Songklanakarin J. Sci. Technol.
44 (4), 1109–1118, Jul. – Aug. 2022

http://www.sjst.psu.ac.th

Original Article

A chain regression exponential type imputation method for mean estimation in the presence of missing data

Kanisa Chodjuntug, and Nuanpan Lawson*

Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bang Sue, Bangkok, 10800 Thailand

Received: 14 March 2022; Revised: 1 June 2022; Accepted: 8 June 2022

Abstract

Imputation methods deal with item nonresponse to solve the missing data problem. A new imputation method and corresponding point estimators for population mean have been proposed under two situations: using the response rate and the constant that gives the minimum mean square error for the estimator. The biases and mean square errors of the proposed estimators are derived. The performance of this method is compared with some existing methods via simulations and an application to fine particulate matter data. The results show that the proposed estimator, which uses the optimum value of a constant, performs the best. It performs the second best when using the response rate in the estimator, which is free of known parameters. The estimated fine particulate matter in Kanchana Phisek Road in Bangkok using the best method is equivalent to 42.22 micrograms per cubic meter with a mean square error of 0.34 micrograms per cubic meter squared.

Keywords: imputation, regression estimator, bias, mean square error, missing data

1. Introduction

Missing data or nonresponse usually occurs in sample surveys, in which some sampling units refuse to respond sometimes, or are unable to participate in the sample surveys. This is a serious problem for researchers. If a dataset contains missing values, it leads to a negative effect on the results obtained from standard statistical methods such as population mean, population total, and population variance estimates. To solve this problem, one of the most popular methods is imputation using available data as a tool for the replacement of missing observations. In the case of a full response, several researchers have worked on estimating the population mean of the study variable Y by utilizing the information on an auxiliary variable X to increase the efficiency of the estimator. For example, Cochran (1940) applied auxiliary information to mean estimation and proposed a ratio estimator under a simple random sampling without replacement (SRSWOR) scheme as follows:

$$\bar{y}_R = \bar{y}_n \frac{\bar{X}}{\bar{x}_n}, \tag{1}$$

where $\bar{y}_n$ is sample mean of Y, and $\bar{X}$ and $\bar{x}_n$ are the population mean and sample mean of X, respectively.

Srivenkatarmana (1980) was the first who proposed the transformation of an auxiliary variable and Bandyopadhyay (1980) suggested the transformation of an auxiliary variable to improve the population mean estimator as the dual to product estimator, for estimating the population mean as follows:

$$\bar{y}_R = \bar{y}_n \frac{\bar{x}}{\bar{X}_n}, \tag{2}$$

where $\bar{X}_n^* = \frac{N\bar{X} - n\bar{x}_n}{N-n}$, and N and n are the sizes of population and sample.

Bahl and Tuteja (1991) proposed a new ratio type exponential method for estimating population mean under the SRSWOR scheme and their method is more efficient than the

*Corresponding author
Email address: nuanpan.n@sci.kmutnb.ac.th

common methods: mean and ratio methods. Motivated by Bahl and Tuteja (1991), Singh and Pal (2015) proposed a chain ratio-ratio type exponential method which is more efficient than the common estimators including the mean, ratio and ratio type exponential estimators. This method replaces a sample mean of Y with $\bar{y}_R$. They defined the method for population mean estimation as follows:

$$\bar{y}_{CR} = \bar{y}_n \frac{\bar{X}}{\bar{x}_n} \exp\left(\frac{\bar{X}-\bar{x}_n}{\bar{X}+\bar{x}_n}\right). \tag{3}$$

The estimators described above cannot be used to estimate the population mean when a dataset contains missing values. In the case of nonresponse, imputation is one of the methods to handle missing data and uses the available data as a source to draw assumptions to make reasonably accurate replacements for the missing observations. Moreover, a corresponding point estimator of population mean is obtained from the imputation method. Consequently, some researchers have investigated the imputation method to improve the efficiency of the estimator obtained from the imputation method. Many statisticians have applied information on an auxiliary variable to develop the imputation method. For example, Singh and Horn (2000) suggested a new imputation method called the compromised imputation method which contains a constant in the linear combination of main information and auxiliary information under the SRSWOR scheme.

