Communications for Statistical Applications and Methods
2022, Vol. 29, No. 2, 263–275

https://doi.org/10.29220/CSAM.2022.29.2.263
Print ISSN 2287-7843 / Online ISSN 2383-4757

A response probability estimation for non-ignorable non-response

Hee Young Chung[a], Key-Il Shin[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract

Use of appropriate technique for non-response occurring in sample survey improves the accuracy of the estimation. Many studies have been conducted for handling non-ignorable non-response and commonly the response probability is estimated using the propensity score method. Recently, post-stratification method to obtain the response probability proposed by Chung and Shin (2017) reduces the effect of bias and gives a good performance in terms of the MSE. In this study, we propose a new response probability estimation method by combining the propensity score adjustment method using the logistic regression model with post-stratification method used in Chung and Shin (2017). The superiority of the proposed method is confirmed through simulation.

Keywords: response probability model, bias estimation, sample distribution, population distribution, post-stratification

1. Introduction

Non-response in sample survey is a common source of non-sampling error that appears when part of the data to be collected is not observed. In the case of missing at random (MAR) in which non-response occurs randomly, many appropriate statistical methods have been developed. On the other hand, in the case of non-ignorable non-response, there are relatively few studies on this subject. Non-ignorable non-response is known to cause bias and so accurate bias estimation is the key to properly handle the non-response. The PSA estimator, propensity-score-adjusted estimator, defined by

$$\widehat{\overline{Y}}_{\mathrm{PSA}} = \frac{1}{N}\sum_{i\in s} w_i \frac{R_i}{\hat{p}_i} y_i, \tag{1.1}$$

is widely used to reduce non-response bias. Here R_i is 1 if response and R_i is 0 if non-response and s is an index set of sample. Also, w_i is the sample weight and $\hat{p}_i$ is the estimated response probability. Kim and Riddles (2012) developed some asymptotic theories of the PSA estimator and suggested minimum variance of the estimator. Kim and Yu (2011) studied a mean estimation method for non-ignorable non-response case in which paper the missing value of the variable of interest is non-parametrically calculated using a kernel estimator. Riddles *et al.* (2016) proposed a propensity score adjustment method for non-ignorable non-response and showed that the proposed method is more efficient than the calibration-weighting method in Chang and Kott (2008) and Kott and Chang (2010).

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT)(NRF-2021R1F1A1045602).

[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin, Gyeonggido 17035, Korea. E-mail: keyshin@hufs.ac.kr

Published 31 March 2022 / journal homepage: http://csam.or.kr

In order to use the PSA estimator, it is necessary to estimate the response probability. Estimating response probabilities relies heavily on the use of model. Iannacchione *et al.* (1991) used a logistic regression model for the response probability estimation and this logistic regression model is widely used. Also Da Silva and Opsomer (2006) and Da Silva and Opsomer (2009) considered nonparametric methods to obtain the response probability. Bethlehem (2020) presented an approximate bias of the sample mean for non-ignorable non-response. Various response probability estimation methods have been addressed in that paper.

In general, the final sample weight considering non-response is used for estimation defined by

$$\widehat{\overline{Y}}_{\text{adj}} = \frac{1}{N}\sum_{i\in s}\widehat{w}_i^F y_i. \tag{1.2}$$

Therefore, the PSA estimator using $\widehat{w}_i^F = w_i/\hat{p}_i$ is one of the non-response adjusted weight estimators. Chung and Shin (2017) suggested a non-response adjusted weight estimation method using a post-stratification which is addressed in Bethlehem (2020). This method reduces the effect of bias and gives a good performance in terms of the MSE.

In this study, we propose a new response probability estimation method that combines the response probability estimation method using logistic regression model with the post-stratification method suggested by Chung and Shin (2017).

The composition of this paper is as follows. In Section 2, the existing methods and a proposed method for the response probability estimation are explained. Section 3 describes the bias corrected PSA estimators using the response probability estimates obtained in Section 2. Section 4 confirms the superiority of the proposed method through simulation studies. There is a conclusion in Section 5.

2. Response probability estimation

2.1. Modeling method

In order to properly handle non-ignorable non-response, the response probability of each data must be appropriately estimated. As mentioned in Bethlehem (2020), various methods can be used for the estimation. The propensity score using the logistic regression model is the most commonly used. Let R_i be the indicator variable representing the response and p_i be the response probability. Also let the auxiliary variable $\mathbf{x_i}$ be a p-dimensional vector and always observable. Usually for non-ignorable non-response, it is assumed that p_i is a function of y_i and $\mathbf{x_i}$. However, in this study, we assume that the response probability p_i is a function of y_i for simplicity.

Then, we can write $p_i = P(R_i = 1|y_i, \mathbf{x_i}) = P(R_i = 1|y_i) = g_2(y_i;\phi)$. Practically, we do not have enough information to estimate p_i and mostly the function is not known. To handle this situation, Kim and Yu (2011) and Kim and Riddles (2012) used some follow-up samples. However, this is not the usual case.

Mostly, the variable of interest is related to the auxiliary variables and therefore many studies consider a super-population model. Let the super-population model be $y_i = g_1(x_i;\beta) + \epsilon_i$ and assume

$$p_i = P\left(R_i = 1|y_i\right) = g_2\left(g_1\left(\mathbf{x_i};\beta\right) + \epsilon_i;\phi\right) \approx m\left(\mathbf{x_i};\theta\right) + \eta_i, \tag{2.1}$$

where θ is a vector of some unknown parameters and η_i is an error related to ϵ_i. Since this study considers non-ignorable non-response, it may not be appropriate to use a function of the auxiliary variable $\mathbf{x_i}$. However, in reality, in order to estimate the response probability, we have no choice but to use the available auxiliary variables. Considering this, equation (2.1) may be useful. However, bias may occur in the result using the response probability obtained in equation (2.1).

