Thailand Statistician
January 2024; 22(1): 76-101
http://statassoc.or.th
Contributed paper

New Variance Estimators in the Presence of Nonresponse Under Unequal Probability Sampling, with Application to Fine Particulate Matter in Thailand

Chugiat Ponkaew [a] and Nuanpan Lawson*[b]

[a] Department of Mathematics and Data Science, Faculty of Science and Technology, Phetchabun Rajabhat University, Phetchabun, Thailand.

[b] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

*Corresponding author; e-mail: nuanpan.n@sci.kmutnb.ac.th

Received: 2 January 2023
Revised: 6 September 2023
Accepted: 13 September 2023

Abstract

Fine particulate matter exacerbates the environmental problem of air pollution in Thailand and contributes to premature deaths. Monitoring of levels of fine particulate matter data is imperative, however, values are seldom complete. One of the most important issues for variance estimation of population total estimators using ratio and generalized regression estimators under unequal probability sampling without replacement is that it requires joint inclusion probability which can be difficult to find. In this paper we solve this problem by proposing new variance estimators for population total in the presence of nonresponse under unequal probability sampling without replacement under the uniform nonresponse mechanism. Two approaches are used to construct the new variance estimators; estimating the joint inclusion probability and free joint inclusion probability. Simulation studies and an application to fine particulate matter in the north of Thailand are considered in the study. The results show that the variance estimators from the latter method give a smaller relative root mean square error and relative bias than the variance estimator from the former one. Nevertheless, the proposed variance estimator with the free joint inclusion probability provides a narrower confidence interval compared to others.

Keywords: Generalized regression estimator, joint inclusion probability, ratio estimator, response probability, variance estimator.

1. Introduction

The levels of fine particulate matter (PM2.5) in cities all over Thailand are perpetually higher than the World Health Organization's (WHO) guidelines. Thailand was ranked as the 45th most polluted country according to the 2021 IQAir World Air Quality Report. However, some data on the amount of PM2.5 may not be completely recorded due to technical difficulties. Missing data can hinder national assessments of health risks and proper management, leading to estimation being crucial to mitigate the prevailing issue of pollution. Horvitz and Thompson's (1952) estimator is a well-known population total estimator under unequal probability sampling without replacement. However, the calculation of the variance of Horvitz and Thompson's estimator requires joint inclusion probabilities or second order inclusion probabilities, which are the probability of two different units in the population to be selected in the sample and sometimes are difficult to be calculated. Under unequal probability sampling with replacement, the formulas of the variance estimators are in their simple forms because they do not require joint inclusion probability, which is different from the variance formula under unequal probability sampling without replacement which requires joint inclusion probabilities. Berger (2003) stated that the size of the joint inclusion probabilities set is equal to $n(n-1)/2$ where n denotes sample size, therefore it might be inconvenient to calculate the variance when n is large.

Furthermore, a computational procedure for determining the value of the joint inclusion probability of units i and j in sample s require first order inclusion probability for all units i in population U, whereas in practice they are often unknown. Hansen and Hurwitz (1943) proposed a new unbiased estimator for estimating population total under unequal probability sampling with replacement in the case of full response. They also discussed variance and its estimator which is in a simple form because it does not require joint inclusion probability. Horvitz and Thompson (1952) suggested an unbiased estimator for estimating population total under unequal probability sampling without replacement. Horvitz and Thompson's (1952) estimator's formula is a function of the ratio of the study variable and first order inclusion probability where the variance estimator requires both first order inclusion and joint inclusion probabilities. When the units in the sample have high inclusion probabilities, the variance estimator proposed by Horvitz and Thompson (1952) may lead to negative values. The alternative formulas of the variance estimators when the sample size is fixed were proposed by Sen (1953) and was implemented by Yates and Grudy (1953). Nevertheless, these alternative variance estimators require the joint inclusion probability similarly to Horvitz and Thompson's (1952) estimator, which is sometimes unknown or complicated in calculation. To address this issue, Hartley and Rao (1962) suggested to approximate the joint inclusion probability under randomized systematic sampling. Hájek (1964) also approximated joint inclusion probability and modified the variance of Horvitz and Thompson's (1952) estimator for large-entropy sampling designs such as rejective sampling. Under stratified sampling, Hájek (1981) used the first inclusion probability to approximate the joint inclusion probability in each stratum and showed that this procedure for the estimation of joint inclusion probability is valid when rejective sampling is used in each stratum. Brewer (2002) proposed two methods to approximate joint inclusion probability based on Hartley and Rao (1962) and Hájek (1964). Brewer and Donadio (2003) also proposed four methods to estimate joint inclusion probability. Sichera (2020) constructed the "jipApprox" package in R to estimate joint inclusion probability based on Hartley and Rao (1962), Hájek (1964), and Brewer and Donadio (2003).

When auxiliary information is available, the generalized regression estimator (GREG) is popular for estimating the population total. The formula of the GREG estimator is in the form of the Horvitz and Thompson (1952) estimator with additional adjustments calculated from an auxiliary variable. There are some papers concerning the optimal GREG estimator. Montanari (1987) proposed an optimal GREG estimator when the true value of the population regression coefficient is available. However, this value is usually unknown in practice because its formula depends on the population's elements therefore an estimated value of regression coefficient from the sample elements is required. Under single stratified sampling, Berger et al. (2003) used an estimated value of the regression coefficient to construct an optimal GREG estimator based on the estimator of Montanari (1987). Nangsue and Berger (2014) also proposed an optimal GREG estimator under stratified two-stage cluster sampling following Berger et al. (2003) and Estevao and Särndal (2006). The variance and associated estimator of the GREG estimator can be obtained by the Taylor linearization approach because the GREG estimator has a form of a nonlinear estimator. Under the Taylor linearization approach, the approximate linear estimator of the GREG estimator is obtained by using the first-order Taylor series approximation then the variance of the GREG estimator is calculated by deriving the variance of the approximate linear estimator. The formula for variance of the GREG estimator under unequal probability sampling without replacement requires joint inclusion probabilities similar to the variance from Horvitz and Thompson (1952). Therefore, the value of the variance estimator of the GREG estimator requires complicated computation because the joint inclusion probabilities in the formula are large. Furthermore, the exact value of joint inclusion probabilities cannot be determined for many sampling designs such as the rejective sampling, Rao-Sampord sampling, and successive sampling, so the true value of the variance estimator is impossible to find (Berger 1998).

In the presence of nonresponse, Särndal and Lundström (2005) have proposed an almost unbiased GREG estimator for estimating population total and also an associated variance estimator under two-phase framework which requires nonresponse propensities. Later, a new almost unbiased GREG estimator for estimating population total was proposed by Lawson and Ponkaew (2019) along with its variance estimators. The GREG estimator and its variance estimators are considered under the uniform nonresponse mechanism where the sampling fraction is negligible which does not require response probabilities. Ponkaew and Lawson (2018) proposed a new ratio estimator for estimating population total under the reverse framework where the sampling fraction is negligible. Their variance estimators require the true value of joint inclusion probabilities. Then, Midzuno's (1952) scheme was used to determine the true value of the joint inclusion probability for computing variance estimators. In 2020, Lawson (2020) proposed a new method for estimating the variance of the population total estimator with free joint inclusion probabilities based on Hájek's (1964) method. In the full response, Berger (2003) used Hájek's (1981) method and weighted least squares regression to investigate the free joint inclusion probability of the variance of the population total estimator under stratified sampling. Berger's (2003) variance estimators are also discussed with three different methods in each stratum, the first one used Hájek's (1981) variance estimator, the second considered simple random sampling, and finally Berwer's (2002) variance estimators were discussed.

This paper aims to propose a new variance estimator using the ratio and GREG estimators under unequal probability sampling without replacement with uniform nonresponse. The new variance estimators do not require the true value of the joint inclusion probability. Two estimation methods have been suggested; free joint inclusion probability following Lawson (2020) and estimated joint inclusion probability. The basic setup is shown in Section 2 .In section 3, the estimators of the joint inclusion probabilities are reviewed. In Section 4, we discuss variance estimation with free joint inclusion probability. In Section 5, the existing estimators under uniform nonresponse are discussed. Then, the proposed variance estimators are given in section 6. In Sections 7 and 8, we lay down the results of the simulation studies and application to Thai maize data in 2018 and 2019. Finally, some conclusions are given in Section 9.

2. Basic Setup

In this paper, we aim to estimate the population total of a study variable y which is defined by $N^{-1}\sum_{i\in U} y_i$ where $U=\{1,2,...,N\}$. Suppose the population information of auxiliary variables $x_1, x_2, ..., x_q$, w and k are known. The auxiliary variables $x_1, x_2, ..., x_q$ were used as calibration variables, we defined $\boldsymbol{x}_i = (1\ x_{i1}\ \cdots\ x_{iq})'$ and $\boldsymbol{X}_N = (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_N)'$ be the $N \times (q+1)$ matrix values of $\boldsymbol{x}_i$ The vector $(w_1\ \ w_2\ \ \cdots\ \ w_N)'$ was defined as the value of the auxiliary variables w for constructing the ratio estimator. The vector of values of auxiliary variables k are $(k_1\ \ k_2\ \ \cdots\ \ k_N)'$ and was used to determine values of first and joint inclusion probabilities under unequal probability sampling without replacement.

Under unequal probability sampling without replacement (UPWOR), a sample s of size n was selected. Let, $\mathscr{F}$ be the set of all possible subsets of U and the sampling design $P(.)$ is the probability measure on the possible s, i.e., $P(s) \geq 0$ for all $s \in \mathscr{F}$. Let $\pi_i = P(i \in s) = \sum_{s \ni i} P(s)$ be the first order inclusion probability and $\pi_{ij} = P(i \wedge j \in s) = \sum_{s \supset \{i,j\}} P(s)$ be the second order inclusion probability. With a sample s of size n assume that the information of $n \times (q+1)$ matrix of values x or $\boldsymbol{X}_n = (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n)'$ is known for all $\boldsymbol{x}_i$ when $i \in s$. We also define $E_S(\bullet)$ and $V_S(\bullet)$ as the expectation and variance operators with respect to the UPWOR sampling design.