Singh *et al.* (2014) proposed an exponential-type compromised imputation method which was motivated by Bahl and Tuteja (1991). They suggested the procedure as follows:

$$y_{.i} = \begin{cases} k\frac{n}{r}y_i + (1-k)\bar{y}_r \exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right); i \in R \\ (1-k)\bar{y}_r \exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right) \qquad ; i \in R^c, \end{cases} \tag{4}$$

where k is a suitably chosen constant, $\bar{x}_r$ and $\bar{y}_r$ are the response mean of X and Y respectively, $\bar{X}$ is the population mean of auxiliary variable X, Y is an observed value of Y for the i^{th} unit, and R and R^c are the sets of responding and non-responding units, respectively.

Under this method, the corresponding point estimator of population mean is defined as follows:

$$\bar{y}_{Exp} = k\bar{y}_r + (1-k)\bar{y}_r \exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right). \tag{5}$$

The mean square error of $\bar{y}_{Exp}$ is given by

$$MSE\left(\bar{y}_{Exp}\right) = \bar{Y}^2\left(\frac{1}{r}-\frac{1}{N}\right)\left[C_Y^2 + \frac{(1-\alpha)^2}{4}C_X^2 - (1-\alpha)\rho C_X C_Y\right], \tag{6}$$

where ρ is the correlation coefficient between X and Y, and C_X and C_Y are the coefficients of variation of X and Y, respectively.

Their research showed the superiority of the Singh *et al.* (2014) method of imputation with $k = 1-2\frac{C_{XY}}{C_X^2}$, $C_{XY} = \rho C_X C_Y$ over the common mean, ratio, and compromised imputation methods under certain conditions.

Recently, Chodjuntug and Lawson (2022) proposed an improved imputation method using the idea of Singh and Pal (2015) to improve Singh *et al.*'s (2014) alternative with the chain ratio estimator. In addition, their research suggested two constants (w_1, w_2) that replace k with w_1 and (1-k) with w_2 in Singh *et al.*'s method. The Chodjuntug and Lawson estimator is defined by

$$y_{.i} = \begin{cases} w_1\frac{n}{r}y_i + w_2\bar{y}_r\frac{\bar{X}}{\bar{x}_r}\exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right); i \in R \\ w_2\bar{y}_r\frac{\bar{X}}{\bar{x}_r}\exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right) \qquad ; i \in R^c, \end{cases} \tag{7}$$

where w_1 and w_2 are suitably chosen constants.

The point estimator of the population mean obtained from Chodjuntug and Lawson (2022) is

$$\bar{y}_{NExp} = w_1\bar{y}_r + w_2\bar{y}_r\frac{\bar{X}}{\bar{x}_r}\exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right). \tag{8}$$

The mean square error of $\bar{y}_{NExp}$ is

$$\begin{aligned} MSE\left(\bar{y}_{NExp}\right) = \bar{Y}^2\Big[&\left(w_1+w_2-1\right)^2 + \left(w_1+w_2\right)^2\psi_{r,N}C_Y^2 \\ &+ \psi_{r,N}\frac{3}{4}w_2C_X\left[3w_2C_X - 4\left(w_1+w_2\right)\rho C_Y\right], \end{aligned} \tag{9}$$

where $\psi_{r,N} = \frac{1}{r}-\frac{1}{N}$.

This method is more efficient than Singh *et al.*'s method of imputation, mean method of imputation, ratio method of imputation, or compromised method of imputation under conditions $w_1 = \frac{1}{A}\left(1-\frac{2}{3}K\right)$, $w_2 = \frac{2K}{3A}$, $A = 1+MC_Y^2\left(1-\rho^2\right)$ and $K = \rho\frac{C_Y}{C_X}$. However, when each constant calculated by the different fomulas, this is complicated to apply for general researchers.

In this study, we propose an improved exponential-type imputation method using the ideas of Singh *et al.* (2014), and Chodjuntug and Lawson (2022) along with a corresponding estimator obtained from the proposed method, which we will consider under two situations; using the response rate and the constant that gives minimum mean square error of the estimator. The biases and mean square errors of the proposed estimators are obtained up to the first

degree of approximation using a Taylor series. The mean square errors are used to compare the performances of the proposed estimators with some existing estimators in simulation studies, and in an application using the fine particulate matter 2.5 data from Kanchana Phisek road, Thailand.

2. Materials and Methods

This section reviews basic steps in our research. The procedures of the proposed imputation method are presented along with a corresponding point estimator, which is obtained from the proposed imputation method. In addition, the properties bias and mean square error of the proposed estimators are derived.

