ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 4, 161-169 (2021)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Estimation of the Population Mean for Incomplete Data by using Information of Simple Linear Relationship Model in Data Set

Juthaphorn Sinsomboonthong[*,1], Saichon Sinsomboonthong[2]

[1]*Department of Statistics, Faculty of Science, Kasetsart University (KU), Bangkok 10900, Thailand*

[2]*Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok 10520, Thailand*

ARTICLE INFO

Article history:
Received: 03 June, 2021
Accepted: 13 July, 2021
Online: 20 July, 2021

Keywords:
Bias
Estimator
Mean Square Error
Missing
Population Mean
Simple Linear Relationship

ABSTRACT

The objective of this research is to propose the estimator of the population mean for incomplete data by using information of simple linear relationship model in the data set. In addition, the factorization of the likelihood function is created to derive the maximum likelihood estimator for the population mean. The simulation study was conducted for 630 situations to compare the efficiency of the proposed estimator with the two population mean estimators, namely pairwise deletion and Anderson estimators. In this study, two criteria—bias and mean square error—of the performances for estimators are examined. It is found that all percentage levels of missing data, the mean square error of the proposed estimator tends to be lower than those of pairwise deletion and Anderson estimators for the large correlation levels between two variables in the data set whatever the sample sizes will be, especially for the large percentage level of missing data. However, for the small correlation between two variables in the data set, the three estimators tend to have the same performances in terms of both two criteria for all sample sizes and all percentage levels of missing data.

1. Introduction

Missing data are frequently found in many fields of research [1,2]. For example, some individuals may refuse to express any attitude for some sensitive questions in an opinion survey. In an experimental research, the experimental units may be leave or die before the experiment is completed. In longitudinal study, the monotone missing data pattern usually occurs. These missing data problems lead to increase an inaccuracy of the inference about the parameters in the population if the researchers ignore about the missing value in the data set. In estimation of the population mean for incomplete data set, imputation technique [3,4] is one of the familiar methods that researchers used it to replace the missing values with substituted values before estimate the population mean by using standard methods. However, the variance of estimator for this technique is underestimated and lead to the wrong inference about the population mean [5–7]. Available cases analysis is another technique that sample mean is used for estimation about the population mean and sometimes this is called pairwise deletion method. Moreover, this method will not suitable for the large amount of missing values because it will give the biased estimator and its standard error will increase [5, 8]. Ignoring missing values from the data set for inferential statistical analysis will affect the reliability of the conclusion about parameter in the population as the studied of [9–13]. Therefore, there are several researchers proposed about the estimators of the population mean for incomplete data set by considering only available cases analysis as follows: the maximum likelihood estimators of parameters for a bivariate normal distribution and case of some observations are missing for one variable were studied by [14]. That is, the factorization of likelihood function approach that proposed by [14] has been mostly used to derive the estimators of parameters for incomplete data set such as the studied of [15] and the research of [16]. Furthermore, these studies were found that the estimators derived by using likelihood function approach have a good performance, especially for a small sample size. Therefore, the proposed estimator of the population mean for incomplete dataset was derived based on a factorization of the likelihood function and using information of a simple linear relationship model in the data set. Moreover, a simulation study was conducted 630 situations to compare the efficiency of the proposed estimator with the two estimators, namely pairwise deletion estimator and Anderson

*Corresponding Author: Juthaphorn Sinsomboonthong, Faculty of Science, Kasetsart University, Thailand, E-mail: fscijps@ku.ac.th

estimator. In this study, the efficiency comparison criteria are bias and mean square error (MSE).

2. Materials and Methods

In this paper, the estimation methods of a population mean for incomplete data set are studied for efficiency comparison as follows:

2.1. Anderson Estimator

In 1957, the maximum likelihood estimators of the parameters of a bivariate normal distribution for incomplete data set with one variable was proposed by [14]. Suppose random variables Y_1 and Y_2 have the bivariate normal distribution with mean vector (μ_1, μ_2) and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. Suppose r observations of Y_1 and Y_2 are bivariate normally distributed with mean vector (μ_1, μ_2) and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. In addition, $n-r$ observations of Y_1 are normally distributed with mean μ_1 and variance σ_1^2. The data are shown in Figure 1.

