Transformed Regression Type Estimators in the Presence of Missing Observations: Case Studies on COVID-19 Incidence in Chiang Mai, Thailand

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: - The transformation technique can be used to modify the shape of the variable to improve the performance of the population mean estimator. In the presence of missing data, before estimating the population mean using standard statistical methods, missing data has to be taken care of. In this study, we focus on new transformed regression type estimators when missing data are present in the study variable under the uniform nonresponse mechanism and assume that the population mean of the auxiliary variable is unavailable which usually occurs in practice. An auxiliary variable can assist by increasing the efficacy of estimating the population mean. The bias and mean square error are investigated up to the first order degree approximation using the Taylor series. A simulation and case studies on COVID-19 incidence in Chiang Mai, Thailand are used to assess the performance of the new transformed estimators. The estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen and the estimated daily confirmed cases of COVID-19 in Chiang Mai from the best proposed estimator are around 17 cases and 118 cases, respectively.

Key-Words: - Transformed variable, missing data, COVID-19, uniformly nonresponse, fine particulate matter, mean imputation, ratio imputation, bias, mean square error

Received: November 2, 2023. Revised: December 13, 2023. Accepted: January 22, 2024. Published: March 21, 2024.

1 Introduction

Changing the shape of the variables can be done by using the transformation method. The well-known transformed variable was suggested by [1], who introduced changing an auxiliary variable for estimating the population mean using the dual-to-ratio estimator under simple random sampling without replacement (SRSWOR). The transformed auxiliary variable in [1], is:

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

and the corresponding sample mean is

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

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

There are a plethora of works related the transformed variables. For example, a linear transformation of the study variable was suggested to improve the ratio estimator, [2]. Based on the work of [1], the author [3], suggested using the transformation technique in [1] to improve the dual to ratio estimators under SRSWOR assuming that some known auxiliary parameters are available, [4], [5].

Missing data is a prominent issue occurring in sample surveys. Ignoring the missing data may lead to bias and high variance. Imputation methods have assisted in dealing with missing data, [6], [7], [8], [9]. The mean imputation method is applied by replacing the missing data with a sample mean of the available information. The point estimator of the mean imputation technique is:

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

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

The bias of $\hat{\bar{Y}}_{\mathrm{S}}$ is

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

The variance of $\hat{\bar{Y}}_{\mathrm{S}}$ is

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

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

The ratio imputation is also another popular imputation method to use when there is a connection between an auxiliary and a study variable. The point estimator for the ratio imputation method is:

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

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

The bias of $\hat{\bar{Y}}_{\mathrm{Rat}}$ is:

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

and MSE of $\hat{\bar{Y}}_{\mathrm{Rat}}$ is

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

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 $\rho = S_{xy} / (S_x S_y).$

The COVID-19 pandemic has afflicted the entire world devastatingly in an abundance of areas from way of life and the economy to global healthcare. In terms of health, the impact of the virus on humans goes much further than many anticipate, as it not only presents with acute respiratory infection symptoms but may continue to linger on in the body in the form of organ damage. Severe infection can permanently alter the immune system. Efforts to produce vaccines and reduce incidence through isolation measures to mitigate the repercussions of the pandemic have been successful. However, the spread of the virus is still ongoing for years and can affect the more susceptible population severely so researchers are delving deeper into the issue and the etiologies to stop the virus. There are innumerable risk factors for severe hospitalization from COVID-19, one interesting aspect is the influence of pollution. Thailand is one of the countries with pollution from fine particulate matter as a prevailing problem for many years. It affects areas all over the country and is mostly caused by burning agricultural waste every year. Research has found a correlation between exposure to fine particulate matter and COVID-19 hospitalization and incidence rates. The pollution can permanently damage the body's immune system, increasing the risk of viral infection and severe symptoms. Fine particulate matter exceeding safe levels in Thailand, especially Chiang Mai, is an obstinate concern that increases the risk for a myriad of non-communicable diseases that are the leading causes of death. Examining these levels is critical for the prevention of further consequences of pollution through national policies and measures, albeit data regarding fine particulate matter are often missing.