2.1 Basic setup

Let $U = \{U_1, U_2, \ldots, U_N\}$ be a finite population of size N, y_i and x_i be values of study variable Y and auxiliary variable X, where $i \in \{1, 2, \ldots, N\}$. Let $\bar{Y} = \sum_{i-1}^{N} y_i / N$ and $\bar{X} = \sum_{i-1}^{N} x_i / N$ be the population means of Y and X respectively. Let R and R^c be the sets of responding units and non-responding units. The value y_i is observed for every $i \in R$, but is missing for every $i \in R^c$. Based on a simple random sampling without replacement, s of size n with paired variables (X, Y) is selected from U and contains both r responding units and $(n\text{-}r)$ non-responding units. Let $\bar{x}_n = \sum_{i=1}^{n} x_i / n$, $\bar{x}_r = \sum_{i=1}^{r} x_i / r$ and $\bar{y}_r = \sum_{i=1}^{r} y_i / r$ be the sample mean of X and the response mean of X and Y, respectively.

2.2 Proposed imputation method

A new exponential-type imputation method is proposed following the ideas of Singh *et al.* (2014) and Chodjuntug and Lawson (2022). We propose to replace the ratio estimator $\bar{y}_r \frac{\bar{X}}{\bar{x}_r}$ in equation (7) with the regression estimator $\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)$ under the condition $w_1 + w_2 = 1.$ The general form of imputation method is as follows:

$$y_{.i} = \begin{cases} w_1 \frac{n}{r} y_i + w_2 \left[\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)\right] \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right); i \in R \\ w_2 \left[\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)\right] \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right) \quad ; i \in R^c, \end{cases} \tag{10}$$

where $b = \frac{s_{xy}}{s_x^2}$, $S_{XY} = \sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)/(N-1)$, $S_X^2 = \sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2/(N-1)$, $\bar{x}_r = \sum_{i=1}^{r} x_i / r$ and $\bar{y}_r = \sum_{i=1}^{r} y_i / r$.

Under the proposed imputation method, the point estimator of the population mean is

$$\bar{y}_{REXp} = w_1 \bar{y}_r + w_2 \left[\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)\right] \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right). \tag{11}$$

Corollary 1. Under simple random sampling without replacement with nonresponse in the study variable, $w_1 + w_2 = 1$ and $w_1 = k$, $w_2 = 1\text{-}k$.

To find the properties of the proposed estimator, we define

$$\bar{y}_r = \bar{Y}\left(1 + e_0\right), \ \bar{x}_r = \bar{X}\left(1 + e_1\right), \ s_{xy} = S_{XY}\left(1 + e_2\right), \ s_x^2 = S_X^2\left(1 + e_3\right).$$

Under SRSWOR, the first and second moments of e_i; $I = 1, 2, 3$ are

$$E\left(e_i\right) = 0; \ i = 1, 2, 3, \ E\left(e_0^2\right) = \psi_{r,N} C_Y^2, \ E\left(e_1^2\right) = \psi_{r,N} C_X^2 \text{ and } E\left(e_0 e_1\right) = \psi_{r,N} C_{XY},$$

where

$$\psi_{r,N} = \frac{1}{r} - \frac{1}{N}, \ C_X = \frac{S_X}{\bar{X}}, \ C_Y = \frac{S_Y}{\bar{Y}}, C_{XY} = \rho C_X C_Y, \ \rho = \frac{S_{XY}}{S_X S_Y} \text{ and } S_Y^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2/(N-1).$$

Next, writing $\bar{y}_{REXp}$ in terms of e_i's, equation (11) takes the following form:

$$\begin{aligned}\bar{y}_{RExp} &= w_1\bar{Y}(1+e_0)+w_2\left(\left[\bar{Y}(1+e_0)+\frac{S_{XY}(1+e_2)}{S_X^2(1+e_3)}\left(-\bar{X}e_1\right)\right]\times\exp\left\{\frac{-\bar{X}e_1}{2\bar{X}+\bar{X}e_1}\right\}\right)\\ &= w_1\bar{Y}(1+e_0)+w_2\left[\bar{Y}(1+e_0)-\beta\bar{X}(e_1+e_1e_2-e_1e_3)-\frac{1}{2}\left[\bar{Y}(e_1+e_0e_1)-\beta\bar{X}e_1^2\right]\right.\\ &\left.+\frac{3}{8}\left[\bar{Y}\left(e_1^2+e_0e_1^2\right)-\beta\bar{X}e_1^3\right]+\ldots\right].\end{aligned} \tag{12}$$