The approximate estimate of p_i can be obtained using the available auxiliary variable $\mathbf{x_i}$. The most common response probability estimation method is to use the parametric logistic regression model. Of course, nonparametric methods can be used as Da Silva and Opsomer (2006) and Da Silva and Opsomer (2009). Da Silva and Opsomer (2009) showed that local polynomial regression results in better practical and theoretical properties for the ignorable case. Recently Bethlehem (2020) studied the non-ignorable non-response and used the logistic regression modeling. Since the logistic regression model is commonly used and simple to implement, we use the logistic regression model in this paper. The logistic regression model is defined by

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi}.$$

Using this model, we can obtain the response probability estimate $\hat{p}_i^P$. Now let $\widehat{w}_i^P = w_i/\hat{p}_i^P$. Then the final adjusted sample weight is obtained by following,

$$\widehat{w}_i^{F(P)} = \widehat{w}_i^P \frac{N}{\sum_{j\in s}\widehat{w}_j^P} \tag{2.2}$$

In Riddles *et al.* (2016), (2.2) is called naive estimator for non-ignorable non-response.

2.2. Post-stratification method

Several methods estimating response probability have been developed. One of them is the post-stratification method. As mentioned in Bethlehem (2020), post-stratification is a well-known and frequently used weighting technique. Usually categorical variables are used. Using these variables, population is divided into a number of non-overlapping subpopulations, called strata. Of course, continuous variables can be used and Bethlehem (2020) used the estimated response probability to construct subpopulations. Although (2.1) is an approximate result, naturally we can use the available auxiliary variable in order to construct strata. Chung and Shin (2017) and Min and Shin (2018) proposed a response probability estimation method using strata. In the proposed method by Min and Shin (2018), population is divided into L strata with boundaries obtained using percentiles of given auxiliary variable.

To construct the strata, we determine the total number of strata, L satisfying about $r/L \geq 10$ where r is the total number of response samples. Then with L, we calculate the percentiles of the auxiliary variables and use the percentiles as boundaries. For instance, with $L = 10$, we use $10, 20, \ldots, 90$ percentiles for boundaries. For univariate case, the percentiles become directly boundaries. However, for multivariate case, we use the Cartesian product with the percentiles of each auxiliary variable to meet the total number of strata L.

Now, let N_h and r_h be the number of population and the number of final response samples in the h^{th} stratum, respectively. Then the sample weight of y_i in the h^{th} stratum using the post-stratification method is defined by

$$\widehat{w}_i^{F(D)} = \sum_{h=1}^{L} \frac{N_h}{r_h} I(i \in s_h), \tag{2.3}$$

where s_h is the index set of h^{th} stratum sample. Since this method does not use any model to estimate the response probability, it is model free and naturally robust of model misspecification.

2.3. Proposed response probability estimation method

The widely used $\widehat{w}_i^{F(P)}$ is obtained using the individual data values. However, since it is based on a model, the results may vary depending on the model used. On the other hand, $\widehat{w}_i^{F(D)}$ has the advantage of being easy to calculate because it uses the number of samples in a given stratum rather than the individual data values. Again, the result may depend on the method of constructing the strata and the optimal boundaries of the strata are determined according to the given data.

A new method combining two methods explained in section 2.1 and 2.2 is proposed. First we calculate $\widehat{w}_i^{F(P)}$ explained in section 2.1. Then we make sum of $\widehat{w}_j^{F(P)}$ in the h^{th} stratum be N_h. That is, the final sample weight is defined by

$$\widehat{w}_i^{F(C)} = \widehat{w}_i^P \sum_{h=1}^{L} \frac{N_h}{\sum_{j\in s_h} \widehat{w}_j^P} I(i \in s_h), \tag{2.4}$$

where N_h and s_h are the same as those defined in (2.3).

3. Bias corrected propensity-score-adjusted (PSA) estimator

In the case of non-ignorable non-response, since the response probability is a function of the variable of interest, it is necessary to estimate the response probability using the variable of interest. However, as mentioned before it is practically hard to obtain the response probability using the variable of interest. In Section 2, the response probability is estimated using the available auxiliary variables and this usually produces bias in estimation. To obtain better estimation results, the bias should be estimated and corrected.

3.1. Bias estimation

For bias estimation, this section considers a super-population model. Let the inclusion probability, π_i be a function of the variable of interest y_i and $f_p(y_i|\mathbf{x_i})$ be a population distribution with a super-population model. Pfeffermann *et al.* (1998) showed $f_s(y_i|\theta^*, \mathbf{x_i}) = f(y_i|i \in s, \mathbf{x_i}) = \{Pr(i \in s|y_i, \mathbf{x_i}) f_p(y_i|\theta, \mathbf{x_i})\} / \{Pr(i \in s|\mathbf{x_i})\}$ where θ^* is a function of θ. Since $Pr(i \in s|y_i, \mathbf{x_i}) = E_p(\pi_i|y_i, \mathbf{x_i})$ and $Pr(i \in s|\mathbf{x_i}) = E_p(\pi_i|\mathbf{x_i})$, finally we have

