variable in the response model under callbacks. We prove the unbiasedness and the efficiency of the proposed estimator under the assumption that we know the true response probability. We also suggest a replication variance estimator of the estimator which satisfies the consistency under infinite sample size.

Since the response probability is usually unknown, however, we consider an estimator using the estimated response probability instead of the true response probability. For the estimation of the response probability, one can refer to Ekholm and Laaksonen (1991) and Iannacchione (2003). We also propose a consistent replication variance estimator of this estimator.

This Section is organized as follows. In Section 2, we introduce an estimator using the true response probability and an auxiliary variable under callbacks and calculate the expectation and the

* *E-mail address:* parkha03@yahoo.co.kr

variance of the estimator. In Section 3, we propose a replication variance estimator of the estimator of Section 2. In Section 4, we consider the estimation of the response probability assuming the logistic response model, where the estimator and its variance estimator corresponding to the estimated response probability are also given. The numerical evaluation of the estimators in this paper is performed through a simulation study in Section 5. Finally, we remark another estimator using the response probability and the imputation under callbacks in Section 6.

2 Estimation using response probability

Let the population total be $Y = \sum_{i=1}^{N} y_i$ and the population mean be $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$, where N is the population size and y_i is the value of the target variable of unit i. Let $\hat{Y}_n$ be an estimator of the population total Y defined by $\hat{Y}_n = \sum_{i\in A} w_i y_i$, where n is the sample size, w_i is the sampling weight of unit i and $A = \{1, 2, \dots, n\}$ is the set of indices of the sample.

We define the response indicator function under the first survey as

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

for $i \in A$. Let $\pi_i = P(R_i = 1|i \in A)$ be the response probability of sample unit i under the first survey. We consider only one-step callback in this paper. We write the response indicator function under callback as

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

for $i \in A_{NR}$, where $A_{NR} = \{i : R_i = 0,\ i \in A\}$ is the set of the indices of nonresponding units. Let $p_i = Pr(T_i = 1|i \in A_{NR})$ be the response probability of sample unit i for $i \in A_{NR}$ under callback. We assume that R_i and T_i are ignorable such that π_i and p_i depend on an auxiliary variable z_i but not on y_i. In this section we also assume that all π_i and p_i are known priori.

We suppose that there is another auxiliary variable x_i that is related with the study variable y_i and can be observed throughout the sample. We now introduce some preliminary estimators. Let $\tilde{Y}_R = \sum_{i=1}^{n} w_i \pi_i^{-1} R_i y_i$ and $\tilde{Y}_T = \sum_{i=1}^{n} w_i (1-\pi_i)^{-1} p_i^{-1} (1-R_i) T_i y_i$. For another auxiliary variable x_i, we define $\tilde{X}_T = \sum_{i=1}^{n} w_i (1-\pi_i)^{-1} p_i^{-1} (1-R_i) T_i x_i$ and $\tilde{X}_{NR} = \sum_{i=1}^{n} w_i (1-\pi_i)^{-1} (1-R_i) x_i$. The estimator $\tilde{Y}_R$ is estimated with data of the first survey and π_i. The estimators $\tilde{Y}_T$ and $\tilde{X}_T$ are estimated using data under callback and the response probabilities. Using the ratio $\tilde{r}_1 = \tilde{X}_T^{-1}\tilde{Y}_T$, we define

$$\tilde{Y}_I = \tilde{Y}_T + \tilde{r}_1(\tilde{X}_{NR} - \tilde{X}_T)$$

and

$$\tilde{\phi} = [Var(\tilde{Y}_R) + Var(\tilde{Y}_I) - 2Cov(\tilde{Y}_R, \tilde{Y}_I)]^{-1}[Var(\tilde{Y}_I) - Cov(\tilde{Y}_R, \tilde{Y}_I)]. \tag{1}$$

We know that the estimator $\tilde{Y}_I$ is regression-type or ratio-type estimator using response probabilities. We also think various estimator except the ratio $\tilde{r}_1 = \tilde{X}_T^{-1}\tilde{Y}_T$.

Then, our proposed estimator is defined as

$$\tilde{Y}_C = \tilde{\phi}\tilde{Y}_R + (1-\tilde{\phi})\tilde{Y}_I.$$

Note that the variance of $k\tilde{Y}_R + (1-k)\tilde{Y}_I$ is minimized at $k = \tilde{\phi}$. We adopt the extended definition of the response indicator function introduced by Fay (1991). Conceptually, the response indicator functions R_i and T_i can be extended to the entire population.

In this section we assume the following conditions:

(A1) A sequence of finite populations and samples are defined as in Isaki and Fuller (1982). The finite populations satisfy that for some $\tau > 0$

$$N^{-1}\sum_{i=1}^{N}\theta_i^{2+\tau} = O(1),$$

where θ_i represents y_i, x_i and z_i. The sampling mechanism satisfies

$$E(\hat{Y}_n) = Y.$$

(A2) For nonnegative constants C_1, C_2 and C_3,

$$C_1 < \pi_i < C_2$$

and

$$p_i > C_3.$$

(A3) The response indicator functions R_i and T_i are mutually independent, respectively, such that

$$P(R_i = 1, R_j = 1) = P(R_i = 1)P(R_j = 1)$$

and

$$P(T_i = 1, T_j = 1) = P(T_i = 1)P(T_j = 1)$$

for different i and j.

(A4) The sampling mechanism satisfies that for nonnegative constants D_1, D_2, D_3 and D_4

$$D_1 < \max_{1\le i\le N}(N^{-1}nw_i) < D_2$$

and

$$D_3 < N^{-2}nVar(\hat{Y}_n) < D_4,$$

where the variance is calculated under the sampling mechanism.

In the following theorem we deal with the expectation and the variance of our proposed estimator $\tilde{Y}_C$.

Theorem 2.1 *Under the assumptions (A1)-(A4),*

$$E(\tilde{Y}_C) = Y + o(n^{-1/2}N) \tag{2}$$

and

$$Var(\tilde{Y}_C) = Var(\hat{Y}_n) + (E_1 + E_2 + 2E_3)^{-1}(E_1E_2 - E_3^2) + o(n^{-1}N^2), \tag{3}$$

where

$$\begin{aligned} E_1 &= E\left[\sum_{i=1}^{n} w_i^2(\pi_i^{-1} - 1)y_i^2\right], \\ E_2 &= E\left[\sum_{i=1}^{n} w_i^2(1-\pi_i)^{-1}(\pi_i y_i^2 + (p_i^{-1} - 1)(y_i - rx_i)^2)\right], \\ E_3 &= E\left[\sum_{i=1}^{n} w_i^2 y_i^2\right] \end{aligned}$$

and $r = X^{-1}Y$ *for* $X = \sum_{i=1}^{N} x_i$.