In the presence of nonresponse, let subscript R and r_i be the nonresponse mechanism and nonresponse indicator variable of y_i which $r_i = 1$ if unit i responds to item y otherwise $r_i = 0$. Let, $\boldsymbol{R} = (r_1\ r_2\ \cdots\ r_N)'$ be the vector of the response indicator and $p_i = p = P(r_i = 1)$ be the response probability under uniform nonresponse. Let $E_R(\bullet)$ and $V_R(\bullet)$ be the expectation and variance operators with respect to the nonresponse mechanism. The linear assisting model ξ of Särndal, Swensson and Wretman (1992) and Särndal (2007) is considered in this study which defines $E_\xi(y_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$ and $V_\xi(y_i) = \sigma_i^2$, where $\boldsymbol{\beta}$ is the population regression coefficient, $\hat{\boldsymbol{\beta}}_r = \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1} \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i}\right)$ is the estimator of $\boldsymbol{\beta}$ and q_i is determined by the linear assisting model ξ. In the presence of nonresponse, the sample set s of size n is classified into the response sample set s_r and nonresponse set s_m. The size of s_r and s_m are n_r and n_m, respectively. Three assumptions are defined; (A_1) the response mechanism is uniform response (A_2) $\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta} = O_p(n_r^{-\frac{1}{2}})$ and (A_3) $V\left(\sum_{i \in s} b_i \pi_i^{-1}\right) \to 0$ as $n \to \infty$ where $b_i = w_i$ or r_i.

3. The Estimators of Joint Inclusion Probabilities

Recall from Section 2, the first and joint inclusion probabilities are given by $\pi_i = \sum_{s \ni i} P(s)$ and $\pi_{ij} = \sum_{s \supset \{i,j\}} P(s)$, let D_{ij} be the function of π_i and π_{ij} and be defined by

$$D_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}. \tag{1}$$

The estimator of D_{ij} is given by

$$\hat{D}_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j}. \tag{2}$$

Let z_i be a function of y_i such as $z_i = f(y_i) = y_i$, $z_i = f(y_i) = y_i - \overline{Y}$ or $z_i = f(y_i) = y_i - \mathrm{x}_i'\beta$, the estimator for estimating $Z = \sum_{i \in U} z_i$ by using Horvitz and Thompson's (1952) procedure is

$$\hat{Z}_{HT} = \sum_{i \in s} \frac{\hat{z}_i}{\pi_i}, \tag{3}$$

where $\hat{z}_i$ is the estimator of z_i if z_i is unknown otherwise $\hat{z}_i = z_i$. The variance of $\hat{Z}_{HT}$ is given by

$$V\left(\hat{Z}_{HT}\right) = \sum_{i \in U} D_i z_i^2 + \sum_{i \in U} \sum_{j \setminus \{i\} \in U} D_{ij} z_i z_j, \tag{4}$$

where $D_i = (1-\pi_i)\pi_i^{-1}$. The estimator of $V\left(\hat{Z}_{HT}\right)$ is defined as follows

$$\hat{V}\left(\hat{Z}_{HT}\right) = \sum_{i \in s} \hat{D}_i \hat{z}_i^2 + \sum_{i \in s} \sum_{j \setminus \{i\} \in s} \hat{D}_{ij} \hat{z}_i \hat{z}_j, \tag{5}$$

where $\hat{D}_i = (1-\pi_i)\pi_i^{-2}$. The joint inclusion probabilities π_{ij} are difficult to calculate under unequal probability sampling without replacement. Then, we consider two approximations of π_{ij} which were proposed by Hájek (1981) and Brewer and Danadio (2003). The details are below.

3.1. Hájek's (1981) approximation

Hájek (1981) proposed an approximation of π_{ij} based on rejective sampling which is defined by

$$\hat{\pi}_{ij(1)} = \pi_i \pi_j \left[1 - \frac{(1-\pi_i)(1-\pi_j)}{c_1} \right], \tag{6}$$

where $c_1 = \sum_{i \in U} c_i$ and $c_i = \pi_i(1-\pi_i)$.

The formula of c_1 in (6) requires all values of the first order inclusion probabilities $\pi_1, \pi_2, ..., \pi_N$ in population U. Hence, $\hat{\pi}_{ij(1)}$ cannot be computed when the first order inclusion probabilities are only known in the sample. Therefore, the estimator of c_1 is,

$$\hat{c}_2 = \sum_{i \in s} \frac{c_i}{\pi_i} = \sum_{i \in s} \frac{\pi_i(1-\pi_i)}{\pi_i} = n - \sum_{i \in s} \pi_i. \tag{7}$$

Then, an alternative approximation of π_{ij} is given by

$$\hat{\pi}_{ij(2)} = \pi_i \pi_j \left[1 - \frac{(1-\pi_i)(1-\pi_j)}{\hat{c}_2} \right]. \tag{8}$$

Alternatively, Hájek (1981) also estimated c_1 by,

$$\hat{c}_3 \approx N n^{-1} \sum_{i \in s} c_i. \tag{9}$$

Then, the estimator of joint inclusion probabilities is given by

$$\hat{\pi}_{ij(3)} = \pi_i \pi_j \left[1 - \frac{(1-\pi_i)(1-\pi_j)}{\hat{c}_3} \right]. \tag{10}$$

Therefore, the general form of $\hat{\pi}_{ij}$ approximation of Hájek (1981) is defined by

$$\hat{\pi}_{ij(l)} = \pi_i \pi_j \left[1 - \frac{(1-\pi_i)(1-\pi_j)}{\hat{c}_l} \right]. \tag{11}$$

where $l = 1,2,3$ and $\hat{c}_1 = c_1$.

3.2. Brewer and Danadio's (2003) approximation

Brewer and Danadio (2003) also proposed formulas to estimate the joint inclusion probabilities based on the first order inclusion probabilities in population U as follows. Let

$$d = \sum_{i \in U} \pi_i^2, \quad \hat{c}_{4i} = \frac{n-1}{n-\pi_i}, \quad \hat{c}_{5i} = \frac{n-1}{n-n^{-1}d}, \quad \hat{c}_{6i} = \frac{n-1}{n-2\pi_i + n^{-1}d},$$

$$\hat{c}_{7i} = \frac{n-1}{n-(2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1}d}. \tag{12}$$

The general formula of joint inclusion probabilities approximation of Brewer and Danadio (2003) is,

$$\hat{\pi}_{ij(l)} = \frac{\pi_i \pi_j (\hat{c}_{li} + \hat{c}_{lj})}{2}, \; l = 4,5,6,7. \tag{13}$$

In Subsection 3.1 and 3.2, two general forms of joint inclusion probability approximation were discussed. One is the general form of Hájek (1981) approximation, and the other is the general form of Brewer and Danadio (2003) approximation. From (11) and (13), the value of $\hat{D}_{ij}$ in (2) can be estimated by

$$\hat{D}_{ij(l)} = \frac{\hat{\pi}_{ij(l)} - \pi_i \pi_j}{\hat{\pi}_{ij(l)} \pi_i \pi_j}, \; l = 1,2,...,7. \tag{14}$$

The formula of $\hat{V}\left(\hat{Z}_{HT}\right)$ in (5) can be written as

$$\hat{V}_l\left(\hat{Z}_{HT}\right) = \sum_{i \in s} \hat{D}_i \hat{z}_i^2 + \sum_{i \in s} \sum_{i \backslash \{j\} \in s} \hat{D}_{ij(l)} \hat{z}_i \hat{z}_j, \; l = 1,2,...,7. \tag{15}$$

4. Variance estimation with free joint inclusion probability

Variance and associated variance estimators under unequal probability sampling without replacement such as the Horvitz and Thompson (1952) estimator require joint inclusion probabilities in the calculation process. Then, Hájek (1964) proposed an alternative variance of the Horvitz and Thompson (1952) estimator in the case of high entropy sampling as follows. Let $c_i = \pi_i(1-\pi_i)$, $z_{io} = 1$ for all $i \in U$, z_i and $\hat{z}_i$ are defined in (3). From (3), we may write

$$\hat{Z}_{HT} = \sum_{i \in s} \frac{\hat{z}_i}{\pi_i} = \sum_{i \in s} \hat{o}_i,$$

where $\hat{o}_i = \frac{\hat{z}_i}{\pi_i}$. The variance of $\hat{Z}_{HT}$ by using the method proposed by Hájek (1964) is given by

$$V\left(\hat{Z}_{HT}\right) = V\left(\sum_{i \in s} \hat{o}_i\right) = \sum_{i \in U} c_i \left\{\left[\hat{o}_i - \left(\sum_{j \in U} c_j z_{j0}^2\right)^{-1} \left(\sum_{i \in U} c_i \hat{o}_i z_{i0}\right)\right]\right\}.$$

Let $\tilde{c}_i = \frac{c_i}{\pi_i}$ then the estimator of $V\left(\hat{Z}_{HT}\right)$ is defined by

$$\hat{V}\left(\hat{Z}_{HT}\right) = \sum_{i \in s} \tilde{c}_i \left\{\left[\hat{o}_i - \left(\sum_{j \in s} \tilde{c}_j z_{j0}^2\right)^{-1} \left(\sum_{i \in s} \tilde{c}_i \hat{o}_i z_{i0}\right)\right]\right\}.$$

In the presence of nonresponse, Ponkaew and Lawson (2019) proposed a population total estimator under the reverse framework with uniform nonresponse which is defined by

$$\hat{Y}_r^{(1)} = \sum_{i\in s}\frac{r_i y_i}{\pi_i p} = \sum_{i\in s}\frac{\hat{z}_i}{\pi_i} = \sum_{i\in s}\hat{o}_i, \tag{16}$$

where $\hat{z}_i = \frac{r_i y_i}{p}$ and $\hat{o}_i = \frac{\hat{z}_i}{\pi_i}$.

