Lawson, Nuanpan

Article

Review of how the generalized regression estimators contribute to estimating the financial and economic data with missing observations under unequal probability sampling

Asian Journal of Economics and Banking (AJEB)

Provided in Cooperation with:
Ho Chi Minh University of Banking (HUB), Ho Chi Minh City

Suggested Citation: Lawson, Nuanpan (2024) : Review of how the generalized regression estimators contribute to estimating the financial and economic data with missing observations under unequal probability sampling, Asian Journal of Economics and Banking (AJEB), ISSN 2633-7991, Emerald, Leeds, Vol. 8, Iss. 3, pp. 445-459, https://doi.org/10.1108/AJEB-12-2023-0134

Asian Journal of Economics and Banking

445

Review of how the generalized regression estimators contribute to estimating the financial and economic data with missing observations under unequal probability sampling

Received 25 December 2023
Revised 7 February 2024
11 April 2024
14 April 2024
Accepted 19 April 2024

Nuanpan Lawson
Department of Applied Statistics,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Abstract

Purpose – Knowing financial and economic information beforehand benefits in planning and developing policies for every country especially for a developing country like Thailand and for other Asian countries. Unfortunately, missing data or non-response plays an essential role in many areas of studies including finance and economics. Eradication of missing data in a proper way before further analysis can gain remarkable outcomes and can be effective for planning policies. This review on the generalized regression estimators for population total can be applied to financial, economic and other data when missing data are present.

Design/methodology/approach – The generalized regression estimators for estimating population total, including the variance estimators under unequal probability sampling without replacement with missing data are explored under the reverse framework. Applications to financial and economic data in Thailand are also reviewed.

Findings – The review of literatures related to the proposed estimator shows the best performance, giving smaller variances in all scenarios.

Originality/value – The generalized regression estimators can assist in estimating financial and economic data that contain missing values with different missing mechanisms and can be used in other applications which help gain more superior estimators.

Keywords Generalized regression estimators, Missing data, Financial data, Economic data,
Unequal probability sampling without replacement

Paper type General review

1. Introduction

Generalized regression (GREG) estimation is optimized for design-based estimations of population totals for survey sampling, which are often used in financial data which are seldom complete, becoming an inherent issue requiring a solution. An opulence of economic advancement is imperative in every country to maintain the country's infrastructure and

JEL Classification — C83

Many thanks to Prof. Sa-Aat Niwitpong and Prof. Hung Nguyen for recommending the *Asian Journal of Economics and Banking*.

Asian Journal of Economics and Banking
Vol. 8 No. 3, 2024
pp. 445-459
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-12-2023-0134

quality of life of citizens which calls for statistical analysis of data, where the problems of missing data and suitable estimators arise. Measures have been placed on a plethora of aspects to ensure economic development in Thailand, as seen in sustainable development plans in "Thailand 4.0", as Thailand is a country highly dependent on revenue from tourism. With this reason, the economy is liable to fluctuations, especially recently due to the coronavirus pandemic. After withdrawal of revenue from foreign tourists, the economy became more focused on citizens' assets, income, and cash flow within the country. A myriad of policies have been enforced to provide stability to individuals' financial stability and capability to manage their assets during a pandemic. Analysis of the population's financial issues is vital for proper repairment of the crisis and instigation of solutions and endorsement for citizens in need throughout the duration of the pandemic. Data on the population's expenses is required for insight on the financial obstacles being faced and to further analyze then address the concerns suitably.

Furthermore, the government has induced many means to stimulate tourism within the country such as the "Thai Travel Together" campaign which allows cash flow within the country and mitigates hardships inflicted upon the economy as a result of the crisis from COVID-19. Moreover, additional facets impact the economy, including unsubstantial investment that afflicts the economy on a large-scale. Sustainable development plans have been enforced to target ten industries and try to resolve production efficiency and competitiveness afflicting Thailand's industrial economic structure.

However, missing data or nonresponse often occurs in real world data which can obscure facts used for decision making in business and economics, so opportunities are lost due to incomplete data. Missing data occurs due to nonresponse or participants choosing not to answer specific questions for instance. Missing data can occur when it does not depend on missing values or observed values, called missing completely at random (MCAR) or uniform nonresponse, or the missingness correlates to the observations but is not related to the missing values and this is called missing at random (MAR). Therefore, resolving nonresponse is imperative for appropriate financial planning. Difficulties in acquiring accurate data can be a result of lack of records or nonresponse derived from surveys. In conclusion, statistical methods that tackle nonresponse are vital measures to solving this problem. The nonresponse issue was first recommended by Hansen and Hurwitz (1946) in the mail survey. They introduced an unbiased estimator for population mean that used data from a sample survey on both respondents and non-respondents under unequal probability sampling without replacement (UPWOR). Horvitz and Thompson (1952) suggested using the weight to create an unbiased population total estimator under unequal probability sampling for with and without replacement. The first order of inclusion probability is used as the weight for correction of the bias. Unfortunately, there is an issue in calculating variance in Horvitz and Thompson due to it requiring joint inclusion probabilities which are hard to find in some complex survey designs. Later, Hajek (1964) proposed a new estimator to correct an issue of the variance estimator which produces less variance with respect to Horvitz and Thompson (1952), but only when there is no relationship between the study variable and the inclusion probabilities. Their new estimator is a ratio estimator, which is the ratio of sample means of two random variables. for estimating population total which is an approximately unbiased ratio estimator.

The GREG estimator is a special type of calibration estimator and improves this method of estimation using auxiliary information. It is in the shape of the Horvitz and Thompson (1952) estimator which integrates with the weighting approach as it can assist in reducing the nonresponse bias. Bethlehem and Keller (1987) introduced to use weights using linear models which is a new weighting method that can be used in person-based estimations. Many works have been done based on GREG to use the benefit of the relationship between the study and auxiliary variables to skyrocket the efficiency of the population total or population mean

estimators and also the variance estimators (see, e.g. Montanari, 1987; Särndal *et al.*, 1992; Estevao and Särndal, 2003; Särndal and Lundström, 2005; Särndal, 2007). The two-phase framework concerns studying the selected sample and nonresponse in the first and second phases, respectively, under nonresponse. It is a popular technique to use to study the GREG estimators' variance (see, e.g. Rao, 1990; Särndal, 1992; Deville and Särndal, 1994; Särndal and Lundström, 2005).