Proof. Note that under (A1), (A3) and (A4),

$$\begin{aligned} E[(\tilde{X}_{NR} - X)^2] &= Var(\hat{X}_n) + E[\sum_{i=1}^{n} w_i^2(\pi_i^{-1} - 1)^{-1}x_i^2] \\ &= O(n^{-1}N^2) \end{aligned}$$

and

$$\begin{aligned} E[(\tilde{X}_T - X)^2] &= Var(\hat{X}_n) + E[\sum_{i=1}^{n} w_i^2(1-\pi_i)^{-1}(\pi_i + p_i^{-1} - 1)x_i^2] \\ &= O(n^{-1}N^2), \end{aligned}$$

where $\hat{X}_n = \sum_{i=1}^{n} w_i x_i$. Similarly, we also obtain that

$$\tilde{Y}_T - Y = O_P(n^{-1/2}N).$$

Then, by Taylor's expansion, we have

$$\begin{aligned} \tilde{r}_1 - r &= X^{-1}[(\tilde{Y}_T - Y) - r(\tilde{X}_T - X)] + o_P(n^{-1/2}) \\ &= O_P(n^{-1/2}) \end{aligned} \tag{4}$$

and

$$\begin{aligned} \tilde{Y}_I &= \tilde{Y}_T + (\tilde{r}_1 - r)(\tilde{X}_{NR} - \tilde{X}_T) + r(\tilde{X}_{NR} - \tilde{X}_T) \\ &= \tilde{Y}_T + r(\tilde{X}_{NR} - \tilde{X}_T) + o_P(n^{-1/2}N). \end{aligned}$$

Observe that

$$\tilde{Y}_C - \hat{Y}_n = \tilde{\phi}(\tilde{Y}_R - \hat{Y}_n) + (1 - \tilde{\phi})[(\tilde{Y}_T - \hat{Y}_n) + r(\tilde{X}_{NR} - \tilde{X}_T)] + o_P(n^{-1/2}N).$$

This, together with (A1), implies (2).

Let $\tilde{Y}_I' = \tilde{Y}_T + r(\tilde{X}_{NR} - \tilde{X}_T)$. By definition of $\tilde{\phi}$, we have

$$\begin{aligned} Var(\tilde{Y}_C) &= [Var(\tilde{Y}_R) + Var(\tilde{Y}_I') - 2Cov(\tilde{Y}_R, \tilde{Y}_I')]^{-1}[Var(\tilde{Y}_R)Var(\tilde{Y}_I') \\ &\quad -Cov(\tilde{Y}_R, \tilde{Y}_I')^2] + o(n^{-1}N^2) \end{aligned}$$

since $\tilde{Y}_I - \tilde{Y}_I' = o_P(n^{-1/2}N)$. Observe that from (A3)

$$Var(\tilde{Y}_R) = Var(\hat{Y}_n) + E[\sum_{i=1}^{n} w_i^2(\pi_i^{-1} - 1)y_i^2],$$

$$Var(\tilde{Y}_I') = Var(\hat{Y}_n) + E[\sum_{i=1}^{n} w_i^2(1-\pi_i)^{-1}(\pi_i y_i^2 + (p_i^{-1} - 1)(y_i - rx_i)^2)]$$

and

$$Cov(\tilde{Y}_R, \tilde{Y}_I') = Var(\hat{Y}_n) - E[\sum_{i=1}^{n} w_i^2 y_i^2].$$

Thus, the result (3) follows immediately. **Q.E.D.**

3 Variance estimation using known response probability

In this section we propose a method of estimating the variance of the estimator when we know the response probability. Note that the variance must be estimated to calculate the efficiency of the proposed estimator. We consider the replication methods such as jackknife, balanced half samples and bootstrap since it is well known that replication methods are good to estimate the variances of complex estimators.

First, we illustrate a replication variance estimator for $\hat{Y}_n$. Let an estimator of $Var(\hat{Y}_n)$ be

$$\hat{V}(\hat{Y}_n) = \sum_{k=1}^{L} c_k(\hat{Y}_n^{(k)} - \hat{Y}_n)^2,$$

where L is the number of replications, c_k is a factor associated with the kth replication determined by the replication method and $\hat{Y}_n^{(k)}$ is the kth estimator of Y based on the observations included in the kth replication, that is,

$$\hat{Y}_n^{(k)} = \sum_{i=1}^{n} w_i^{(k)} y_i,$$

where $w_i^{(k)}$ is the replication weight for the ith unit in the kth replication. For example, if inclusion probability is $N^{-1}n$ and $w_i = n^{-1}N$, then the standard jackknife variance estimator $\hat{V}(\hat{Y}_n)$ is defined by $L = n$, $c_k = (1 - N^{-1}n)n^{-1}(n-1)$, $w_i^{(k)} = (n-1)^{-1}nw_i$ for $i \neq k$ and $w_k^{(k)} = 0$.

We suggest a replication estimator for the variance of $\tilde{Y}_C$ by

$$\hat{V}(\tilde{Y}_C) = \sum_{k=1}^{L} c_k(\tilde{Y}_C^{(k)} - \tilde{Y}_C)^2,$$

where

$$\tilde{Y}_C^{(k)} = \tilde{\phi}\tilde{Y}_R^{(k)} + (1-\tilde{\phi})\tilde{Y}_I^{(k)}.$$

Here, analogously as before, $\tilde{Y}_R^{(k)} = \sum_{i=1}^{n} w_i^{(k)}\pi_i{}^{-1}R_i y_i$ and $\tilde{Y}_I^{(k)} = \tilde{Y}_T^{(k)} + \tilde{r}_1^{(k)}(\tilde{X}_{NR}^{(k)} - \tilde{X}_T^{(k)})$ for $\tilde{r}_1^{(k)} = \tilde{Y}_T^{(k)}/\tilde{X}_T^{(k)}$, $\tilde{Y}_T^{(k)} = \sum_{i=1}^{n} w_i^{(k)}(1-\pi_i)^{-1}p_i{}^{-1}(1-R_i)T_i y_i$, $\tilde{X}_T^{(k)} = \sum_{i=1}^{n} w_i^{(k)}(1-\pi_i)^{-1}p_i{}^{-1}(1-R_i)T_i x_i$ and $\tilde{X}_{NR}^{(k)} = \sum_{i=1}^{n} w_i^{(k)}(1-\pi_i)^{-1}(1-R_i)x_i$. Note that (k) denotes the kth replication.

We assume that the variance of a linear estimator of the total is a quadratic function of y, that is,

$$N^{-2}nVar(\hat{Y}_n) = \sum_{i=1}^{N}\sum_{j=1}^{N}\Omega_{ij}y_iy_j, \tag{5}$$

where the coefficients Ω_{ij} satisfy

$$\max_{1\leq i\leq N}\Omega_{ii} = O(N^{-1}) \tag{6}$$

and

$$\sum_{i=1}^{N}|\Omega_{ij}| = O(N^{-1}). \tag{7}$$

For example, the simple random sampling with $w_i = n^{-1}N$ satisfies (6) and (7) because

$$\Omega_{ij} = \begin{cases} N^{-1}(1-N^{-1}n) & \text{if } i=j \\ -N^{-1}(N-1)^{-1}(1-N^{-1}n) & \text{if } i\neq j \ . \end{cases}$$

In order to establish the consistency of $\hat{V}(\tilde{Y}_C)$, we first prove the consistency of the variance estimator of $\tilde{Y}_R$ in the following lemma.