Many researchers investigated the connection between COVID-19 and air pollution data. The studies indicate that there is a positive correlation between COVID-19 and air pollution data such as PM2.5, [10], [11], [12], [13], [14]. Missing data occur in many real data including COVID-19 and fine particulate matter and as a result the proper statistical techniques should be applied to deal with these data.

In this study, a class of regression type estimators utilizing the transformation of an auxiliary variable has been proposed using simple random sampling (SRSWOR). The uniform nonresponse mechanism is considered in this study and it is assumed the population mean of the auxiliary variable is unknown. This study investigates the bias and mean square error of the proposed estimators. A simulation study and applications to COVID-19 incidence in Chiang Mai, Thailand are studied using the proposed transformed estimators.

2 Proposed Estimator

Inspired by [3], assuming that the study variable is missing under the uniform nonresponse mechanism and the population mean of an auxiliary variable is unknown, a class of regression type estimators for estimating population mean utilizing the transformation of an auxiliary variable is proposed as below.

$$\hat{\hat{Y}}_{\mathrm{N}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{A\bar{x}^* + D}{A\bar{x}_n + D}\right), \quad (7)$$

where

$$\bar{x}^* = \frac{n\bar{x}_n - r\bar{x}_r}{n-r} = \left(1+\pi\right)\bar{x}_n - \pi\bar{x}_r,\ \pi = \frac{r}{n-r},$$

$b = \frac{s_{xy}}{s_x^2}$ is the sample regression coefficient, $\left(A \neq 0, D\right)$ are real numbers or functions of the auxiliary variable.

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

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

$$\bar{x}_r = \left(1+\varepsilon_1\right)\bar{X}, \varepsilon_2 = \frac{\bar{x}_n - \bar{X}}{\bar{X}}, \bar{x}_n = \left(1+\varepsilon_2\right)\bar{X},$$

$$E\left(\varepsilon_0\right) = E\left(\varepsilon_1\right) = E\left(\varepsilon_2\right) = 0,$$

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

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

$$E\left(\varepsilon_1\varepsilon_2\right) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2,\ K = \frac{\bar{X}}{\bar{Y}}.$$

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

$$\hat{\hat{Y}}_{\mathrm{N}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{A\bar{x}^* + D}{A\bar{x}_n + D}\right)$$

$$= \left[\left(1+\varepsilon_0\right)\bar{Y} + b\left(\varepsilon_1 - \varepsilon_2\right)\pi\bar{X}\right] \times \left(\frac{\left(A\bar{X} + D\right) + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)A\bar{X}}{\left(A\bar{X} + D\right) + \varepsilon_2 A\bar{X}}\right).$$

Let $\theta = \frac{A\bar{X}}{A\bar{X} + D}$, will get

$$\hat{\hat{Y}}_{\mathrm{N}} = \left[\left(1+\varepsilon_0\right)\bar{Y} + b\left(\varepsilon_1 - \varepsilon_2\right)\pi\bar{X}\right] \times \left(\frac{\frac{A\bar{X}}{\theta} + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)A\bar{X}}{\frac{A\bar{X}}{\theta} + \varepsilon_2 A\bar{X}}\right)$$

$$= \bar{Y}\left[1 + \varepsilon_0 + \pi bK\varepsilon_1 - \pi bK\varepsilon_2\right] \left(1 + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)\theta\right)\left(1 + \varepsilon_2\theta\right)^{-1}.$$

Using the Taylor series approximation, we get:

$$\hat{\hat{Y}}_{\mathrm{N}} = \bar{Y}\left[1 + \varepsilon_0 + \pi bK\varepsilon_1 - \pi bK\varepsilon_2\right] \left(1 + \left(\varepsilon_2 + \pi\varepsilon_2 - \pi\varepsilon_1\right)\theta\right)\left(1 - \varepsilon_2\theta + \varepsilon_2^2\theta^2 - \ldots\right)$$