To find $Bias\left(\bar{y}_{RExp}\right)$, subtract $\bar{Y}$ from equation (12), expanding and neglecting the higher order terms, and on taking expectations we have

$$\begin{aligned}Bias\left(\bar{y}_{RExp}\right) &= E\left[\bar{y}_{RExp}-\bar{Y}\right]\\ &= E\left[\left(w_1+w_2\right)\bar{Y}(1+e_0)+w_2\left(-\beta\bar{X}(e_1+e_1e_2-e_1e_3)-\frac{1}{2}\left(\bar{Y}(e_1+e_0e_1)-\beta\bar{X}e_1^2\right)\right.\right.\\ &\left.\left.+\frac{3}{8}\left(\bar{Y}\left(e_1^2+e_0e_1^2\right)-\beta\bar{X}e_1^3\right)+\ldots\right)-\bar{Y}\right]\\ &\approx (1-k)\left[-\beta\bar{X}E(e_1e_2-e_1e_3)-\frac{1}{2}\bar{Y}E(e_0e_1)+\left(\frac{\beta}{2}\bar{X}+\frac{3}{8}\bar{Y}\right)E\left(e_1^2\right)\right]\\ &= (1-k)\left[-\beta\bar{X}\psi_{r,N}\left(\frac{\mu_{21}}{\bar{X}S_{XY}}-\frac{\mu_{30}}{\bar{X}S_{XY}}\right)-\frac{1}{2}\bar{Y}\psi_{r,N}C_{YX}+\left(\frac{\beta}{2}\bar{X}+\frac{3}{8}\bar{Y}\right)\psi_{r,N}C_X^2\right].\end{aligned}$$

Therefore, we get

$$Bias\left(\bar{y}_{RExp}\right)\approx(1-k)\left[-\beta\bar{X}\psi_{r,N}\left(\frac{\mu_{21}}{\bar{X}S_{XY}}-\frac{\mu_{30}}{\bar{X}S_X^2}\right)+\frac{3}{8}\psi_{r,N}\bar{Y}C_X^2\right], \tag{13}$$

where $\psi_{r,N}=\frac{1}{r}-\frac{1}{N}$, $\beta=\frac{S_{XY}}{S_X^2}$, $\mu_{21}=E\left[\left(x-\bar{X}\right)^2\left(y-\bar{Y}\right)\right]$ and $\mu_{30}=E\left[\left(x-\bar{X}\right)^3\right]$.

To find $MSE\left(\bar{y}_{RExp}\right)$, subtract $\bar{Y}$ from equation (12) along with squaring, expanding and retaining the terms up to the first degree of approximation, and on taking expectations we have

$$\begin{aligned}MSE\left(\bar{y}_{RExp}\right) &= E\left[\bar{y}_{RExp}-\bar{Y}\right]^2\\ &= E\left[\left(w_1+w_2\right)\bar{Y}(1+e_0)+w_2\left(-\beta\bar{X}(e_1+e_1e_2-e_1e_3)\right.\right.\\ &\left.\left.-\frac{1}{2}\left(\bar{Y}(e_1+e_0e_1)-\beta\bar{X}e_1^2\right)+\frac{3}{8}\left(\bar{Y}\left(e_1^2+e_0e_1^2\right)-\beta\bar{X}e_1^3\right)+\ldots\right)-\bar{Y}\right]^2\\ &\approx E\left[\bar{Y}e_0+(1-k)\left(-\beta\bar{X}e_1-\frac{\bar{Y}}{2}e_1\right)\right]^2\\ &= \bar{Y}^2E\left(e_0^2\right)-2\bar{Y}(1-k)\left(\beta\bar{X}+\frac{\bar{Y}}{2}\right)E(e_0e_1)\\ &\quad+(1-k)^2\left(\beta\bar{X}+\frac{\bar{Y}}{2}\right)^2E\left(e_1^2\right)\\ &= \bar{Y}^2\psi_{r,N}C_Y^2-2\bar{Y}(1-k)\left(\beta\bar{X}+\frac{\bar{Y}}{2}\right)\psi_{r,N}C_{YX}\\ &\quad+\left[(1-k)\left(\beta\bar{X}+\frac{\bar{Y}}{2}\right)\right]^2\psi_{r,N}C_X^2.\end{aligned}$$

Therefore, we get

$$MSE\left(\bar{y}_{RExp}\right) \approx \psi_{r,N}\left(S_Y^2 - 2A\rho S_Y S_X + A^2 S_X^2\right), \tag{14}$$

where $A = (1-k)\left(\frac{RT}{2} + \beta\right)$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

From corollary 1, the constant $w_1 = k$ is unknown. We consider it under the two situations below.