$y_{11}, \cdots, y_{1r}, y_{1,r+1}, \cdots, y_{1n}$
$y_{21}, \cdots, y_{2r}$

Figure 1: Missing data pattern of the bivariate normal distribution

From data pattern in Figure 1, the likelihood function of vector parameter $\underline{\theta}^* = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$ can be written in the formula of equation (1).

$$L(\underline{\theta}^* | Y_{Obs}) = \prod_{j=1}^{n} f_{Y_1}(y_{1j}|\mu_1, \sigma_1^2) \prod_{j=1}^{r} f_{Y_2|Y_1}(y_{2j}|\beta_0 + \beta_1 y_{1j}, \sigma_{2|1}^2) \quad (1)$$

where $\beta_0 = \mu_2 - \beta_1 \mu_1$, $\beta_1 = \rho \frac{\sigma_2}{\sigma_1}$ and $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$.

The maximum likelihood estimators of $\mu_1, \sigma_1^2, \sigma_{2|1}^2, \beta_1$ and β_0 are as follows:

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}, \quad \hat{\sigma}_1^2 = \frac{1}{n}\sum_{j=1}^{n}(y_{1j} - \bar{y}_1)^2, \quad \hat{\sigma}_{2|1}^2 = s_2'^2 - \frac{s_{12}'^2}{s_1'^2},$$

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{r}(y_{1j} - \bar{y}_1')(y_{2j} - \bar{y}_2')}{\sum_{j=1}^{r}(y_{1j} - \bar{y}_1')^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y}_2' - \hat{\beta}_1 \bar{y}_1'$$

where, $s_1'^2 = \frac{1}{r}\sum_{j=1}^{r}(y_{1j} - \bar{y}_1')^2$, $\bar{y}_2' = \frac{1}{r}\sum_{j=1}^{r} y_{2j}$, $\bar{y}_1' = \frac{1}{r}\sum_{j=1}^{r} y_{1j}$

$s_2'^2 = \frac{1}{r}\sum_{j=1}^{r}(y_{2j} - \bar{y}_2')^2$ and $s_{12}' = \frac{1}{r}\sum_{j=1}^{r}(y_{1j} - \bar{y}_1')(y_{2j} - \bar{y}_2')$.

Moreover, the maximum likelihood estimators of μ_2 and σ_2^2 are given by $\hat{\mu}_2 = \bar{y}_2' - \hat{\beta}_1(\bar{y}_1' - \bar{y}_1)$ and $\hat{\sigma}_2^2 = \hat{\sigma}_{2|1}^2 + \hat{\beta}_1^2 \hat{\sigma}_1^2 = \hat{\beta}_1 \frac{\hat{\sigma}_1}{\hat{\sigma}_2}$, respectively.

2.2. Pairwise Deletion Estimator

In this study, pairwise deletion estimator is the estimation of the population mean for incomplete data set based on complete data or available-cases analysis [5], even if the values for the same individual on other variables are missing. Suppose three variables Y_1, Y_2 and Y_3 are trivariate normally distributed in the population and n observations of Y_1 are completely observed for all individuals, but Y_2 and Y_3 are not completely observed for all individuals or they have missing data occurrence. That is, r observations of Y_2 are observed whereas $n-r$ observations of Y_3 are observed. Available cases analysis for the population means μ_1, μ_2 and μ_3 can be written in the forms of equation (2).

$$\hat{\mu}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}, \quad \hat{\mu}_2 = \frac{1}{r}\sum_{j=1}^{r} y_{2j} \quad \text{and} \quad \hat{\mu}_3 = \frac{1}{n-r}\sum_{j=r+1}^{n} y_{3j} \quad (2)$$

Under MCAR [5] of the missing data mechanism, pairwise deletion method will yield consistent and unbiased estimators in a large sample size [5].

