A Combined Transformed Variable for Population Mean Estimators When Missing Data Occur with an Application to COVID-19 Incidence

NATTHAPAT THONGSAK[1], NUANPAN LAWSON[2*]
[1]State Audit Office of the Kingdom of Thailand,
Bangkok, 10400,
THAILAND

[2]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800,
THAILAND

Abstract: - COVID-19 has killed many people and continues to be a major problem in all countries around the world. Estimating COVID-19 data in advance is helpful for the World Health Organization and governments in countries all over the globe to prepare the necessary resources. However, some of this information may be missing and needs to be dealt with before processing to estimation. The transformation method of an auxiliary variable can assist by increasing the performance of estimating the population mean. A combined transformed variable is suggested for estimating population mean when a study variable contains some missing values with uniform nonresponse, and it is applied in an application to data on COVID-19 incidence. The bias and mean square error of the suggested estimator are investigated and the performance is compared with existing estimators via a simulation study and an application to COVID-19 data. The results show that the suggested combined transformed estimators overtake existing estimators in terms of higher efficiency which yields the estimated value of total deaths of COVID-19 equal to 29497 cases.

Key-Words: - Combined transformed variable, missing data, COVID-19, uniformly nonresponse, population mean

Received: October 3, 2023. Accepted: November 7, 2023. Published: November 15, 2023.

1 Introduction

Human lives are harmed by the severe virus called COVID-19 or coronavirus pandemic which emerged in Wuhan, China at the end of 2019. After that, the world has been changed by the pandemic due to it killing a dramatic number of lives and afflicting human respiratory systems. Not only did it affect human being's lives in terms of health but it also affected the world's economy in various ways. An abundance of sectors were stopped. No investments can be made in the country or around the world, no traveling can happen so there are no businesses and tourists. Although vaccinations were invented to help stop the virus from taking human life and assist in reducing the number of deaths and active cases, there are still a significant amount of deaths and active cases nowadays. To help every country around the world including Thailand to prepare for unexpected situations that may arise due to the increasing numbers of deaths and active cases that could occur due to new mutations of COVID-19. Estimating COVID-19 incidence could benefit by dealing with this issue. Some of the data may be lost for example the variation of the patient population or struggles to collect clinical data during the collection process and therefore these missing values should be dealt with in suitable ways before processing to the policy planning process.

The imputation method is one of the techniques that are used to cope with missing data by replacing the possible values. There are numerous single imputation techniques including the mean imputation method, ratio imputation method, regression imputation method, and compromised imputation method, [1], [2], [3]. The study, [4], suggested three ratio estimators for estimating the population mean when the study variable contains missing values under simple random sampling without replacement (SRSWOR) assisting with the correlation coefficient between the study and auxiliary variables under uniform nonresponse. The results found that [4] estimators can improve the performance of the population mean estimator

through numerical studies. Likewise, [5] proposed an exponential method of imputation for estimating population mean under SRSWOR. The study [5] also found that the suggested estimators outperformed the existing ones.

The transformation technique can be helpful in sample surveys to ameliorate the efficiency of the population mean or population total estimator. A popular transformation technique was invented by, [6], who suggested transforming an auxiliary variable in the dual-to-ratio estimator which utilizes the benefit of the connection between the auxiliary variable and study variable to improve the efficiency of the estimator. In the study, [6], the transformed auxiliary variable under SRSWOR is defined by

$$x_i^* = \left(1+\pi_{\text{SRS}}\right)\bar{X} - \pi_{\text{SRS}} x_i \ ; i = 1,2,3,\ldots,N, \quad (1)$$

and the corresponding sample mean x_i^* is

$$\bar{x}_{\text{SRS}}^* = \left(1+\pi_{\text{SRS}}\right)\bar{X} - \pi_{\text{SRS}}\bar{x}, \quad (2)$$

where $\bar{X} = \sum_{i=1}^{N} x_i / N$ is the population mean of X and $\bar{x} = \sum_{i=1}^{n} x_i / \mathrm{n}$ is a sample mean of X, $\pi_{\text{SRS}} = n/N - n$ and n is a sample size drawn from a population of size N.