1) Use the response rate as a weight. We define $k = \frac{r}{n} = k_r$, r is the number of responses and n is the sample size. The proposed estimator $\bar{y}_{RExp}$ from equation (11) becomes

$$\bar{y}_{RExp.kr} = \frac{r}{n}\bar{y}_r + \left(1 - \frac{r}{n}\right)\left[\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)\right]\exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right). \tag{15}$$

The $MSE\left(\bar{y}_{RExp.kr}\right)$ under $k = \frac{r}{n} = k_r$ is

$$MSE\left(\bar{y}_{RExp.kr}\right) \approx \psi_{r,N}\left(S_Y^2 - 2A_1\rho S_Y S_X + A_1^2 S_X^2\right), \tag{16}$$

where $A_1 = \left(1 - \frac{r}{n}\right)\left(\frac{RT}{2} + \beta\right)$.

2) Find the optimum constant k, k_{opt} which gives the minimum mean square error. To obtain the constant k, differentiate equation (14) with respect to k and equate the derivative to zero.

$$\frac{dMSE(\bar{y}_{RExp})}{dk} = 0$$

$$\frac{d\psi_{r,N}\left(S_Y^2 - 2(1-k)(\frac{RT}{2} + \beta)S_{XY} + ((1-k)(\frac{RT}{2} + \beta))^2 S_X^2\right)}{dk} = 0$$

$$\psi_{r,N}\left(2a\rho S_Y S_X - 2a^2 S_X^2 + 2ka^2 S_X^2\right) = 0$$

The optimum value of k is

$$k = 1 - \frac{\rho S_Y}{aS_X} = k_{opt}, \tag{17}$$

where $a = \left(\frac{RT}{2} + \beta\right)$.

The proposed estimator $\bar{y}_{RExp}$ from equation (11) becomes

$$\bar{y}_{RExp.kopt} = k_{opt}\bar{y}_r + \left(1 - k_{opt}\right)\left[\bar{y}_r + b\left(\bar{X} - \bar{x}_r\right)\right]\exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right). \tag{18}$$

We get the minimum mean square error of estimator $\bar{y}_{RExp.kopt}$ by replacing constant k with $k_{opt} = 1 - \frac{\rho S_Y}{aS_X}$ in equation (14).

Therefore $MSE\left(\bar{y}_{RExp.kopt}\right)$ is

$$MSE\left(\bar{y}_{RExp.kopt}\right) \approx \psi_{r,N}\left(S_Y^2 - 2A_2 S_Y S_X + A_2^2 S_X^2\right), \tag{19}$$

where $A_2 = \frac{\rho S_Y}{aS_X}\left(\frac{RT}{2} + \beta\right)$.

Note, that we can use $s_y^2 = \frac{\sum_{i=1}^{r}(y_i - \bar{y}_r)^2}{r-1}$, $s_x^2 = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_r)^2}{r-1}$ and $s_{xy} = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_r)(y_i - \bar{y}_r)}{r-1}$ to estimate S_Y^2, S_X^2, S_{XY} when the parameters are unknown. Therefore, we suggest the estimator of k_{opt} as follows:

$$\hat{k}_{opt} = 1 - \frac{s_{xy}}{\left(\frac{\bar{y}_r}{2\bar{x}_r} + b\right)s_x^2}. \tag{20}$$

2.3 Efficiency comparison of the proposed estimators

In this section, the performances of the proposed estimators using k as the response rate and using k with the optimum value are compared with some existing estimators: $\bar{y}_{Exp}$ and $\bar{y}_{NExp}$ using the mean square error in order to derive the conditions for when the proposed estimators $\bar{y}_{RExp.kr}$ and $\bar{y}_{RExp.kopt}$ are better than other mentioned estimators under the condition $w_1 + w_2 = 1$.

2.3.1 Comparison of proposed estimator $\bar{y}_{RExp.kr}$ with estimator $\bar{y}_{Exp}$

We have $MSE\left(\bar{y}_{Exp}\right) > MSE\left(\bar{y}_{RExp.kr}\right)$

$$\text{if } \rho - \frac{(A_1 + C)S_X}{2S_Y} > 0, \tag{21}$$

where $A_1 = \left(1 - \frac{r}{n}\right)\left(\frac{RT}{2} + \beta\right)$, $C = \frac{1}{2}\left(1 - \frac{r}{n}\right)RT$, $\beta = \frac{S_{XY}}{S_X^2}$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

When the condition in equation (21) is satisfied, $\bar{y}_{RExp.kr}$ is more efficient than $\bar{y}_{Exp}$.