However, the variance estimator of $\hat{Y}_r^{(1)}$ in (16) requires the joint inclusion probability. Then, Lawson (2020) investigated the variance proposed by Ponkaew and Lawson (2019). Lawson (2020) suggested to apply Hájek's (1964) method under the reverse framework and when the sampling fraction is negligible which is defined by

$$\begin{aligned} V_L\left(\hat{Y}_r^{(1)}\right) &\approx E_R\left[V_S\left\{\hat{Y}_r^{(1)}\middle|\boldsymbol{R}\right\}\right] = E_R\left[V_S\left\{\sum_{i\in s}\hat{o}_i\middle|\boldsymbol{R}\right\}\right] \\ &= \sum_{i\in U}c_i\left\{\left[E_R(\hat{o}_i) - \left(\sum_{j\in U}c_j z_{j0}^2\right)^{-1}\left(\sum_{i\in U}c_i E_R(\hat{o}_i)z_{i0}\right)\right]\right\} = \sum_{i\in U}c_i\left\{\left[\tilde{y}_i - \left(\sum_{j\in U}c_j z_{j0}^2\right)^{-1}\left(\sum_{i\in U}c_i\tilde{y}_i z_{i0}\right)\right]\right\}, \end{aligned}$$

where $\tilde{y}_i = \frac{y_i}{\pi_i}$. The estimator of $V\left(\hat{Y}_r^{(1)}\right)$ is defined by

$$\hat{V}_L\left(\hat{Y}_r^{(1)}\right) = \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_i - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\sum_{i\in s}\tilde{c}_i\hat{o}_i z_{i0}\right)\right]\right\}, \tag{17}$$

where $\hat{o}_i = \frac{\hat{z}_i}{\pi_i}, \hat{z}_i = \frac{r_i y_i}{p}$.

5. The existing estimators under uniform nonresponse

In this section, we discuss the existing ratio estimators proposed by Ponkaew and Lawson (2018) and also the GREG estimators proposed by Lawson and Ponkaew (2019) along with their variance estimators. The details are as follows.

5.1. The ratio estimator of Ponkaew and Lawson (2018)

The ratio estimator is a popular method for estimating population total when the information on an auxiliary variable is known for all units in a population U. For full response, Bacanli and Kadilar (2008) and Särndal and Lundström (2005) introduced a ratio estimator under equal probability sampling without replacement. Later, Ponkaew and Lawson (2018) extended their estimator for estimating population total with the uniform nonresponse mechanism and is defined by,

$$\hat{Y}_R^* = \frac{\sum_{i\in s}\frac{r_i y_i}{\pi_i p}}{\sum_{i\in s}\frac{w_i}{\pi_i}}\sum_{i\in U}w_i = \frac{\hat{Y}_r^*}{\hat{w}_{HT}}W, \tag{18}$$

where $\hat{Y}_r^* = \sum_{i\in s}\frac{r_i y_i}{\pi_i p}$, $\hat{w}_{HT} = \sum_{i\in s}\frac{w_i}{\pi_i}$, $W = \sum_{i\in U}w_i$. Under the reverse framework the variance of $\hat{Y}_R^*$ is defined by

$$V(\hat{Y}_R^*) = E_R V_S(\hat{Y}_R^* \mid \boldsymbol{R}) + V_R E_S(\hat{Y}_R^* \mid \boldsymbol{R}) = V_1 + V_2, \tag{19}$$

where $V_1 = E_R V_S(\hat{Y}_R^* \mid \boldsymbol{R})$ and $V_2 = V_R E_S(\hat{Y}_R^* \mid \boldsymbol{R})$. We note that the second term in (19) can be omitted when the sampling fraction (nN^{-1}) is negligible then the variance of $\hat{Y}_R^*$ is approximately by

$$V(\hat{Y}_R^*) \approx V_1 = E_R V_S \left(\frac{\sum_{i\in s} \frac{r_i y_i}{\pi_i p}}{\sum_{i\in s} \frac{w_i}{\pi_i}} \sum_{i\in U} w_i \middle| \boldsymbol{R} \right).$$

However, $\hat{Y}_R^*$ is in the form of a nonlinear estimator so Ponkaew and Lawson (2018) proposed to use the Taylor linearization approach to transform $\hat{Y}_R^*$ to a linear estimator and is defined by

$$\hat{Y}_R^* \approx \sum_{i\in s} \frac{z_i}{\pi_i}, \tag{20}$$

where $z_i = \frac{r_i y_i}{p} - x_i R_r$ and $R_r = \frac{\sum_{i\in U} \frac{r_i y_i}{p}}{\sum_{i\in U} w_i}$. Then,

$$V(\hat{Y}_R^*) \approx V_1 \approx E_R V_S \left(\sum_{i\in s} \frac{z_i}{\pi_i} \middle| \boldsymbol{R} \right) = \sum_{i\in U} D_{2i} y_i^2 + \sum_{i\in U} \sum_{j\backslash\{i\}\in U} D_{ij} e_i e_j - \sum_{i\in U} D_i (2y_i - Rw_i) R w_i,$$

where $D_{2i} = \frac{1-\pi_i}{\pi_i p}$, $e_i = y_i - Rw_i$, and $R = \frac{\overline{Y}}{\overline{X}}$.

The variance of the ratio estimator of Ponkaew and Lawson (2018) is discussed under a situation where the sampling fraction is large and the second component cannot be omitted. Let us consider the second term in (19),

$$V_2 = V_R E_S(\hat{Y}_R^* \mid \boldsymbol{R}) = V_R E_S \left(\frac{\sum_{i\in s} \frac{r_i y_i}{\pi_i p}}{\sum_{i\in s} \frac{w_i}{\pi_i}} \sum_{i\in U} w_i \middle| \boldsymbol{R} \right) \approx \sum_{i\in U} \frac{(1-p) y_i^2}{p}.$$

The estimator of V_2 is given by

$$\hat{V}_2 \approx \frac{(1-p')}{(p')^2} \sum_{i\in s} r_i y_i^2, \tag{21}$$

where $p' = p$ if p is known otherwise $p' = \hat{p} = \sum_{i\in s} \frac{r_i}{\pi_i} \left(\sum_{i\in s} \frac{1}{\pi_i} \right)^{-1}$.

Then, a new variance estimator is shown as follows.

$$V(\hat{Y}_R^*) \approx \sum_{i\in U} D_{2i} y_i^2 + \sum_{i\in U} \sum_{j\backslash\{i\}\in U} D_{ij} e_i e_j - \sum_{i\in U} D_i (2y_i - Rw_i) R w_i + \sum_{i\in U} \frac{(1-p) y_i^2}{p}.$$

The estimator of variance of Ponkaew and Lawson (2018) under a large sampling fraction is given by

$$\hat{V}(\hat{Y}_R^*) \approx \frac{1}{p'} \sum_{i\in s} r_i \hat{D}_{2i} y_i^2 + \frac{1}{(p')^2} \sum_{i\in s} \sum_{j\backslash\{i\}\in s} r_i r_j \hat{D}_{ij} e_i e_j - \frac{1}{p'} \sum_{i\in s} r_i \hat{D}_i (2y_i - \hat{R}_r w_i) \hat{R}_r w_i + \frac{(1-p')}{(p')^2} \sum_{i\in s} r_i y_i^2,$$

$$= \frac{1}{p'} \sum_{i\in s} r_i \left[\hat{D}_{2i} y_i^2 - \hat{D}_i (2y_i - \hat{R} w_i) \hat{R} w_i + \frac{(1-p')}{p'} y_i^2 \right] + \frac{1}{(p')^2} \sum_{i\in s} \sum_{j\backslash\{i\}\in s} r_i r_j \hat{D}_{ij} \hat{e}_i \hat{e}_j, \tag{22}$$

where $p' = p$ if p is known otherwise $p' = \hat{p} = \sum_{i\in s}\frac{r_i}{\pi_i}\left(\sum_{i\in s}\frac{1}{\pi_i}\right)^{-1}$, $\hat{D}_{2i} = \frac{1-\pi_i}{\pi_i^2 p}$ and

$\hat{R}_r = \sum_{i\in s}\frac{r_i y_i}{\pi_i p'}\left(\sum_{i\in s}\frac{w_i}{\pi_i}\right)^{-1}$.

5.2. The GREG estimator of Lawson and Ponkaew (2019)

Under the reverse framework with uniform nonresponse Lawson and Ponkaew (2019) proposed the GREG estimator for estimating population mean given by

$$\hat{\bar{Y}}_{GREG.LP} = \frac{\sum_{i\in s}\frac{r_i y_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}} + \left(\bar{\boldsymbol{X}} - \frac{\sum_{i\in s}\frac{r_i \boldsymbol{x}_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)'\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i}\right) = \hat{\bar{Y}}_r + \left(\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}}_r\right)'\hat{\boldsymbol{\beta}}_r. \tag{23}$$

The estimator of population total can be obtained by

$$\hat{Y}_{GREG.LP} = N\hat{\bar{Y}}_{GREG.LP} = N\left[\hat{\bar{Y}}_r + \left(\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}}_r\right)'\hat{\boldsymbol{\beta}}_r\right], \tag{24}$$

where $\hat{\bar{Y}}_r = \sum_{i\in s}\frac{r_i y_i}{\pi_i}\Big/\sum_{i\in s}\frac{r_i}{\pi_i}$, $\hat{\bar{\boldsymbol{X}}}_r = \sum_{i\in s}\frac{r_i \boldsymbol{x}_i}{\pi_i}\Big/\sum_{i\in s}\frac{r_i}{\pi_i}$, $\hat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i}\right)$, and $\bar{\boldsymbol{X}} = \frac{1}{N}\sum_{i\in U}\boldsymbol{x}_i$.