Fay (1991) invented an alternative to the two-phase measure, the reverse framework. The name comes from the order of studies being reversed, nonresponse is a candidate in the first phase and the sampling shown in the second phase (see, e.g. Shao and Steel, 1999; Haziza and Rao, 2006; Haziza, 2010). Under this reverse method, the population total estimators and the GREG estimators along with their variance estimators were investigated within the MCAR and MAR nonresponse mechanisms and under different assumptions for the response probabilities and the sampling fractions (Lawson, 2017; Lawson and Ponkaew, 2019; Lawson and Siripanich, 2022; Ponkaew and Lawson, 2023).

In this paper, the GREG estimators under the reverse framework will be reviewed. The structure of this paper is as follows. The literature review is shown in section 2. The basic setup and the generalized regression estimators with missing data are reviewed in sections 3 and 4, respectively. Examples of the application related to financial and economic data in Bangkok, Thailand are displayed in section 5. Lastly, some conclusions and discussions are presented in section 6.

2. Literature review

First of all, let's see how the generalized regression estimators have been developed and can be useful for estimating financial, economic, and other data. The generalized regression estimator can estimate the population mean or total. It is in the shape of Horvitz and Thompson's (1952), a very well-known population total estimator under unequal probability sampling for both including and not including replacement. Nevertheless, the Horvitz and Thompson's variance estimator is facing issues as it calls for the known joint inclusion probabilities, also known as the second order inclusion probabilities. They are the probabilities of two different units of populations selected in the sample. These values are difficult to find in complex survey designs and therefore the Horvitz and Thompson estimator is not easy to use in practice. Sometimes they are difficult to be calculated. Under unequal probability sampling using replacement, the formulas of the variance estimators are in their simple forms because these probability values, which is different from the variance formula under UPWOR which requires joint inclusion probabilities.

Some researchers also made an effort to solve this issue in the estimation of variance (Sen, 1953; Yates and Grundy, 1953) but still face the same issue requiring joint inclusion probability which is not known or hard to find. Therefore, some methods have been suggested in estimating the joint inclusion probability (Hartley and Rao, 1962; Hajek, 1964, 1981; Brewer, 2002; Brewer and Donadio, 2003).

The GREG estimators assist in finding population mean and total when there is information based on the related auxiliary variable to the study variable. The formula of the GREG estimator is in the structure of the Horvitz and Thompson (1952) estimator with additional adjustments calculated from an auxiliary variable. Optimal GREG estimators were developed using the known value of the regression coefficient in the population (Montanari, 1987; Berger *et al.*, 2003) under different sampling plans such as stratified two-stage cluster sampling. The Taylor linearization method is used to study the variance and associated variance of the GREG estimator which is in a nonlinear form and therefore it needs to be transformed to a linear one. A drawback of the GREG variance estimator under this situation is that it requires complex methods in calculating the variance under UPWOR due to the

requirement of the known joint inclusion probabilities as same as Horvitz and Thompson's (1952) method. With nonresponse, Särndal and Lundström (2005) have introduced an almost unbiased GREG estimator for estimating population total and a variance estimator under the two-phase framework which requires nonresponse propensities. Under the reverse framework, some literatures explored GREG estimators including missing data. A GREG estimator based on the population total estimator when unit nonresponse appears within the study variable with a negligible sampling fraction under an unstratified, one-stage sample, with probability being unequal has been suggested when the nonresponse mechanism is MCAR. This is quite a restrictive assumption where the response probability is constant and tend to not occur in practice and also the estimator is in a nonlinear form (Lawson and Ponkaew, 2019). However, they proposed to use the modified automated linearization method to deal with this problem and showed that their estimator is unbiased and response probability is not essential. Recently in 2023, under the same assumptions of the previous work, the ratio method of estimation is applied to create the new GREG estimators (Ponkaew and Lawson, 2023). Their estimators are more efficient than the previous work in terms of giving smaller relative bias and root mean square errors as the criterions. We can also see from the application results that were applied to the Thai maize agricultural industry in Thailand in 2019 based on the data from the Office of the Agricultural Economics that their estimators provide a smaller variance in estimating the estimate values of total yield of maize in Thailand which could help in planning for policies for the economics part of Thailand's agriculture in the future.

Under a more flexible nonresponse mechanism such as MAR to allow for more practicality to use in realistic situations, an approximately unbiased GREG estimator and its variance under UPWOR has been suggested in less controlled circumstances, with the response probabilities both known and unknown and the nonresponse mechanism is non-uniform, with both a small sampling fraction or any sampling fraction. This type of nonresponse mechanism can be called MAR or the ignorable nonresponse mechanism. The less restrictive situations in this estimator can assist by acquiring vital data imperative for financial and economic projects in many areas where missingness happens in the study variable. For example, to study farm profitability and resilience, which brings in revenue for the country can be investigated using the GREG estimators by estimating liabilities and net worth using some variables for instance farm type, farm size, region, tenure, and economic performance. Nevertheless, economic data, e.g. the agricultural industry such as total yield, total profit, and total income can be applied using the GREG estimator to find out these values in advance for planning for effective decision making which can develop economic wealth for the whole nation. Handling missingness appropriately can benefit the reliability of the data that is utilized for planning in Thailand and other countries around the world (Lawson and Siripanich, 2022).

3. Basic setup

The notations and the basic notions under the reverse framework will be introduced. Let y be a study variable and a population total of the y variable is $Y = \sum_{i \in U} y_i$ where $U = \{1, 2, ..., N\}$ and N is a population size. Let x be an auxiliary variable and the population total of the x variable is $X = \sum_{i \in U} x_i$. The order of the paired ith values of the study variable y and auxiliary variable x is (y_i, x_i), $i = 1, 2, ..., N$. For the ratio estimator, the variable x is an auxiliary variable. The auxiliary variables k and w are used to define the first and joint inclusion probabilities under UPWOR and utilized to construct the ratio estimator respectively. A sample s of size n is drawn using UPWOR. For selecting the population unit i in U, the known and nonzero probability is represented by $P_i = X_i/X$ where $\sum_{i=1}^{N} P_i = 1$. Let, $\pi_i = P(i \in s) = \sum_{i \in s} P(s)$ be the first order

inclusion probability and $\pi_{ij} = P(i \wedge j \in s) = \sum_{\{i,j\} \subset s} P(s)$ be the second order inclusion probability. Assume that the information of $n \times (q+1)$ matrix of values x or $\boldsymbol{X}_n = (\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n)'$ is known for all $\boldsymbol{x}_i$ when $i \in s$. The expectation and variance according to UPWOR sampling are defined as E_S and V_S respectively.