2.3. The Proposed Estimator of the Population Mean for Incomplete Data Set

In this section, the estimator of the population mean for incomplete data set is proposed. This proposed estimator is derived using the factorization of the likelihood function [5,14] and a procedure of finding the usual maximum likelihood estimator is applied. Suppose dependent variable Y_1 is assumed to have the linear relationship with independent variable X_1 and its relationship model is given by equation (3).

$$y_{1j} = \delta_0 + \delta_1 x_{1j} + \varepsilon_{1j}, \quad j = 1, 2, ..., n \quad (3)$$

where δ_0 and are random ε_{1j} and , are unknown parameters δ_1 errors that have the normal distribution with mean 0 and variance σ_1^2. Then the mean and variance of Y_1 can be written as $E(Y_1) = \delta_0 + \delta_1 X_1 = \mu_1$ and $V(Y_1) = \sigma_1^2$, respectively. Further, ε_{1j} can be written in the form of equation (4).

$$\varepsilon_{1j} = y_{1j} - \delta_0 - \delta_1 x_{1j}, \quad j = 1, 2, ..., n \quad (4)$$

Let Y_2
μ_2 and variance σ_2^2. In addition, r observations of Y_1 and Y_2

$\underline{\mu} = (\delta_0 + \delta_1 X_1 \ , \ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.

The $n-r$ observations of Y_1 are normally distributed with mean $\delta_0 + \delta_1 X_1$ and variance σ_1^2. The study data pattern is shown in Figure 2.

Observations	X_1	Y_1	Y_2
1	x_{11}	y_{11}	y_{21}
2	x_{12}	y_{12}	y_{22}
⋮	⋮	⋮	⋮
r	x_{1r}	y_{1r}	y_{2r}
$r+1$	$x_{1,r+1}$	$y_{1,r+1}$	
⋮	⋮	⋮	
n	x_{1n}	y_{1n}	

Figure 2: Missing data pattern of the proposed study

Let E_1 be a random variable that have the relationship of Y_1 and X_1 in the form of $E_1 = Y_1 - \delta_0 - \delta_1 X_1$. Then two random variables E_1 and Y_2 are bivariate normally distributed with mean vector $\underline{\mu} = (0, \ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. Additionally, the missing data pattern of E_1 and Y_2 are shown in Figure 3.

Observations	E_1	Y_2
1	ε_{11}	y_{21}
2	ε_{12}	y_{22}
⋮	⋮	⋮
r	ε_{1r}	y_{2r}
$r+1$	$\varepsilon_{1,r+1}$	
⋮	⋮	
n	ε_{1n}	

Figure 3: Random error and missing data pattern of Y_2

Lemma 1 Let $E_1 = Y_1 - \delta_0 - \delta_1 X_1$, Y_1 and Y_2 be the random variables where δ_0, δ_1 are unknown parameters and X_1 be independent variable. Suppose E_1 and Y_2 are bivariate normally distributed with mean vector $\underline{\mu} = (0, \ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. Then, $Y_2 \,|E_1 = \varepsilon_1$ is normally distributed with mean $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ and variance $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$ where $\varepsilon_1 = y_1 - \delta_0 - \delta_1 x_1$,. $\underline{\theta}_{2|1} = (\delta_0, \delta_1, \sigma_{2|1}^2, \tau_{12})$ and $\tau_{12} = \dfrac{\rho\sigma_2}{\sigma_1}$

Proof Let $E_1 = Y_1 - \delta_0 - \delta_1 X_1$ and Y_2 be bivariate normally distributed with mean vector $\underline{\mu} = (0, \ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. Then, the joint probability density function of E_1 and Y_2 is given by equation (5).