The GREG estimator for estimating population mean of Lawson and Ponkaew (2019) is in the form of a nonlinear estimator so the modified automated linearization approach used to transform $\hat{\bar{Y}}_{GREG.LP}$. $\hat{\bar{Y}}_{GREG.LP}$ is given by

$$\hat{\bar{Y}}_{GREG.LP} \approx \bar{\boldsymbol{X}}'\boldsymbol{\beta} + \frac{1}{\sum_{i\in s}\frac{r_i}{\pi_i}}\sum_{i\in s}\frac{r_i e_i}{\pi_i}, \tag{25}$$

where $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. However, the GREG estimator $\hat{\bar{Y}}_{GREG.LP}$ is still in a nonlinear form so Lawson and Ponkaew (2019) suggested to use two methods to adjust $\hat{\bar{Y}}_{GREG.LP}$ as follows:

Method 1: Replace $\sum_{i\in s}\frac{r_i}{\pi_i}$ by $\sum_{i\in U} r_i$

Let $\hat{\bar{Y}}_{GREG.LP(1)} \approx \bar{\boldsymbol{X}}'\boldsymbol{\beta} + \frac{1}{\sum_{i\in U} r_i}\sum_{i\in s}\frac{r_i e_i}{\pi_i}$. From the assumption (A_3) $V\left(\sum_{i\in s}\frac{r_i}{\pi_i}\right) \to 0$ as $n \to \infty$ then $\sum_{i\in s}\frac{r_i}{\pi_i}$ converges to $\sum_{i\in U} r_i$ in probability. Therefore, the linear estimator of $\hat{\bar{Y}}_{GREG.LP}$ can be approximated by $\hat{\bar{Y}}_{GREG.LP(1)}$ because $\hat{\bar{Y}}_{GREG.LP}$ converges to $\hat{\bar{Y}}_{GREG.LP(1)}$ in probability.

Method 2: The Taylor linearization approach

$\hat{\bar{Y}}_{GREG.LP}$ from (25) can be rewritten as,

$$\hat{\bar{Y}}_{GREG.LP} \approx g(\hat{\boldsymbol{t}}_s) = \bar{\boldsymbol{X}}'\boldsymbol{\beta} + \frac{\hat{t}_1}{\hat{t}_2}, \tag{26}$$

where $g(.)$ is a smooth function of $\hat{\boldsymbol{t}}_s$, $\hat{t}_1 = \sum_{i \in s} \frac{r_i e_i}{\pi_i}$, $\hat{t}_2 = \sum_{i \in U} r_i$, $\hat{\boldsymbol{t}}_s = [\hat{t}_1 \;\; \hat{t}_2]'$. Let $t_1 = \sum_{i \in U} r_i e_i, t_2 = \sum_{i \in U} r_i$, $\boldsymbol{t} = [t_1 \;\; t_2]'$, $g(\boldsymbol{t}) = \bar{\boldsymbol{X}}'\boldsymbol{\beta} + \frac{\sum_{i \in U} r_i e_i}{\sum_{i \in U} r_i}$. Then, the linear estimator of $\hat{\hat{Y}}_{GREG.LP}$ can be obtained by using the Taylor linearization approach and is defined by

$$\hat{\hat{Y}}_{GREG.LP(2)} \approx g(\boldsymbol{t}_s) + \sum_{j=1}^{2} \left. \frac{\partial g(\hat{\boldsymbol{t}}_s)}{\partial \hat{t}_j} \right|_{\hat{\boldsymbol{t}} = \boldsymbol{t}} (\hat{t}_j - t_j), \tag{27}$$

Therefore, the variance of $\hat{\hat{Y}}_{GREG.LP}$ under the reverse framework and a negligible sampling fraction can be approximated by

$$V_m(\hat{\hat{Y}}_{GREG.LP}) = V_1 \approx E_R \left[V_S \left(\frac{1}{\sum_{i \in U} r_i} \sum_{i \in s} \frac{z_{mi}}{\pi_i} \middle| \boldsymbol{R} \right) \right], \tag{28}$$

where $m = 1, 2$, $z_{1i} = r_i e_i, z_{2i} = r_i(e_i - \bar{e}_r)$, $e_i = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$ and $\bar{e}_r = \sum_{i \in U} r_i e_i \Big/ \sum_{i \in U} r_i$. The variance of $\hat{Y}_{GREG.LP}$ can be approximated from

$$V_m(\hat{Y}_{GREG.LP}) = V_1 \approx E_R \left[V_S \left(\frac{N}{\sum_{i \in U} r_i} \sum_{i \in s} \frac{z_{mi}}{\pi_i} \middle| \boldsymbol{R} \right) \right], \tag{29}$$

where $V_m(\hat{Y}_{GREG.LP})$, $m = 1, 2$ are the first components of variance under the reverse framework and using method m to transform the GREG estimator to a linear estimator.

Then, the variances of the GREG estimator for estimating population total are given by,

$$V_1(\hat{Y}_{GREG.LP}) \approx \frac{1}{p} \sum_{i \in U} D_i e_i^2 + \sum_{i \in U} \sum_{j \backslash \{i\} \in U} D_{ij} e_i e_j,$$

$$V_2(\hat{Y}_{GREG.LP}) \approx \frac{1}{p} \sum_{i \in U} D_i (e_i - \bar{e})^2 + \sum_{i \in U} \sum_{j \backslash \{i\} \in U} D_{ij} (e_i - \bar{e})(e_j - \bar{e}),$$

where $D_i = (1 - \pi_i)\pi_1^{-1}$; $D_{ij} = (\pi_i \pi_j - \pi_{ij})(\pi_i \pi_j)^{-1}$; $e_i = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$, $\bar{e} = \frac{1}{N} \sum_{i \in U} e_i$.

Next, we extended the variances of $\hat{Y}_{GREG.LP}$ to a situation where the sampling fraction was not negligible and are defined by

$$V_1(\hat{Y}_{GREG.LP}) \approx \frac{1}{p} \sum_{i \in U} D_i e_i^2 + \sum_{i \in U} \sum_{j \backslash \{i\} \in U} D_{ij} e_i e_j + \frac{(1-p)}{p} \sum_{i \in U} (e_i - \bar{e})^2, \tag{30}$$

$$V_2(\hat{Y}_{GREG.LP}) \approx \frac{1}{p} \sum_{i \in U} \frac{(1 - \pi_i)}{\pi_i} (e_i - \bar{e})^2 + \sum_{i \in U} \sum_{j \backslash \{i\} \in U} D_{ij} (e_i - \bar{e})(e_j - \bar{e}) + \frac{(1-p)}{p} \sum_{i \in U} (e_i - \bar{e})^2, \tag{31}$$

where $\bar{e} = N^{-1} \sum_{i \in U} e_i$. We note that, the third component in (30) and (31) is the second term of variance of $\hat{Y}_{GREG.LP}$ under the reverse framework that is,

$$V_2 = V_R \left[E_S \left(\frac{N}{\sum_{i \in U} r_i} \sum_{i \in s} \frac{z_{mi}}{\pi_i} \middle| \boldsymbol{R} \right) \right] \approx \frac{(1-p)}{p} \sum_{i \in U} (e_i - \bar{e})^2. \tag{32}$$

The estimator of V_2 is given by

$$\hat{V}_2 \approx \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2. \tag{33}$$

where $p' = p$ if p is known otherwise $p' = \hat{p} = \sum_{i\in s}\frac{r_i}{\pi_i}\left(\sum_{i\in s}\frac{1}{\pi_i}\right)^{-1}$, $\hat{e}_i = (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_r)$ and $\hat{\bar{e}} = \frac{\sum_{i\in s} r_i\hat{e}_i/\pi_i}{N}$.

Therefore, the estimators of $V_1(\hat{Y}_{GREG.LP})$ and $V_2(\hat{Y}_{GREG.LP})$ are obtained respectively by,

$$\hat{V}_1(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i\hat{e}_i^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij}r_i\hat{e}_i r_i\hat{e}_j\right] + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2. \tag{34}$$

$$\hat{V}_2(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i(\hat{e}_i-\hat{\bar{e}}_r)^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij}r_i(\hat{e}_i-\hat{\bar{e}}_r)^2 r_j(\hat{e}_j-\hat{\bar{e}}_r)^2\right]$$

$$+\frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2, \tag{35}$$

where $\hat{D}_i = (1-\pi_i)\pi_1^{-2}$, $\hat{D}_{ij} = (\pi_i\pi_j - \pi_{ij})(\pi_{ij}\pi_i\pi_j)^{-1}$; $\hat{e}_i = (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_r)$, $\hat{\bar{e}}_r = \frac{\sum_{i\in s} r_i\hat{e}_i/\pi_i}{\sum_{i\in s} r_i/\pi_i}$.

6. Proposed variance estimators

The formulas of the variance estimators of the ratio and GREG estimators in Section 5 require joint inclusion probability because the first component of total variance is in the form of Horvitz and Thompson's (1952) formula. In this section, we propose variance estimators under two approaches by estimating the joint inclusion probability or free joint inclusion probability as follows.