The population total GREG estimator is

$$\widehat{Y}_{GREG} = \sum_{i \in s} \frac{y_i}{\pi_i} + \left[\sum_{i \in U} \boldsymbol{x}_i - \sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i}\right]' \left(\sum_{i \in s} \frac{q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1} \left(\sum_{i \in s} \frac{q_i \boldsymbol{x}_i y_i}{\pi_i}\right) = \widehat{Y}_{HT} + \left[\boldsymbol{X} - \widehat{\boldsymbol{X}}_{HT}\right]' \widehat{\boldsymbol{\beta}}_r$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im})', i = 1, 2, \ldots, n$, are the column vectors of the auxiliary variable with $m \geq 1$, $\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$, $\widehat{\boldsymbol{X}}_{HT} = \sum_{i \in s} \frac{\boldsymbol{x}_i}{\pi_i}$, $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$, $\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i \in s} \frac{q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1} \left(\sum_{i \in s} \frac{q_i \boldsymbol{x}_i y_i}{\pi_i}\right)$ and q_i are calculated by the linear assisting model ξ: $E_\xi(y_i) = \boldsymbol{\beta}' \boldsymbol{x}_i$ and $V_\xi(y_i) = \sigma_i^2$ that is $q_i = 1/\sigma_i^2$.

Under nonresponse, R and r_i denote the response mechanism and the y_i response indicator variable, respectively.

$$r_i = \begin{cases} 1, \text{if } y_i \text{ is observed} \\ 0, \text{if } y_i \text{ is missing.} \end{cases}$$

Let p_i be the response probability shown as $p_i = P(r_i = 1)$. Let E_R and V_R be the expectation and variance operators according to the response mechanism, and E and V be the overall expectation and variance operators, respectively. Therefore, $E_R(r_i) = P(r_i = 1) = p$ and $V_R(r_i) = p(1-p)$.

The GREG estimator $\widehat{Y}_{GREG}$ variance from the reverse framework is

$$V\left(\widehat{Y}_{GREG}\right) = E_R V_S\left(\widehat{Y}_{GREG} \middle| \boldsymbol{R}\right) + V_R E_S\left(\widehat{Y}_{GREG} \middle| \boldsymbol{R}\right),$$

4. Generalized regression estimators with missing data

Numerous works have investigated the GREG estimators with missing data under the two-phase framework to study the GREG estimators' variance where in the first phase only the interested sample is examined and in the second phase only the nonresponse is contemplated. Under the two-phase framework, the GREG estimator and variance were studied in the presence of nonresponse (Särndal and Lundström, 2005). They also recommended an automated linearization method in finding the variance of the GREG estimator where the partial derivatives are not obligatory as in the Taylor series linearization (see, e.g. Estevao and Särndal, 2003; Särndal and Lundström, 2005; Särndal, 2007).

A GREG estimator for population total with nonresponse using the two-phase framework is (Särndal and Lundström, 2005)

$$\begin{aligned} \widehat{Y}_{GREG.SL} &= \sum_{i \in s} \frac{r_i y_i}{\pi_i p_i} + \left[\boldsymbol{X} - \sum_{i \in s} \frac{r_i \boldsymbol{x}_i}{\pi_i p_i}\right]' \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1} \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right) \\ &= \widehat{Y}_r + \left[\boldsymbol{X} - \widehat{\boldsymbol{X}}_r\right]' \widehat{\boldsymbol{\beta}}_r, \end{aligned}$$

where $\widehat{Y}_r = \sum_{i \in s} \frac{r_i y_i}{\pi_i p_i}$, $\widehat{\boldsymbol{X}}_r = \sum_{i \in s} \frac{r_i \boldsymbol{x}_i}{\pi_i p_i}$, $\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1} \left(\sum_{i \in s} \frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right)$,

The variance of $\widehat{Y}_{GREG.SL}$ is

$$V\left(\widehat{Y}_{GREG.SL}\right) = \sum_{i\in U} D_i e_i^2 + \sum_{i\in U}\sum_{j\setminus\{i\}\in U} D_{ij} e_i + \sum_{i\in U} \frac{(1-p_i)}{p_i} e_i^2,$$

where $e_i = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \boldsymbol{\beta} = \left(\sum_{i\in U} q_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}\left(\sum_{i\in U} q_i \boldsymbol{x}_i y_i\right), D_i = \frac{1-\pi_i}{\pi_i}, D_{ij} = \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}.$

When p_i is known for all $i \in s$ under the reverse framework, $V(\widehat{Y}_{GREG.SL})$ is

$$\widehat{V}\left(\widehat{Y}_{GREG.SL}\right) = \sum_{i\in s} \widehat{D}_i \frac{r_i \widehat{e}_i^2}{p_i} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s} \widehat{D}_{ij} \frac{r_i \widehat{e}_i}{p_i} \frac{r_j \widehat{e}_j}{p_j} + \sum_{i\in s} \frac{(1-p_i)}{\pi_i p_i^2} \widehat{e}_i^2,$$

where $\widehat{e}_i = (y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r), \widehat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s} \frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1}\left(\sum_{i\in s} \frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right), \widehat{D}_i = \frac{1-\pi_i}{\pi_i^2}, \widehat{D}_{ij} = \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}.$

When p_i is unknown for all $i \in s$, let $\widehat{p}_i$ be the estimator of p_i, then the estimator of $V(\widehat{Y}_{GREG.SL})$ is

$$\widehat{V}\left(\widehat{Y}_{GREG.SL}\right) = \sum_{i\in s} \widehat{D}_i \frac{r_i \widehat{e}_i^2}{\widehat{p}_i} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s} \widehat{D}_{ij} \frac{r_i \widehat{e}_i}{\widehat{p}_i} \frac{r_j \widehat{e}_j}{\widehat{p}_j} + \sum_{i\in s} \frac{\left(1-\widehat{p}_i\right)}{\pi_i \widehat{p}_i^2} \widehat{e}_i^2,$$

where $\widehat{\boldsymbol{e}}_{\boldsymbol{i}} = (y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r), \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} = \left(\sum_{i\in s} \frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i \widehat{p}_i}\right)^{-1}\left(\sum_{i\in s} \frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i \widehat{p}_i}\right)$