2.3.2 Comparison of proposed estimator $\bar{y}_{RExp.kr}$ with estimator $\bar{y}_{NExp}$

We have $MSE\left(\bar{y}_{NExp}\right) > MSE\left(\bar{y}_{RExp.kr}\right)$

$$\text{if } \rho - \frac{(A_1 + D)S_X}{2S_Y} > 0, \tag{22}$$

where $A_1 = \left(1 - \frac{r}{n}\right)\left(\frac{RT}{2} + \beta\right)$, $D = \frac{3}{2}\left(1 - \frac{r}{n}\right)RT$, $\beta = \frac{S_{XY}}{S_X^2}$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

When the condition in equation (22) is satisfied, $\bar{y}_{RExp.kr}$ is more efficient than $\bar{y}_{NExp}$.

2.3.3 Comparison of proposed estimator $\bar{y}_{RExp.kopt}$ with estimator $\bar{y}_{Exp}$

We have $MSE\left(\bar{y}_{Exp}\right) > MSE\left(\bar{y}_{RExp.kopt}\right)$

$$\text{if } \rho - \frac{(A_2 + C)S_X}{2S_Y} > 0, \tag{23}$$

where $A_2 = \frac{\rho S_Y}{aS_X}\left(\frac{RT}{2} + \beta\right)$, $C = \frac{1}{2}\left(1 - \frac{r}{n}\right)RT$, $\beta = \frac{S_{XY}}{S_X^2}$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

When the condition in equation (23) is satisfied, $\bar{y}_{RExp.kopt}$ is more efficient than $\bar{y}_{Exp}$.

2.3.4 Comparison of proposed estimator $\bar{y}_{RExp.kopt}$ with estimator $\bar{y}_{NExp}$

We have $MSE\left(\bar{y}_{NExp}\right) > MSE\left(\bar{y}_{RExp.kopt}\right)$

if $\rho - \frac{(A_2 + D)S_X}{2S_Y} > 0,$ (24)

where $A_2 = \frac{\rho S_Y}{aS_X}\left(\frac{RT}{2} + \beta\right)$, $D = \frac{3}{2}\left(1 - \frac{r}{n}\right)RT$, $\beta = \frac{S_{XY}}{S_X^2}$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

When the condition in equation (24) is satisfied, $\bar{y}_{RExp.kopt}$ is more efficient than $\bar{y}_{NExp}$.

2.3.5 Comparison of proposed estimator $\bar{y}_{RExp.kopt}$ with estimator $\bar{y}_{RExp.kr}$

We have $MSE\left(\bar{y}_{RExp.kr}\right) > MSE\left(\bar{y}_{RExp.kopt}\right)$

if $\rho - \frac{(A_1 + A_2)S_X}{2S_Y} > 0,$ (25)

where $A_1 = \left(1 - \frac{r}{n}\right)\left(\frac{RT}{2} + \beta\right)$, $A_2 = \frac{\rho S_Y}{aS_X}\left(\frac{RT}{2} + \beta\right)$, $\beta = \frac{S_{XY}}{S_X^2}$ and $RT = \frac{\bar{Y}}{\bar{X}}$.

When the condition in equation (25) is satisfied, $\bar{y}_{RExp.kopt}$ is more efficient than $\bar{y}_{RExp.kr}$.

3. Results and Discussion

To see the performances of the corresponding estimator obtained from the proposed method for estimating population mean, we consider both simulation studies and an application to fine particulate matter data in Thailand using the R program (R Core Team (2021)). The details are as follows.

3.1. Simulation studies

In the simulation studies, we compare the performance of the new imputation method with the existing methods which are used to estimate the population mean in the presence of missing data, to support the theoretical findings. A paired (*X, Y*) dataset is generated from bivariate normal distribution with parameters $\bar{X} = 50$, $\bar{Y} = 200$, $S_x = 5$, $S_y = 100$ and ρ = 0.3, 0.5, 0.8 where all the conditions in equations (21)-(25) are satisfied. Random samples of sizes *n*(*n* = 100, 300, 600, 1000) are drawn from a population of size *N*=2,000 by the SRSWOR scheme. The study variable *Y* is missing completely at random at the three levels 30, 20, and 10%, respectively. The simulation is repeated 10,000 times. We calculated the $\bar{y}$ of each estimator, and then calculated the percentage relative efficiencies (PRE) of the estimators with respect to $\bar{y}_{Exp}$. The results are shown in Table 1.