$$f_s(y_i|\mathbf{x_i}) = \frac{E_p(\pi_i|y_i, \mathbf{x_i}) f_p(y_i|\mathbf{x_i})}{E_p(\pi_i|\mathbf{x_i})}, \tag{3.1}$$

where $f_s(y_i|\mathbf{x_i})$ is a sample distribution, $E_p(\pi_i|y_i, \mathbf{x_i})$ is the sample inclusion probability given $\mathbf{x_i}$, y_i and $E_p(\pi_i|\mathbf{x_i})$ is the sample inclusion probability given $\mathbf{x_i}$. Whenever $E_p(\pi_i|y_i, \mathbf{x_i}) = E_p(\pi_i|\mathbf{x_i})$, then population distribution and sample distribution are the same. Therefore, with known inclusion probability and population distribution we can obtain the sample distribution and finally, we can calculate the bias. Bias estimation using various inclusion probabilities and population distributions have been obtained in previous studies.

In this study we consider a non-ignorable non-response with a non-informative sampling. Actually the final inclusion probability π_i is obtained by combining the inclusion probability at the time of sampling design and the response probability. As a non-informative sampling design, we use simple random sampling and so w_i in (1.1) is constant. Therefore π_i used in (3.1) is defined by $\pi_i = p_i/w_i = p_i/w$ where $w = N/n$. Since p_i is a function of the variable of interest y_i, we have $f_s(y_i|\mathbf{x_i}) \neq f_p(y_i|\mathbf{x_i})$ and so we can apply the bias estimation method developed in the informative sampling design to non-ignorable non-response.

We consider a linear inclusion probability model whose bias is easily estimated theoretically. This is because, if the linear inclusion probability model is effective, it can be expected to be effective when using other inclusion probability models. The linear inclusion probability model considered is as follows,

$$E_p\left(\pi_i|y_i, \mathbf{x_i}\right) = E_p\left(\pi_i|y_i\right) = b_0 + b_1 y_i. \tag{3.2}$$

Then simply we have

$$E_p\left(\pi_i|\mathbf{x_i}\right) = E\left(E_p\left(\pi_i|y_i, \mathbf{x_i}\right)|\mathbf{x_i}\right) = E\left(b_0 + b_1 y_i|\mathbf{x_i}\right) = b_0 + b_1 E_p\left(y_i|\mathbf{x_i}\right), \tag{3.3}$$

$$\frac{E_p\left(\pi_i|y_i, \mathbf{x_i}\right)}{E_p\left(\pi_i|\mathbf{x_i}\right)} = \frac{b_0 + b_1 y_i}{b_0 + b_1 E_p\left(y_i|\mathbf{x_i}\right)}. \tag{3.4}$$

Let $E_s(y_i|\mathbf{x_i}) = \mu_i^{(s)}$. Then plugging (3.4) into (3.1) we have,

$$\mu_i^{(s)} = \int y_i \left(\frac{b_0 + b_1 y_i}{b_0 + b_1 E_p\left(y_i|\mathbf{x_i}\right)}\right) f_p\left(y_i|\mathbf{x_i}\right) dy_i = \frac{b_0\mu_i + b_1 E_p\left(y_i^2|\mathbf{x_i}\right)}{b_0 + b_1\mu_i}, \tag{3.5}$$

where $\mu_i = E_p(y_i|\mathbf{x_i})$. Since $E_p(y_i^2|\mathbf{x_i}) = \mathrm{Var}_p(y_i|\mathbf{x_i}) + \mu_i^2$, we have $\mu_i^{(s)} = \mu_i + b_1/b_0 + b_1\mu_i \times \mathrm{Var}_p(y_i|\mathbf{x_i})$ and finally the bias is obtained as following,

$$\frac{b_1}{b_0 + b_1\mu_i} \times \mathrm{Var}_p\left(y_i|\mathbf{x_i}\right). \tag{3.6}$$

3.2. Parameter estimation for inclusion probability model

Bias is estimated based on the super-population model and the inclusion probability model established.

In (3.2), $E_p(\pi_i|y_i, \mathbf{x_i})$ can be calculated using known parameters b_0, b_1 and given y_i for the informative sampling. However, since we consider the non-ignorable non-response situation with non-informative sampling scheme, this situation usually can not happen in practice. The parameters b_0, b_1 should be estimated from the data.

In this study we use the linear probability model (3.2) defined by $E_p(\pi_i|y_i, \mathbf{x_i}) = b_0 + b_1 y_i$. Since $E_p(\pi_i|y_i, \mathbf{x_i}) = 1/E_s(w_i|y_i, \mathbf{x_i})$ and $E_s(w_i|y_i, \mathbf{x_i}) \approx w_i$ from Pffeffermann and Sverchkov (2003) we have $1/w_i \approx b_0 + b_1 y_i$. Therefore by plugging the estimated final sample weight $\widehat{w}_i^F$ obtained in Section 2.1 into w_i, we have the following model.

$$\frac{1}{\widehat{w}_i^F} \approx b_0 + b_1 y_i. \tag{3.7}$$

So using simple regression analysis we have the estimates of b_0, b_1.

3.3. Parameter estimation for super-population model

Three super-population models are considered in this study.

1. Normal distribution

$$f_p\left(y_i|\mathbf{x_i}\right) = N\left(\beta_0 + \beta_1^T \mathbf{x_i}, \sigma^2\right). \tag{3.8}$$

That is, $y_i = \beta_0 + \beta_1^T \mathbf{x_i} + \epsilon_i$ where β_1^T is a p-dimensional vector of parameters, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\mathrm{Var}_p(y_i|\mathbf{x_i}) = \sigma^2$.