6.1. The proposed variance estimators of a ratio estimator

6.1.1 The proposed variance estimators of a ratio estimator with estimated joint inclusion probability

From (14) the formula of $\hat{D}_{ij(k)}$ is defined by

$$\hat{D}_{ij(k)} = \frac{\hat{\pi}_{ij(k)} - \pi_i\pi_j}{\hat{\pi}_{ij(k)}\pi_i\pi_j},\ k = 1,2,\ldots,7. \tag{36}$$

Let us consider the variance estimator of the ratio estimator from (22),

$$\hat{V}(\hat{Y}_R^*) \approx \frac{1}{p'}\sum_{i\in s}r_i\left[\hat{D}_{2i}y_i^2 - \hat{D}_i(2y_i - \hat{R}w_i)\hat{R}w_i + \frac{(1-p')}{p'}y_i^2\right] + \frac{1}{(p')^2}\sum_{i\in s}\sum_{i\setminus\{j\}\in s} r_i r_j\hat{D}_{ij}\hat{e}_i\hat{e}_j,$$

Substitute $\hat{D}_{ij}$ by $\hat{D}_{ij(k)}$ then the proposed variance estimators of ratio estimator are defined by

$$\hat{V}'^{(k)}(\hat{Y}_R^*) \approx \frac{1}{p'}\sum_{i\in s}r_i\left[\hat{D}_{2i}y_i^2 - \hat{D}_i(2y_i - \hat{R}w_i)\hat{R}w_i + \frac{(1-p')}{p'}y_i^2\right] + \frac{1}{(p')^2}\sum_{i\in s}\sum_{i\setminus\{j\}\in s} r_i r_j\hat{D}_{ij(k)}\hat{e}_i\hat{e}_j, \tag{37}$$

where $k = 1,2,\ldots,7$.

6.1.2 The proposed variance estimators of the ratio estimator with free joint inclusion probability

In this subsection we investigate the free joint inclusion probability of the variance estimator by using Lawson's (2018) method as follows. From (20), the linear form of the ratio estimator is

$$\hat{Y}_R^* \approx \sum_{i\in s}\frac{z_i}{\pi_i}, \tag{38}$$

where $z_i = \dfrac{r_i y_i}{p} - x_i R_r$. From (38), we may write

$$\hat{Y}_R^* \approx \sum_{i\in s}\frac{z_i}{\pi_i} = \sum_{i\in s}\frac{z_i}{\pi_i} = \sum_{i\in s} o_i,$$

where $o_i = \dfrac{z_i}{\pi_i}$. By using Lawson's (2020) method,

$$\hat{V}_1 = \hat{V}_L(\hat{Y}_R^*) \approx \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_i - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\sum_{i\in s}\tilde{c}_i\hat{o}_i z_{i0}\right)\right]\right\},$$

where $\hat{o}_i = \dfrac{\hat{z}_i}{\pi_i}$, $\hat{z}_i = \dfrac{r_i y_i}{p'} - x_i\hat{R}_r$, $p' = \hat{p} = \sum_{i\in s}\dfrac{r_i}{\pi_i}\left(\sum_{i\in s}\dfrac{1}{\pi_i}\right)^{-1}$ if p is unknown otherwise $p' = p$. From (21), $\hat{V}_2$ is given by

$$\hat{V}_2 \approx \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i y_i^2}{\pi_i}.$$

Then, the free joint inclusion probability of the variance estimator of the ratio estimator is defined as

$$\hat{V}''(\hat{Y}_R^*) = \hat{V}_1 + \hat{V}_2 = \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_i - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\sum_{i\in s}\tilde{c}_i\hat{o}_i z_{i0}\right)\right]\right\} + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i y_i^2}{\pi_i}. \tag{39}$$

6.2 The proposed variance estimators of the GREG estimator

6.2.1 The proposed variance estimators of the GREG estimator with estimated joint inclusion probability

In Subsection 5.2, we used two methods to transform the GREG estimator to a linear estimator and discussed variance estimators in (34) and (35). In this subsection, we propose new variance estimators of the GREG estimator by estimating the joint inclusion probability in (34) and (35) as follows.

From (34), the variance estimator under Method 1 is defined by

$$\hat{V}_1(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i\hat{e}_i^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij} r_i\hat{e}_i r_i\hat{e}_j\right] + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i - \hat{\bar{e}})^2, \tag{40}$$

where $\hat{D}_i = (1-\pi_i)\pi_1^{-2}$, $\hat{D}_{ij} = (\pi_i\pi_j - \pi_{ij})(\pi_{ij}\pi_i\pi_j)^{-1}_{,}$ $\hat{e}_i = (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_r)$, $\hat{\bar{e}}_r = \dfrac{\sum_{i\in s} r_i\hat{e}_i/\pi_i}{\sum_{i\in s} r_i/\pi_i}$ and

$p' = p$ if p is known otherwise $p' = \hat{p} = \sum_{i\in s}\frac{r_i}{\pi_i}\left(\sum_{i\in s}\frac{1}{\pi_i}\right)^{-1}.$

The proposed variance estimators of the GREG estimator with Method 1 can be obtained by substituting $\hat{D}_{ij}$ in (37) by $\hat{D}_{ij(k)}$. Therefore,

$$\hat{V}_1'^{(k)}(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i \hat{e}_i^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij(k)} r_i\hat{e}_i r_i\hat{e}_j\right] + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i - \hat{\bar{e}})^2. \quad (41)$$

From (35), the variance estimator under Method 2 is defined by,

$$\hat{V}_2(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i(\hat{e}_i - \hat{\bar{e}}_r)^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij} r_i(\hat{e}_i - \hat{\bar{e}}_r)^2 r_j(\hat{e}_j - \hat{\bar{e}}_r)^2\right]$$
$$+\frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i - \hat{\bar{e}})^2. \quad (42)$$

Substitute $\hat{D}_{ij}$ in (42) with $\hat{D}_{ij(k)}$, the proposed variance estimator by Method 2 is given by,

$$\hat{V}_2'^{(k)}(\hat{Y}_{GREG.LP}) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2\left[\sum_{i\in s}\hat{D}_i r_i(\hat{e}_i - \hat{\bar{e}}_r)^2 + \sum_{i\in s}\sum_{j/\{i\}\in s}\hat{D}_{ij(k)} r_i(\hat{e}_i - \hat{\bar{e}}_r)^2 r_j(\hat{e}_j - \hat{\bar{e}}_r)^2\right]$$
$$+\frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i - \hat{\bar{e}})^2, \quad (43)$$

where $k = 1, 2, ..., 7$.

6.2.2 The proposed variance estimators of the GREG estimator with free joint inclusion probability

We investigate the estimator of variance of the GREG estimator by using Lawson's (2020) method. From (29), we may write

$$V_{1m} \approx E_R\left[V_S\left(\frac{N}{\sum_{i\in U} r_i}\sum_{i\in s}\frac{z_{mi}}{\pi_i}\middle|\boldsymbol{R}\right)\right] = E_R\left[V_S\left(\sum_{i\in s} o_{mi}\middle|\boldsymbol{R}\right)\right],$$

where $m = 1, 2,$ $o_{mi} = \frac{N}{\sum_{i\in U} r_i}\frac{z_{mi}}{\pi_i},$ $z_{1i} = r_i e_i,$ $z_{2i} = r_i(e_i - \bar{e}_r),$ $e_i = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$ and $\bar{e}_r = \sum_{i\in U} r_i e_i \Big/ \sum_{i\in U} r_i$.

By using Lawson's (2020) method, the estimator of V_{1m} is defined by

$$\hat{V}_{1m} \approx \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_{mip'} - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\sum_{i\in s}\tilde{c}_i\hat{o}_{mip'} z_{i0}\right)\right]\right\}, \quad (44)$$

where $m = 1, 2,$ $\hat{o}_{mip'} = \frac{N}{\hat{N}_r}\frac{\hat{z}_{mip'}}{\pi_i},$ $\hat{z}_{1i} = r_i\hat{e}_i,$ $z_{2i} = r_i(\hat{e}_i - \hat{\bar{e}}_r),$ $e_i = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_r)$ and $\hat{\bar{e}}_r = \frac{\sum_{i\in s}\frac{r_i\hat{e}_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}}.$

Recall from (33) $\hat{V}_2$ is defined by

$$\hat{V}_2 \approx \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2.$$

Then, the free joint inclusion probability of the variance estimator of the GREG estimator can be obtained by

$$\hat{V}_m''(\hat{Y}_{GREG.LP}) = \hat{V}_{1m} + \hat{V}_2$$

$$\approx \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_{mip'} - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\left(\sum_{i\in s}\tilde{c}_i\hat{o}_{mip'}z_{i0}\right)\right)\right]\right\} + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2, \quad (45)$$

where $m = 1,2$. The proposed variance estimator with free joint inclusion probability under Method 1 is defined by

$$\hat{V}_1''(\hat{Y}_{GREG.LP}) = \hat{V}_{11} + \hat{V}_2$$

$$\approx \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_{1ip'} - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\left(\sum_{i\in s}\tilde{c}_i\hat{o}_{1ip'}z_{i0}\right)\right)\right]\right\} + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2. \quad (46)$$

Moreover, the proposed variance estimator with free joint inclusion probability under Method 2 is given by

$$\hat{V}_2''(\hat{Y}_{GREG.LP}) = \hat{V}_{12} + \hat{V}_2$$

$$\approx \sum_{i\in s}\tilde{c}_i\left\{\left[\hat{o}_{2ip'} - \left(\sum_{j\in s}\tilde{c}_j z_{j0}^2\right)^{-1}\left(\left(\sum_{i\in s}\tilde{c}_i\hat{o}_{2ip'}z_{i0}\right)\right)\right]\right\} + \frac{(1-p')}{(p')^2}\sum_{i\in s}\frac{r_i}{\pi_i}(\hat{e}_i-\hat{\bar{e}})^2. \quad (47)$$

7. Simulation studies

To compare the efficiency of the proposed variance estimators, we apply Sichera's (2020) model to generate a population consisting of y, a study variable, x and w are auxiliary variables and k is a size variable which is from the model from package "jipApprox" in R program (R Core Team 2021). The Sichera model is used to investigate the estimate values of the joint inclusion probabilities under unequal probability sampling which we also considered in this study. The model is $y_i = 0.2x_i + 0.01w_i + 2k + 3.7k^{\frac{1}{2}} + \varepsilon_i$, where $k_i \sim gamma(10,5), w_i \sim gamma(5,10), x_i \sim gamma(5,5)$, $\varepsilon_i \sim N(0,1)$, $i = 1,2,,...,N$ and the population size is $N = 1,000$. We consider four levels of sample sizes that is $n = 50$, $n = 100$, $n = 150$ and $n = 300$. The sample sizes were selected using unequal probability sampling without replacement. Two levels of response probabilities; 70% and 85% were considered in the simulation studies and repeated in the study 10,000 times (M =10,000). The relative root mean square error (RRMSE), the relative bias (RB), and the coverage rate (CR) are used to compare the efficiencies of the proposed variance estimators in this study and the formulas are as follows.