$$f_{12}(\varepsilon_1, y_2; \underline{\theta}_{12}) = \frac{1}{2\pi\sqrt{(1-\rho^2)\sigma_1^2\sigma_2^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{\varepsilon_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{\varepsilon_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right\}} \quad (5)$$

where $-\infty < \varepsilon_1 < \infty$, $-\infty < y_2 < \infty$ and $\underline{\theta}_{2|1} = (\delta_0, \delta_1, \sigma_{2|1}^2, \tau_{12})$. Moreover, the probability density function of E_1 is given by equation (6).

$$f_1(\varepsilon_1; \underline{\theta}_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{\varepsilon_1}{\sigma_1}\right)^2} \quad (6)$$

where $-\infty < \varepsilon_1 < \infty$ and $\underline{\theta}_1 = (\delta_0, \delta_1, \sigma_1^2)$.

Hence, a conditional probability density function of Y_2 given $E_1 = \varepsilon_1$ can be written as follows:

$$\begin{aligned} f_{2|1}(y_2 \mid \varepsilon_1; \underline{\theta}_{2|1}) &= \frac{f_{12}(\varepsilon_1, y_2; \underline{\theta}_{12})}{f_1(\varepsilon_1; \underline{\theta}_1)} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{y_2-\mu_2}{\sigma_2}\right) - \rho\left(\frac{\varepsilon_1}{\sigma_1}\right)\right\}^2} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left\{ y_2 - \mu_2 - \frac{\rho\sigma_2}{\sigma_1}\varepsilon_1 \right\}^2} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\{ y_2 - \mu_2 - \tau_{12}\varepsilon_1 \}^2} \ ; \ \tau_{12} = \frac{\rho\sigma_2}{\sigma_1} \\ &= \frac{1}{\sqrt{2\pi\sigma_{2|1}^2}} e^{-\frac{1}{2\sigma_{2|1}^2}\{ y_2 - \mu_{2|1} \}^2} \end{aligned} \quad (7)$$

where $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ and $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$.

From Equation (7), this is the probability density function of a normal distribution with mean $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ and variance $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$. Therefore, a random variable $Y_2 \,|E_1 = \varepsilon_1$ is normally distributed with mean $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ and variance $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$ where $\varepsilon_1 = y_1 - \delta_0 - \delta_1 x_1$ and $\tau_{12} = \dfrac{\rho\sigma_2}{\sigma_1}$.

Lemma 2 For $j = 1, 2, ..., r$, the two random variables E_{1j} and Y_{2j} are assumed to have the bivariate normal distribution with a mean vector $\underline{\mu} = (0,\ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. For $j = r+1,\ r+2,\ ...,\ n,$ the random variable E_{1j} is assumed to have a normal distribution with a mean 0 and variance σ_1^2 where $E_{1j} = Y_{1j} - \delta_0 - \delta_1 X_{1j}$; δ_0 and δ_1 are unknown parameters and X_{1j} be independent variable. Let $\underline{W} = [E_{11}\ E_{12} ... E_{1n}\ Y_{21}\ Y_{22} ... Y_{2r}]'$ be a random vector. Then, the likelihood function of parameter vector $\underline{\theta} = (\delta_0, \delta_1, \sigma_1^2, \sigma_{2|1}^2, \tau_{12})$ is denoted by equation (8).

$$L(\underline{\theta} \mid \underline{w}) = \left(\left(2\pi\sigma_1^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n}\varepsilon_{1j}^2} \right) \left(\left(2\pi\sigma_{2|1}^2\right)^{-\frac{r}{2}} e^{-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j} - \mu_{2|1}\right)^2} \right) \quad (8)$$

where $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$ and $\tau_{12} = \dfrac{\rho\sigma_2}{\sigma_1}$, $\varepsilon_{1j} = y_{1j} - \delta_0 - \delta_1 x_{1j}$

Proof For $j = 1, 2, ..., r$, the two random variables E_{1j} and Y_{2j} are assumed to have a bivariate normal distribution with mean vector $\underline{\mu} = (0,\ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. For $j = r+1,\ r+2,\ ...,\ n,$ the random variable E_{1j} is assumed to have a normal distribution with a mean 0 and variance σ_1^2.