2. Gamma distribution with log-linear model

$$f_p(y_i|\alpha_i, \mu_i) \propto y_i^{\alpha-1} \exp\left(-\alpha \frac{y_i}{\mu_i}\right), \tag{3.9}$$

where $\mu_i = E_p(y_i|\mathbf{x_i})$ and $\ln(\mu_i) = \beta_0 + \beta_1^T \mathbf{x_i}$. That is, we use $\Gamma(\alpha, \beta^*)$ with $\beta^* = \mu_i/\alpha$ and so we have $\mathrm{Var}_p(y_i|\mathbf{x_i}) = \mu_i^2/\alpha$.

3. Log-normal distribution with log-linear model

$$f_p(y_i|\mathbf{x_i}) = \mathrm{LN}\left(\beta_0 + \beta_1 \mathbf{x_i}, \sigma^2\right), \tag{3.10}$$

where $\mu_i = E_p(y_i|\mathbf{x_i}) = \exp(\beta_0 + \beta_1^T \mathbf{x_i} + \sigma^2/2)$ and $\mathrm{Var}_p(y_i|\mathbf{x_i}) = \mu_i^2(\exp(\sigma^2) - 1)$.

As used in Riddles *et al.* (2016), using the obtained data $(\mathbf{x_i}, y_i)$ and the known super-population model, we can estimate μ_i and $\mathrm{Var}_p(y_i|\mathbf{x_i})$.

3.4. Bias corrected propensity-score-adjusted (PSA) estimator

Bias corrected PSA estimator can be obtained based on the bias estimates and the PSA estimator defined in (1.2). That is, the bias corrected PSA estimator is defined by

$$\widehat{\overline{Y}}^{\mathrm{BC}} = \frac{1}{N} \sum_{i \in s} \widehat{w}_i^F \left(y_i - \widehat{\mathrm{bias}}_i\right). \tag{3.11}$$

In (3.11) we can use the final sample weights, $\widehat{w}_i^{F(P)}$, $\widehat{w}_i^{F(D)}$ and $\widehat{w}_i^{F(C)}$ defined in Section 2 and $\widehat{\mathrm{bias}}_i$ estimated according to the super-population model. Therefore, we have three bias corrected PSA estimators depending on the final sample weight,

$$\begin{aligned}
\widehat{\overline{Y}}_P^{\mathrm{BC}} &= \frac{1}{N} \sum_{i \in s} \widehat{w}_i^{F(P)} \left(y_i - \widehat{\mathrm{bias}}_i\right), \\
\widehat{\overline{Y}}_D^{\mathrm{BC}} &= \frac{1}{N} \sum_{i \in s} \widehat{w}_i^{F(D)} \left(y_i - \widehat{\mathrm{bias}}_i\right), \\
\widehat{\overline{Y}}_C^{\mathrm{BC}} &= \frac{1}{N} \sum_{i \in s} \widehat{w}_i^{F(C)} \left(y_i - \widehat{\mathrm{bias}}_i\right).
\end{aligned}$$

4. Simulation studies

To investigate the finite sample properties of the proposed method, we perform simulation studies. In simulation, for $x_{1i} \overset{iid}{\sim} \mathrm{Unif}(100, 200)$ and $x_{2i} \overset{iid}{\sim} \mathrm{Unif}(30, 500)$ we generate a variable of interest y_i according to the following error distribution of the super-population model,

- Normal distribution: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, 900)$.

Here we use $\beta_0 = 10, \beta_1 = 5$ for univariate auxiliary variable case and $\beta_0 = 10, \beta_1 = 5, \beta_2 = 3$ for bivariate auxiliary variable case.

- Gamma distribution: $y_i \overset{iid}{\sim} \Gamma(10, \mu_i/10)$.

  Here, we use $\mu_i = \exp(0.01 + 0.03x_{1i})$ for univariate auxiliary variable case and $\mu_i = \exp(0.01 + 0.03x_{1i} + 0.002x_{2i})$ for bivariate auxiliary variable case.

- Log-normal distribution: $y_i \overset{iid}{\sim} \mathrm{LN}(\mu_i, 0.1)$ Here we use $\mu_i = 0.01 + 0.03x_{1i}$ for univariate auxiliary variable case and $\mu_i = 0.01 + 0.03x_{1i} + 0.002x_{2i}$ for bivariate auxiliary variable case.

From $N = 10,000$ population data, $n = 500$ samples are drawn using simple random sampling. In order to remove the influence of outliers, 10,100 data are generated and then 100 largest valued data are deleted in the samples of the gamma distribution and log-normal distribution.
We consider three response probability models,

- exponential response: $p_i = \exp(a_0 + a_1 y_i)$
- linear response: $p_i = b_0 + b_1 y_i$
- logistic response: $\ln\left(\frac{p_i}{1-p_i}\right) = c_0 + c_1 y_i$

For the linear response probability model, $p_i = b_0 + b_1 y_i$, $p_i \in [0, 1]$, let $\pi_y^{\min}$ and $\pi_y^{\max}$ denote the response probabilities corresponding to the minimum and maximum values of y_i, respectively. Using given response probabilities, for instance $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.3)$, we calculate b_0, b_1 and then calculate p_i. We consider four cases of response probabilities,

- Case I $(\pi_y^{\min}, \pi_y^{\max}) = (0.9, 0.3)$,
- Case II : $(\pi_y^{\min}, \pi_y^{\max}) = (0.8, 0.3)$,
- Case III : $(\pi_y^{\min}, \pi_y^{\max}) = (0.3, 0.8)$,
- Case IV : $(\pi_y^{\min}, \pi_y^{\max}) = (0.3, 0.9)$.