Lemma 3.1 *Suppose that the conditions (A1)-(A4) are satisfied. We assume that for any y with bounded fourth moment,*

$$E[(Var(\hat{Y}_n)^{-1}\hat{V}(\hat{Y}_n)-1)^2] = o(1), \tag{8}$$

where the expectation is calculated under the given sampling mechanism. Assume also that

$$N^{-1}n = o(1). \tag{9}$$

Then,

$$\hat{V}(\tilde{Y}_R) = \sum_{k=1}^{L}c_k(\tilde{Y}_R^{(k)}-\tilde{Y}_R)^2 = Var(\tilde{Y}_R) + o_P(n^{-1}N^2). \tag{10}$$

Proof. From (A4) and (8),

$$\hat{V}(\tilde{Y}_R) = Var(\tilde{Y}_R|R_1,\ldots,R_N) + o_P(n^{-1}N^2).$$

From (9),

$$Var[E(\tilde{Y}_R|R_1,\ldots,R_N)] = o(n^{-1}N^2).$$

Then, (10) follows immediately if we prove

$$Var[N^{-2}nVar(\tilde{Y}_R|R_1,\ldots,R_N)] = o(1). \tag{11}$$

It is observed that from (5)

$$\begin{aligned} & Var[N^{-2}nVar(\tilde{Y}_R|R_1,\ldots,R_N)] \\ = & \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{m=1}^{N}\Omega_{ij}\Omega_{km}Cov(y_{\pi i}y_{\pi j},y_{\pi k}y_{\pi m}), \end{aligned}$$

where $y_{\pi i} = \pi_i^{-1} R_i y_i$ and the covariances are taken with respect to R_i's. Since R_i's are independent, we can see that

$$\begin{aligned} & Var[N^{-2} n Var(\tilde{Y}_R | R_1, \ldots, R_N)] \\ = & \sum_{i=1}^{N} \sum_{j=1}^{N} (\Omega_{ij}^2 + \Omega_{ij}\Omega_{ji}) Var(y_{\pi i} y_{\pi j}) \\ \leq & 2 \max_{1 \leq i,j \leq N} Var(y_{\pi i} y_{\pi j}) \max_{1 \leq i,j \leq N} |\Omega_{ij}| \sum_{i=1}^{N} \sum_{j=1}^{N} |\Omega_{ij}|. \end{aligned}$$

From (A1) with $\tau > 2$ and (A2), $\max_{1 \leq i,j \leq N} Var(y_{\pi i} y_{\pi j}) = O(1)$. By (6) and the nonnegative definiteness of $\Omega = (\Omega_{ij})$, $\max_{1 \leq i,j \leq N} |\Omega_{ij}| = O(N^{-1})$. These, together with (7), imply (11). **Q.E.D.**

Secondly, we deal with the variance estimator for $\tilde{Y}_I$. Here, an additional condition is listed.

(A5) For the replication factors

$$\max_{1 \leq k \leq L} c_k^{-1} = O(L)$$

and

$$E[\{c_k(\hat{Y}_n^{(k)} - \hat{Y}_n)^2\}^2] < C_y L^{-2} \{Var(\hat{Y}_n)\}^2$$

for some constant C_y

Lemma 3.2 *Assume the conditions (A1)-(A5), (8) and (9). Then,*

$$\hat{V}(\tilde{Y}_I) = \sum_{k=1}^{L} c_k (\tilde{Y}_I^{(k)} - \tilde{Y}_I)^2 = Var(\tilde{Y}_I) + o_P(n^{-1} N^2). \tag{12}$$

Proof. Split $\tilde{Y}_I^{(k)} - \tilde{Y}_I$ as

$$\tilde{Y}_I^{(k)} - \tilde{Y}_I = (\tilde{Y}_I^{(k)} - \tilde{Y}_I'^{(k)}) + (\tilde{Y}_I'^{(k)} - \tilde{Y}_I') + (\tilde{Y}_I' - \tilde{Y}_I),$$

where $\tilde{Y}_I'^{(k)} = \tilde{Y}_T^{(k)} + r(\tilde{X}_{NR}^{(k)} - \tilde{X}_T^{(k)})$. Observe that

$$\begin{aligned} & (\tilde{Y}_I^{(k)} - \tilde{Y}_I'^{(k)}) + (\tilde{Y}_I' - \tilde{Y}_I) \\ = & (\tilde{r}_1^{(k)} - \tilde{r}_1)\{(\tilde{X}_{NR}^{(k)} - \tilde{X}_{NR}) - (\tilde{X}_T^{(k)} - \tilde{X}_T)\} \\ & + (\tilde{r}_1 - r)\{(\tilde{X}_{NR}^{(k)} - \tilde{X}_{NR}) - (\tilde{X}_T^{(k)} - \tilde{X}_T)\} + (\tilde{r}_1^{(k)} - \tilde{r}_1)(\tilde{X}_{NR} - \tilde{X}_T) \\ = & o_P(n^{-1/2} N) \end{aligned}$$

since $\tilde{X}_{NR}^{(k)} - \tilde{X}_{NR} = O_P(n^{-1/2} N)$, $\tilde{X}_T^{(k)} - \tilde{X}_T = O_P(n^{-1/2} N)$ and

$$\begin{aligned} \tilde{r}_1^{(k)} - \tilde{r}_1 &= \tilde{X}_T^{-1} \{(\tilde{Y}_T^{(k)} - \tilde{Y}_T) - \tilde{r}(\tilde{X}_T^{(k)} - \tilde{X}_T)\} + o_P(n^{-1/2}) \\ &= O_P(n^{-1/2}). \end{aligned} \tag{13}$$

which are consequences of (A3)-(A5), (4) and the fact that $\tilde{X}_{NR} - \tilde{X}_T = O_P(n^{-1/2}N)$. Then, from (8) and (A5),

$$\begin{aligned}\hat{V}(\tilde{Y}_I) &= \sum_{k=1}^{L} c_k(\tilde{Y}_I'^{(k)} - \tilde{Y}_I')^2 + o_P(n^{-1}N^2) \\ &= Var(\tilde{Y}_I'|T_1, \ldots, T_N, R_1, \ldots, R_N) + o_P(n^{-1}N^2).\end{aligned}$$

If we prove

$$Var[E(\tilde{Y}_I'|T_1, \ldots, T_N, R_1, \ldots, R_N)] = o(n^{-1}N^2) \tag{14}$$

and

$$Var[N^{-2}nVar(\tilde{Y}_I'|T_1, \ldots, T_N, R_1, \ldots, R_N)] = o(1), \tag{15}$$

then we have

$$Var(\tilde{Y}_I'|T_1, \ldots, T_N, R_1, \ldots, R_N) = Var(\tilde{Y}_I') + o_P(n^{-1}N^2),$$

which implies (12) since $\tilde{Y}_I - \tilde{Y}_I' = o_P(n^{-1/2}N)$. Note that (14) results from (9).