Let $\underline{w} = [\varepsilon_{11}\ \varepsilon_{12}\ ...\ \varepsilon_{1n}\ \ y_{21}\ \ y_{22}\ ...\ y_{2r}]'$ be a vector of value for the random vector $\underline{W} = [E_{11}\ E_{12} ... E_{1n}\ Y_{21}\ Y_{22} ... Y_{2r}]'$. Then, the likelihood function of $\underline{\theta} = (\delta_0, \delta_1, \sigma_1^2, \sigma_{2|1}^2, \tau_{12})$ can be written as follows:

$$\begin{aligned} L(\underline{\theta} \mid \underline{w}) &= \prod_{j=1}^{r} f_{12}(\varepsilon_{1j}, y_{2j}; \underline{\theta}_{12}) \prod_{j=r+1}^{n} f_1(\varepsilon_{1j}; \underline{\theta}_1) \\ &= \left(\prod_{j=1}^{r} f_1(\varepsilon_{1j}; \underline{\theta}_1) \times f_{2|1}(y_{2j} \mid \varepsilon_{1j}; \underline{\theta}_{2|1}) \right) \left(\prod_{j=r+1}^{n} f_1(\varepsilon_{1j}; \underline{\theta}_1) \right) \\ &= \prod_{j=1}^{n} f_1(\varepsilon_{1j}; \underline{\theta}_1) \prod_{j=1}^{r} f_{2|1}(y_{2j} \mid \varepsilon_{1j}; \underline{\theta}_{2|1}) \end{aligned} \quad (9)$$

From Lemma 1, the likelihood function $L(\underline{\theta} \mid \underline{w})$ in equation (9) can be written as

$$\begin{aligned} &L(\underline{\theta} \mid \underline{w}) \\ &= \left(\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{\varepsilon_{1j}}{\sigma_1}\right)^2} \right) \left(\prod_{j=1}^{r} \frac{1}{\sqrt{2\pi\sigma_{2|1}^2}} e^{-\frac{1}{2\sigma_{2|1}^2}\left\{ y_{2j} - \mu_{2|1} \right\}^2} \right) \\ &= \left(\left(2\pi\sigma_1^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\varepsilon_{1j}^2} \right) \left(\left(2\pi\sigma_{2|1}^2\right)^{-\frac{r}{2}} e^{-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j} - \mu_{2|1}\right)^2} \right) \end{aligned}$$

Theorem 1 For $j = 1, 2, ..., r$, the two random variables E_{1j} and Y_{2j} are assumed to have a bivariate normal distribution with mean vector $\underline{\mu} = (0,\ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. For $j = r+1,\ r+2,\ ...,\ n,$ the random variable E_{1j} is assumed to have a normal distribution with mean 0 and variance σ_1^2 where $E_{1j} = Y_{1j} - \delta_0 - \delta_1 X_{1j}$; δ_0 and δ_1 are unknown parameters and X_{1j} be independent variable. Let $\underline{W} = [E_{11}\ E_{12} ... E_{1n}\ Y_{21}\ Y_{22} ... Y_{2r}]'$ be a random vector. Then, the factorization maximum likelihood estimator of μ_2 is given in equation (10).

$$\hat{\mu}_{2_{\text{Proposed}}} = \overline{y}_2' - \hat{\tau}_{12}\overline{e}_1' \quad (10)$$

where $\hat{\delta}_1 = \dfrac{\sum_{j=1}^{n} x_{1j}y_{1j} - n\overline{x}_1\overline{y}_1}{\sum_{j=1}^{n} x_{1j}^2 - n(\overline{x}_1)^2}$,, $\overline{y}_1 = \dfrac{1}{n}\sum_{j=1}^{n} y_{1j}$, $\overline{x}_1 = \dfrac{1}{n}\sum_{j=1}^{n} x_{1j}$