Research based on the transformation technique proposed by, [6], has been investigated by many researchers. For example, [7] proposed a general class of ratio estimators based on the transformed auxiliary variable assisting with some known parameters of the auxiliary variable under SRSWOR and the results found that [7] estimators gave better performances compared to the existing estimators. The authors in, [8], studied the bias and mean square error (MSE) of the ratio estimators that were invented by transformation of the ratio estimators. The authors in, [8], found that the transformed estimators gave better performances in terms of bias and MSE which could be reduced by a minimum of 70 percent with respect to the untransformed ones.

Nevertheless, some researchers applied the transformation method when missing data occurred in the variables. The authors in, [9], proposed two ratio estimators for estimating population mean under SRSWOR owing to the transformation of an auxiliary variable. Likewise, [10], suggested an exponential class of population mean estimators under SRSWOR utilizing the transformation of the auxiliary and study variables and the help of the known parameter of the auxiliary varaible to gain more efficiency for the population mean estimator in case of missing data. Both, [9], [10], gave a superior performance compared to the considered estimators when nonresponse is uniformly nonresponse.

In this study, a combined transformed variable, when there are some missing values on the study variable is investigated under SRSWOR and the uniform nonresponse mechanism. The properties of the proposed combined estimator are studied by simulation studies and an application to data on COVID-19 incidence.

2 Existing Estimators for Missing Data

Let (X , Y) be the pair of the auxiliary and study variables, r and be the number of responding units out of a sample (n units) that is obtained from a population (N units) under the SRSWOR scheme.

2.1 Mean Imputation Method

The point estimator for estimating population mean under the mean imputation method is

$$\hat{\bar{Y}}_{\text{S}} = \bar{y}_r, \quad (3)$$

where $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$ is a sample mean of the response variable of Y,

The bias and variance of $\hat{\bar{Y}}_{\text{S}}$ are

$$Bias\left(\hat{\bar{Y}}_{\text{S}}\right) = 0, \quad (4)$$

$$V\left(\hat{\bar{Y}}_{\text{S}}\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2, \quad (5)$$

where $C_y = S_y / \bar{Y}$, $S_y^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 / (\mathrm{N}-1)$.

2.2 Ratio Imputation Method

The point estimator for estimating the population mean under the ratio imputation method is

$$\hat{\bar{Y}}_{\text{Rat}} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}, \quad (6)$$

where $\bar{x}_n = \sum_{i=1}^{n} x_i / n$, and $\bar{x}_r = \sum_{i=1}^{r} x_i / r$.

The bias and MSE of $\hat{\bar{Y}}_{\text{Rat}}$ are

$$Bias\left(\hat{\bar{Y}}_{\text{Rat}}\right) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(C_x^2 - \rho C_x C_y\right), \quad (7)$$

$$MSE\left(\hat{\bar{Y}}_{\text{Rat}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right), \quad (8)$$

where $\bar{Y} = \sum_{i=1}^{N} y_i / N$, $C_x = S_x / \bar{X}$,

$S_x^2 = \sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2 / (\mathrm{N}-1)$, and $\rho = S_{xy} / (S_x S_y)$.

3 Proposed Estimator

Using the transformed auxiliary variable can improve the efficiency of the estimator. For that reason, a class of population mean estimator utilizing the transformed auxiliary variable in the case of missing values of the study variable is proposed. The proposed class estimator is

$$\hat{\bar{Y}}_{\mathrm{N}} = \alpha \bar{y}_r \left(\frac{A\bar{x}^* + D}{A\bar{X} + D}\right) + (1-\alpha)\left[\bar{y}_r + b\left(\bar{X} - \bar{x}^*\right)\right]\left(\frac{G\bar{x}^* + H}{G\bar{X} + H}\right), \quad (9)$$

where α is a selected constant which minimizes the MSE of the proposed estimator.