Now we deal with (15). Observe that

$$\tilde{Y}_I' = \sum_{i=1}^{n} w_i u_i + \sum_{i=1}^{n} w_i y_{ri} := \hat{U}_n + \hat{Y}_{rn},$$

where $u_i = (1-\pi_i)^{-1}(1-R_i)(p_i^{-1}T_i - 1)(y_i - rx_i)$ and $y_{ri} = (1-\pi_i)^{-1}(1-R_i)y_i$. It suffices to show

$$Var[N^{-2}nCov(\hat{W}_n, \hat{W}_n'|T_1, \ldots, T_N, R_1, \ldots, R_N)] = o(1), \tag{16}$$

for three cases of $\hat{W}_n = \hat{W}_n' = \hat{U}_n$, $\hat{W}_n = \hat{W}_n' = \hat{Y}_{rn}$ and $\hat{W}_n = \hat{U}_n$, $\hat{W}_n' = \hat{Y}_{rn}$. Here, we note that similar arguments in Lemma 3.1 and the conditions assumed in this lemma lead to (16) in each case. **Q.E.D.**

In the following theorem, the consistency of the variance estimator $\hat{V}(\tilde{Y}_C)$ is established.

Theorem 3.1 *Under the same conditions as in Lemma 3.2,*

$$\hat{V}(\tilde{Y}_C) = Var(\tilde{Y}_C) + o_P(n^{-1}N^2). \tag{17}$$

Proof. Observe that

$$\begin{aligned}&\sum_{k=1}^{L} c_k(\tilde{Y}_C^{(k)} - \tilde{Y}_C)^2 \\ &= \tilde{\phi}^2 \sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} - \tilde{Y}_R)^2 + (1-\tilde{\phi})^2 \sum_{k=1}^{L} c_k(\tilde{Y}_I^{(k)} - \tilde{Y}_I)^2 \\ &\quad +2\tilde{\phi}(1-\tilde{\phi}) \sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} - \tilde{Y}_R)(\tilde{Y}_I^{(k)} - \tilde{Y}_I).\end{aligned}$$

Using the similar arguments in Lemma 3.2, we obtain that

$$\begin{aligned} & \sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} + \tilde{Y}_I^{(k)} - \tilde{Y}_R - \tilde{Y}_I)^2 \\ & = \sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} + \tilde{Y}_I'^{(k)} - \tilde{Y}_R - \tilde{Y}_I')^2 + o_P(n^{-1}N^2). \end{aligned}$$

By (A4) and (8),

$$\begin{aligned} & \sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} + \tilde{Y}_I'^{(k)} - \tilde{Y}_R - \tilde{Y}_I')^2 \\ & = Var(\tilde{Y}_R + \tilde{Y}_I' | T_1, \ldots, T_N, R_1, \ldots, R_N) + o_P(n^{-1}N^2) \\ & = Var(\tilde{Y}_R + \tilde{Y}_I) + o_P(n^{-1}N^2), \end{aligned}$$

where the same steps in Lemmas 3.1 and 3.2 and the fact that $\tilde{Y}_I - \tilde{Y}_I' = o_P(n^{-1/2}N)$ have been used for the last equality. This, together with (10) and (12), implies that

$$\sum_{k=1}^{L} c_k(\tilde{Y}_R^{(k)} - \tilde{Y}_R)(\tilde{Y}_I^{(k)} - \tilde{Y}_I) = Cov(\tilde{Y}_R, \tilde{Y}_I) + o_P(n^{-1}N^2),$$

which implies (17). **Q.E.D.**

4 Estimation using estimated response probability

In most practical situations, it is impossible to know the response probability and the weight. In this section, we estimate the response probability and the weight. We assume the parametric logistic model for the response probability:

$$\pi_i = \pi(z_i; \alpha) = (1 + \exp(-z_i^T \alpha))^{-1},$$

where z_i is the value of an auxiliary variable for unit i and $\alpha = (\alpha_1, \ldots, \alpha_p)^T$. The response probability p_i under callback is assumed to be a constant p. Let $\hat{\pi}_i = \pi(z_i; \hat{\alpha})$ be the estimated response probability, where $\hat{\alpha}$ satisfies

$$n^{1/2}(\hat{\alpha} - \alpha) = n^{-1/2} \sum_{i=1}^{n} H(z_i, R_i; \alpha) + o_P(1), \tag{18}$$

where $E[H(z_i, R_i; \alpha)] = 0$ and $E[H(z_i, R_i; \alpha) H(z_i, R_i; \alpha)^T]$ is positive definite (cf. Kim and Park, 2006). For example, the logistic regression model defined by $\pi_i = \{1 + \exp(-\alpha_1 - \alpha_2 z_i)\}^{-1}$ satisfies

$$H(z_i, R_i; \alpha) = n \{I(\alpha_1, \alpha_2)\}^{-1} (R_i - \pi_i)(1, z_i)^T$$

and

$$I(\alpha_1, \alpha_2) = E\left\{ \sum_{i=1}^{n} \pi_i (1 - \pi_i)(1, z_i)^T (1, z_i) \right\}.$$

The response probability p_i is estimated by $\hat{p} = R^{*-1}T^*$, where $R^* = \sum_{i=1}^n w_i(1-R_i)$ and $T^* = \sum_{i=1}^n w_i(1-R_i)T_i$.

The estimators with the estimated response probabilities plugged-in are defined analogously as before. We define $\hat{Y}_R = \sum_{i=1}^n w_i \hat{\pi}_i^{-1} R_i y_i$ and $\hat{Y}_I = \hat{Y}_T + \hat{r}_1(\hat{X}_{NR} - \hat{X}_T)$, where $\hat{Y}_T = \sum_{i=1}^n w_i(1-\hat{\pi}_i)^{-1}\hat{p}^{-1}(1-R_i)T_i y_i$, $\hat{X}_T = \sum_{i=1}^n w_i(1-\hat{\pi}_i)^{-1}\hat{p}^{-1}(1-R_i)T_i x_i$, $\hat{X}_{NR} = \sum_{i=1}^n w_i(1-\hat{\pi}_i)^{-1}(1-R_i)x_i$ and $\hat{r}_1 = \hat{X}_T^{-1}\hat{Y}_T$. We discuss the asymptotic properties of $\hat{Y}_R$ and $\hat{Y}_I$ in the following Lemma.