$e_{1j} = y_{1j} - \hat{\delta}_0 - \hat{\delta}_1 x_{1j}$, $\hat{\delta}_0 = \overline{y}_1 - \hat{\delta}_1\overline{x}_1$ for, $j = 1, 2, ..., r$

$\hat{\tau}_{12} = \dfrac{\sum_{j=1}^{r} e_{1j}y_{2j} - r\overline{e}_1'\overline{y}_2'}{\sum_{j=1}^{r} e_{1j}^2 - r(\overline{e}_1')^2}$, $\overline{y}_2' = \dfrac{1}{r}\sum_{j=1}^{r} y_{2j}$ and $\overline{e}_1' = \dfrac{1}{r}\sum_{j=1}^{r} e_{1j}$.

Proof Let $\underline{W} = [E_{11}\ E_{12} ... E_{1n}\ Y_{21}\ Y_{22} ... Y_{2r}]'$ be a random vector. From Lemma 2, we known that the likelihood function of $\underline{\theta} = (\delta_0, \delta_1, \sigma_1^2, \sigma_{2|1}^2, \tau_{12})$ is denoted by equation (8). Then, the log-likelihood function can be written in the form of equation (11).

$$\begin{aligned} \ln L(\underline{\theta} \mid \underline{w}) &= -\frac{n}{2}\ln\left(2\pi\sigma_1^2\right) - \frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\varepsilon_{1j}^2 - \frac{r}{2}\ln\left(2\pi\sigma_{2|1}^2\right) \\ &\quad - \frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j} - \mu_{2|1}\right)^2 \end{aligned} \quad (11)$$

From Lemma 1, the random variable $Y_2 \mid E_1 = \varepsilon_1$ is normally distributed with mean $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ and variance $\sigma_{2|1}^2 = (1-\rho^2)\sigma_2^2$ where $\varepsilon_1 = y_1 - \delta_0 - \delta_1 x_1$ and $\tau_{12} = \dfrac{\rho\sigma_2}{\sigma_1}$.

Then, the log-likelihood function as shown in equation (11) need to maximize and achieve the maximum likelihood estimators of μ_2, δ_0, δ_1 and are as follows: τ_{12}

$$\frac{\partial}{\partial\delta_0}\ln L(\underline{\theta}\,|\,\underline{w}) = \frac{\partial}{\partial\delta_0}\left[-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\varepsilon_{1j}^2\right] = 0$$

$$= \frac{\partial}{\partial\delta_0}\left[-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\left(y_{1j}-\delta_0-\delta_1 x_{1j}\right)^2\right] = 0$$

$$= \sum_{j=1}^{n} y_{1j} - n\delta_0 - \delta_1\sum_{j=1}^{n} x_{1j} = 0 \qquad (12)$$

$$\frac{\partial}{\partial\delta_1}\ln L(\underline{\theta}\,|\,\underline{w}) = \frac{\partial}{\partial\delta_1}\left[-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\varepsilon_{1j}^2\right] = 0$$

$$= \frac{\partial}{\partial\delta_1}\left[-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\left(y_{1j}-\delta_0-\delta_1 x_{1j}\right)^2\right] = 0$$

$$= \sum_{j=1}^{n} x_{1j}y_{1j} - \delta_0\sum_{j=1}^{n} x_{1j} - \delta_1\sum_{j=1}^{n} x_{1j}^2 = 0 \qquad (13)$$

Equation (12) is multiplied by $\sum_{j=1}^{n} x_{1j}$, then it will give the form in equation (14).

$$\sum_{j=1}^{n} x_{1j}\sum_{j=1}^{n} y_{1j} - n\delta_0\sum_{j=1}^{n} x_{1j} - \delta_1\left(\sum_{j=1}^{n} x_{1j}\right)^2 = 0 \qquad (14)$$

Equation (13) is multiplied by *n*, then it will give the form in equation (15).