The following notations are used to investigate the bias and MSE of the proposed estimator.

$$\varepsilon_0 = \frac{\bar{y}_r - \bar{Y}}{\bar{Y}}, \bar{y}_r = (1+\varepsilon_0)\bar{Y}, \quad \varepsilon_1 = \frac{\bar{x}_r - \bar{X}}{\bar{X}}, \bar{x}_r = (1+\varepsilon_1)\bar{X},$$

$$\varepsilon_2 = \frac{\bar{x}_n - \bar{X}}{\bar{X}}, \bar{x}_n = (1+\varepsilon_2)\bar{X}, \; E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0,$$

$$E\left(\varepsilon_0^2\right) = \left(\frac{1}{r} - \frac{1}{N}\right)C_y^2, \; E\left(\varepsilon_1^2\right) = \left(\frac{1}{r} - \frac{1}{N}\right)C_x^2, \; E\left(\varepsilon_2^2\right) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2,$$

$$E(\varepsilon_0\varepsilon_1) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho C_x C_y, \; E(\varepsilon_0\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho C_x C_y, \; E(\varepsilon_1\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2.$$

Rewriting $\hat{\bar{Y}}_{\mathrm{N}}$ in terms of $e_i's, i = 0,1,2$, we have

$$\hat{\bar{Y}}_{\mathrm{N}} = \alpha(1+\varepsilon_0)\bar{Y}\left(\frac{A\bar{X} + D - \pi\varepsilon_2 A\bar{X}}{A\bar{X} + D}\right) + (1-\alpha)\left[(1+\varepsilon_0)\bar{Y} + \pi b\varepsilon_2\bar{X}\right]\left(\frac{G\bar{X} + H - \pi\varepsilon_2 G\bar{X}}{G\bar{X} + H}\right)$$

Let $\theta_1 = \dfrac{A\bar{X}}{A\bar{X} + D}$ and $\theta_2 = \dfrac{G\bar{X}}{G\bar{X} + H}$, then

$$\begin{aligned}\hat{\bar{Y}}_{\mathrm{N}} &= \alpha\bar{Y}(1+\varepsilon_0)(1-\pi\theta_1\varepsilon_2) + (1-\alpha)\bar{Y}(1+\varepsilon_0+\pi bK\varepsilon_2)\\ &\quad (1-\pi\theta_2\varepsilon_2)\\ &= \alpha\bar{Y}(1+\varepsilon_0-\pi\theta_1\varepsilon_2-\pi\theta_1\varepsilon_0\varepsilon_2) + (1-\alpha)\bar{Y}\\ &\quad (1+\varepsilon_0+\pi bK\varepsilon_2-\pi\theta_2\varepsilon_2-\pi\theta_2\varepsilon_0\varepsilon_2-\pi^2 bK\theta_2\varepsilon_2^2)\end{aligned}$$

The estimation error of $\hat{\bar{Y}}_{\mathrm{N}}$ is

$$\begin{aligned}\hat{\bar{Y}}_{\mathrm{N}} - \bar{Y} &= \alpha\bar{Y}(\varepsilon_0 - \pi\theta_1\varepsilon_2 - \pi\theta_1\varepsilon_0\varepsilon_2) + (1-\alpha)\bar{Y}\\ &\quad (\varepsilon_0 + \pi bK\varepsilon_2 - \pi\theta_2\varepsilon_2 - \pi\theta_2\varepsilon_0\varepsilon_2 - \pi^2 bK\theta_2\varepsilon_2^2)\end{aligned}$$

Then the bias of $\hat{\bar{Y}}_{\mathrm{N}}$ is

$$\begin{aligned}Bias\left(\hat{\bar{Y}}_{\mathrm{N}}\right) &\cong E\left(\alpha\bar{Y}(\varepsilon_0 - \pi\theta_1\varepsilon_2 - \pi\theta_1\varepsilon_0\varepsilon_2)\right.\\ &\quad \left.+(1-\alpha)\bar{Y}(\varepsilon_0 + \pi bK\varepsilon_2 - \pi\theta_2\varepsilon_2 - \pi\theta_2\varepsilon_0\varepsilon_2 - \pi^2 bK\theta_2\varepsilon_2^2)\right)\\ &= -\left(\frac{1}{n} - \frac{1}{N}\right)\pi\bar{Y}\left[(1-\alpha)\pi\beta K\theta_2 C_x^2 + \left(\alpha\theta_1 + (1-\alpha)\theta_2\right)\rho C_x C_y\right]. \quad (10)\end{aligned}$$