(i) The relative root mean square error, $RRMSE\left(\hat{V}(\hat{Y})\right) = \dfrac{\sqrt{\dfrac{1}{M-1}\sum_{m=1}^{M}(\hat{V}_m\left(\hat{Y}\right) - V(\hat{Y})^2}}{V(\hat{Y})}$, where $\hat{V}(\hat{Y})$ is the proposed variance estimator and $V(\hat{Y})$ is the true value of variance.

(ii) The relative bias, $RB\left(\hat{V}(\hat{Y})\right)=\dfrac{E\left(\hat{V}\left(\hat{Y}\right)\right)-V(\hat{Y})}{V(\hat{Y})}$, where $E\left(\hat{V}\left(\hat{Y}\right)\right)=\dfrac{1}{M}\sum_{m=1}^{M}\hat{V}_m\left(\hat{Y}\right)$,

(iii) The coverage rate for a nominal 95% level, $CR\left[\hat{V}(\hat{Y})\right]=\dfrac{1}{M}\sum_{m=1}^{M}I_{(m)}$, where $I_{(m)}=1$ if $Y\in[l_m,u_m]$ otherwise $I_{(m)}=0$ and $l_m=\hat{Y}-1.960\sqrt{\hat{V}(\hat{Y})}$, and $l_m=\hat{Y}+1.960\sqrt{\hat{V}(\hat{Y})}$.

In this study, we consider only the case where the response probability is unknown because it is more practical in real life. The proposed variance estimators can be classified into two cases; estimated joint inclusion probability variance estimators and free joint inclusion probability variance estimators.

Table 1 shows the relative root mean square error of the variance estimator of the ratio estimator. The result showed that the free joint inclusion probability variance estimators from (39) gave a smaller relative root mean square error than the estimated joint inclusion probability for all situations. Table 2 and Table 3 shows the relative root mean square error of the variance estimator of the GREG estimator under Methods 1 and 2, respectively. The results from Tables 2 and 3 are similar to Table 1, that is the free joint inclusion probability variance estimators from (46) and (47) respectively gave a smaller relative root mean square error than the estimated joint inclusion probability variance estimators.

Table 1 The relative root mean square error of the variance estimators of the ratio estimator

Response rate (%)	*RRMSE*	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RRMSE\left(\hat{V}'^{(1)}(\hat{Y}_R^*)\right)$	0.82372	0.81767	0.81115	0.78965
	$RRMSE\left(\hat{V}'^{(2)}(\hat{Y}_R^*)\right)$	0.82372	0.81767	0.81115	0.78965
	$RRMSE\left(\hat{V}'^{(3)}(\hat{Y}_R^*)\right)$	0.82395	0.81777	0.81121	0.78961
	$RRMSE\left(\hat{V}'^{(4)}(\hat{Y}_R^*)\right)$	0.82373	0.81765	0.81105	0.78882
	$RRMSE\left(\hat{V}'^{(5)}(\hat{Y}_R^*)\right)$	0.82376	0.81768	0.81110	0.78890
	$RRMSE\left(\hat{V}'^{(6)}(\hat{Y}_R^*)\right)$	0.82370	0.81761	0.81100	0.78873
	$RRMSE\left(\hat{V}'^{(7)}(\hat{Y}_R^*)\right)$	0.82370	0.81761	0.81100	0.78873
	$RRMSE\left(\hat{V}''(\hat{Y}_R^*)\right)$	**0.47344**	**0.42117**	**0.31730**	**0.29570**
0.85	$RRMSE\left(\hat{V}'^{(1)}(\hat{Y}_R^*)\right)$	0.73648	0.72061	0.71201	0.67729
	$RRMSE\left(\hat{V}'^{(2)}(\hat{Y}_R^*)\right)$	0.73648	0.72061	0.71201	0.67729
	$RRMSE\left(\hat{V}'^{(3)}(\hat{Y}_R^*)\right)$	0.73669	0.72071	0.71205	0.67724
	$RRMSE\left(\hat{V}'^{(4)}(\hat{Y}_R^*)\right)$	0.73650	0.72057	0.71188	0.67643
	$RRMSE\left(\hat{V}'^{(5)}(\hat{Y}_R^*)\right)$	0.73653	0.72058	0.71188	0.67641
	$RRMSE\left(\hat{V}'^{(6)}(\hat{Y}_R^*)\right)$	0.73647	0.72056	0.71187	0.67644
	$RRMSE\left(\hat{V}'^{(7)}(\hat{Y}_R^*)\right)$	0.73647	0.72056	0.71187	0.67644
	$RRMSE\left(\hat{V}''(\hat{Y}_R^*)\right)$	**0.38682**	**0.31888**	**0.30396**	**0.28637**

Table 2 The relative root mean square error of the variance estimators of the GREG estimator under Method 1

Response rate (%)	$RRMSE$	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RRMSE\left(\hat{V}_1'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.79349	0.73205	0.65174	0.59526
	$RRMSE\left(\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.79349	0.73204	0.65172	0.59523
	$RRMSE\left(\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.79056	0.73077	0.65131	0.59547
	$RRMSE\left(\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.79272	0.73238	0.65388	0.60068
	$RRMSE\left(\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.79164	0.73125	0.65259	0.59967
	$RRMSE\left(\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.79381	0.73350	0.65518	0.60170
	$RRMSE\left(\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.79383	0.73351	0.65518	0.60171
	$RRMSE\left(\hat{V}_1''(\hat{Y}_{GREG.LP})\right)$	**0.48835**	**0.33498**	**0.26434**	**0.18547**
0.85	$RRMSE\left(\hat{V}_1'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.60449	0.51598	0.51045	0.46472
	$RRMSE\left(\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.60451	0.51596	0.51044	0.46465
	$RRMSE\left(\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.60227	0.51562	0.50955	0.46609
	$RRMSE\left(\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.60347	0.51794	0.51090	0.48353
	$RRMSE\left(\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.60219	0.51639	0.50994	0.48208
	$RRMSE\left(\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.60475	0.51950	0.51186	0.48500
	$RRMSE\left(\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.60478	0.51951	0.51187	0.48500
	$RRMSE\left(\hat{V}_1''(\hat{Y}_{GREG.LP})\right)$	**0.43639**	**0.26975**	**0.25404**	**0.16953**

Table 3 The relative root mean square error of the variance estimators of the GREG estimator under Method 2

Response rate (%)	*RRMSE*	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RRMSE\left(\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.52824	0.36853	0.31136	0.26758
	$RRMSE\left(\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.52823	0.36869	0.31157	0.26772
	$RRMSE\left(\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.61558	0.43684	0.36607	0.31053
	$RRMSE\left(\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.52184	0.35741	0.29557	0.24719
	$RRMSE\left(\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.51625	0.34694	0.28027	0.22712
	$RRMSE\left(\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.52749	0.36812	0.31142	0.26833
	$RRMSE\left(\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.52761	0.36823	0.31152	0.26844
	$RRMSE\left(\hat{V}_2''(\hat{Y}_{GREG.LP})\right)$	**0.48835**	**0.33498**	**0.26434**	**0.18547**
0.85	$RRMSE\left(\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.51392	0.40042	0.37936	0.33858
	$RRMSE\left(\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.51398	0.40041	0.37943	0.33862
	$RRMSE\left(\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.53580	0.42345	0.39738	0.34569
	$RRMSE\left(\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.51005	0.39136	0.36585	0.31241
	$RRMSE\left(\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.50639	0.38216	0.35148	0.27913
	$RRMSE\left(\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.51376	0.40070	0.38056	0.34693
	$RRMSE\left(\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.51384	0.40080	0.38066	0.34704
	$RRMSE\left(\hat{V}_2''(\hat{Y}_{GREG.LP})\right)$	**0.43639**	**0.26975**	**0.25404**	**0.16953**

Table 4 shows the relative bias of the variance estimators of the ratio estimator while Tables 5 and 6 show the relative bias of the variance estimators of the GREG estimator with methods 1 and 2 respectively. We can see similar results for Tables 4-6 where the free joint inclusion probability variance estimators gave the best results compared to others for all scenarios with smaller biases when the sample size and response rate are increased.