Table 1 shows that both proposed estimators give higher percentage relative efficiencies than the existing estimators. We can see big improvements in the proposed estimators when the correlation between *Y* and *X* is equal to 0.8 at all levels of missing values. The proposed estimator using the value of optimum *k*, $\bar{y}_{RExp.kopt}$ performs the best and is followed by the proposed estimator using the response rate as the value of *k*, which is also an alternative estimator to use when some parameters are unknown.

3.2. Case study

To assess the performance of the proposed estimator, fine particulate matter $PM_{2.5}$($\mu g/m^3$) and carbon monoxide *CO* (ppm) data from Kanchana Phisek Road in Bangkok, Thailand, in January 2020, are used in this study. The data were collected as averages for every hour, and were obtained from the website of the Pollution Control Department of Thailand. $PM_{2.5}$ is considered the study variable *Y* and *CO* is considered the auxiliary variable *X*. The parameters of the data are as follows: *N* = 708, $\bar{Y} = 46.4$, $\bar{X} = 3.0$, S_Y = 20.2, S_X = 0.3, and ρ = 0.7 where all the conditions in equations (21)-(25) are satisfied. Random sampling is used to select sample sizes *n*(*n* = 30, 50, 80) from the population. The data contain 20% of missing values of $PM_{2.5}$. The results are shown in Figure 1 and Table 2.

Table 2 shows similar results to Table 1 in that the proposed estimator $\bar{y}_{RExp.kopt}$ outperforms other existing estimators, followed by $\bar{y}_{RExp.kr}$. The proposed estimators works well with the $PM_{2.5}$ data set. Therefore, it can be used to impute missing fine particulate matter and to estimate fine particulate matter in Kanchana Phisek Road in Bangkok, which is 42.22 micrograms per cubic meter with a mean squared error of 0.34 micrograms per cubic meter squared.

Table 1. The percentage relative efficiencies of the estimators at different levels of sample size

ρ	n	Estimator	Percentage relative efficiency		
			Missing 30%	Missing 20%	Missing 10%
0.30	100	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	103.02	102.34	101.20
		$\bar{y}_{RExp.kr}$	104.64	103.63	101.88
		$\bar{y}_{RExp.kopt}$	**107.97**	**110.07**	**110.07**
	300	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	103.23	102.30	101.24
		$\bar{y}_{RExp.kr}$	104.83	103.62	101.98
		$\bar{y}_{RExp.kopt}$	**108.87**	**109.71**	**110.62**
	600	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	103.25	102.25	101.22
		$\bar{y}_{RExp.kr}$	104.90	103.56	101.95
		$\bar{y}_{RExp.kopt}$	**108.76**	**109.88**	**110.47**
	1,000	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	103.18	102.22	101.15
		$\bar{y}_{RExp.kr}$	104.87	103.50	101.84
		$\bar{y}_{RExp.kopt}$	**108.75**	**109.03**	**108.94**
0.50	100	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	105.75	104.08	102.04
		$\bar{y}_{RExp.kr}$	115.02	110.77	105.39
		$\bar{y}_{RExp.kopt}$	**130.94**	**135.59**	**135.53**
	300	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	105.96	104.05	102.07
		$\bar{y}_{RExp.kr}$	115.35	110.69	105.52
		$\bar{y}_{RExp.kopt}$	**132.39**	**133.27**	**134.78**
	600	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	105.97	103.99	102.05
		$\bar{y}_{RExp.kr}$	115.36	110.57	105.46
		$\bar{y}_{RExp.kopt}$	**130.83**	**133.73**	**133.99**
	1,000	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	105.90	103.96	101.98

Table 1. Continued.

ρ	n	Estimator	Percentage relative efficiency		
			Missing 30%	Missing 20%	Missing 10%
0.50	1000	$\bar{y}_{RExp.kr}$	115.17	110.39	105.26
		$\bar{y}_{RExp.kopt}$	**130.14**	**129.95**	**129.13**
0.80	100	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	110.26	106.83	103.32
		$\bar{y}_{RExp.kr}$	150.95	131.88	114.54
		$\bar{y}_{RExp.kopt}$	**275.13**	**290.69**	**289.50**
	300	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	110.42	106.83	103.35
		$\bar{y}_{RExp.kr}$	151.36	131.66	114.63
		$\bar{y}_{RExp.kopt}$	**273.40**	**271.52**	**273.16**
	600	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	110.45	106.76	103.33
		$\bar{y}_{RExp.kr}$	151.01	131.31	114.53
		$\bar{y}_{RExp.kopt}$	**256.04**	**266.28**	**260.32**
	1,000	$\bar{y}_{Exp}$	100.00	100.00	100.00
		$\bar{y}_{NExp}$	110.36	106.71	103.27
		$\bar{y}_{RExp.kr}$	150.07	130.76	114.16
		$\bar{y}_{RExp.kopt}$	**241.69**	**234.19**	**225.03**

