Jouínal of Modeín Applied Statistical

Methods

Volume 24 ||Issue 1 Aíticle 10

Efficient Utilization of Multiple Auxiliary Variables for Nonresponse Problem in Estimating the Population Mean Under Sub-sampling Technique

Napattchan Dansawad

Department of Mathematics,
Faculty of Science, King Mongkut's University of Technology Thonburi,
Bangkok, Thailand.

Journal of Modern Applied Statistical Methods
2025, Vol. 24, No. 1,
Doi: 10.56801/Jmasm.V24.i1.10

ISSN 1538 – 9472

Efficient Utilization of Multiple Auxiliary Variables for Nonresponse Problem in Estimating the Population Mean Under Sub-sampling Technique

Napattchan Dansawad

Department of Mathematics,
Faculty of Science, King Mongkut's University of Technology Thonburi,
Bangkok, Thailand.

The main key objective of this paper is to address the nonresponse problems by adapting Hansen and Hurwitz's technique (1964) and Saini *et al.*'s estimator (2022) to propose a novel estimator of population mean under sub-sampling technique using multiple auxiliary variables. A comparative analysis of the proposed novel estimator's efficacy has been performed through theoretical and numerical studies. The results of this paper confirm that our estimator is more effective than others under the same situation.

Keywords: Multiple auxiliary variables, Nonresponse, Sub-Sampling, Survey sampling.

1. Introduction

Generally, the main causes of many fields of surveys, such as agricultural, educational, meteorology, biomedical, engineering, and so on, are the researcher collected incomplete information, lack of cooperation from data sources, or refusal of the respondents, including insufficient time to survey, which creates problems of nonresponse. Nonresponse has been a significant challenge in nearly all sample surveys, and its rate is likely to rise, particularly in sensitive matters. For various statistical tasks, various estimators are created to estimate the population parameters of interest, such as the mean, and nonresponse problems will diminish the accuracy of these estimators and cause the estimator's bias and mean square error (MSE) to increase. Therefore, these estimators are inapplicable in nonresponse or have missing data on different variables. A crucial way to deal with these problems is to employ the sub-sampling technique, first suggested by Hansen and Hurwitz (1946), by selecting a sub-sample from a group of respondents who lack cooperation before collecting data through personal interviews.

In this technique, the whole population $\mathrm{J} = (\mathrm{J}_1, \mathrm{J}_2, \ldots, \mathrm{J}_N)$ of size N is portioned into the responding units (N_1), and not responding units (N_2). Suppose that the sample of size n is twitched with no return from the population J, which is portioned into two groups composed of n_1 units of the responding and n_2, $(n_2 = n - n_1)$ units of the not responding. In addition, the values of the study and auxiliary variables for the i th units of the population J are defined as y_i and x_i, respectively. However, a sub-sample of size s, $s = n_2 m^{-1}$ is twitched by making an extra effort from the not responding units n_2, where m, $(m > 1)$ is the inverse sampling rate for the first sample of size n. Therefore, the population mean of the study variable can be estimated by using $n_1 + s$ units substituted for the sample of size n.

In addition to suggesting a sub-sampling technique, Hansen and Hurwitz (1964) presented an unbiased estimator along with variance to estimate the population's mean in the case of nonresponse. The formula of this estimator are given as, respectively

$$t_1 = \varphi_1 \bar{y}_1 + \varphi_2 \bar{y}_{2(s)} \tag{1}$$

and

$$V(t_1) = \left(\frac{1}{n} - \frac{1}{N}\right) C_y^2 + \frac{N_2}{N} \frac{(m-1)}{n} C_{y(2)}^2 \tag{2}$$

where $\bar{y}_1 = \sum_{i=1}^{n_1} y_i / n_1$ and $\bar{y}_{2(s)} = \sum_{i=1}^{s} y_i / s$ are the sample means of the study variable contingent on n_1 and s, respectively. $\varphi_1 = n_1 / n$ and $\varphi_2 = n_2 / n$ are the proportion of units of the responding and not responding of the first sample of size n. For other symbols can be shown as follow: $C_y^2 = S_y^2 / \bar{Y}^2$, $C_{y(2)}^2 = S_{y(2)}^2 / \bar{Y}^2$, $\bar{Y} = \sum_{i=1}^{N} y_i / N$, $S_y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N-1)$, and $S_{y(2)}^2 = \sum_{i=1}^{N_2} (y_i - \bar{Y}_2)^2 / (N_2 - 1)$.