Response data are obtained according to the calculated response probability p_i. Similarly, the exponential response probability model and the logistic response probability model are used to generate response data.

Three PSA estimators, $\widehat{\overline{Y}}_P$, $\widehat{\overline{Y}}_D$ and $\widehat{\overline{Y}}_C$ are calculated according to the sample weight estimation method defined in (2.2), (2.3) and (2.4). In (2.3) and (2.4) for univariate case, the number of strata $L = 20$ is used. For bivariate case, we use $L_1 = 5$ for x_{1i} and $L_2 = 4$ for x_{2i}. Also three bias corrected PSA estimators, $\widehat{\overline{Y}}_P^{\mathrm{BC}}$, $\widehat{\overline{Y}}_D^{\mathrm{BC}}$ and $\widehat{\overline{Y}}_C^{\mathrm{BC}}$ defined in section 3.4 are calculated. The performance of the estimators is compared using the following comparison statistics; bias, absolute relative bias (ARB),

Table 1: Results of PSA estimator with Normal distribution

Case	Estimator	Exponential			Linear			Logistic		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P$	–0.319	0.007	6.608	–0.889	0.007	6.597	–1.233	0.007	6.632
	$\widehat{\overline{Y}}_D$	–1.557	0.003	2.518	–1.408	0.003	2.332	–1.521	0.003	2.370
	$\widehat{\overline{Y}}_C$	–1.476	0.003	2.468	–1.360	0.002	2.304	–1.443	0.003	2.320
II	$\widehat{\overline{Y}}_P$	–0.530	0.007	6.606	–0.972	0.007	6.602	–1.096	0.007	6.629
	$\widehat{\overline{Y}}_D$	–1.394	0.003	2.473	–1.301	0.003	2.334	–1.331	0.003	2.336
	$\widehat{\overline{Y}}_C$	–1.323	0.003	2.433	–1.233	0.002	2.297	–1.261	0.002	2.297
III	$\widehat{\overline{Y}}_P$	0.989	0.007	6.638	1.455	0.007	6.670	1.585	0.007	6.715
	$\widehat{\overline{Y}}_D$	1.436	0.003	2.446	1.427	0.003	2.374	1.468	0.003	2.389
	$\widehat{\overline{Y}}_C$	1.365	0.003	2.407	1.360	0.002	2.336	1.398	0.003	2.348
IV	$\widehat{\overline{Y}}_P$	0.772	0.007	6.626	1.321	0.007	6.687	1.655	0.007	6.694
	$\widehat{\overline{Y}}_D$	1.636	0.003	2.523	1.564	0.003	2.418	1.590	0.003	2.393
	$\widehat{\overline{Y}}_C$	1.554	0.003	2.473	1.489	0.003	2.371	1.513	0.003	2.345

and root mean squared error (RMSE) defined by

$$\begin{aligned}
\text{Bias} &= \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\overline{Y}}_r - \overline{Y}_r\right), \\
\text{ARB} &= \frac{1}{R}\sum_{r=1}^{R}\frac{\left|\widehat{\overline{Y}}_r - \overline{Y}_r\right|}{\overline{Y}_r}, \\
\text{RMSE} &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\overline{Y}}_r - \overline{Y}_r\right)^2}.
\end{aligned}$$

Here, $R = 1,000$ is used and the comparison statistics are calculated by generating a new population for each iteration. This is to reduce the influence of the specific population generated, so the true value of the r^{th} repeat population mean is denoted as $\bar{Y}_r$.

Tables 1 to 3 contain the results of the PSA estimators. Table 1 shows the results when the super-population model is linear and the error distribution is normal with three response probability models. Even though there are some differences of response rates depending on the response probability cases, in Case I and IV, the response rate of about $53-65\%$ is obtained, and in Case II and III, the response rate of about $50-56\%$ is obtained. The bias of $\widehat{\overline{Y}}_P$ in Case I and II shows smaller than that in the other cases and shows the smallest in the exponential response model and the largest in the logistic response model. Also we have similar results of $\widehat{\overline{Y}}_P$ in terms of ARB and RMSE regardless of the response model.

Comparing $\widehat{\overline{Y}}_P$, $\widehat{\overline{Y}}_D$ and $\widehat{\overline{Y}}_C$ in terms of bias, in some cases $\widehat{\overline{Y}}_P$ gives better results than the others, however $\widehat{\overline{Y}}_D$ and $\widehat{\overline{Y}}_C$ have good results in terms of ARB and RMSE with a very large difference. Also comparing $\widehat{\overline{Y}}_D$ and $\widehat{\overline{Y}}_C$, $\widehat{\overline{Y}}_C$ is superior to $\widehat{\overline{Y}}_D$, although it is not a big difference in the results of ARB and RMSE. In particular, $\widehat{\overline{Y}}_C$ is superior to $\widehat{\overline{Y}}_D$ in the bias results. Therefore, based on the results