Lemma 4.1 *Assume the same conditions as in Lemma 3.2. Let us define* $\Gamma_R = \sum_{i=1}^N \pi_i(\partial \pi_i^{-1}/\partial\alpha) y_i$, $\Gamma_Y = \sum_{i=1}^N p y_i$, $\Gamma_X = \sum_{i=1}^N p x_i$ *and* $\Gamma_T = \sum_{i=1}^N (1-\pi_i)(\partial(1-\pi_i)^{-1}/\partial\alpha) y_i$. *Then,*

$$\hat{Y}_R - \tilde{Y}_R = (\hat{\alpha}-\alpha)^T \Gamma_R + o_P(n^{-1/2}N) \tag{19}$$

and

$$\begin{aligned} \hat{Y}_I - \tilde{Y}_I &= \bar{T}^{-1}[(R^* - \bar{R}) - p^{-1}(T^* - \bar{T})](\Gamma_Y - r\Gamma_X) \\ &\quad + (\hat{\alpha}-\alpha)^T \Gamma_T + o_P(n^{-1/2}N), \end{aligned} \tag{20}$$

where $\bar{T} = \sum_{i=1}^N (1-\pi_i)p$, $\bar{R} = \sum_{i=1}^N (1-\pi_i)$.

Proof. By (18) and Taylor's expansion,

$$\hat{\pi}_i^{-1} - \pi_i^{-1} = (\hat{\alpha}-\alpha)^T(\partial \pi_i^{-1}/\partial\alpha) + o_P(n^{-1/2}).$$

Then, using (A1), (A3) and (A4), we obtain (19). Observe that

$$\begin{aligned} \hat{Y}_I - \tilde{Y}_I &= \hat{Y}_T - \tilde{Y}_T + \hat{r}_1[(\hat{X}_{NR} - \tilde{X}_{NR}) - (\hat{X}_T - \tilde{X}_T)] \\ &\quad + (\hat{r}_1 - \tilde{r}_1)(\tilde{X}_{NR} - \tilde{X}_T). \end{aligned}$$

By Taylor's expansion and the assumed conditions,

$$(1-\hat{\pi}_i)^{-1} - (1-\pi_i)^{-1} = (\hat{\alpha}-\alpha)^T(\partial(1-\pi_i)^{-1}/\partial\alpha) + o_P(n^{-1/2})$$

and

$$\hat{p}^{-1} - p^{-1} = \bar{T}^{-1}[R^* - \bar{R} - p^{-1}(T^* - \bar{T})] + o_P(n^{-1/2}).$$

Then, we have

$$\begin{aligned} &\hat{Y}_T - \tilde{Y}_T \\ &= \bar{T}^{-1}[R^* - \bar{R} - p^{-1}(T^* - \bar{T})]\Gamma_Y + (\hat{\alpha}-\alpha)^T\Gamma_T + o_P(n^{-1/2}N) \end{aligned}$$

and

$$\begin{aligned} &\hat{X}_T - \tilde{X}_T \\ &= \bar{T}^{-1}[R^* - \bar{R} - p^{-1}(T^* - \bar{T})]\Gamma_X + (\hat{\alpha}-\alpha)^T\Gamma_{TX} + o_P(n^{-1/2}N), \end{aligned}$$

where $\Gamma_{TX} = \sum_{i=1}^N (1-\pi_i)(\partial(1-\pi_i)^{-1}/\partial\alpha)x_i$. We also obtain that

$$\hat{X}_{NR} - \tilde{X}_{NR} = (\hat{\alpha}-\alpha)^T\Gamma_{TX} + o_P(n^{-1/2}N)$$

and

$$\hat{r}_1 - \tilde{r}_1 = \tilde{X}_T^{-1}[(\hat{Y}_T - \tilde{Y}_T) - \tilde{r}(\hat{X}_T - \tilde{X}_T)] + o_P(n^{-1/2}).$$

These, together with (4) and the fact that $\tilde{X}_{NR} - \tilde{X}_T = O_P(n^{-1/2}N)$, immediately imply (20). **Q.E.D.**

Let $\hat{\pi}_i^{(k)} = \pi(z_i; \hat{\alpha}^{(k)})$ be the k-th replicate of $\hat{\pi}_i$, where $\hat{\alpha}^{(k)} = (\hat{\alpha}_1^{(k)}, \cdots, \hat{\alpha}_p^{(k)})^T$ is the k-th replicate of $\hat{\alpha}$ satisfying

$$\sum_{k=1}^{L} c_k(\hat{\alpha}^{(k)} - \hat{\alpha})(\hat{\alpha}^{(k)} - \hat{\alpha})^T = \Sigma_\alpha + o_P(n^{-1}) \tag{21}$$

for $\Sigma_\alpha = E[(\hat{\alpha} - E(\hat{\alpha}))(\hat{\alpha} - E(\hat{\alpha}))^T]$. Note that the replication variance estimators can be defined analogously as before, where the response probabilities are replaced by the k-th replicates of the estimated ones. For example,

$$\hat{V}(\hat{Y}_R) = \sum_{k=1}^{L} c_k(\hat{Y}_R^{(k)} - \hat{Y}_R)^2,$$

where $\hat{Y}_R^{(k)} = \sum_{i=1}^{n} w_i^{(k)} \hat{\pi}_i^{(k)-1} R_i y_i$. The covariance of $\hat{Y}_R$ and $\hat{Y}_I$ is estimated by

$$\hat{C}(\hat{Y}_R, \hat{Y}_I) = \sum_{k=1}^{L} c_k(\hat{Y}_R^{(k)} - \hat{Y}_R)(\hat{Y}_I^{(k)} - \hat{Y}_I),$$

where $\hat{Y}_I^{(k)} = \hat{Y}_T^{(k)} + \hat{r}_1^{(k)}(\hat{X}_{NR}^{(k)} - \hat{X}_T^{(k)})$ with $\hat{Y}_T^{(k)}$, $\hat{X}_T^{(k)}$, $\hat{X}_{NR}^{(k)}$ and $\hat{r}_1^{(k)}$ analogously defined using the k-th replicates of the estimated response probabilities. Finally, the proposed estimator is given by

$$\hat{Y}_C = \hat{\phi}\hat{Y}_R + (1 - \hat{\phi})\hat{Y}_I$$

for $\hat{\phi} = [\hat{V}(\hat{Y}_R) + \hat{V}(\hat{Y}_I) - 2\hat{C}(\hat{Y}_R, \hat{Y}_I)]^{-1}[\hat{V}(\hat{Y}_I) - \hat{C}(\hat{Y}_R, \hat{Y}_I)]$.

We discuss the consistency of $\hat{V}(\hat{Y}_R)$ in the following Lemma.