Apart from the two-phase framework, the reverse framework by Fay (1991) is also studied to investigate the GREG estimators variance with the order of the selected sample and nonresponse reversed in the phases of sampling. Again, the same issue arises in the variance estimator which is in a nonlinear form and as a result it needs to be transformed to a linear function. Under the reverse framework, a new GREG estimator has been suggested MCAR or the uniform nonresponse mechanism where the response probability is constant. Most researchers (Lawson and Ponkaew, 2019; Ponkaew and Lawson, 2023) considered it under this assumption due to simplicity. A new GREG estimator for nonresponse under UPWOR was developed based on Lawson's (2017) concept, a nonlinear estimator for population total/mean and is an almost unbiased estimator with probability being proportional to size sampling consisting of replacement. The benefit of the Lawson estimator is that the response probability is not required in the estimation but is under the assumption that the probabilities of response are the same for all units and the sampling fraction can be omitted. Lawson's (2017) population mean estimator is

$$\widehat{\overline{Y}}_r = \frac{\sum_{i\in s} \frac{r_i y_i}{\pi_i \widehat{p}_i}}{\sum_{i\in s} \frac{r_i}{\pi_i \widehat{p}_i}}.$$

When $p_i = p$ for all units i in U, then

$$\widehat{\overline{Y}}_r = \frac{\sum_{i\in s} \frac{r_i y_i}{\pi_i p}}{\sum_{i\in s} \frac{r_i}{\pi_i p}} = \frac{\sum_{i\in s} \frac{r_i y_i}{\pi_i}}{\sum_{i\in s} \frac{r_i}{\pi_i}}.$$

Additionally, the Lawson (2017) estimator for estimating the population total is

$$\widehat{Y}_r = N\widehat{\overline{Y}}_r = N\frac{\sum_{i\in s}\frac{r_i y_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}}.$$

The associated variance estimator for $\widehat{Y}_r$ is

$$V\left(\widehat{\overline{Y}}_r\right) = \frac{1}{N^2 p^2}\left[n\sum_{i\in U}\frac{1}{\pi_i^2}p\left(y_i-\overline{Y}\right)^2 P_i - \frac{1}{n}\left(\sum_{i\in U}\frac{p}{\pi_i}\left(y_i-\overline{Y}\right)P_i\right)^2\right].$$

The estimated variance of $V(\widehat{\overline{Y}}_r)$ is

$$\widehat{V}\left(\widehat{\overline{Y}}_r\right) = \frac{1}{\left(\sum_{i\in s}\frac{r_i}{\pi_i}\right)^2}\frac{n}{n-1}\sum_{i\in U}\frac{r_i}{\pi_i^2}\left(y_i-\widehat{\overline{Y}}_r\right)^2.$$

The associated variance estimator for the $\widehat{Y}_r$ is

$$V\left(\widehat{Y}_r\right) = \frac{1}{p^2}\left[n\sum_{i\in U}\frac{1}{\pi_i^2}p\left(y_i-\overline{Y}\right)^2 P_i - \frac{1}{n}\left(\sum_{i\in U}\frac{p}{\pi_i}\left(y_i-\overline{Y}\right)P_i\right)^2\right], \tag{3.6}$$

and the estimated variance of $V(\widehat{Y}_r)$ is

$$\widehat{V}\left(\widehat{Y}_r\right) = \frac{N^2}{\left(\sum_{i\in s}\frac{r_i}{\pi_i}\right)^2}\frac{n}{n-1}\sum_{i\in U}\frac{r_i}{\pi_i^2}\left(y_i-\widehat{\overline{Y}}_r\right)^2.$$

Under the same assumptions where the nonresponse mechanism is MCAR, the sampling fraction is can be omitted under UPWOR, based on the Lawson (2017) estimator, a new GREG estimator has been suggested as follows (Lawson and Ponkaew, 2019).

$$\widehat{\overline{Y}}_{GREG.LP} = \widehat{\overline{Y}}_r + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r.$$

where $\widehat{\overline{Y}}_r = \frac{\sum_{i\in s}\frac{r_i y_i}{\pi_i p}}{\sum_{i\in s}\frac{r_i}{\pi_i p}} = \frac{\sum_{i\in s}\frac{r_i y_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}}$, $\widehat{\overline{\boldsymbol{X}}}_r = \frac{\sum_{i\in s}\frac{r_i \boldsymbol{x}_i}{\pi_i}}{\sum_{i\in s}\frac{r_i}{\pi_i}}$, $\overline{\boldsymbol{X}} = \sum_{i\in U}\boldsymbol{x}_i/N$,

$$\widehat{\beta}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p}\right) = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i}\right) \text{ and } q_i = 1\Big/\sigma_i^2. \tag{3.9}$$

When the population size N is known, the population total GREG estimator is

$$\widehat{Y}_{GREG.LP} = N\widehat{\overline{Y}}_{GREG.LP} = N\left[\widehat{\overline{Y}}_r + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r\right].$$

They also assumed that $\widehat{\boldsymbol{\beta}}_r - \boldsymbol{\beta} = O_p\left(n_r^{-\frac{1}{2}}\right)$ and $r_n \to 0$ as $n \to \infty$, where $\{r_n\}$ is a sequence consisting of positive real numbers. For the GREG estimators' variance, they considered two situations; replace $\sum_{i\in s}\frac{r_i}{\pi_i}$ by $\sum_{i\in U} r_i$, then $V_1(\widehat{Y}_{GREG.LP}) \approx \frac{1}{p}\sum_{i\in U}\frac{(1-\pi_i)}{\pi_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \sum_{i\in U}\sum_{j\neq i\in U} D_{ij}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})(y_j - \boldsymbol{x}_j'\boldsymbol{\beta})$ and using the Taylor linearization approach, then $V_2(\widehat{Y}_{GREG.LP}) \approx \frac{1}{p}\sum_{i\in U}\frac{(1-\pi_i)}{\pi_i}(e_i - \bar{e})^2 + \sum_{i\in U}\sum_{j\neq i\in U} D_{ij}(e_i - \bar{e})(e_j - \bar{e})$. The estimated variances of these estimators are respectively,

$$\widehat{V}_1\left(\widehat{Y}_{GREG.LP}\right) \approx \left(\frac{N}{\sum_{i\in s}\frac{r_i}{\pi_i}}\right)^2 \left[\sum_{i\in s}\frac{(1-\pi_i)}{\pi_i^2} r_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r\right)^2 + \sum_{i\in s}\sum_{j\neq i\in s}\breve{D}_{ij} r_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r\right) r_j\left(y_j - \boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}_r\right)\right].$$

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

where $\widehat{e}_i = (y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r)$.

They also studied in theory that $\widehat{V}_1(\widehat{Y}_{GREG.LP})$ and $\widehat{V}_2(\widehat{Y}_{GREG.LP})$ are almost unbiased estimators.