In the same background as mentioned above, the unbiased estimator in the case of nonresponse of population mean $(\bar{X})$ of the auxiliary variable x along with variance can be defined as

$$t_2 = \varphi_1 \bar{x}_1 + \varphi_2 \bar{x}_{2(s)} \tag{3}$$

and

$$V(t_2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2 + \frac{N_2}{N} \frac{(m-1)}{n} C_{x(2)}^2 \tag{4}$$

where $\bar{x}_1 = \sum_{i=1}^{n_1} x_i / n_1$, $\bar{x}_{2(s)} = \sum_{i=1}^{s} x_i / s$, $C_x^2 = S_x^2 / \bar{X}^2$, $C_{x(2)}^2 = S_{x(2)}^2 / \bar{X}^2$, $\bar{X} = \sum_{i=1}^{N} x_i / N$,
$S_x^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2 / (N-1)$, and $S_{x(2)}^2 = \sum_{i=1}^{N_2} (x_i - \bar{X}_2)^2 / (N_2 - 1)$.

Following the pioneering work of Hansen and Hurwitz (1946), many researchers and academics have utilized the benefits of auxiliary data along with Hansen and Hurwitz's (1946) estimator to improve their interest estimators , such as the population's mean. Bouza-Herrera and Subzar (2019), Vishwakarma *et al.* (2019), Sanaullah and Hanif (2020), Ünal and Kadilar (2021), Jaiswal *et al.* (2022), Ahmadini *et al.* (2022), Tiwari and Sharma (2023), etc. are examples of researchers and academics who proposed their estimators in the situation of nonresponse under two well-known cases. Firstly, nonresponse occurred only on the study variable. Secondly, nonresponse occurred on both the variables of the study and the auxiliary.

However, for proposing the mean estimator of the population, using auxiliary data is an alternative to compensate for data for many researchers and academics in the situation where the group of samples fails to provide enough responses, including in the case of population units missing out of the sampling frame. Because auxiliary data can help increase their estimators' precision or efficiency. For example, the use of two population means of auxiliary variables (denoted as $\bar{X}_1$ and $\bar{X}_2$) in creating the estimator for $\bar{Y}$ has been recently proceeded by Saini *et al.* (2022) as follows:

$$t_3 = \frac{\upsilon_1 \bar{y} + \upsilon_2(\bar{X}_1 - \bar{x}_1) + \upsilon_3(\bar{X}_2 - \bar{x}_2)}{4} \left(\frac{\bar{X}_1}{\bar{x}_1} + \frac{\bar{x}_1}{\bar{X}_1} \right) \left(\frac{\bar{X}_2}{\bar{x}_2} + \frac{\bar{x}_2}{\bar{X}_2} \right) \tag{5}$$

where υ_1 , υ_2 , and υ_3 are any constants.

Getting inspiration from Hansen and Hurwitz's (1946) and Saini *et al.* (2022) work, when nonresponse occurs on both the study variable y and the auxiliary variable x, this present paper aims to study estimating a population mean by using multiple auxiliary variables under sub-sampling of nonresponse. Some properties of the new estimator will be examined. The remainder of this study is an efficiency comparison of the new proposed estimator using theoretical and numerical analysis using two numerical examples under the percent relative efficiencies (PRE) criterion.

2. The Estimator

Following Saini *et al.* (2022), one adapt the estimator in equation (5) to a new estimator for the population mean of $\bar{Y}$ by using multiple auxiliary variables under the sub-sampling of nonresponse. The new estimator is given as follows:

$$t_4 = \frac{\upsilon_1 \bar{y}^* + \upsilon_2(\bar{X}_1 - \bar{x}_1^*) + \upsilon_3(\bar{X}_2 - \bar{x}_2^*)}{4}\left(\frac{\bar{X}_1}{\bar{x}_1^*} + \frac{\bar{x}_1^*}{\bar{X}_1}\right)\left(\frac{\bar{X}_2}{\bar{x}_2^*} + \frac{\bar{x}_2^*}{\bar{X}_2}\right) \tag{6}$$

To find out some properties of the new estimator, such as bias and MSE, one will consider $\bar{y}^* = \bar{Y}(1+e_0^*)$, $\bar{x}_1^* = \bar{X}_1(1+e_1^*)$, and $\bar{x}_2 = \bar{X}_2(1+e_2)$. Then, $E(e_0^*) = E(e_1^*) = E(e_2) = 0$,
$E(e_0^{*2}) = \phi C_y^2 + \phi^* C_{y(2)}^2$, $E(e_1^{*2}) = \phi C_{x1}^2 + \phi^* C_{x(1)}^2$, $E(e_2^2) = \phi C_{x2}^2$,
$E(e_0^* e_1^*) = \phi\rho_{yx1} C_y C_{x1} + \phi^* \rho_{yx(1)} C_y C_{x(1)}$, $E(e_0^* e_2) = \phi\rho_{yx2} C_y C_{x2}$, and $E(e_1^* e_2) = \phi\rho_{x1x2} C_{x1} C_{x2}$.
where $\phi = (N-n)/Nn$, $\phi^* = \frac{N_2}{N}\frac{(k-1)}{n}$