To find the MSE of the proposed estimator, consider

$$\begin{aligned}MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) &= E\left(\alpha\bar{Y}(\varepsilon_0 - \pi\theta_1\varepsilon_2 - \pi\theta_1\varepsilon_0\varepsilon_2)\right.\\ &\quad \left.+(1-\alpha)\bar{Y}(\varepsilon_0 + \pi bK\varepsilon_2 - \pi\theta_2\varepsilon_2 - \pi\theta_2\varepsilon_0\varepsilon_2 - \pi^2 bK\theta_2\varepsilon_2^2)\right)^2\end{aligned}$$

Under the assumption that terms of ε involving powers more than two are negligibly small,

$$\begin{aligned}MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) &\cong E\left[\alpha\bar{Y}(\varepsilon_0 - \pi\theta_1\varepsilon_2) + (1-\alpha)\bar{Y}(\varepsilon_0 + \pi bK\varepsilon_2 - \pi\theta_2\varepsilon_2)\right]^2\\ &= \bar{Y}^2 E\left[\varepsilon_0^2 + \left((1-\alpha)(\theta_2 - bK) + \alpha\theta_1\right)^2\pi^2\varepsilon_2^2\right.\\ &\quad \left.+2\left((1-\alpha)(\theta_2 - bK) + \alpha\theta_1\right)\pi\varepsilon_0\varepsilon_2\right]\\ &= \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\left(\left((1-\alpha)(\theta_2 - \beta K) + \alpha\theta_1\right)^2\pi^2 C_x^2\right.\\ &\quad \left.-2\left((1-\alpha)(\theta_2 - \beta K) + \alpha\theta_1\right)\pi\rho C_x C_y\right). \quad (11)\end{aligned}$$

Seeking the optimum value α to obtain the minimum MSE of the estimator, taking a partial derivative of MSE with respect to α and equating it to zero. The MSE of the proposed estimator $\hat{\bar{Y}}_{\mathrm{N}}$ is minimized for

$$\alpha^{\mathrm{opt}} = \frac{(\theta_2 - \beta K)\pi C_x - \rho C_y}{(\theta_2 - \beta K - \theta_1)\pi C_x}. \quad (12)$$

The minimum MSE of $\hat{\bar{Y}}_{\mathrm{N}}$ is

$$MSE_{\min}\left(\hat{\bar{Y}}_{\mathrm{N}}^{\mathrm{opt}}\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 - \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2\rho^2 C_y^2. \quad (13)$$

Some members of the proposed estimator are shown in Table 1.