Later, a new GREG estimator derived from the ratio method has been proposed based on the work of Lawson and Ponkaew (2019) using the same assumptions where the nonresponse mechanism is MCAR and they stretched it to cover the situation where the sampling fraction is also large and therefore it cannot be neglected. They also developed to cases where the response probabilities are known and unknown assisting with the benefit of the known auxiliary variable with nonresponse. Usually under the reverse framework the second part of the variance component is omitted but they considered the case that the variance component in this part cannot be ignored (Ponkaew and Lawson, 2023). Therefore, $V_2 = V_R E_S(\widehat{Y}_{GREG.LP}\big|\boldsymbol{R})$. Again, they considered the automated linearization approach in the transformation of the $\widehat{Y}_{GREG.LP}$ into a less complex form. They assumed three assumptions in their study; the response mechanism is MCAR, $\widehat{\boldsymbol{\beta}}_r - \boldsymbol{\beta} = O_p\left(n_r^{-\frac{1}{2}}\right)$, and $V\left(\sum_{i\in s}\frac{b_i}{\pi_i}\right) \to 0$ as $n \to \infty$ where $b_i = w_i$ or r_i.

Their GREG estimators for population mean and total are respectively,

$$\widehat{\overline{Y}}^*_{GREG.R} = \widehat{\overline{Y}}^*_R + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r.$$

$$\widehat{Y}^*_{GREG.R} = N\widehat{\overline{Y}}^*_{GREG.R} = N\left[\widehat{\overline{Y}}^*_R + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r\right].$$

where $\widehat{\overline{Y}}_R^* = \frac{\widehat{\overline{Y}}_r^{(1)}}{\widehat{\overline{w}}_{HT}}\overline{W}$, $\widehat{\overline{\boldsymbol{X}}}_r = \sum_{i\in s}\frac{r_i\boldsymbol{x}_i}{\pi_i}/\sum_{i\in s}\frac{r_i}{\pi_i}$, $\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i}{\pi_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i}\right)$,

$$\overline{\boldsymbol{X}} = \frac{1}{N}\sum_{i\in U}\boldsymbol{x}_i.$$

Under the reverse framework the $V(\widehat{\overline{Y}}_{GREG.R}^*)$ can be gained by,

$$V\left(\widehat{\overline{Y}}_{GREG.R}^*\right) = E_R V_S\left(\widehat{\overline{Y}}_{GREG.R}^*\middle|\boldsymbol{R}\right) + V_R E_S\left(\widehat{\overline{Y}}_{GREG.R}^*\middle|\boldsymbol{R}\right) = V_1 + V_2.$$

where $V_1 = E_R V_S(\widehat{\overline{Y}}_{GREG.R}^*|\boldsymbol{R})$, $V_2 = V_R E_S(\widehat{\overline{Y}}_{GREG.R}^*|\boldsymbol{R})$.

The variance of Ponkaew and Lawson (2023) are

(1) $V_1(\widehat{Y}_{GREG.R}^*)$ is

$$V_1\left(\widehat{Y}_{GREG.R}^*\right) \approx \sum_{i\in U}\left(D_i(Npe_i)^2 + (1-p)p^{-1}\left(e_i + \overline{\boldsymbol{X}}'\boldsymbol{\beta}\right)^2\right) + (Np)^2\sum_{i\in U}\sum_{j\setminus\{i\}\in U}D_{ij}e_i e_j.$$

(2) $V_2(\widehat{Y}_{GREG.R}^*)$ is

$$V_2\left(\widehat{Y}_{GREG.R}^*\right) \approx \sum_{i\in U}\left(D_i(e_i - o_i)^2 + (1-p)p^{-1}\left(e_i + \overline{\boldsymbol{X}}'\boldsymbol{\beta}\right)^2\right) + \sum_{i\in U}\sum_{j\setminus\{i\}\in U}D_{ij}(e_i - o_i)(e_j - o_i).$$

(1) The estimators of $V_1(\widehat{Y}_{GREG.R}^*)$ are

$$\widehat{V}_1\left(\widehat{Y}_{GREG.R}^*\right) = \begin{cases} \widehat{E}_{1p} + \dfrac{(1-p)}{p^2}\displaystyle\sum_{i\in s}\frac{r_i}{\pi_i}\left(\widehat{e}_i + \frac{1}{Np}\sum_{i\in s}\frac{r_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right)^2, when\ p\ is\ known \\ \widehat{E}_{1\widehat{p}} + \dfrac{\left(1-\widehat{p}\right)}{\widehat{p}^2}\displaystyle\sum_{i\in s}\frac{r_i}{\pi_i}\left(\widehat{e}_i + \frac{1}{N\widehat{p}}\sum_{i\in s}\frac{r_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right)^2, when\ p\ is\ unknown \end{cases}.$$

where $\widehat{p} = \sum_{i\in s}\frac{r_i}{\pi_i}\left(\sum_{i\in s}\frac{1}{\pi_i}\right)^{-1}$, $\widehat{Z}_{1ip} = r_i\left(\frac{\widehat{N}_r y_i}{Np} - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r\right)$, $\widehat{Z}_{1i\widehat{p}} = r_i\left(\frac{\widehat{N}_r y_i}{N\widehat{p}} - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r\right)$,

$$\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r, \widehat{N}_r = \sum_{i\in s}\frac{r_i}{\pi_i}, \widehat{E}_{1p} = N^2\left[\sum_{i\in s}\widehat{D}_i\widehat{Z}_{1ip}^2 + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\widehat{Z}_{1ip}\widehat{Z}_{1jp}\right].$$

$$\widehat{E}_{1\widehat{p}} = N^2\left[\sum_{i\in s}\widehat{D}_i\widehat{Z}_{1i\widehat{p}}^2 + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\widehat{Z}_{1i\widehat{p}}\widehat{Z}_{1j\widehat{p}}\right].$$