Table 4 The relative bias of the variance estimators of the ratio estimator

Response rate (%)	RB	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RB\left(\hat{V}'^{(1)}(\hat{Y}_R^*)\right)$	0.6977	0.6953	0.6800	0.6566
	$RB\left(\hat{V}'^{(2)}(\hat{Y}_R^*)\right)$	0.6977	0.6953	0.6800	0.6566
	$RB\left(\hat{V}'^{(3)}(\hat{Y}_R^*)\right)$	0.6983	0.6956	0.6801	0.6565
	$RB\left(\hat{V}'^{(4)}(\hat{Y}_R^*)\right)$	0.6978	0.6953	0.6798	0.6547
	$RB\left(\hat{V}'^{(5)}(\hat{Y}_R^*)\right)$	0.6979	0.6955	0.6799	0.6550
	$RB\left(\hat{V}'^{(6)}(\hat{Y}_R^*)\right)$	0.6977	0.6952	0.6796	0.6543
	$RB\left(\hat{V}'^{(7)}(\hat{Y}_R^*)\right)$	0.6977	0.6952	0.6796	0.6543
	$RB\left(\hat{V}''(\hat{Y}_R^*)\right)$	0.3964	0.3885	0.3800	0.3646
0.85	$RB\left(\hat{V}'^{(1)}(\hat{Y}_R^*)\right)$	0.7476	0.747	0.7326	0.7132
	$RB\left(\hat{V}'^{(2)}(\hat{Y}_R^*)\right)$	0.7476	0.7475	0.7326	0.7132
	$RB\left(\hat{V}'^{(3)}(\hat{Y}_R^*)\right)$	0.7482	0.7478	0.7328	0.7130
	$RB\left(\hat{V}'^{(4)}(\hat{Y}_R^*)\right)$	0.7476	0.7474	0.7323	0.7107
	$RB\left(\hat{V}'^{(5)}(\hat{Y}_R^*)\right)$	0.7477	0.7476	0.7325	0.7110
	$RB\left(\hat{V}'^{(6)}(\hat{Y}_R^*)\right)$	0.7475	0.7473	0.7322	0.7105
	$RB\left(\hat{V}'^{(7)}(\hat{Y}_R^*)\right)$	0.7475	0.7473	0.7322	0.7105
	$RB\left(\hat{V}''(\hat{Y}_R^*)\right)$	0.1773	0.1732	0.1638	0.1629

Table 5 The relative bias of the variance estimators of the GREG estimator under Method 1

Response rate (%)	RB	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RB\left(\hat{V}_1'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.4841	0.3525	0.3503	0.1917
	$RB\left(\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.4841	0.3525	0.3503	0.1917
	$RB\left(\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.4853	0.3517	0.3484	0.1878
	$RB\left(\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.5031	0.3551	0.3510	0.1915
	$RB\left(\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.5023	0.3545	0.3504	0.1911
	$RB\left(\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.5039	0.3557	0.3516	0.1920
	$RB\left(\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.5039	0.3557	0.1262	0.1920
	$RB\left(\hat{V}_1''(\hat{Y}_{GREG.LP})\right)$	0.1308	0.1262	0.0932	0.0903
0.85	$RB\left(\hat{V}_1'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.3551	0.296	0.2438	0.1542
	$RB\left(\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.3551	0.2960	0.2438	0.1542
	$RB\left(\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.3568	0.2953	0.2421	0.1504
	$RB\left(\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.3793	0.2995	0.2449	0.1541
	$RB\left(\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.3786	0.2989	0.2443	0.1537
	$RB\left(\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.3800	0.3000	0.2454	0.1545
	$RB\left(\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.3800	0.3000	0.2454	0.1545
	$RB\left(\hat{V}_1''(\hat{Y}_{GREG.LP})\right)$	0.0382	0.0207	0.0078	0.0013

Table 6 The relative bias of the variance estimators of the GREG estimator under Method 2

Response rate (%)	RB	Sample size (n)			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$RB\left(\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.4841	0.3525	0.3503	0.1917
	$RB\left(\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.4841	0.3525	0.3503	0.1917
	$RB\left(\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.4853	0.3517	0.3484	0.1878
	$RB\left(\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.5031	0.3551	0.3510	0.1915
	$RB\left(\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.5023	0.3545	0.3504	0.1911
	$RB\left(\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.5039	0.3557	0.3516	0.1920
	$RB\left(\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.5039	0.3557	0.1262	0.1920
	$RB\left(\hat{V}_2''(\hat{Y}_{GREG.LP})\right)$	0.1308	0.1262	0.0932	0.0903
0.85	$RB\left(\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})\right)$	0.3551	0.296	0.2438	0.1542
	$RB\left(\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})\right)$	0.3551	0.2960	0.2438	0.1542
	$RB\left(\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})\right)$	0.3568	0.2953	0.2421	0.1504
	$RB\left(\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})\right)$	0.3793	0.2995	0.2449	0.1541
	$RB\left(\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})\right)$	0.3786	0.2989	0.2443	0.1537
	$RB\left(\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})\right)$	0.3800	0.3000	0.2454	0.1545
	$RB\left(\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})\right)$	0.3800	0.3000	0.2454	0.1545
	$RB\left(\hat{V}_2''(\hat{Y}_{GREG.LP})\right)$	0.0382	0.0207	0.0078	0.0013

Table 7 shows the simulation coverage rate for a nominal 95% of the ratio estimator while Tables 8 and 9 show the simulation coverage rate for a nominal 95% of the GREG estimator with Methods 1 and 2, respectively. Similarly, to the results found for RRMSE and RB, in terms of the coverage rate the proposed estimator with the free joint inclusion probability performed the best with the coverage rate close to 0.95 which outperformed other estimators.

Table 7 The simulation coverage rate for a nominal 95% of the ratio estimator

Response rate (%)	The estimator formula	The simulation coverage rate for a nominal 95%			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$\hat{V}'^{(1)}(\hat{Y}_R^*)$	0.9025	0.9068	0.9216	0.9378
	$\hat{V}'^{(2)}(\hat{Y}_R^*)$	0.8988	0.9047	0.9180	0.9383
	$\hat{V}'^{(3)}(\hat{Y}_R^*)$	0.8997	0.9044	0.9169	0.9387
	$\hat{V}'^{(4)}(\hat{Y}_R^*)$	0.8994	0.9021	0.9200	0.9355
	$\hat{V}'^{(5)}(\hat{Y}_R^*)$	0.9003	0.9033	0.9183	0.9383
	$\hat{V}'^{(6)}(\hat{Y}_R^*)$	0.8987	0.9052	0.9197	0.9336
	$\hat{V}'^{(7)}(\hat{Y}_R^*)$	0.9020	0.9022	0.9212	0.9378
	$\hat{V}''(\hat{Y}_R^*)$	**0.9120**	**0.9225**	**0.9400**	**0.9475**
0.85	$\hat{V}'^{(1)}(\hat{Y}_R^*)$	0.9104	9.2516	0.9308	0.9354
	$\hat{V}'^{(2)}(\hat{Y}_R^*)$	0.9090	9.2534	0.9314	0.9357
	$\hat{V}'^{(3)}(\hat{Y}_R^*)$	0.9102	9.2494	0.9328	0.9370
	$\hat{V}'^{(4)}(\hat{Y}_R^*)$	0.9127	9.2492	0.9311	0.9371
	$\hat{V}'^{(5)}(\hat{Y}_R^*)$	0.9112	9.2492	0.9280	0.9365
	$\hat{V}'^{(6)}(\hat{Y}_R^*)$	0.9112	9.2536	0.9331	0.9357
	$\hat{V}'^{(7)}(\hat{Y}_R^*)$	0.9131	9.2523	0.9280	0.9356
	$\hat{V}''(\hat{Y}_R^*)$	**0.9291**	**0.9298**	**0.9398**	**0.9410**

Table 8 The simulation coverage rate for a nominal 95% of the GREG estimator under Method 1

Response rate (%)	The estimator formula	The simulation coverage rate for a nominal 95%			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$\hat{V}_1^{(1)}(\hat{Y}_{GREG.LP})$	0.9109	9.2530	0.9340	0.9405
	$\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})$	0.9135	9.2546	0.9340	0.9409
	$\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})$	0.9121	9.2537	0.9336	0.9408
	$\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})$	0.9112	9.2523	0.9340	0.9402
	$\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})$	0.9124	9.2538	0.9337	0.9409
	$\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})$	0.9112	9.2522	0.9339	0.9404
	$\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})$	0.9128	9.2521	0.9330	0.9402
	$\hat{V}_1''(\hat{Y}_{GREG.LP})$	**0.9203**	**9.3538**	**0.9436**	**0.9470**
0.85	$\hat{V}_1^{(1)}(\hat{Y}_{GREG.LP})$	0.9220	0.9357	0.9389	0.9479
	$\hat{V}_1'^{(2)}(\hat{Y}_{GREG.LP})$	0.9211	0.9357	0.9389	0.9477
	$\hat{V}_1'^{(3)}(\hat{Y}_{GREG.LP})$	0.9212	0.9363	0.9385	0.9479
	$\hat{V}_1'^{(4)}(\hat{Y}_{GREG.LP})$	0.9212	0.9360	0.9385	0.9475
	$\hat{V}_1'^{(5)}(\hat{Y}_{GREG.LP})$	0.9220	0.9360	0.9389	0.9477
	$\hat{V}_1'^{(6)}(\hat{Y}_{GREG.LP})$	0.9209	0.9359	0.9390	0.9477
	$\hat{V}_1'^{(7)}(\hat{Y}_{GREG.LP})$	0.9232	0.9356	0.9386	0.9480
	$\hat{V}_1''(\hat{Y}_{GREG.LP})$	**0.9358**	**0.9390**	**0.9488**	**0.9495**

Table 9 The simulation coverage rate for a nominal 95% of the GREG estimator under Method 2

Response rate (%)	The estimator formula	The simulation coverage rate for a nominal 95%			
		$n=50$	$n=100$	$n=150$	$n=300$
0.7	$\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})$	0.9109	9.2530	0.9340	0.9405
	$\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})$	0.9135	9.2546	0.9340	0.9409
	$\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})$	0.9121	9.2537	0.9336	0.9408
	$\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})$	0.9112	9.2523	0.9340	0.9402
	$\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})$	0.9124	9.2538	0.9337	0.9409
	$\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})$	0.9112	9.2522	0.9339	0.9404
	$\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})$	0.9128	9.2521	0.9330	0.9402
	$\hat{V}_2''(\hat{Y}_{GREG.LP})$	**0.9203**	**9.3538**	**0.9436**	**0.9470**
0.85	$\hat{V}_2'^{(1)}(\hat{Y}_{GREG.LP})$	0.9220	0.9357	0.9389	0.9479
	$\hat{V}_2'^{(2)}(\hat{Y}_{GREG.LP})$	0.9211	0.9357	0.9389	0.9477
	$\hat{V}_2'^{(3)}(\hat{Y}_{GREG.LP})$	0.9212	0.9363	0.9385	0.9479
	$\hat{V}_2'^{(4)}(\hat{Y}_{GREG.LP})$	0.9212	0.9360	0.9385	0.9475
	$\hat{V}_2'^{(5)}(\hat{Y}_{GREG.LP})$	0.9220	0.9360	0.9389	0.9477
	$\hat{V}_2'^{(6)}(\hat{Y}_{GREG.LP})$	0.9209	0.9359	0.9390	0.9477
	$\hat{V}_2'^{(7)}(\hat{Y}_{GREG.LP})$	0.9232	0.9356	0.9386	0.9480
	$\hat{V}_2''(\hat{Y}_{GREG.LP})$	**0.9358**	**0.9390**	**0.9488**	**0.9495**

8. Application to Real Data

To apply the proposed estimators to the fine particulate matter data in the north of Thailand where the PM2.5 dust level is one of the highest in the world. The data are from the air quality and noise management bureau, the Pollution Control Department of Thailand during October and November 2022. The Midzuno (1952) scheme is applied to select a sample of size 13 stations out of 23 stations (http://air4thai.pcd.go.th/webV2/history). The PM2.5 on 28 Novemeber 2002 is used as a study variable y (micrograms per cubic meter) and the average PM2.5 and the air quality index average on October 2002 are considered as the auxiliary variables x and w, respectively. The correlation coefficients between y and w and y and x are equal to 0.91 and 0.66 respectively. The ratio estimator is created using w while the GREG estimator is created using x and only the case where the response probability is unknown is considered. The maximum value of PM2.5 on October 2002 is considered as the size variable k. The nonresponse rate is 8.7% in this study. The results are displayed in Table 10.