Table 1. Some members of the proposed estimator

Estimator	A or G	D or H
$\hat{\bar{Y}}_{N1}=\alpha_{N1}^{opt}\bar{y}_r\left(\frac{\bar{x}^*}{\bar{X}}\right)+\left(1-\alpha_{N1}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\bar{x}^*}{\bar{X}}\right)$	1	0
$\hat{\bar{Y}}_{N2}=\alpha_{N2}^{opt}\bar{y}_r\left(\frac{\bar{x}^*+Q_3}{\bar{X}+Q_3}\right)+\left(1-\alpha_{N2}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\bar{x}^*+Q_3}{\bar{X}+Q_3}\right)$	1	Q_3
$\hat{\bar{Y}}_{N3}=\alpha_{N3}^{opt}\bar{y}_r\left(\frac{\bar{x}^*+Q_r}{\bar{X}+Q_r}\right)+\left(1-\alpha_{N3}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\bar{x}^*+Q_r}{\bar{X}+Q_r}\right)$	1	Q_r
$\hat{\bar{Y}}_{N4}=\alpha_{N4}^{opt}\bar{y}_r\left(\frac{\bar{x}^*+Q_d}{\bar{X}+Q_d}\right)+\left(1-\alpha_{N4}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\bar{x}^*+Q_d}{\bar{X}+Q_d}\right)$	1	Q_d
$\hat{\bar{Y}}_{N5}=\alpha_{N5}^{opt}\bar{y}_r\left(\frac{\bar{x}^*+Q_a}{\bar{X}+Q_a}\right)+\left(1-\alpha_{N5}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\bar{x}^*+Q_a}{\bar{X}+Q_a}\right)$	1	Q_a
$\hat{\bar{Y}}_{N6}=\alpha_{N6}^{opt}\bar{y}_r\left(\frac{\beta_1\bar{x}^*+\beta_2}{\beta_1\bar{X}+\beta_2}\right)+\left(1-\alpha_{N6}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\beta_1\bar{x}^*+\beta_2}{\beta_1\bar{X}+\beta_2}\right)$	β_1	β_2
$\hat{\bar{Y}}_{N7}=\alpha_{N7}^{opt}\bar{y}_r\left(\frac{\beta_2\bar{x}^*+\beta_1}{\beta_2\bar{X}+\beta_1}\right)+\left(1-\alpha_{N7}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\beta_2\bar{x}^*+\beta_1}{\beta_2\bar{X}+\beta_1}\right)$	β_2	β_1
$\hat{\bar{Y}}_{N8}=\alpha_{N8}^{opt}\bar{y}_r\left(\frac{C_x\bar{x}^*+Q_1}{C_x\bar{X}+Q_1}\right)+\left(1-\alpha_{N8}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{C_x\bar{x}^*+Q_1}{C_x\bar{X}+Q_1}\right)$	C_x	Q_1
$\hat{\bar{Y}}_{N9}=\alpha_{N9}^{opt}\bar{y}_r\left(\frac{\beta_2\bar{x}^*+Q_2}{\beta_2\bar{X}+Q_2}\right)+\left(1-\alpha_{N9}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\beta_2\bar{x}^*+Q_2}{\beta_2\bar{X}+Q_2}\right)$	β_2	Q_2
$\hat{\bar{Y}}_{N10}=\alpha_{N10}^{opt}\bar{y}_r\left(\frac{\rho\bar{x}^*+Q_3}{\rho\bar{X}+Q_3}\right)+\left(1-\alpha_{N10}^{opt}\right)\left[\bar{y}_r+b\left(\bar{X}-\bar{x}^*\right)\right]\left(\frac{\rho\bar{x}^*+Q_3}{\rho\bar{X}+Q_3}\right)$	ρ	Q_3

where Q_1 and Q_3 are the first and the third quartiles of the auxiliary variable, respectively, $Q_r=Q_3-Q_1$ is the interquartile range of the auxiliary variable, $Q_d=\left(Q_3-Q_1\right)/2$ is the semi-quartile range of the auxiliary variable, $Q_a=\left(Q_3+Q_1\right)/2$ is the quartile mean of the auxiliary variable, β_1 and β_2 is the coefficient of skewness and kurtosis of auxiliary variable, respectively.

4 Efficiency Comparison

The efficiency comparison of the proposed estimator and the existing estimators; mean imputation estimator ($\hat{\bar{Y}}_S$), ratio imputation estimator ($\hat{\bar{Y}}_{Rat}$), and, [9], [10], estimators ($\hat{\bar{Y}}_R$, $\hat{\bar{Y}}_{Reg}$) by using the MSEs as a criterion is shown.