After that, one will change equation (6) in terms of e_0^* and e_1^* before retaining only the terms that do not exceed the second degree of the error terms and then subtracting $\bar{Y}$ on both sides of this equation. So, the new equation can be expressed as follows:

$$t_4 = (\upsilon_1 - 1)\bar{Y} + \upsilon_1 \bar{Y} e_0^* - \upsilon_2 \bar{X}_1 e_1^* - \upsilon_3 \bar{X}_2 e_2 + \frac{1}{2}\upsilon_1 \bar{Y} e_1^{*2} + \frac{1}{2}\upsilon_1 \bar{Y} e_2^2 \tag{7}$$

After taking the expectation on both sides of equation (7), one will get the term of bias of the new estimator as follows:

$$\begin{aligned} Bias(t_4) &= E(t_3 - \bar{Y}) \\ &\cong \bar{Y}\left[(\upsilon_1 - 1) + \frac{1}{2}\upsilon_1(\phi C_{x1}^2 + \phi^* C_{x(1)}^2) + \frac{1}{2}\upsilon_1 \phi C_{x2}^2\right] \end{aligned} \tag{8}$$

The MSE of t_4 can be obtained from squaring and taking the expectation on both sides of equation (7), one get

$$\begin{aligned} MSE(t_4) &= E(t_4 - \bar{Y})^2 \\ &\cong \bar{Y}^2\left[(\upsilon_1 - 1)^2 + \upsilon_1^2(\phi C_y^2 + \phi^* C_{y(2)}^2) + (\upsilon_1 - 1)\upsilon_1(\phi C_{x1}^2 + \phi^* C_{x(1)}^2) + (\upsilon_1 - 1)\upsilon_1 \phi C_{x2}^2\right] \\ &\quad - 2\phi\upsilon_1 \bar{Y} C_y\left[\upsilon_2 \bar{X}_1 \rho_{yx1} C_{x1} + \upsilon_3 \bar{X}_2 \rho_{yx2} C_{x2}\right] \\ &\quad + \phi\left[\upsilon_2^2 \bar{X}_1^2 C_{x1}^2 + \upsilon_3^2 \bar{X}_2^2 C_{x2}^2 + 2\upsilon_2\upsilon_3 \bar{X}_1 \bar{X}_2 \rho_{x1x2} C_{x1} C_{x2}\right] \\ &\quad + \phi^*\left[\upsilon_2^2 \bar{X}_1^2 C_{x(1)}^2 - 2\upsilon_2\upsilon_3 \overline{YX}_1 \rho_{yx(1)} C_y C_{x(1)}\right] \end{aligned} \tag{9}$$

The equation (9) is minimum when

$$\upsilon_1 = \frac{OP[2 + \mathrm{P} + \phi C_{x2}^2]}{2OP[1 + Q + P + \phi C_{x2}^2] - 2C_y^2 M[MO - \phi\rho_{yx2}\rho_{x1x2} C_{x1} P + \phi\rho_{x1x2}^2 C_{x1}^2 M] - PC_y^2[2\rho_{yx2}^2 P + \rho_{yx2}\rho_{x1x2} C_{x1} M]}$$

$$\upsilon_2 = \frac{C_y \bar{Y} M[2+P+\phi C_{x2}^2][P-\phi\rho_{yx2}\rho_{x1x2}P+\phi\rho_{x1x2}^2 C_{x1}^2 M]}{\bar{X}_1[2OP(1+Q+P+\phi C_{x2}^2)-2C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)-PC_y^2(2\rho_{yx2}^2 P+\rho_{yx2}\rho_{x1x2}C_{x1}M)]} \tag{10}$$

$$\upsilon_3 = \frac{C_y \bar{Y}[OP(2+P+\phi C_{x2}^2)][\rho_{yx2}P-\rho_{x1x2}C_{x1}M]}{\bar{X}_1 C_{x2}[2OP(1+Q+P+\phi C_{x2}^2)-2C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)-PC_y^2(2\rho_{yx2}^2 P+\rho_{yx2}\rho_{x1x2}C_{x1}M)][P-\phi\rho_{x1x2}^2 C_{x1}^2]}$$

where $M=\phi\rho_{yx1}C_{x1}+\phi^*\rho_{yx(1)}C_{x(1)}$, $O=\phi C_{x1}^2(1-\rho_{x1x2}^2)+\phi^* C_{x(1)}^2$, $P=\phi C_{x1}^2+\phi^* C_{x(1)}^2$