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

$$Bias\left(\hat{\hat{Y}}_{\mathrm{N}}\right) = E(\hat{\hat{Y}}_{\mathrm{N}} - \bar{Y})$$

$$= \bar{Y}E\left(\begin{pmatrix} \varepsilon_0 + \left(-\theta_2 + bK\right)\pi\varepsilon_1 + \left(\theta_2 - bK\right)\pi\varepsilon_2 + \left(-bK\theta_2\right)\pi^2\varepsilon_1^2 \\ + \left(-\pi\theta_2^2 - \pi^2 bK\theta_2\right)\varepsilon_2^2 + \left(-\pi\theta_2\right)\varepsilon_0\varepsilon_1 + \left(\pi\theta_2\right)\varepsilon_0\varepsilon_2 \\ + \left(\pi\theta_2^2 + 2\pi^2 bK\theta_2\right)\varepsilon_1\varepsilon_2 \end{pmatrix}\right)$$

$$= -\pi\theta\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(\pi\beta KC_x^2 + \rho C_x C_y\right) \quad (8)$$

Under the assumption that the terms of ε involving the powers more than two are negligibly small, the mean square error of $\hat{\hat{Y}}_{\mathrm{N}}$ is:

$$MSE\left(\hat{\hat{Y}}_{\mathrm{N}}\right) \cong E\left[\bar{Y}\left(\varepsilon_0 + \left(-\theta + bK\right)\pi\varepsilon_1 + \left(\theta - bK\right)\pi\varepsilon_2\right)\right]^2$$

$$= \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2 \left(\left(\theta - \beta K\right)^2\pi^2 C_x^2 - 2\left(\theta - \beta K\right)\pi\rho C_x C_y\right). \quad (9)$$

Some proposed estimators are shown in Table 1.

Table 1. Some members of the proposed estimator

Estimator	A	D
$\hat{\hat{Y}}_{\mathrm{N1}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\bar{x}^*}{\bar{x}_n}\right)$	1	0
$\hat{\hat{Y}}_{\mathrm{N2}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\bar{x}^* + \beta_2}{\bar{x}_n + \beta_2}\right)$	1	β_2
$\hat{\hat{Y}}_{\mathrm{N3}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\bar{x}^* + C_x}{\bar{x}_n + C_x}\right)$	1	C_x
$\hat{\hat{Y}}_{\mathrm{N4}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{C_x\bar{x}^* + \beta_1}{C_x\bar{x}_n + \beta_1}\right)$	C_x	β_1
$\hat{\hat{Y}}_{\mathrm{N5}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{Q_1\bar{x}^* + Q_a}{Q_1\bar{x}_n + Q_a}\right)$	Q_1	Q_a
$\hat{\hat{Y}}_{\mathrm{N6}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{\beta_2\bar{x}^* + Q_2}{\beta_2\bar{x}_n + Q_2}\right)$	β_2	Q_2
$\hat{\hat{Y}}_{\mathrm{N7}} = \left[\bar{y}_r + b\left(\bar{x}_n - \bar{x}^*\right)\right]\left(\frac{Q_d\bar{x}^* + C_x}{Q_d\bar{x}_n + C_x}\right)$	Q_d	C_x

where β_1 and β_2 are the coefficient of skewness and kurtosis of the auxiliary variable, respectively, Q_1 , Q_2, and Q_3 are the first, second, and third quartiles of the auxiliary variable, respectively, $Q_a = (Q_3 + Q_1)/2$ and $Q_d = (Q_3 - Q_1)/2$ is the quartile mean and quartile deviation of the auxiliary variable, respectively.

3 Efficiency Comparison

The proposed estimator's ($\hat{\bar{Y}}_{\mathrm{N}}$) efficiency is compared with the existing estimators; mean imputation estimator ($\hat{\bar{Y}}_{\mathrm{S}}$), and ratio imputation estimator ($\hat{\bar{Y}}_{\mathrm{Rat}}$) based on the MSE.

1) $\hat{\bar{Y}}_{\mathrm{N}}$ performs better than $\hat{\bar{Y}}_{\mathrm{S}}$ if:

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) < MSE\left(\hat{\bar{Y}}_{\mathrm{S}}\right)$$

$$\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left((\theta-\beta K)^2\pi^2C_x^2-2(\theta-\beta K)\pi\rho C_xC_y\right)$$
$$<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2$$

$$(\theta-\beta K)^2\pi^2C_x^2-2(\theta-\beta K)\pi\rho C_xC_y<0$$

$$\rho>\frac{\pi(\theta-\beta K)C_x}{2C_y}. \tag{10}$$

2) $\hat{\bar{Y}}_{\mathrm{N}}$ performs better than $\hat{\bar{Y}}_{\mathrm{Rat}}$ if

$$MSE\left(\hat{\bar{Y}}_{\mathrm{N}}\right) < MSE\left(\hat{\bar{Y}}_{\mathrm{Rat}}\right)$$

$$\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left((\theta-\beta K)^2\pi^2C_x^2-2(\theta-\beta K)\pi\rho C_xC_y\right)$$

$$<\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2C_y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left(C_x^2-2\rho C_xC_y\right)$$

$$\pi^2(\theta-\beta K)^2C_x^2-2\pi(\theta-\beta K)\rho C_xC_y<C_x^2-2\rho C_xC_y$$

$$\rho<\frac{\left(1+\pi(\theta-\beta K)\right)C_x}{2C_y}. \tag{11}$$

4 A Simulation Study

To assess the performance of the proposed estimators, the data was generated from bivariate normal distribution with the following parameters; $N=4,000,\ \bar{X}=150,\ \bar{Y}=280, C_x=2.2,\ C_y=1.1,\ \rho=0.8$

Then 40% of the study variable values were randomly assigned as missing and we randomly selected a sample of 25% units from a population of size $N=4,000$ using the SRSWOR scheme.

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

$$Bias\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left|\hat{\bar{Y}}_i-\bar{Y}\right|, \tag{12}$$

$$MSE\left(\hat{\bar{Y}}\right)=\frac{1}{10,000}\sum_{i=1}^{10,000}\left(\hat{\bar{Y}}_i-\bar{Y}\right)^2. \tag{13}$$

Table 2. Biases and MSEs of the estimators

Estimator	Bias	MSE
Mean imputation ($\hat{\bar{Y}}_{\mathrm{S}}$)	10.82	179.34
Ratio imputation ($\hat{\bar{Y}}_{\mathrm{Rat}}$)	9.66	147.54
Proposed ($\hat{\bar{Y}}_{\mathrm{N1}}$)	8.34	110.57
Proposed ($\hat{\bar{Y}}_{\mathrm{N2}}$)	**8.27**	**108.59**
Proposed ($\hat{\bar{Y}}_{\mathrm{N3}}$)	8.28	108.98
Proposed ($\hat{\bar{Y}}_{\mathrm{N4}}$)	8.34	110.58
Proposed ($\hat{\bar{Y}}_{\mathrm{N5}}$)	8.39	112.20
Proposed ($\hat{\bar{Y}}_{\mathrm{N6}}$)	8.49	111.87
Proposed ($\hat{\bar{Y}}_{\mathrm{N7}}$)	8.34	110.56

According to Table 2, the proposed estimators performed superior to the existing estimators. Both the bias and MSE of the proposed estimators are smaller than with the mean and ratio imputation methods. The best estimator is $\hat{\bar{Y}}_{\mathrm{N2}}$ that utilized the coefficient of kurtosis of the auxiliary variable which gave the smallest bias and MSE.

5 Applications to COVID-19 Incidence

In this section, the COVID-19 dataset collected from Chiang Mai province, [15] and daily PM2.5 concentration, [16], between 1 April 2022 and 31 July 2022 (population size $N=122$) were applied to illustrate the efficiency of the proposed estimators.

Population I: we assigned the number of COVID-19 patients who have pneumonia and require high-flow oxygen as the study variable, and the PM2.5 concentration (micrograms per cubic meters) as the auxiliary variable. The population parameters are described as follows:

$N = 122,\ \bar{X} = 18.16,\ \bar{Y} = 17.11,\ C_x = 0.85,\ C_y = 0.60,\ \rho = 0.62$

Population II: we assigned the daily confirmed cases of COVID-19 as the study variable, and the PM2.5 concentration (micrograms per cubic meter) as the auxiliary variable. The population parameters are described as follows:

$N = 122,\ \bar{X} = 18.16,\ \bar{Y} = 117.88,\ C_x = 0.85,\ C_y = 1.19,\ \rho = 0.79$

A sample of size $n = 36$ is acquired from the population of size $N = 122$ using SRSWOR with 30% and 25% missing in the study variable for population I and II, respectively.