Lemma 4.2 *Assume the conditions of Lemma 3.2. Assume also that*

$$\sum_{k=1}^{L} c_k(\hat{Y}_n^{(k)} - \hat{Y}_n)(\hat{\alpha}_l^{(k)} - \hat{\alpha}_l) = Cov(\hat{Y}_n, \hat{\alpha}_l) + o_P(n^{-1}N) \tag{22}$$

for $1 \leq l \leq p$. *Then,*

$$\hat{Y}_R^{(k)} - \hat{Y}_R = \tilde{Y}_R^{(k)} - \tilde{Y}_R + (\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_R + o_P(n^{-1/2}N) \tag{23}$$

and

$$\hat{V}(\hat{Y}_R) = Var(\hat{Y}_R) + o_P(n^{-1}N^2). \tag{24}$$

Proof. Write $\hat{Y}_R^{(k)} - \hat{Y}_R$ as

$$\hat{Y}_R^{(k)} - \hat{Y}_R = (\hat{Y}_R^{(k)} - \tilde{Y}_R^{(k)}) + (\tilde{Y}_R^{(k)} - \tilde{Y}_R) + (\tilde{Y}_R - \hat{Y}_R).$$

From (18) and (21),

$$\hat{\pi}_i^{(k)-1} - \hat{\pi}_i^{-1} = (\hat{\alpha}^{(k)} - \hat{\alpha})^T (\partial \pi_i^{-1} / \partial \alpha) + o_P(n^{-1/2})$$

which, together with the assumed conditions, implies that

$$(\hat{Y}_R^{(k)} - \tilde{Y}_R^{(k)}) + (\tilde{Y}_R - \hat{Y}_R) = (\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_R + o_P(n^{-1/2} N)$$

and, hence, (23). Observe that by (A5)

$$\begin{aligned} \sum_{k=1}^{L} c_k (\hat{Y}_R^{(k)} - \hat{Y}_R)^2 &= \sum_{k=1}^{L} c_k (\tilde{Y}_R^{(k)} - \tilde{Y}_R)^2 + \sum_{k=1}^{L} c_k [(\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_R]^2 \\ &\quad + 2 \sum_{k=1}^{L} c_k (\tilde{Y}_R^{(k)} - \tilde{Y}_R)[(\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_R] + o_P(n^{-1} N^2). \end{aligned}$$

From (18) and (22),

$$\begin{aligned} &\sum_{k=1}^{L} c_k (\tilde{Y}_R^{(k)} - \tilde{Y}_R)(\hat{\alpha}_l^{(k)} - \hat{\alpha}_l) \Gamma_{Rl} \\ &\quad = Cov(\tilde{Y}_R, n^{-1} \Gamma_{Rl} \sum_{i=1}^{n} H_{li} | R_1, \cdots, R_N) + o_P(n^{-1} N^2) \end{aligned}$$

where Γ_{Rl} is the lth element of Γ_R for $l = 1, \ldots, p$. From (9), we have

$$Cov[E(\tilde{Y}_R | R_1, \cdots, R_N), E(n^{-1} \Gamma_{Rl} \sum_{i=1}^{n} H_{li} | R_1, \cdots, R_N)] = o(n^{-1} N^2).$$

The similar arguments in Lemma 1 lead to

$$\sum_{k=1}^{L} c_k (\tilde{Y}_R^{(k)} - \tilde{Y}_R)[(\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_R] = Cov(\tilde{Y}_R, (\hat{\alpha} - \alpha)^T \Gamma_R) + o_P(n^{-1} N^2),$$

which, together with (10), (19) and (21), implies (24). **Q.E.D.**

We state the following lemma regarding the consistency of $\hat{V}(\hat{Y}_I)$ without proof since the technical details resembles that in Lemma 4.2.

Lemma 4.3 *Assume the same conditions as in Lemma 4.2. Then,*

$$\begin{aligned} &\hat{Y}_I^{(k)} - \hat{Y}_I \\ &= (\tilde{Y}_I^{(k)} - \tilde{Y}_I) + \bar{T}^{-1}[(R^{*(k)} - R^*) - p^{-1}(T^{*(k)} - T^*)](\Gamma_Y - r\Gamma_X) \\ &\quad + (\hat{\alpha}^{(k)} - \hat{\alpha})^T \Gamma_T + o_P(n^{-1/2} N) \end{aligned} \quad (25)$$

and

$$\hat{V}(\hat{Y}_I) = \sum_{k=1}^{L} c_k (\hat{Y}_I^{(k)} - \hat{Y}_I)^2 = Var(\hat{Y}_I) + o_P(n^{-1} N^2), \quad (26)$$

where $R^{*(k)} = \sum_{i=1}^{n} w_i^{(k)} (1 - R_i)$ *and* $T^{*(k)} = \sum_{i=1}^{n} w_i^{(k)} (1 - R_i) T_i$.

The variance estimator for $\hat{Y}_C$ is defined by

$$\hat{V}(\hat{Y}_C) = \sum_{k=1}^{L} c_k(\hat{Y}_C^{(k)} - \hat{Y}_C)^2,$$

where $\hat{Y}_C^{(k)} = \hat{\phi}\hat{Y}_R^{(k)} + (1-\hat{\phi})\hat{Y}_I^{(k)}$. We deal with the consistency of the variance estimator $\hat{V}(\hat{Y}_C)$ in the following Theorem.

Theorem 4.1 *Under the same conditions as in Lemma 4.2,*

$$E(\hat{Y}_C) = Y + o(n^{-1/2}N) \tag{27}$$

and

$$\hat{V}(\hat{Y}_C) = Var(\hat{Y}_C) + o_P(n^{-1}N^2). \tag{28}$$

Proof. Using the same techniques of Theorem 3.1 and Lemma 4.2, we can obtain that

$$\sum_{k=1}^{L} c_k(\hat{Y}_R^{(k)} + \hat{Y}_I^{(k)} - \hat{Y}_R - \hat{Y}_I)^2 = Var(\hat{Y}_R + \hat{Y}_I) + o_P(n^{-1}N^2)$$

which, together with (24) and (26), implies that

$$\hat{C}(\hat{Y}_R, \hat{Y}_I) = Cov(\hat{Y}_R, \hat{Y}_I) + o_p(n^{-1}N^2)$$

and

$$\hat{\phi} - \phi = o_P(1)$$

for $\phi = [Var(\hat{Y}_R) + Var(\hat{Y}_I) - 2Cov(\hat{Y}_R, \hat{Y}_I)]^{-1}[Var(\hat{Y}_I) - Cov(\hat{Y}_R, \hat{Y}_I)]$. We can write that

$$\hat{Y}_C = (\hat{\phi} - \phi)(\hat{Y}_R - Y) + [(1-\hat{\phi}) - (1-\phi)](\hat{Y}_I - Y) + \phi\hat{Y}_R + (1-\phi)\hat{Y}_I.$$

Then, using (19) and (20), we obtain that

$$\hat{Y}_C = \phi\hat{Y}_R + (1-\phi)\hat{Y}_I + o_P(n^{-1/2}N),$$

which implies (27). Furthermore, it is easily obtained from (23) and (25) that

$$\hat{Y}_C^{(k)} - \hat{Y}_C = \phi(\hat{Y}_R^{(k)} - \hat{Y}_R) + (1-\phi)(\hat{Y}_I^{(k)} - \hat{Y}_I) + o_P(n^{-1/2}N).$$

Finally, using (24) and (26), we obtain (28). **Q.E.D.**

5 Simulation results

In this section, we provide the results of a limited simulation study performed to test our theory. In the simulation study, $B = 1,000$ samples of size $n = 100$ are generated by

$$y_i = \beta x_i + \sqrt{x_i}\epsilon_i,$$

where $x_i \sim Uniform(0,1)$, $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$, and x_i and ϵ_i are independent. We simulate various σ^2 for $\beta = 4, 8$. For the response probability under the first survey, we use the