(2) The estimators of $V_2(\widehat{Y}^*_{GREG.R})$ are

$$\widehat{V}_2\left(\widehat{Y}^*_{GREG.R}\right) = \begin{cases} \widehat{E}_{1p} + \dfrac{(1-p)}{p^2}\sum\limits_{i\in s}\dfrac{r_i}{\pi_i}\left(\widehat{e}_i + \dfrac{1}{N}\sum\limits_{i\in s}\dfrac{r_i \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right)^2, \textit{when } p \textit{ is known} \\ \widehat{E}_{2\widehat{p}} + \dfrac{\left(1-\widehat{p}\right)}{\widehat{p}^2}\sum\limits_{i\in s}\dfrac{r_i}{\pi_i}\left(\widehat{e}_i + \dfrac{1}{N}\sum\limits_{i\in s}\dfrac{r_i \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right)^2, \textit{when } p \textit{ is unknown} \end{cases}.$$

where $\widehat{p} = \sum_{i\in s}\frac{r_i}{\pi_i}\left(\sum_{i\in s}\frac{1}{\pi_i}\right)^{-1}$, $\widehat{Z}_{2ip} = \frac{1}{N}\left(\frac{r_i y_i}{p} - \frac{\frac{1}{N}\sum_{i\in s}\frac{r_i y_i}{\pi_i}}{\overline{W}}w_i\right) - \frac{r_i}{\widehat{N}_r}\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} - \frac{1}{N_r}\sum_{i\in s}\frac{r_i \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right)$,

$$\widehat{Z}_{2i\widehat{p}} = \frac{1}{N}\left(\frac{r_i y_i}{\widehat{p}} - \frac{\frac{1}{N}\sum_{i\in s}\frac{r_i y_i}{\pi_i}}{\overline{W}}w_i\right) - \frac{r_i}{\widehat{N}_r}\left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} - \frac{1}{\widehat{N}_r}\sum_{i\in s}\frac{r_i \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r}{\pi_i}\right), \widehat{E}_{2p}$$

$$= N^2\left(\sum_{i\in s}\widehat{D}_i\widehat{Z}^2_{2ip} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\widehat{Z}_{2ip}\widehat{Z}_{2jp}\right), \widehat{E}_{2\widehat{p}} = N^2\left(\sum_{i\in s}\widehat{D}_i\widehat{Z}^2_{2i\widehat{p}} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\widehat{Z}_{2i\widehat{p}}\widehat{Z}_{2j\widehat{p}}\right).$$

Unfortunately, the works we mentioned above are considered under a strong assumption when the nonresponse mechanism is MCAR where the response probability is constant only. The novel GREG estimators for population mean and total under a more flexible situation where nonresponse occurs under missing at random or MAR, which is a more practical situation, were proposed based on the previous works when the auxiliary variable is known to improve the efficiency of the estimators (Lawson and Siripanich (2022). In their study, they assumed that, C_1: $r_n \to 0$ as $n \to \infty$, where $\{r_n\}$ is a sequence of positive real numbers and $C_2 : \widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} - \boldsymbol{\beta} = O_p\left(n_r^{-\frac{1}{2}}\right)$ and $V\left(\sum_{i\in s}\frac{r_i}{\pi_i p_i}\right) \to 0$ as $n \to \infty$ and the sampling fraction is negligible and non-negligible.

The Lawson and Siripanich (2022) estimator are

$$\widehat{\overline{Y}}^*_{GREG.LS} = \frac{\sum_{i\in s} r_i y_i \Big/ \pi_i p_i}{\sum_{i\in s} r_i \Big/ \pi_i p_i} + \left[\overline{\boldsymbol{X}} - \frac{\sum_{i\in s} r_i \boldsymbol{x}_i \Big/ \pi_i p_i}{\sum_{i\in s} r_i \Big/ \pi_i p_i}\right]' \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right)$$

$$= \widehat{\overline{Y}}_r + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r,$$

where $\widehat{\overline{Y}}_r = \sum_{i\in s}\frac{r_i y_i}{\pi_i p_i}\Big/\sum_{i\in s}\frac{r_i}{\pi_i p_i}$, $\widehat{\overline{\boldsymbol{X}}}_r = \sum_{i\in s}\frac{r_i \boldsymbol{x}_i}{\pi_i p_i}\Big/\sum_{i\in s}\frac{r_i}{\pi_i p_i}$, $\overline{\boldsymbol{X}} = \sum_{i\in U}\boldsymbol{x}_i/N$,

$$\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right).$$

$$\widehat{Y}^{*}_{GREG.LS} = N\widehat{\overline{Y}}^{*}_{GREG.LS} = N\left[\widehat{\overline{Y}}_r + \left(\overline{\boldsymbol{X}} - \widehat{\overline{\boldsymbol{X}}}_r\right)'\widehat{\boldsymbol{\beta}}_r\right].$$

In variance estimation due to the nonlinear estimator, they suggested two estimation techniques called the modified automated linearization approaches to deal with this issue. They suggested to replace $\sum_{i\in s}\frac{r_i}{\pi_i p_i}$ by $\sum_{i\in U}\frac{r_i}{p_i}$ in their estimators and used the Taylor linearization approach to transform nonlinear estimator to linear form.

Their variance estimators are

$$V_1\left(\widehat{Y}^{*}_{GREG.LS}\right) \approx \sum_{i\in U}\frac{1}{p_i}\left(D_i e_i^2 + (1-p_i)\left(e_i - \overline{e}\right)^2\right) + \sum_{i\in U}\sum_{j\setminus\{i\}\in U} D_{ij} e_i e_j.$$

$$V_2\left(\widehat{Y}^{*}_{GREG.LS}\right) \approx \sum_{i\in U}\frac{1}{\pi_i p_i}\left(e_i - \overline{e}\right)^2 + \sum_{i\in U}\sum_{j\setminus\{i\}\in U} D_{ij} e_i e_j.$$

The estimators of $V_1(\widehat{Y}^{*}_{GREG.LS})$ are

$$\widehat{V}_1\left(\widehat{Y}^{*}_{GREG.LS}\right) = \begin{cases} \dfrac{N^2}{\left(\sum_{i\in s}\frac{r_i}{\pi_i p_i}\right)^2}\widehat{E}_{1p_i} + \sum_{i\in s}\dfrac{(1-p_i)}{\pi_i p_i^2} r_i\left(\widehat{e}_i - \widehat{\overline{e}}_r\right)^2, \text{when } p_i \text{ is known for all } i\in s \\ \dfrac{N^2}{\left(\sum_{i\in s}\frac{r_i}{\pi_i \widehat{p}_i}\right)^2}\widehat{E}_{1\widehat{p}_i} + \sum_{i\in s}\dfrac{\left(1-\widehat{p}_i\right)}{\pi_i \widehat{p}_i^2} r_i\left(\widehat{e}_i - \widehat{\overline{e}}_r\right)^2, \text{when } p_i \text{ is unknown for all } i\in s \end{cases},$$

where $\widehat{E}_{1p_i} = \sum_{i\in s}\widehat{D}_i\frac{r_i\widehat{e}_i^2}{p_i^2} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\frac{r_i\widehat{e}_i}{p_i}\frac{r_j\widehat{e}_j}{p_j}$, $\widehat{E}_{1\widehat{p}_i} = \sum_{i\in s}\widehat{D}_i\frac{r_i\widehat{e}_i^2}{\widehat{p}_i^2} + \sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\frac{r_i\widehat{e}_i}{\widehat{p}_i}\frac{r_j\widehat{e}_j}{\widehat{p}_j}$,