1) $\hat{\bar{Y}}_N$ is more efficient than $\hat{\bar{Y}}_S$ if

$$MSE\left(\hat{\bar{Y}}_N\right)<MSE\left(\hat{\bar{Y}}_S\right)$$
$$\left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2\rho^2C_y^2>0$$
$$\rho^2>0$$

2) $\hat{\bar{Y}}_N$ is more efficient than $\hat{\bar{Y}}_{Rat}$ if

$$MSE\left(\hat{\bar{Y}}_N\right)<MSE\left(\hat{\bar{Y}}_{Rat}\right)$$
$$\frac{\frac{1}{n}-\frac{1}{N}}{\frac{1}{r}-\frac{1}{n}}>\frac{2\rho C_xC_y-C_x^2}{\rho^2C_y^2}$$
$$\frac{r(N-n)}{N(n-r)}>\frac{2\rho C_xC_y-C_x^2}{\rho^2C_y^2}$$

5 Simulation Studies

The efficiency of the proposed estimators with respect to the existing estimators is also supported by the simulation studies. The data are generated from a bivariate normal distribution with the following parameters; $N=5,000,\ \bar{X}=60,$ $\bar{Y}=200, C_x=1.1,\ C_y=2.0,$ and $\rho=0.6,\ 0.8$. Two levels of missing values; 5% and 20% in the study variable and the sampling fractions at $f=n/N=5\%,\ 10\%,$ and 30% are considered under SRSWOR. The simulation is repeated 10,000 times using the R program, [11].

The biases and MSEs of the proposed and existing estimators are represented in Table 2, Table 3, and Table 4, where

$$Bias\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left|\hat{\bar{Y}}_i-\bar{Y}\right|, \quad (14)$$

$$MSE\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left(\hat{\bar{Y}}_i-\bar{Y}\right)^2. \quad (15)$$

According to Table 2, Table 3, and Table 4, the proposed combined estimators performed superior to the existing estimators in terms of smaller biases and MSEs for all levels of correlation, percentage of missing data, and sampling fraction. Increasing the percentage of missing values gave higher biases and MSEs. On the other hand, increasing levels of sampling fractions and the correlation coefficient between X and Y lead to smaller biases and MSEs. All proposed combined estimators using different known parameters gave similar biases and MSEs in this scenario and performed a lot better than the existing estimators.

Table 2. Biases and MSEs of the estimators when $\rho = 0.6$ and percent of missing = 5%

Estimator	f =5%		f =10%		f =30%	
	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\bar{Y}}_{S}$	19.94	632.25	13.82	300.16	7.09	78.84
$\hat{\bar{Y}}_{Rat}$	19.79	622.10	13.73	297.02	7.06	78.26
$\hat{\bar{Y}}_{N1}$	16.23	416.53	11.19	196.56	5.76	52.24
$\hat{\bar{Y}}_{N2}$	16.23	416.66	11.19	196.64	5.76	52.30
$\hat{\bar{Y}}_{N3}$	16.23	416.65	11.19	196.64	5.76	52.30
$\hat{\bar{Y}}_{N4}$	16.23	416.62	11.19	196.62	5.76	52.28
$\hat{\bar{Y}}_{N5}$	16.23	416.63	11.19	196.63	5.76	52.29
$\hat{\bar{Y}}_{N6}$	16.23	416.84	11.20	196.72	5.76	52.30
$\hat{\bar{Y}}_{N7}$	16.23	416.53	11.19	196.56	5.76	52.24
$\hat{\bar{Y}}_{N8}$	16.23	416.57	11.19	196.59	5.76	52.26
$\hat{\bar{Y}}_{N9}$	16.23	416.58	11.19	196.60	5.76	52.27
$\hat{\bar{Y}}_{N10}$	16.23	416.68	11.19	196.65	5.76	52.30

Table 3. Biases and MSEs of the estimators when $\rho = 0.6$ and percent of missing = 20%