Table 1: $(\beta = 4, \sigma^2 = 1) MSE(\bar{Y}_D)/MSE(\tilde{Y}_C)$ and $MSE(\bar{Y}_D)/MSE(\hat{Y}_C)$

π_i	p 0.3	0.5	0.7	0.9
0.28	1.228	1.140	1.003	0.913
$(0, -2)$	1.282	1.169	1.029	0.935
0.44	1.288	1.241	1.139	1.060
$(-0.5, 0.5)$	1.252	1.181	1.081	1.006
0.5	1.207	1.174	1.095	1.019
$(-0.5, 1)$	1.198	1.144	1.063	0.995
0.69	1.069	1.076	1.028	0.973
$(1.3, -1)$	1.099	1.094	1.051	1.001

logistic model $\pi_i = [1 + \exp(-\alpha_1 - \alpha_2 z_i)]^{-1}$, where $z_i \sim Uniform(0, 1)$ and the value of α is assumed to be (α_1, α_2) =(0,-2), (-0.5,0.5), (-0.5,1) and (1.3,-1). Thus, the overall response rate becomes 0.28, 0.44, 0.50 and 0.69, respectively. For the response probability p under callback, we use constants 0.3, 0.5, 0.7 and 0.9.

We use the maximum likelihood method to estimate α and compute the value iteratively using the Newton-Rapshon method. The response probability p_i under callback is estimated as the response rate among nonresponses. For the variance estimator, we use the standard jackknife method, where c_k is $n^{-1}(n-1)$ and $w_i^{(k)}$ is defined as $(n-1)^{-1} n w_i$ for $i \neq k$ and 0 for $i = k$.

To survey the properties of the variance estimator, we calculate the relative mean and t-statistic. The relative mean of the variance estimator is the empirical mean of the variance estimator divided by the empirical variance of the point estimator. The t-statistic for the variance estimator is the empirical bias of variance estimator divided by the empirical standard error of the empirical bias, which was considered by Kim (2004).

Using B samples of $\{(y_i, x_i, \epsilon_i, R_i, T_i);\ i = 1, ..., n\}$ and $w_i = n^{-1}$, we computed the empirical values of $MSE(\bar{Y}_D)/MSE(\tilde{Y}_C)$ and $MSE(\bar{Y}_D)/MSE(\hat{Y}_C)$, where $MSE(\bar{Y}_D)$ is the mean square error of the Deming's estimator. We also computed the relative means and t-statistics for $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$. Each cell in Tables 1, 3 and 5 contain $MSE(\bar{Y}_D)/MSE(\tilde{Y}_C)$ and $MSE(\bar{Y}_D)/\ MSE(\hat{Y}_C)$ in this order for various response probabilities π_i and p. Each cell in Tables 2, 4 and 6 show the relative means (t-statistics in parentheses) of $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$ in this order for varying response probabilities.

In real survey, we can sometimes find that response rate after callbacks is small. The data come from an experimental sampling of fruit orchards in North Carolina in 1946. Three successive mailings of the same questionnaire were sent to growers. Response rate of first mailing is 10%, response rate of second mailing is 17% and response rate of third mailing is 14%, that is, nonresponse rate after three mailings is 59%(cf. Finkner, 1950).

As anticipated, it is observed in Tables 1, 3 and 5 that $MSE(\tilde{Y}_C)$ and $MSE(\hat{Y}_C)$ are smaller than $MSE(\bar{Y}_D)$ for various small response probabilities π_i and p. We see that the efficiency of $\tilde{Y}_C$ and $\hat{Y}_C$ in Table 3 is better than that of $\tilde{Y}_C$ and $\hat{Y}_C$ in Table 5, because explanation of regression

Table 2: ($\beta = 4, \sigma^2 = 1$)Relative mean (t-statistic) of $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$

π_i	p			
	0.3	0.5	0.7	0.9
0.28	1.069 (1.625)	1.001 (0.018)	1.011 (0.270)	1.022 (0.550)
$(0, -2)$	1.003 (0.065)	0.931 (-1.567)	0.936 (-1.468)	0.939 (-1.423)
0.44	1.078 (1.631)	1.018 (0.400)	1.019 (0.422)	1.020(0.452)
$(-0.5, 0.5)$	0.991 (-0.189)	0.964 (-0.793)	0.972 (-0.634)	0.976(-0.563)
0.5	1.086 (1.731)	1.028 (0.610)	1.029 (0.654)	1.034(0.741)
$(-0.5, 1)$	0.989 (-0.227)	0.967 (-0.733)	0.975 (-0.575)	0.985(-0.339)
0.69	1.063 (1.437)	0.995 (-0.110)	0.999 (-0.020)	1.002(0.053)
$(1.3, -1)$	0.970 (-0.678)	0.973 (-0.609)	0.997 (-0.065)	1.010(0.241)

Table 3: ($\beta = 8, \sigma^2 = 1$)$MSE(\bar{Y}_D)/MSE(\tilde{Y}_C)$ and $MSE(\bar{Y}_D)/MSE(\hat{Y}_C)$

π_i	p			
	0.3	0.5	0.7	0.9
0.28	1.481	1.236	1.021	0.900
$(0, -2)$	1.595	1.321	1.095	0.965
0.44	1.553	1.356	1.192	1.075
$(-0.5, 0.5)$	1.468	1.285	1.133	1.021
0.5	1.435	1.272	1.141	1.029
$(-0.5, 1)$	1.392	1.247	1.121	1.015
0.69	1.184	1.125	1.038	0.967
$(1.3, -1)$	1.206	1.155	1.063	0.993

Table 4: ($\beta = 8, \sigma^2 = 1$)Relative mean (t-statistic) of $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$

π_i	p			
	0.3	0.5	0.7	0.9
0.28	1.030 (0.725)	1.002 (0.042)	1.008 (0.191)	1.012 (0.293)
$(0, -2)$	0.983 (-0.390)	0.942 (-1.328)	0.946 (-1.249)	0.946 (-1.239)
0.44	1.049 (1.053)	1.015 (0.345)	1.016 (0.358)	1.015 (0.345)
$(-0.5, 0.5)$	0.989 (-0.240)	0.969 (-0.708)	0.973 (-0.627)	0.973 (-0.623)
0.5	1.068 (1.423)	1.033 (0.717)	1.033 (0.735)	1.034 (0.763)
$(-0.5, 1)$	0.996 (-0.099)	0.981 (-0.442)	0.984 (-0.376)	0.989 (-0.258)
0.69	1.018 (0.427)	0.986 (-0.351)	0.987 (-0.326)	0.987 (-0.301)
$(1.3, -1)$	0.983 (-0.389)	0.982 (-0.404)	0.984 (-0.365)	0.988 (-0.288)