Therefore, the resulting minimum mean squared error (MMSE) of t_4 can be shown as follows:

$$MMSE(t_4) = \frac{\phi\bar{Y}^2[4L(\phi C_y^2+\phi^* C_{y(2)}^2)-2C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)^2]}{4[A+L(\phi C_y^2+\phi^* C_{y(2)}^2)+2C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)]} \tag{11}$$

3. Efficiency Comparison

For a theoretical comparison, one will confirm that the proposed estimator t_4 will be more efficient than the Hansen and Hurwitz (1964) estimator if the equation (12) is true.

$V(t_1) > MMSE(t_4)$ if and only if

$$[\phi C_y^2+\phi^* C_{y(2)}^2] > \frac{2\phi[L(\phi C_y^2+\phi^* C_{y(2)}^2)-Z^2}{2P[1+W]} \tag{12}$$

where $W=2C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)]/(A+L(\phi C_y^2+\phi^* C_{y(2)}^2))$,
$Z=\phi C_y^2 M(MO-\phi\rho_{yx2}\rho_{x1x2}C_{x1}P+\phi\rho_{x1x2}^2 C_{x1}^2 M)$

4. Numerical Study

After theoretical comparisons, one also used the following two real datasets from Khare and Sinha (2007) and Khare and Sinha (2014) to validate the efficiency of the estimator t_4 compared with the efficiency of Hansen and Hurwitz (1964) estimator (t_1) by using the following formula as:

$$PRE(t_4,t_1) = \frac{V(t_1)}{MMSE(t_4)} \times 100 \tag{13}$$

The details of two real data sets are presented as follows:

Dataset 1: This dataset was presented by Khare and Sinha (2007) and is related to the physical development of upper-class children of Indian ancestry from 95 schools around the Varanasi district of Uttar Pradesh recorded by the Indian Council of Medical Research. The study variable y was taken as children's weights (in kilograms), whereas the skull and chest circumference (in centimeters) were taken as the auxiliary variables x_1 and x_2, respectively. Assume the first 25% of all data will be the nonresponse group. For this dataset, one have

$N=95$, $n=35$, $N_2/N=0.25$, $\bar{Y}=19.4968$, $\bar{X}_1=51.1726$, $\bar{X}_2=55.8611$, $C_y=0.1561$, $C_{y(2)}=0.1208$, $C_{x1}=0.0301$, $C_{x(1)}=0.0248$, $C_{x2}=0.0586$, $\rho_{yx1}=0.3280$, $\rho_{yx(1)}=0.4770$, $\rho_{yx2}=0.8460$, $\rho_{y(2)}=0.7290$, $\rho_{x1x2}=0.2970$

Dataset 2: One considered Khare and Sinha's study (2014). This dataset is related to the population of 109 towns in urban areas around the Baria and Tahasil-Champua police stations in the Kendujhar district of Odisha state in India. For this dataset, the last 25% of all data will be the nonresponse group. The number of laborers in the town was assumed as the study variable y. In contrast, the town's number of non-laborers and cultivators was considered an auxiliary variable (denoted as x_1 and x_2). The details of this dataset are given as follows:

$N=109$, $n=70$, $N_2/N=0.25$, $\bar{Y}=165.2661$, $\bar{X}_1=259.0826$, $\bar{X}_2=100.5505$, $C_y=0.6828$, $C_{y(2)}=0.0035$, $C_{x1}=0.7645$, $C_{x(1)}=0.5429$, $C_{x2}=0.7314$, $\rho_{yx1}=0.8160$, $\rho_{yx(1)}=0.8711$, $\rho_{yx2}=0.9460$, $\rho_{y(2)}=0.9050$, $\rho_{x1x2}=0.7320$

When using the datasets mentioned above, the efficiency of the proposed estimator t_4 can be compared to t_1, and the results can be found in the following Table.

Table 1. PRE of the estimator t_2 compared to t_1

m	Dataset 1		Dataset 2	
	Estimators		Estimators	
	t_1	t_4	t_1	t_4
2	100.0000 (0.0005439)	212.8767 (0.0002555)	100.0000 (0.0023831)	140.9284 (0.0016910)
3	100.0000 (0.0006482)	197.6822 (0.0003279)	100.0000 (0.0024314)	134.3463 (0.0018098)
4	100.0000 (0.0007524)	187.9121 (0.0004004)	100.0000 (0.0026732)	125.5790 (0.0021287)
5	100.0000 (0.0008566)	181.1760 (0.0004728)	100.0000 (0.0027418)	112.0245 (0.0024475)

* Figures in parentheses indicate the MSE.