Table 2: Results of PSA estimator with Gamma distribution

Case	Estimator	Exponential			Linear			Logistic		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P$	–5.459	0.050	7.434	–5.288	0.048	7.161	–5.483	0.049	7.208
	$\widehat{\overline{Y}}_D$	–4.626	0.038	5.398	–4.262	0.035	4.995	–4.219	0.035	4.896
	$\widehat{\overline{Y}}_C$	–4.569	0.038	5.350	–4.217	0.035	4.957	–4.181	0.035	4.864
II	$\widehat{\overline{Y}}_P$	–5.394	0.050	7.483	–5.207	0.048	7.201	–5.415	0.049	7.334
	$\widehat{\overline{Y}}_D$	–4.166	0.035	5.081	–3.927	0.033	4.781	–3.993	0.034	4.825
	$\widehat{\overline{Y}}_C$	–4.121	0.035	5.045	–3.890	0.033	4.751	–3.958	0.033	4.797
III	$\widehat{\overline{Y}}_P$	5.717	0.053	8.004	5.262	0.050	7.579	5.363	0.050	7.638
	$\widehat{\overline{Y}}_D$	4.380	0.037	5.472	4.228	0.036	5.244	4.359	0.037	5.338
	$\widehat{\overline{Y}}_C$	4.338	0.037	5.439	4.182	0.036	5.206	4.311	0.036	5.299
IV	$\widehat{\overline{Y}}_P$	6.225	0.056	8.378	5.601	0.051	7.766	5.486	0.050	7.608
	$\widehat{\overline{Y}}_D$	4.885	0.041	5.852	4.676	0.039	5.543	4.794	0.039	5.568
	$\widehat{\overline{Y}}_C$	4.839	0.040	5.813	4.625	0.038	5.499	4.738	0.039	5.519

Table 3: Results of PSA estimator with Log-normal distribution

Case	Estimator	Exponential			Linear			Logistic		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P$	–5.720	0.050	7.771	–5.474	0.048	7.445	–5.646	0.048	7.472
	$\widehat{\overline{Y}}_D$	–4.646	0.037	5.527	–4.257	0.034	5.071	–4.199	0.033	4.947
	$\widehat{\overline{Y}}_C$	–4.587	0.036	5.478	–4.211	0.033	5.033	–4.162	0.033	4.915
II	$\widehat{\overline{Y}}_P$	–5.618	0.050	7.751	–5.380	0.048	7.459	–5.611	0.049	7.574
	$\widehat{\overline{Y}}_D$	–4.154	0.034	5.198	–3.893	0.032	4.882	–3.976	0.032	4.907
	$\widehat{\overline{Y}}_C$	–4.106	0.033	5.160	–3.855	0.032	4.852	–3.941	0.032	4.879
III	$\widehat{\overline{Y}}_P$	5.767	0.051	8.147	5.298	0.048	7.712	5.428	0.048	7.775
	$\widehat{\overline{Y}}_D$	4.521	0.037	5.715	4.390	0.036	5.485	4.526	0.036	5.567
	$\widehat{\overline{Y}}_C$	4.477	0.037	5.680	4.341	0.035	5.445	4.475	0.036	5.525
IV	$\widehat{\overline{Y}}_P$	6.323	0.054	8.511	5.736	0.050	7.991	5.605	0.049	7.814
	$\widehat{\overline{Y}}_D$	5.067	0.040	6.099	4.928	0.039	5.885	5.052	0.039	5.912
	$\widehat{\overline{Y}}_C$	5.018	0.040	6.057	4.873	0.039	5.838	4.992	0.039	5.860

of Table 1, we can conclude that the proposed response probability estimation method gives the best results.

Table 2 shows the results when the super-population model is a log-linear model and the error distribution is a gamma distribution. Also three response probability models are considered. In Case I and II, the response rate of about $65-80\%$ is obtained, and in Case III and IV, the response rate of about $40-45\%$ is obtained. This result comes from the asymmetry of the gamma distribution. Investigating the results of Table 2, one can see that large biases occur in all estimators and the bias has a very large effect on the RMSE. As shown in Table 1, $\widehat{\overline{Y}}_C$ is the best through all comparison statistics results in Table 2.

Table 3 shows the results where the super-population model is a log-linear model and the error distribution is log-normal with three response probability models. The response rate is very similar

Table 4: Results of bias corrected PSA estimator using linear response probability model

Case	Estimator	Normal			Gamma			Log-normal		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P^{BC}$	0.459	0.007	6.569	–2.442	0.037	5.614	–1.577	0.035	5.601
	$\widehat{\overline{Y}}_D^{BC}$	–0.150	0.002	1.936	–1.320	0.021	3.216	–0.168	0.021	3.386
	$\widehat{\overline{Y}}_C^{BC}$	–0.100	0.002	1.933	–1.261	0.020	3.196	–0.103	0.021	3.390
II	$\widehat{\overline{Y}}_P^{BC}$	0.251	0.007	6.546	–2.691	0.038	5.836	–1.926	0.037	5.804
	$\widehat{\overline{Y}}_{D^{BC}}$	–0.162	0.002	2.007	–1.270	0.021	3.282	–0.179	0.022	3.587
	$\widehat{\overline{Y}}_C^{BC}$	–0.091	0.002	2.004	–1.221	0.021	3.268	–0.126	0.022	3.593
III	$\widehat{\overline{Y}}_P^{BC}$	0.233	0.007	6.539	2.893	0.039	6.139	1.416	0.035	5.683
	$\widehat{\overline{Y}}_D^{BC}$	0.284	0.002	1.971	1.831	0.023	3.656	0.513	0.021	3.481
	$\widehat{\overline{Y}}_C^{BC}$	0.213	0.002	1.966	1.781	0.023	3.633	0.458	0.021	3.475
IV	$\widehat{\overline{Y}}_P^{BC}$	–0.020	0.007	6.589	2.947	0.039	6.087	1.452	0.034	5.643
	$\widehat{\overline{Y}}_D^{BC}$	0.293	0.002	1.919	1.954	0.023	3.620	0.577	0.021	3.394
	$\widehat{\overline{Y}}_C^{BC}$	0.214	0.002	1.913	1.899	0.023	3.592	0.516	0.021	3.385