Figure 1 shows the scatter plot between PM2.5 concentration and the number of COVID-19 patients who had pneumonia and required high-flow oxygen. Figure 2 shows the scatter plot between PM2.5 concentration and the daily confirmed cases of COVID-19. The estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen, estimated daily confirmed cases of COVID-19, and the PREs of the estimators are calculated using the R program, [17], which are presented in Table 3.

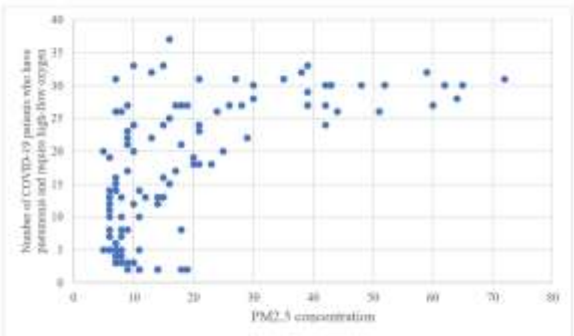

Fig. 1: The scatter plot between PM2.5 concentration and the number of COVID-19 patients who had pneumonia and required high-flow oxygen

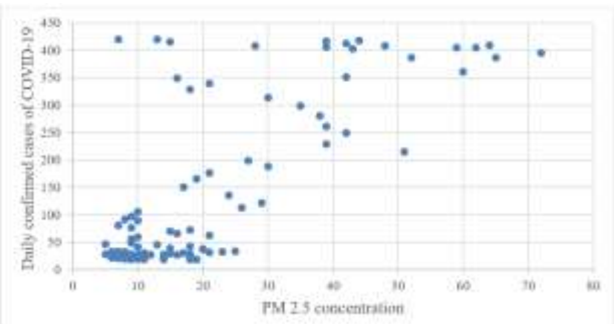

Fig. 2: The scatter plot between PM2.5 concentration and the daily confirmed cases of COVID-19

Table 3. Estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen, estimated daily confirmed cases of COVID-19, and PREs of the estimators

Population I		
Estimator	Estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen	PRE
Mean imputation ($\hat{\bar{Y}}_S$)	15.17	100.00
Ratio imputation ($\hat{\bar{Y}}_{Rat}$)	18.89	120.37
Proposed ($\hat{\bar{Y}}_{N1}$)	17.44	3662.29
Proposed ($\hat{\bar{Y}}_{N2}$)	16.43	804.96
Proposed ($\hat{\bar{Y}}_{N3}$)	**17.23**	**29480.82**
Proposed ($\hat{\bar{Y}}_{N4}$)	16.99	23578.54
Proposed ($\hat{\bar{Y}}_{N5}$)	16.97	19108.33
Proposed ($\hat{\bar{Y}}_{N6}$)	16.93	11003.23
Proposed ($\hat{\bar{Y}}_{N7}$)	17.41	4481.27
Population II		
Estimator	Estimated daily confirmed case of COVID-19	PRE
Mean imputation ($\hat{\bar{Y}}_S$)	142.52	100.00
Ratio imputation ($\hat{\bar{Y}}_{Rat}$)	155.50	42.89
Proposed ($\hat{\bar{Y}}_{N1}$)	125.30	1103.22
Proposed ($\hat{\bar{Y}}_{N2}$)	**118.93**	**54344.66**
Proposed ($\hat{\bar{Y}}_{N3}$)	123.99	1624.85
Proposed ($\hat{\bar{Y}}_{N4}$)	122.48	2867.37
Proposed ($\hat{\bar{Y}}_{N5}$)	122.39	2981.08
Proposed ($\hat{\bar{Y}}_{N6}$)	122.11	3392.37
Proposed ($\hat{\bar{Y}}_{N7}$)	125.10	1163.02

Figure 1 and Figure 2 indicate that both the number of COVID-19 patients who have pneumonia and require high-flow oxygen and daily confirmed cases of COVID-19 have a positive relation with PM2.5 concentration and the correlation coefficient is 0.62 and 0.79, respectively.