Based on the numerical results in above Table, it is clear that the estimator t_1 by Hansen and Hurwitz (1964) is less efficient than the proposed estimator t_4 in every dataset. However, looking at PRE and MSE values for each estimator, one finds that the proposed estimator t_4 has a larger PRE than the estimator t_1, despite having lower MSE values in the same datasets. It is also noted that when the value of the nonresponse rate (m) increases, the efficiencies of the proposed estimator t_4 decrease. Therefore, our proposed estimator t_4 is more justifiable in practical applications than previous similar work.

5. Conclusions

The nonresponse of collected data, especially missing data, has been a significant challenge in nearly all sample surveys. The nonresponse poses considerable challenges for researchers, and increasing the sample size will not solve this issue. This phenomenon of nonresponse will diminish the accuracy of estimators of interest and introduce bias in estimates, leading to a higher mean square error (MSE) and ultimately reducing their efficiency. An important way to cope with these problems is to apply the subsampling technique introduced by Hansen and Hurwitz (1964).

Therefore, this paper aims to address these problems by adapting Hansen and Hurwitz's technique (1964) and Saini *et al.*'s estimator (2022) to propose a novel estimator for nonresponse problems in estimating the mean of the population using multiple auxiliary variables under the situation of nonresponse occurs on both the study and auxiliary variables. The efficiency of the novel estimator against the other ones is compared through two numerical analyses and two statistics, namely, mean square error (MSE) and minimum mean squared error (MMSE) under the criterion of percent relative efficiencies (PRE). Results of the two numerical analyses demonstrated that our novel estimator consistently outperforms the estimator of Hansen and Hurwitz (1964), which has relatively fewer MSE values and a relatively high value of PRE. Thus, one proposes using our novel estimator, which utilizes multiple auxiliary variables for a more precise estimation of the population mean under the same situation described in this paper.

References

Ahmadini, A.A.H., Yadev, T., Yadav S.K. & A.S.M. Al Luhayb, 2022. Restructured searls family of estimators of population mean in the presence of nonresponse. Frontiers in Applied Mathematics and Statistics, 8, 969068. https://www.doi.org/10.3389/fams.2022.969068.

Bouza-Herrera, C.N., 2019. Subsampling rules for item nonresponse of an estimator based on the combination of regression and ratio. Journal of King Saud University-Science, 31, 171-176.

Hansen, M.H. & W.N. Hurwitz (1946). The problem of nonresponse in sample surveys. Journal of the American Statistical Association, 41, 517-529.

Jaiswal, A.K., Singh, G.N. & A.K. Pandey, (2022). Improved procedures for mean estimation under nonresponse. Alexandria Engineering Journal, 61(12), 12813-12828.

Khare, B.B. & R.R. Sinha, 2007. Estimation of the ratio of the two populations means using multi-auxiliary characters in the presence of nonresponse. Statistical Techniques in Life Testing, Reliability, Sampling Theory and Quality Control, edited by B. N. Pandey, Narosa Publishing House, New House, New Delhi, 163-171.

Khare, B.B. & R.R. Sinha, 2014. A class of two-phase sampling estimator for ratio of two populations means using multi-auxiliary characters in the presence of nonresponse. STATISTICS IN TRANSITION new series, 15(3), 389-402.

Saini, M., Jitendrakumar B.R. & A. Kumar (2022). Optimum estimator in simple random sampling using two auxiliary attributes with application in agriculture, fisheries and education sectors. *MethodsX, 101915.*

Sanaullah, A. & M. Hanif, 2020. Generalized chain exponential-type estimators under stratified two-phase sampling with subsampling the nonrespondents. Journal of Statistical Theory and Applications, 19(2), 185-195.

Tiwari, K.K. & V. Sharma, 2023. Efficient estimation of population mean in the presence of nonresponse and measurement error. STATISTICS IN TRANSITION new series, 24(3), 95-116.

Ünal, C. & C. Kadilar, 2021. A new family of exponential type estimators in the presence of nonresponse. Journal of Mathematical and Fundamental Sciences, 53(1), 1-15.

Vishwakarma, G.K., Singh, N. & A. Kumar, 2019. Computing the combined effect of measurement errors and nonresponse using factor chain-type class of estimator. The Philippine Statistician, 68(1), 27-39.