Table 2. Mean square errors of the proposed estimators and the existing estimators

Estimator	Mean square error		
	n = 30	n = 50	n = 80
$\bar{y}_{Exp}$	15.89	9.31	5.61
$\bar{y}_{NExp}$	14.87	8.72	5.25
$\bar{y}_{RExp.kr}$	12.93	7.57	4.57
$\bar{y}_{RExp.kopt}$	7.91	4.64	2.79

4. Conclusions

The chain regression exponential-type imputation method has been proposed for estimating population mean when nonresponse occurs in the study variable, using simple random sampling without replacement. The corresponding point estimators for population mean were also obtained from the proposed method of imputation under two situations. We suggested two alternatives to estimate the value of k; one is to use the response rate that is available on hand; and one is to use the optimum value of the constant that makes the mean square error its minimum. The properties of the proposed estimators, such as bias and mean squared error, were derived. We performed simulation studies and an application to fine particular matter observed in Bangkok to see the efficiency of the proposed estimator compared to other existing estimators. The results from both simulation studies and from the case study on fine particualte matter showed that the proposed method using the constant that makes mean squared error its optimum performed the best, and the second best was the one using the response rate, which is free from parameters. Therefore, this is an alternative approach to employ when

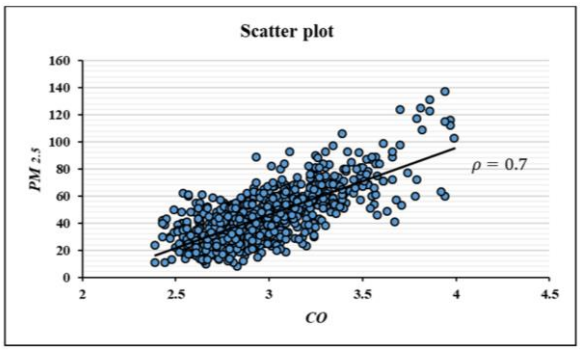

Figure 1. The scatter plot between $PM_{2.5}$ and CO concentration data

some parameters are not available in the study.

The proposed imputation method is an alternative method to handle missing data in the real world. It can be used to impute missing observations in order to create a completed data set and then to estimate the population mean or population total of the study variable. The obtained bias and mean square error formulas are easy to implement for researchers in order to measure the estimator's efficiency. The proposed methods can be applied in complex survey designs such as two-phase sampling, stratified sampling, and cluster sampling, and can be extended to the case where both the study and auxiliary variables are missing. Other aspects in terms of population parameters can also be estimated such as total, proportion, and variance, that can be considered in future research.

Acknowledgements

This research was funded by King Mongkut's University of Technology North Bangkok, Contract no. KMUTNB-65-BASIC-44. We would like to thank all unknown referees for their helpful comments.

References

Badyopadhyay, S. (1980). Improved ratio and product estimators. *Sankhya series C, 42*, 45-49.

Bahl, S., & Tuteja, R. K. (1991). Ratio and product - type exponential estimator. *Journal of Information and Optimization Sciences*, *12*(1), 159-163. doi:10.1080/02522667.1991.10699058

Chodjuntug, K., & Lawson, N. (2022). Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. *Mathematical Population Studies*. doi:10.1080/08898480.2021.1997466

Cochran, W. G. (1940). The estimation of yield of cereal experiments by sampling for the ratio of gain to total produce. *Journal of Agricultural Science*, *30*(2), 262-275. doi:10.1017/S0021859600048012

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Singh, A. K., Singh, P., & Singh, V. (2014). Exponential-type compromised imputation in survey sampling. *Journal of Statistics Applications and Probability*, *3*(2), 211–217. doi:10.12785/jsap/030211.

Singh, S., & Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, *51*(3), 267-276. doi:10.1007/s001840000054

Singh, H. P., & Pal, S. K. (2015). A new chain ratio-ratio type exponential estimator using auxiliary information in sample surveys. *International Journal of Mathematics and its Applications*, *3*(4), 37-46.

Srivenkataramana, T. (1980). A dual to ratio estimator in sample surveys. *Biometrika*, *67*(1), 199-204. doi:10.2307/2335334