Table 10 The estimated total and variance estimates for the total yield

Estimator	Population total estimates	The formula of variances	Variance estimates	95% confidence interval
The ratio estimator ($\hat{Y}_R^*$)	585.95	$\hat{V}'^{(1)}(\hat{Y}_R^*)$	18741.47	(317.63, 854.27)
		$\hat{V}'^{(2)}(\hat{Y}_R^*)$	19166.98	(314.60, 857.30)
		$\hat{V}'^{(3)}(\hat{Y}_R^*)$	18409.36	(320.02, 851.88)
		$\hat{V}'^{(4)}(\hat{Y}_R^*)$	18331.30	(320.58, 851.32)
		$\hat{V}'^{(5)}(\hat{Y}_R^*)$	18127.38	(322.06, 849.84)
		$\hat{V}'^{(6)}(\hat{Y}_R^*)$	18558.42	(318.94, 852.96)
		$\hat{V}'^{(7)}(\hat{Y}_R^*)$	18578.39	(318.80, 853.10)
		$\hat{V}''(\hat{Y}_R^*)$	12726.33	(364.84, 807.06)

Estimator	Population total estimates	The formula of variances	Method 1		Method 2	
			Variance estimates	95% confidence interval	Variance estimates	95% confidence interval
The GREG estimator ($\hat{Y}_{GREG.LP}$)	770.75	$\hat{V}_m'^{(1)}(\hat{Y}_{GREG.LP})$	1210.57	(702.56, 838.94)	1210.57	(702.56, 838.94)
		$\hat{V}_m'^{(2)}(\hat{Y}_{GREG.LP})$	1210.70	(702.55, 838.95)	1210.70	(702.55, 838.95)
		$\hat{V}_m'^{(3)}(\hat{Y}_{GREG.LP})$	1210.43	(702.56, 838.94)	1210.43	(702.56, 838.94)
		$\hat{V}_m'^{(4)}(\hat{Y}_{GREG.LP})$	1236.20	(701.84, 839.66)	1236.20	(701.84, 839.66)
		$\hat{V}_m'^{(5)}(\hat{Y}_{GREG.LP})$	1242.45	(701.66, 839.84)	1242.45	(701.66, 839.84)
		$\hat{V}_m'^{(6)}(\hat{Y}_{GREG.LP})$	1229.24	(702.03, 839.47)	1229.24	(702.03, 839.47)
		$\hat{V}_m'^{(7)}(\hat{Y}_{GREG.LP})$	1228.62	(702.05, 839.45)	1228.62	(702.05, 839.45)
		$\hat{V}_m''(\hat{Y}_{GREG.LP})$	895.81	(712.09, 829.41)	895.81	(712.09, 829.41)

From Table 10, we can see similar results to the simulation studies where the proposed variance estimators using the free joint inclusion probability performed the best with the fine particulate matter data in all situations. The proposed free joint inclusion probability variance estimators gave a narrower confidence interval compared to others which result in better precision in estimating the population total. The estimated variance for the GREG estimators on both Methods 1 and 2 showed the same results due to the small value of $\hat{\bar{e}}_r$ which makes similar results for the variance estimators from Methods 1 and 2 for the estimated joint inclusion probability variance estimators and free joint inclusion probability variance estimators from (40) and (42) and (46) and (47) respectively in this situation. The estimated total PM2.5 from the ratio estimator is 585.95 micrograms per cubic meter which is less than the estimated total PM2.5 from the GREG estimator that is equal to 770.75 micrograms per cubic meter but has a smaller variance.

9. Conclusions

The ratio and GREG estimators are potent for estimating population total when information on the population of an auxiliary variable is known and is highly correlated with the study variable. Many works estimate population total when data on the population of an auxiliary variable exists, but the study variable includes nonresponse. However, the variance estimator under unequal probability sampling without replacement is difficult to compute because it requires joint inclusion probability.

Therefore, we proposed new variance estimators of the ratio and GREG estimators proposed by Ponkaew and Lawson (2019) and Lawson and Ponkaew (2019), respectively. The proposed variance estimators were investigated under two different methods consisting of free joint inclusion probability and estimated joint inclusion probability. In the simulation studies the variance estimators with free joint inclusion probability performed better than the variance estimators with the estimated joint inclusion probability. The application to the fine particulate matter data in the north of Thailand presented results alike to the simulation results. The most superior method to estimate variance allowing practical convenience is using free joint inclusion probability, as it does not require the value of joint inclusion probability. It gave a narrower confidence interval compared to others leading to higher precision of the population total. In future works, other techniques to estimate the joint inclusion probability can be used to see the performance of the proposed estimators.

Acknowledgements

This research was funded by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok Contract no. KMUTNB-FF-66-56. We would like to thank the unknown referees for the comments.

References

Bacanli S, Kadilar C. Ratio estimators with unequal probability designs. Pakistan J Stat. 2008; 24(3): 167-172.

Berger YG. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. J. Stat. Plan. 1998; 67 (2): 209-226.

Berger YG. A simple variance estimator for unequal probability sampling without replacement. J Appl Stat. 2003; 31: 305-315.

Berger YG, Tirari MEH, Tille Y. Towards optimal regression estimation in sample surveys. Aust N Z J Stat. 2003; 45(3): 319-329. https://doi:10.1111/1467-842X.00286.

Brewer KRW. Combined survey sampling inference: weighing Basu's elephants, London: Arnold; 2002.

Brewer KRW, Donadio ME. The high entropy variance of the Horvitz-Thompson estimator. Surv Methodol. 2003; 29 (2): 189-196.

Estevao VM, and Särndal CE. Survey estimates by calibration on complex auxiliary information. Int Stat Rev. 2006; 74 (2): 127-147.

Hajek J. Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann Math Stat. 1964. 35 (4):1491-1523.

Hajek J. Sampling from finite population. New York: Marcel Dekker; 1981.

Hansen MH, Hurwitz WN. On the theory of sampling from finite populations. J Am Stat Assoc. 1943;14(4): 333-362.

Hartley HO, Rao JNK. Sampling with unequal probability and without replacement. J Am Stat Assoc. 1962; 33(2): 350-374.

Horvitz DF, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952; 47(260): 663-685.

IQAir World Air Quality Report. Air quality and pollution city ranking. 2021 [cited 2023 Sep 6]. Available from: https://www.iqair.com/th-en/world-air-quality-ranking.

Lawson N. Variance estimation in the presence of nonresponse with free joint Inclusion probability under unequal probability sampling without replacement. Proceedings of the 3rd International Conference on Big Data and Education; 2020 Apr 1; pp. 50-54.

Lawson N, Ponkaew C. New generalized regression estimator in the presence of nonresponse under unequal probability sampling. Commun Stat Theory Methods. 2019; 48(10): 2483-2498.

Midzuno H. On the sampling system with probability proportional to sum of sizes. Ann Inst Stat Math. 1952; 553: 99-107.

Montanari GE. Post sampling efficient qr-prediction in large sample survey. Int Stat Rev. 1987; 55(2): 191-202.

Nangsue N, Berger YG. Optimal regression estimator for stratified two-stage sampling. In: Fulvia M, Luigi CP, Giovanna RM, editors. Contributions to Sampling Statistics. Cham: Springer; 2014; p. 167-177.

Pollution Control Department. Thailand's air quality and situation reports. Bangkok, Thailand. 2022. [cited 2023 Sep 6]. Available from: http://air4thai.pcd.go.th/webV2/history/.

Ponkaew C, Lawson N. A new ratio estimator for population total in the presence of nonresponse under unequal probability sampling without replacement. Thai J. Math.: Special Issue (ACFPTO2018) on: Advances in Fixed Point Theory towards Real World Optimization Problem. 2018; 417-429.

Ponkaew C, Lawson N. Estimating variance in the presence of nonresponse under unequal probability sampling. Suranaree J Sci Technol. 2019; 26 (3): 293-302.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021 [cited 2023 Sep 6]. Available from: https://www.R-project.org.

Särndal CE, Lundström S. Estimation in surveys with nonresponse. New York: John Wiley & Sons; 2005.

Sen AR. On the estimate of the variance in sampling with varying probabilities, J Indian Soc Agri Stat. 1953; 5: 119-127.

Sichera R. Approximate inclusion probabilities for survey sampling. R package version 0.1.2. 2020. 2020 [cited 2023 Sep 6]. Available from: URL https://CRAN.project.org/package=jipApprox.

Yates F, Grundy PM. Selection without replacement from within strata with probability proportional to size. J R Stat Soc Series B. 1953; 15: 235-261.