Estimator	f =5%		f =10%		f =30%	
	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\bar{Y}}_{S}$	21.81	751.40	15.16	359.17	7.67	92.09
$\hat{\bar{Y}}_{Rat}$	21.02	696.43	14.60	334.20	7.39	85.93
$\hat{\bar{Y}}_{N1}$	18.37	531.04	12.69	251.46	6.41	64.31
$\hat{\bar{Y}}_{N2}$	18.37	531.20	12.69	251.54	6.42	64.34
$\hat{\bar{Y}}_{N3}$	18.37	531.19	12.69	251.54	6.42	64.34
$\hat{\bar{Y}}_{N4}$	18.37	531.15	12.69	251.52	6.42	64.33
$\hat{\bar{Y}}_{N5}$	18.37	531.17	12.69	251.53	6.42	64.34
$\hat{\bar{Y}}_{N6}$	18.37	531.43	12.69	251.65	6.42	64.36
$\hat{\bar{Y}}_{N7}$	18.37	531.04	12.69	251.46	6.41	64.31
$\hat{\bar{Y}}_{N8}$	18.37	531.09	12.69	251.49	6.41	64.32
$\hat{\bar{Y}}_{N9}$	18.37	531.10	12.69	251.49	6.41	64.32
$\hat{\bar{Y}}_{N10}$	18.37	531.22	12.69	251.55	6.42	64.35

6 Application to COVID-19 Data

The COVID-19 dataset from, [12], is used to illustrate the execution of the proposed estimators in practice. The total deaths and total cases that are collected from a population of size $N = 231$ are assigned as the study and auxiliary variables, respectively. Among 231 countries, 2% of the data for the study variable is missing. The population characteristics are summarized as follows:

$$N = 231,\ \bar{X} = 2{,}998{,}167,\ \bar{Y} = 30{,}548.57,$$

$$C_x = 3.25,\ C_y = 3.52,\ \rho = 0.88$$

Then, a sample of size n =70 countries is randomly selected from the population of size $N = 231$ using SRSWOR. The PREs of estimators with respect to mean imputation estimator are calculated by

$$PRE\left(\hat{\bar{Y}}, \hat{\bar{Y}}_{S}\right) = \frac{V\left(\hat{\bar{Y}}_{S}\right)}{MSE\left(\hat{\bar{Y}}\right)} \times 100. \qquad (16)$$

Table 4. Biases and MSEs of the estimators when $\rho = 0.8$ and percent of missing = 20%

Estimator	f =5%		f =10%		f =30%	
	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\bar{Y}}_{S}$	21.81	751.41	15.16	359.17	7.67	92.09
$\hat{\bar{Y}}_{Rat}$	20.47	661.18	14.22	317.10	7.19	81.44
$\hat{\bar{Y}}_{N1}$	15.28	366.41	10.50	172.50	5.32	44.02
$\hat{\bar{Y}}_{N2}$	15.28	366.51	10.51	172.54	5.32	44.04
$\hat{\bar{Y}}_{N3}$	15.28	366.50	10.51	172.54	5.32	44.04
$\hat{\bar{Y}}_{N4}$	15.28	366.48	10.50	172.53	5.32	44.04
$\hat{\bar{Y}}_{N5}$	15.28	366.49	10.51	172.53	5.32	44.04
$\hat{\bar{Y}}_{N6}$	15.28	366.65	10.51	172.60	5.32	44.04
$\hat{\bar{Y}}_{N7}$	15.28	366.41	10.50	172.50	5.32	44.02
$\hat{\bar{Y}}_{N8}$	15.28	366.44	10.50	172.51	5.32	44.03
$\hat{\bar{Y}}_{N9}$	15.28	366.45	10.50	172.51	5.32	44.03
$\hat{\bar{Y}}_{N10}$	15.28	366.52	10.51	172.55	5.32	44.04

Table 5. Estimated deaths and PREs of the estimators with respect to the mean imputation estimator when applied to COVID-19 data

Estimator	Estimated deaths	PRE
$\hat{\bar{Y}}_{S}$	20089.45	100.00
$\hat{\bar{Y}}_{Rat}$	19802.79	94.74
$\hat{\bar{Y}}_{N1}$	28858.56	3830.13
$\hat{\bar{Y}}_{N2}$	29457.11	9182.85
$\hat{\bar{Y}}_{N3}$	29450.30	9069.36
$\hat{\bar{Y}}_{N4}$	29227.26	6265.92
$\hat{\bar{Y}}_{N5}$	29237.97	6368.72
$\hat{\bar{Y}}_{N6}$	28858.57	3830.16
$\hat{\bar{Y}}_{N7}$	28858.56	3830.13
$\hat{\bar{Y}}_{N8}$	28864.29	3856.23
$\hat{\bar{Y}}_{N9}$	28861.03	3841.36
$\hat{\bar{Y}}_{N10}$	29497.32	9898.82