$\widehat{p}_i$ is the estimator of p_i for all $i\in s$, $\widehat{e}_i = (y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_r)$, $\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i p_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i p_i}\right)$ if p_i is known for all $i\in s$ otherwise $\widehat{\boldsymbol{\beta}}_r = \left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\pi_i \widehat{p}_i}\right)^{-1}\left(\sum_{i\in s}\frac{r_i q_i \boldsymbol{x}_i y_i}{\pi_i \widehat{p}_i}\right)$, $\widehat{\overline{e}} = N^{-1}\sum_{i\in s}\frac{r_i\widehat{e}_i}{\widehat{p}_i}$ if p_i is known for all $i\in s$ otherwise $\widehat{\overline{e}} = N^{-1}\sum_{i\in s}\frac{r_i\widehat{e}_i}{\widehat{p}_i}$.

The estimators of $V_2(\widehat{\bar{Y}}^{*}_{GREG.LS})$ are

$$\widehat{V}_2\left(\widehat{\bar{Y}}^{*}_{GREG.LS}\right)=\begin{cases}\dfrac{N^2}{\left(\sum\limits_{i\in s}\dfrac{r_i}{\pi_i p_i}\right)^2}\widehat{E}_{2p_i}+\sum\limits_{i\in s}\dfrac{(1-p_i)}{\pi_i p_i^2}r_i\left(\widehat{e}_i-\widehat{\bar{e}}_r\right)^2, \text{when } p_i \text{ is known for all } i\in s\\ \dfrac{N^2}{\left(\sum\limits_{i\in s}\dfrac{r_i}{\pi_i \widehat{p}_i}\right)^2}\widehat{E}_{2\widehat{p}_i}+\sum\limits_{i\in s}\dfrac{\left(1-\widehat{p}_i\right)}{\pi_i \widehat{p}_i^2}r_i\left(\widehat{e}_i-\widehat{\bar{e}}_r\right)^2, \text{when } p_i \text{ is unknown for all } i\in s\end{cases},$$

where $\widehat{E}_{2p_i}=\sum\limits_{i\in s}\widehat{D}_i\frac{r_i(\widehat{e}_i-\widehat{\bar{e}}_r)^2}{p_i^2}+\sum\limits_{i\in s}\sum\limits_{j\setminus\{i\}\in s}\widehat{D}_{ij}\frac{r_i(\widehat{e}_i-\widehat{\bar{e}}_r)}{p_i}\frac{r_i(\widehat{e}_i-\widehat{\bar{e}}_r)}{p_j}$,

$$\widehat{E}_{2\widehat{p}_i}=\sum_{i\in s}\widehat{D}_i\frac{r_i\left(\widehat{e}_i-\widehat{\bar{e}}_r\right)^2}{\widehat{p}_i^2}+\sum_{i\in s}\sum_{j\setminus\{i\}\in s}\widehat{D}_{ij}\frac{r_i\left(\widehat{e}_i-\widehat{\bar{e}}_r\right)}{\widehat{p}_i}\frac{r_j\left(\widehat{e}_j-\widehat{\bar{e}}_r\right)}{\widehat{p}_j}.$$

These GREG estimators can be calculated using any statistical packages, e.g. R program which was used in the reviewed studies. Due to these new GREG estimators are new estimators under the presence of missing data under unequal probability sampling and so unfortunately there is no function in R that can be used straight away. Although they are not that complex to use in the estimation.

5. Examples of application to financial and economic data

The GREG estimator was applied to estimate the total monthly household income from five communities in Bang Sue district, Bangkok, Thailand (Lawson and Siripanich, 2022). The results were based on a sample of size 195 households that was drawn using UPWOR with Midzuno's (1952) scheme out of 1,181 households which consists of 30% nonresponse in the monthly income. The monthly expenditure, age and work in hours per week were considered as the auxiliary variables to assist in estimating the total income and the variance. The logistic regression model was used to find the unknown response probability using the age variable.

Their results showed that their suggested GREG estimator gave the estimated total income for all households equal to 36,068,543 baht and smaller variances in regards to the Särndal and Lundström (2005) estimator.

Data on total monthly income in households is the key to understanding a core part of a country's economy. Information on the financial status of citizens contributes to money flow in the economy and provides invaluable insights for strategizing policies to overcome economic inequalities. Estimation of these statistics allow policymakers to identify income disparities within the nation, integrate measures to assert equality and stabilize the economy, leading to the amelioration of quality of life on a myriad of aspects.

Another example was found in studying Thailand's agriculture which is one of the sources of income that support Thailand's economy (Ponkaew and Lawson, 2023). The Thai maize of Thailand in 2019 from the Office of the Agricultural Economics was studied based on a sample size of 25 provinces being selected using the UPWOR method by Midzuno (1952) out of 63 provinces. The data contained a 30% nonresponse rate. The total yield of maize estimates for all provinces in Thailand in 2019 was found using their suggested GREG

estimator and cultivated area and the harvest area in 2019 were considered as the auxiliary variables along with the cultivated area in 2018 as the size variable. The estimates of total yield of maize for all provinces in Thailand was 525,124 with the smallest variance with respect to the existing estimator.

Statistical estimation of agricultural yield is imperative for agricultural countries such as Thailand and a large part of Asia. These nations' histories have all consisted of agriculture as their geography and climate incline toward successful growing of crops. In prevailing times, export plays an inherent role as one of the major income sources, and an opulence of land is recruited for farming. These farmers are often short on resources and must go through many lengths to save on time and money, to ensure that their yields bring in profit and not losses. The prediction of crop yields can help policymakers working with farmers to anticipate food shortages leading to losses, and potential risks of farming strategies. As many countries are dependent on agriculture, estimation of accurate yields is an essential component of their economies.

6. Conclusions and discussions

We can see that the GREG estimators can be useful to estimate financial and economic data in Thailand and also other countries. Most of these data contain nonresponse which could occur usually during the collection process and as a result it needs to be take care of to gain more accuracy. Many reviewed works based on the GREG estimators under missing data studied under the reverse framework could benefit in the estimation process where we can apply them to real data, e.g. household income, revenue for business, and inflation and unemployment rate.