Table 5: Results of PSA estimator using linear response probability model with bivariate auxiliary variable

Case	Estimator	Normal			Gamma			Log-normal		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P$	1.586	0.010	19.203	–9.213	0.049	12.677	–9.476	0.048	13.207
	$\widehat{\overline{Y}}_D$	–4.158	0.004	6.968	–8.382	0.040	9.918	–8.657	0.040	10.389
	$\widehat{\overline{Y}}_C$	–0.528	0.003	5.591	–7.198	0.035	8.967	–7.417	0.035	9.415
II	$\widehat{\overline{Y}}_P$	0.545	0.010	18.977	–9.179	0.050	13.058	–9.274	0.048	13.174
	$\widehat{\overline{Y}}_D$	–4.149	0.004	7.283	–7.679	0.038	9.559	–7.875	0.037	9.750
	$\widehat{\overline{Y}}_C$	–0.940	0.003	6.005	–6.734	0.034	8.836	–6.898	0.033	8.997
III	$\widehat{\overline{Y}}_P$	0.504	0.010	19.206	9.513	0.052	13.776	9.879	0.051	14.212
	$\widehat{\overline{Y}}_D$	3.866	0.004	7.225	9.326	0.045	11.43	9.674	0.044	11.864
	$\widehat{\overline{Y}}_C$	0.723	0.003	6.123	8.061	0.040	10.383	8.343	0.039	10.775
IV	$\widehat{\overline{Y}}_P$	–0.444	0.010	19.037	10.069	0.053	14.055	10.624	0.054	14.866
	$\widehat{\overline{Y}}_D$	4.652	0.004	7.478	10.249	0.048	12.032	10.894	0.049	12.930
	$\widehat{\overline{Y}}_C$	1.078	0.003	5.911	8.828	0.042	10.823	9.378	0.043	11.640

to the result of gamma distribution. Also in Table 3, $\widehat{\overline{Y}}_C$ is the best through all comparison statistics results.

Table 4 shows the results of the bias corrected PSA estimator using the linear response probability model. To obtain the bias corrected PSA estimator, known response probability model and known super-population model are required. In this simulation, we use that the error distribution of the linear super-population model is normal and the response probability model is linear. Also we consider gamma and log-normal distributions as the error distribution of the log-linear super-population model.

Through the results, $\widehat{\overline{Y}}_P^{BC}$ gives the worst results in terms of ARB and RMSE and even the results

Table 6: Results of bias corrected PSA estimator using linear response probability model with bivariate auxiliary variable

Case	Estimator	Normal			Gamma			Log-normal		
		BIAS	ARB	RMSE	BIAS	ARB	RMSE	BIAS	ARB	RMSE
I	$\widehat{\overline{Y}}_P^{BC}$	2.064	0.010	19.253	–3.554	0.037	9.936	–3.053	0.036	10.351
	$\widehat{\overline{Y}}_D^{BC}$	–3.728	0.004	6.728	–2.883	0.026	6.851	–2.369	0.026	7.366
	$\widehat{\overline{Y}}_C^{BC}$	–0.079	0.003	5.583	–1.256	0.024	6.535	–0.588	0.026	7.331
II	$\widehat{\overline{Y}}_P^{BC}$	0.980	0.010	19.001	–4.258	0.040	10.670	–3.670	0.037	10.613
	$\widehat{\overline{Y}}_{D^{BC}}$	–3.769	0.004	7.073	–2.698	0.026	6.975	–2.125	0.026	7.414
	$\widehat{\overline{Y}}_C^{BC}$	–0.542	0.003	5.964	–1.398	0.025	6.746	–0.698	0.025	7.516
III	$\widehat{\overline{Y}}_P^{BC}$	0.076	0.010	19.200	5.367	0.042	11.326	3.483	0.037	10.535
	$\widehat{\overline{Y}}_D^{BC}$	3.482	0.004	7.031	5.319	0.032	8.640	3.518	0.027	7.824
	$\widehat{\overline{Y}}_C^{BC}$	0.322	0.003	6.102	3.950	0.029	7.851	2.033	0.026	7.273
IV	$\widehat{\overline{Y}}_P^{BC}$	–0.915	0.010	19.062	5.297	0.041	11.100	3.627	0.038	10.786
	$\widehat{\overline{Y}}_D^{BC}$	4.228	0.004	7.228	5.566	0.032	8.628	4.063	0.029	8.152
	$\widehat{\overline{Y}}_C^{BC}$	0.635	0.003	5.862	4.035	0.029	7.735	2.387	0.026	7.429

are worse than those of $\widehat{\overline{Y}}_C$ in the linear response probability model of Table 2. Comparing $\widehat{\overline{Y}}_D^{BC}$ and $\widehat{\overline{Y}}_C^{BC}$, we can see that $\widehat{\overline{Y}}_D^{BC}$ and $\widehat{\overline{Y}}_C^{BC}$ have similar results in terms of ARB and RMSE. However, the bias of $\widehat{\overline{Y}}_C^{BC}$ is smaller than that of $\widehat{\overline{Y}}_D^{BC}$. Also comparing $\widehat{\overline{Y}}_C$ and $\widehat{\overline{Y}}_C^{BC}$, we have that $\widehat{\overline{Y}}_C^{BC}$ have better results with a very large difference. Therefore, it is very reasonable to use a bias corrected PSA estimator when the error distribution is known and the response probability model is linear. Of course, the bias corrected PSA estimator can be applied if bias is known.