Table 5: $(\beta = 8, \sigma^2 = 4.5) MSE(\bar{Y}_D)/MSE(\tilde{Y}_C)$ and $MSE(\bar{Y}_D)/MSE(\hat{Y}_C)$

π_i	p			
	0.3	0.5	0.7	0.9
0.28	1.203	1.129	1.001	0.914
$(0, -2)$	1.254	1.153	1.022	0.932
0.44	1.262	1.228	1.133	1.058
$(-0.5, 0.5)$	1.233	1.169	1.075	1.004
0.5	1.184	1.162	1.089	1.018
$(-0.5, 1)$	1.181	1.133	1.057	0.992
0.69	1.058	1.070	1.026	0.973
$(1.3, -1)$	1.089	1.087	1.048	1.001

Table 6: $(\beta = 8, \sigma^2 = 4.5)$Relative mean (t-statistic) of $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$

π_i	p			
	0.3	0.5	0.7	0.9
0.28	1.073 (1.709)	1.001 (0.018)	1.011 (0.277)	1.024 (0.579)
$(0, -2)$	1.004 (0.089)	0.930 (-1.587)	0.936 (-1.484)	0.939 (-1.431)
0.44	1.080 (1.668)	1.019 (0.403)	1.019 (0.427)	1.021 (0.464)
$(-0.5, 0.5)$	0.992 (-0.181)	0.964 (-0.804)	0.972 (-0.634)	0.976 (-0.555)
0.5	1.087 (1.741)	1.028 (0.595)	1.029 (0.642)	1.033 (0.737)
$(-0.5, 1)$	0.989 (-0.231)	0.965 (-0.765)	0.974 (-0.596)	0.985 (-0.353)
0.69	1.066 (1.514)	0.997 (-0.084)	1.001 (0.015)	1.004 (0.095)
$(1.3, -1)$	0.968 (-0.710)	0.971 (-0.652)	0.997 (-0.060)	1.012 (0.273)

is weak for large σ^2. Because the coefficient of determination is proportional to an estimator of β, $MSE(\tilde{Y}_C)/MSE(\bar{Y}_D)$ and $MSE(\hat{Y}_C)/MSE(\bar{Y}_D)$ in table 3 are larger than those in table 1. The result of Table 2, 4 and 6 shows that all the variance estimators are asymptotically unbiased. Note that variances of the variance estimators are asymptotically zero. In Tables 2, 4 and 6, we can see that $\hat{V}(\tilde{Y}_C)$ and $\hat{V}(\hat{Y}_C)$ are consistent estimators for $Var(\tilde{Y}_C)$ and $Var(\hat{Y}_C)$ for various response probabilities π_i and p , respectively.

6 concluding and remarks

We also propose an estimator using the response probability and the ratio imputation in the response model under callbacks. Using constants W_1 and W_2 with $W_1 = W_2 = 1/2$, we define

$$\tilde{Y}_J = (W_1 \tilde{X}_T^{-1} \tilde{Y}_T + W_2 \tilde{X}_R^{-1} \tilde{Y}_R) \tilde{X}_{NR}$$

and

$$\tilde{\psi} = [Var(\tilde{Y}_R) + Var(\tilde{Y}_J) - 2Cov(\tilde{Y}_R, \tilde{Y}_J)]^{-1}[Var(\tilde{Y}_J) - Cov(\tilde{Y}_R, \tilde{Y}_J)], \tag{29}$$

where $\tilde{X}_R = \sum_{i=1}^n w_i \pi_i^{-1} R_i x_i$.

The estimator $(\tilde{X}_T^{-1} \tilde{Y}_T) \tilde{X}_{NR}$ is estimated by callback and the estimator $(\tilde{X}_R^{-1} \tilde{Y}_R) \tilde{X}_{NR}$ is estimated by imputation. If all π_i are equal, then $\tilde{X}_R^{-1} \tilde{Y}_R x_i$ is the ratio imputation used by Rao and Sitter (1995) and by Rao (1996).

Then, our proposed estimator is defined as

$$\tilde{Y}_M = \tilde{\psi} \tilde{Y}_R + (1 - \tilde{\psi}) \tilde{Y}_J.$$

Note that the variance of $k\tilde{Y}_R + (1-k)\tilde{Y}_J$ is minimized at $k = \tilde{\psi}$. We suggest an estimator using estimated response probability and a variance estimator. We also can find similar properties in $\tilde{Y}_J$ as in $\tilde{Y}_C$.

References

[1] Deming, W.E., 1953. On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse. J. Amer. Statist. Assoc. 48, 743-772.

[2] Ekholm, A., Laaksonen, S., 1991. Weighting via Response Modeling in the Finnish Household Budget Survey. J. Offic. Statist. 7, 325-337.

[3] Elliott, M.R., Little, R.J.A., Lewitzky, S., 2000. Subsampling Callbacks to Improve Survey Efficiency. J. Amer. Statist. Assoc. 95, 730-738.

[4] Fay, R.E., 1991. A Design-based Perspective on Missing Data Variance, In the ASA Proceedings of Bureau of the Census Annual Research Conference, U.S. Bureau of the Census, Washington, D.C. 429-440.

[5] Fuller, W.A., 1996. Introduction to Statistical Time Series, 2nd ed. Wiley, New York.

[6] Groves, R.M., 1989. Survey Errors and Survey Costs. New York: Wiley.

[7] Hansen, M.H., Hurwitz, W.N., 1946. The Problem of Nonresponse in Sample Surveys. J. Amer. Statist. Assoc. 41, 517-529.
Isaki, C.T., Fuller, W.A., 1982. Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. 77, 89-96.

[8] Iannacchione, V.G., 2003. Sequential Weight Adjustment for Location and Cooperation Propensity for the 1995 National Survey of Family Growth. J. Offic. Statist. 19, 31-43.

[9] Kim, J.K., 2004. Finite Sample Properties of Multiple Imputation Estimators. The Ann. Statist. 32, 766-783.

[10] Kim,J.K., Park, H., 2006. Imputation using response probability. The Canadian J. Statist. 34, 171-182.

[11] Lipsitz, S.R., Ibrahim, J.G., Zhao, L.P., 1999. A Weighted Estimating Equation for Missing Covariate Data with Properties similar to Maximum Likelihood. J. Amer. Statist. Assoc. 94, 1147-1160.

[12] Rao, J.N.K., Sitter, R.R., 1995. Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. Biometrika 82, 453-60.

[13] Rao, J.N.K., Shao, J., 1992. Jackknife Variance Estimation with Survey Data under Hot Deck Imputation. Biometrika 79, 811-822.

[14] Rosenbaum, P.R., 1987. Model-Based Direct Adjustment. J. Amer. Statist. Assoc. 82, 387-394.

[15] Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of Regression Coefficients when some Regressors are not always observed. J. Amer. Statist. Assoc. 89, 846-866.

[16] Shao, J., Steel, P., 1999. Variance Estimation for Survey Data with Composite Imputation and Non-Negligible Sampling Fractions. J. Amer. Statist. Assoc. 94, 254-265.