The GREG estimators are studied under the MCAR and MAR nonresponse mechanisms where both the sampling fractions are small and therefore it can be negligible or either large and cannot be omitted. These GREG estimators are also almost unbiased estimators with reduced variance regarding the existing estimators. The GREG estimators' variance estimators are useful to help in estimating the boundary of the variable of interest to see the lower bound and upper bound for these possible values based on survey sampling. Smaller variance from the GREG estimators can benefit in creating more accuracy for the confidence interval for financial and economic data.

The GREG estimators can assist in estimating these data and therefore knowing these data can be helpful in planning in order to define policies of countries to increase the value of business and finance in the future. The integral concept of economic stability can only be enforced by the support of accurate statistical estimation of financial and economic data through policies and efficient decisions. Flexible statistics can monitor and predict situations such as economic trends, employment figures, and inflation rates, which benefit policymakers, economists, and investors. Most crucial being introducing suitable policies to tackle the nation's financial issues and fill in economic niches, for the well-being of the population through sustainable economic growth.

The GREG estimators can be applied to further studies in any survey designs other than UPWOR for instance, stratified cluster sampling, cluster samplings where nonresponse happens in the study variable and can assist in any application to real data.

References

Berger, Y.G., Tirari, E.H.M. and Till', Y. (2003), "Towards optimal regression estimation in sample surveys", *Australian and New Zealand Journal of Statistics*, Vol. 45 No. 3, pp. 319-329, doi: 10.1111/1467-842x.00286.

Bethlehem, J.G. and Keller, W.J. (1987), "Linear weighting of sample survey data", *Journal of Official Statistics*, Vol. 3 No. 2, pp. 141-153.

Brewer, K.R.W. (2002), *Combined Survey Sampling Inference: Weighing Basu's Elephants*, Arnold, London.

Brewer, K.R.W. and Donadio, M.E. (2003), "The high entropy variance of the Horvitz-Thompson estimator", *Survey Methodology*, Vol. 29 No. 2, pp. 189-196.

Deville, J.C. and Särndal, C.E. (1994), "Variance estimation for the regression imputed Horvitz Thompson estimator", *Journal of Official Statistics*, Vol. 10 No. 4, pp. 381-394.

Estevao, V.M. and Särndal, C.E. (2003), "A new perspective on calibration estimators", *JSM- Section on Survey Research Methods*, pp. 1346-1356.

Fay, R.E. (1991), "A design-based perspective on missing data variance", *Proceedings of the 1991 Annual Research Conference*, US Bureau of the Census, pp. 429-440.

Hajek, J. (1964), "Asymptotic theory of rejective sampling with varying probabilities from a finite population", *Annals of Mathematical Statistics*, Vol. 35 No. 4, pp. 1491-1523, doi: 10.1214/aoms/1177700375.

Hajek, J. (1981), *Sampling from Finite Population*, Marcel Dekker, New York.

Hansen, M.H. and Hurwitz, W.N. (1946), "The problem of nonresponse in sample surveys", *Journal of the American Statistical Association*, Vol. 41 No. 236, pp. 517-529, doi: 10.1080/01621459.1946.10501894.

Hartley, H.O. and Rao, J.N.K. (1962), "Sampling with unequal probability and without replacement", *The Annals of Mathematical Statistics*, Vol. 33 No. 2, pp. 350-374, doi: 10.1214/aoms/1177704564.

Haziza, D. (2010), "Resampling methods for variance estimation in the presence of missing survey data", *Proceedings of the Annual Conference of the Italian Statistical Society*.

Haziza, D. and Rao, J.N.K. (2006), "A nonresponse model approach to inference under imputation for missing survey data", *Survey Methodology*, Vol. 32 No. 1, pp. 53-64.

Horvitz, D.F. and Thompson, D.J. (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, Vol. 47 No. 260, pp. 663-685, doi: 10.1080/01621459.1952.10483446.

Lawson, N. (2017), "Variance estimation in the presence of nonresponse under probability proportional to size sampling", *Proceedings of the 6th Annual International Conference on Computational Mathematics, Computational Geometry and Statistics (CMCGS 2017)*, Singapore, doi: 10.5176/2251-1911_cmcgs17.32.

Lawson, N. and Ponkaew, C. (2019), "New generalized regression estimator in the presence of nonresponse under unequal probability sampling", *Communications in Statistics -Theory and Methods*, Vol. 48 No. 10, pp. 2483-2498, doi: 10.1080/03610926.2018.1465091.

Lawson, N. and Siripanich, P. (2022), "A new generalized regression estimator and variance estimation for unequal probability sampling without replacement for missing data", *Communications in Statistics -Theory and Methods*, Vol. 51 No. 18, pp. 6296-6318, doi: 10.1080/03610926.2020.1860224.

Midzuno, H. (1952), "On the sampling system with probability proportional to sum of sizes", *Annals of the Institute of Statistical Mathematics*, Vol. 55 No. 3, pp. 99-107.

Montanari, G. (1987), "Post sampling efficient qr-prediction in large sample survey", *International Statistics*, Vol. 55 No. 2, pp. 191-202, doi: 10.2307/1403195.

Ponkaew, C. and Lawson, L. (2023), "New generalized regression estimators using a ratio method and its variance estimation for unequal probability sampling without replacement in the presence of nonresponse", *Current Applied Science and Technology*, Vol. 23 No. 2, doi: 10.55003/cast.2022.02.23.007.

Rao, J.N.K. (1990), "Variance estimation under imputation for missing data", Technical report, Statistics Canada, Ottawa, pp. 599-608.

Särndal, C.E. (1992), "Method for estimating the precision of survey estimateswhen imputation has been used", *Survey Methodology*, Vol. 18, pp. 241-252.

Särndal, C.E. (2007), "The calibration approach in survey theory and practice", *Survey Methodology*, Vol. 33 No. 2, pp. 99-119.

Särndal, C.E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley & Sons, New York.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer- Verlag, New York.

Sen, A.R. (1953), "On the estimate of the variance in sampling with varying probabilities", *Journal of the Indian Society of Agricultural Statistics*, Vol. 5, pp. 119-127.

Shao, J. and Steel, P. (1999), "Variance estimation for survey data with composite imputation and nonnegligible sampling fractions", *Journal of the American Statistical Association*, Vol. 94 No. 445, pp. 254-265, doi: 10.2307/2669700.

Yates, F. and Grundy, P.M. (1953), "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society: Series B*, Vol. 15 No. 2, pp. 235-261, doi: 10.1111/j.2517-6161.1953.tb00140.x.

Corresponding author
Nuanpan Lawson can be contacted at: nuanpan.n@sci.kmutnb.ac.th