Table 5 and Table 6 show the results of bivariate auxiliary variable case. The response rates are similar to the univariate variable cases. In Table 5, the result of the PSA estimator, it can be seen that the bias of $\widehat{\overline{Y}}_P$ is relatively small in the normal super-population result. However, the biases of $\widehat{\overline{Y}}_P$ in the gamma distribution and log-normal distribution are larger than those of other estimators. Also, in terms of ARB and RMSE, $\widehat{\overline{Y}}_P$ gives inferior results when compared to other estimators, so it is not good to use $\widehat{\overline{Y}}_P$ for a bivariate auxiliary variable case. Comparing $\widehat{\overline{Y}}_D$ and $\widehat{\overline{Y}}_C$ in the normal super-population result, it can be seen that the bias of $\widehat{\overline{Y}}_C$ is smaller. In addition, $\widehat{\overline{Y}}_C$ shows good results in terms of ARB and RMSE. This pattern of results also can be seen in the gamma distribution and log-normal distribution. Therefore, we conclude that $\widehat{\overline{Y}}_C$ gives the best result when using PSA estimator in the bivariate auxiliary variable case.

Table 6 shows the results of the bias corrected PSA estimator. As in the case of univariate auxiliary variable, $\widehat{\overline{Y}}_C^{BC}$ gives the best result in terms of all comparison statistics in the case of bivariate auxiliary variable. Therefore, it is appropriate to use $\widehat{\overline{Y}}_C^{BC}$ when using the bias corrected PSA estimator.

By comparing the results of Table 5 and Table 6, the effect of bias correction can be confirmed. The effect of bias correction of $\widehat{\overline{Y}}_P^{BC}$ can be seen in the gamma distribution and the log-normal distribution

except for the normal distribution. On the other hand, in the results of $\widehat{\overline{Y}}_D^{\mathrm{BC}}$ and $\widehat{\overline{Y}}_C^{\mathrm{BC}}$, the bias correction effect can be seen in all distributions. Also the bias correction effect of $\widehat{\overline{Y}}_C^{\mathrm{BC}}$ is seen for all distributions, especially the gamma and log-normal distributions. Therefore, it can be confirmed through simulation studies that the bias corrected PSA estimator gives good results when a known super-population model is used in the linear response probability model.

5. Conclusion

Recently, a lot of non-ignorable non-response has occurred in sample survey and several studies have been conducted to properly deal with it. Non-ignorable non-response is not easy to deal with because it causes bias. In particular, in order to properly handle non-ignorable non-responses, accurate estimation of response probability plays an important role. However, since the response probability is a function of the variable of interest, it is not easy to accurately estimate the response probability in practice. Therefore, practically the method of estimating the response probability should use available auxiliary variables. In this study we propose a new response probability estimation method using available auxiliary variables. It is confirmed that the proposed method provides better result than the existing methods and the bias corrected PSA estimator produces significantly better results.

Here we note that the results in this study are obtained using a linear response probability model with known super population models. A model misspecification of the response probability model is a big concern and the linear model may be neither well motivated nor appropriate in practice. Also the super population model plays an important role and a model missspecification of the super population model is also concerned. Therefore, it is necessary to study a method that can be used for an arbitrary unknown response probability model and an arbitrary unknown super-population model.

References

Bethlehem J (2020). Working with response probabilities, *Journal of Official Statistics*, **36**, 647–674.

Chang T and Kott PS (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95**, 555–571.

Chung HY and Shin KI (2017). Estimation using informative sampling technique when response probability follows exponential function of variable of interest, *Korean Journal of Applied Statistics*, **30**, 993–1004.

Da Silva DN and Opsomer JD (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys, *Canadian Journal of Statistics*, **34**, 563–579.

Da Silva DN and Opsomer JD (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression, *Survey Methodology*, **35**, 165–176.

Iannacchinoe VG, Milne JG, and Folsom RE (1991). Response probability weight adjustment using logistic regression. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637–642.

Kim JK and Riddles MK (2012). Some theory for propensity-score-adjustment estimators in survey sampling, *Survey Methodology*, **38**, 157–165.

Kim JK and Yu CL (2011). A semiparametric estimation of mean functionals with nonignorable missing data, *Journal of the American Statistical Association*, **106**, 157–165.

Kott PS and Chang T (2010). Using Calibration weighting to adjust for nonignorable unit nonresponse, *American Statistical Association*, **105**, 1265–1275.

Min JW and Shin KI (2018). A study on the determination of substrata using the information of

exponential response rate by simulation studies, *Korean Journal of Applied Statistics*, **31**, 621–636.

Pfeffermann D, Krieger AM, and Rinott Y (1998). Parametric distributions of complex survey data under informative probability sampling, *Statistica Sinica*, **8**, 1087–1114.

Pfeffermann D and Sverchkov M (2003). Small area estimation under informative sampling, Proceedings of the Survey Research Method Section, *American Statistical Asscociation*, 3284–3295.

Riddles MK, Kim JK, and Im J (2016). A propensity-score-adjustment method for nonignorable nonresponse, *Journal of Survey Statistics and Methodology*, **4**, 215–245.

Received October 25, 2021; Revised February 20, 2022; Accepted February 25, 2022