Table 3 indicates the proposed estimators gave better results in terms of PREs in comparison to the mean imputation method. The proposed estimators produced both the estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen and the estimated daily confirmed cases of COVID-19 closer to the population mean than other estimators and the PREs of the proposed estimators are higher than the mean and ratio imputation estimators, especially $\hat{\hat{Y}}_{N3}$ that utilized the coefficient of variation and $\hat{\hat{Y}}_{N2}$ that utilized the coefficient of kurtosis of the auxiliary variable to increase the precision of the estimator for population mean for population I and II, respectively. The estimated number of COVID-19 patients who have pneumonia and require high-flow oxygen and the estimated daily confirmed cases of COVID-19 in Chiang Mai from the best proposed estimator are around 17 cases and 118 cases, respectively.

6 Conclusion

Transformed estimators have been introduced in the presence of missing data with SRSWOR to improve the performance of the population mean estimator. Employing the transformation method can support altering the form of the variable which results in increasing the performance of the population mean estimator by assuming the auxiliary variable's population mean is not known which usually occurs in practice. As a result, it is going to be helpful in practice. The bias and mean square error of the transformed estimators are investigated. The results illustrated the newly transformed estimators gave the least bias and mean square error compared to others and gave closer estimated values of COVID-19 incidence to the population values. Especially the ones using the coefficient of variation and the coefficient of kurtosis of the auxiliary variable gave a high improvement in terms of highest PREs concerning other estimators. For future work, the suggested estimators may be applicable to assist with other survey designs e.g. stratified random sampling, double sampling, and cluster sampling, and in more flexible nonresponse mechanisms. Moreover, the estimators can be extended to cover the case that the missing data appears in the auxiliary variable or both study and auxiliary variables. The proposed estimators are very useful in practice in estimating the variable of interest in real data when nonresponse occurs in the study.

Acknowledgement:
We are thankful for all the helpful comments from the unknown referees to improve the paper.

References:

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

[2] Singh, A., Mourya, K.K. and Sisodia, B.V.S., Comparison of some ratio estimators using linear transformation, *International of Current Microbiology and Applied Sciences,* Special issue – 9, 2019, pp. 57-67.

[3] Thongsak, N. and Lawson, N., Classes of dual to modified ratio estimators for estimating population mean in simple random sampling, *Proceedings of the 2021 Research, Invention and Innovation Congress, Bangkok, Thailand*, September, 1-2, 2021, pp. 211-215.

[4] Upadhyaya, L.N. and Singh, H.P., Use of transformed auxiliary variable in estimating the finite population mean, *Biometrical Journal*, Vo. 41, 1999, pp. 627–636.

[5] 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.

[6] Nangsue, N., Adjusted ratio and regression type estimators for estimation of population mean when some observations are missing, *International Scholarly and Scientific Research & Innovation*, Vo.3, No.5, 2009, pp. 334-337.

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

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

[9] 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.

[10] Jiang, Y., Wu, X.-J., Guan, Y.-J, Effect of ambient air pollutants and meteorological variables on COVID-19 incidence, *Infection Control Hospital Epidemiology*, Vol. 41, No. 9, 2020, pp.1011–1015.

[11] Chan, S., Chu, J.Zhang, Y, Nadarajah, S, Count regression models for COVID-19, *Physica A*, Vol. 563, 2019, 125460.

[12] Martelletti, L., Martelletti, P, Air pollution and the novel covid-19 disease: a putative disease risk factor. *Sn*

comprehensive Clinical Medicine, Vol. 2, No. 4, 2020, pp. 383–387.

[13] Beloconi, A. and Vounatsou, P, Long-term air pollution exposure and COVID-19 case-severity: An analysis of individual-level data from Switzerland. *Environmental Research*, Vol. 216, 2023, pp. 114481.

[14] Austin, W., Carattini, S, Gomez-Mahecha, J. and Pesko, M.F., The effects of contemporaneous air pollution on COVID-19 morbidity and mortality. *Journal of Environmental Economics and Management*, Vol. 119, 2023, pp. 102815.

[15] Chiang Mai, Covid-19 situation in Chiang Mai Province, (2023), [Online]. https://www.chiangmai.go.th/covid19// (Accessed Date: October 20, 2023).

[16] Pollution Control Department, Daily PM2.5 concentration, (2023), [Online]. http://air4thai.pcd.go.th/webV3/#/History (Accessed Date: October 20, 2023).

[17] 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).

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

The authors equally contributed to 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.