The results in Table 5 showed that the performance of the proposed class of estimators was more outstanding than the mean imputation and ratio imputation when applied to the COVID-19 dataset which also supports the results found in the simulation studies. The proposed combined estimator $\hat{\bar{Y}}_{N10}$ using the benefit of the known Q_3 and ρ gave the highest PREs which yields the estimated values of total deaths equal to 29497 cases.

7 Conclusion

The transformation technique assists in increasing the efficiency of the population mean estimator when missing data occur in the study variable through the proposed class of combined estimators. This technique is suggested for application in the presence of missing data under the uniform nonresponse mechanism in the study variable in this study. The results showed that the proposed transformed estimators gave smaller biases and MSEs through simulation results and an application to COVID-19 data which are recommended to be applied using the available Q_3 and ρ to receive the highest PREs and gave closer estimated values to the population parameter. Due to simplicity, this study investigated under the uniform nonresponse mechanism, and therefore in future work, the proposed estimators can be extended to missing at random or non-ignorable missing at random and also in more complex survey designs e.g. double sampling, stratified random sampling, cluster sampling. Available parameters based on the auxiliary variable can also assist in improving the efficiency of the suggested estimators. Nonetheless, the combined transformed estimators can be applied to all real-world problems in the presence of missing data.

Acknowledgement:
We appreciate all comments from the referees to help in improving the paper.

References:

[1] Singh, S., and Horn, S., Compromised imputation in survey sampling, *Metrika*, Vol.51, No. 3, 2000, pp. 267–276.

[2] Singh, S. and Deo, B., Imputation by power transformation, *Statistical. Papers,* Vol. 44, 2003, pp. 555–579.

[3] Norazian, M. N., Shukri, Y. A., Azam, R. N., and Al Bakri, A.M.M., Estimation of missing values in air pollution data using single imputation techniques, *Science Asia*, Vol.34, No. 3, 2008, pp. 341–345.

[4] Al-Omari, A.I., Bouza, C.N. and Herrera, C., Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient, *Qual Quant,* Vol.47, 2013, pp.353- 365.

[5] Singh, A. K., Singh, P., and Singh, V., Exponential-type compromised imputation in survey sampling, *Journal of Statistics Applications & Probability,* Vol.3, No.2, 2014, pp.211- 217.

[6] Srivenkataramana, T., A dual to ratio estimator in sample surveys, *Biometrika*, Vol. 67, No. 1, 1980, pp.199-204.

[7] Onyeka, A.C., Nlebedim, V.U. and Izunobi, C.H., A Class of estimators for population ratio in simple random sampling using variable transformation, *Open Journal of Statistics,* Vol.4, 2014, pp.284-291.

[8] Thongsak, N. and Lawson, N., Bias and mean square error reduction by changing the shape of the distribution of an auxiliary variable: application to air pollution data in Nan, Thailand, *Mathematical Population Studies*, Vol. 30, No. 3, 2023, pp.180-194.

[9] Khare, B.B. and Srivastava, S, Transformed ratio type estimators for the population mean in the presence of nonresponse, *Communications in Statistics-Theory and Methods,* Vol. 26, No. 7, 1997, pp.1779-1791.

[10] Sharma, V. and Kumar, S., Estimation of population mean using transformed auxiliary variable and non-response, *Revista Investigacion Operacional,* Vol. 41, No. 3, 2020, pp.438-444.

[11] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021, [Online], https://www.R-project.org/ (Accessed Date: November 5, 2023)

[12] Worldometer, COVID-19 Coronavirus pandemic, (2023), [Online], https://www.worldometers.info/coronavirus/ (Accessed Date: November 5, 2023)

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

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-67-B-43.

Conflict of Interest

The author has no conflicts of interest to declare.