$$n\sum_{j=1}^{n} x_{1j}y_{1j} - n\delta_0\sum_{j=1}^{n} x_{1j} - n\delta_1\sum_{j=1}^{n} x_{1j}^2 = 0 \qquad (15)$$

Subtraction equation (14) from equation (15), then it will give the form in equation (16).

$$n\sum_{j=1}^{n} x_{1j}y_{1j} - n\delta_1\sum_{j=1}^{n} x_{1j}^2 - \sum_{j=1}^{n} x_{1j}\sum_{j=1}^{n} y_{1j} + \delta_1\left(\sum_{j=1}^{n} x_{1j}\right)^2 = 0 \qquad (16)$$

That is, $= 0\ n\sum_{j=1}^{n} x_{1j}y_{1j} - \sum_{j=1}^{n} x_{1j}\sum_{j=1}^{n} y_{1j} - \left[n\sum_{j=1}^{n} x_{1j}^2 - \left(\sum_{j=1}^{n} x_{1j}\right)^2\right]\delta_1$

Additionally, the value of δ_1 that maximize the log-likelihood function is denoted by

$$\delta_1 = \frac{n\sum_{j=1}^{n} x_{1j}y_{1j} - \sum_{j=1}^{n} x_{1j}\sum_{j=1}^{n} y_{1j}}{n\sum_{j=1}^{n} x_{1j}^2 - \left(\sum_{j=1}^{n} x_{1j}\right)^2} \quad \text{or} \quad \delta_1 = \frac{\sum_{j=1}^{n} x_{1j}y_{1j} - n\bar{x}_1\bar{y}_1}{\sum_{j=1}^{n} x_{1j}^2 - n(\bar{x}_1)^2}$$

Therefore, the maximum likelihood estimator of δ_1 is given by

$$\hat{\delta}_1 = \frac{\sum_{j=1}^{n} x_{1j}y_{1j} - n\bar{x}_1\bar{y}_1}{\sum_{j=1}^{n} x_{1j}^2 - n(\bar{x}_1)^2} \quad \text{for } \bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{1j} \text{ and } \bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}.$$

From equation (12), the form of this equation can be written as $n\delta_0 = \sum_{j=1}^{n} y_{1j} - \delta_1\sum_{j=1}^{n} x_{1j}$ or Then, the maximum . $\delta_0 = \bar{y}_1 - \delta_1\bar{x}_1$

likelihood estimator of δ_0 is given by $\hat{\delta}_0 = \bar{y}_1 - \hat{\delta}_1\bar{x}_1$.

From Lemma 1, we know that $\mu_{2|1} = \mu_2 + \tau_{12}\varepsilon_1$ then the maximum likelihood estimator of parameter τ_{12} can be derived as follows:

$$\frac{\partial}{\partial\tau_{12}}\ln L(\underline{\theta}\,|\,\underline{w}) = \frac{\partial}{\partial\tau_{12}}\left[-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j}-\mu_{2|1}\right)^2\right] = 0$$

$$= \frac{\partial}{\partial\tau_{12}}\left[-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j}-\mu_2-\tau_{12}\varepsilon_{1j}\right)^2\right] = 0$$

$$= \sum_{j=1}^{r}\varepsilon_{1j}y_{2j} - \mu_2\sum_{j=1}^{r}\varepsilon_{1j} - \tau_{12}\sum_{j=1}^{r}\varepsilon_{1j}^2 = 0 \qquad (17)$$

$$\frac{\partial}{\partial\mu_2}\ln L(\underline{\theta}\,|\,\underline{w}) = \frac{\partial}{\partial\mu_2}\left[-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j}-\mu_{2|1}\right)^2\right] = 0$$

$$= \frac{\partial}{\partial\mu_2}\left[-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left(y_{2j}-\mu_2-\tau_{12}\varepsilon_{1j}\right)^2\right] = 0$$

$$= \sum_{j=1}^{r} y_{2j} - r\mu_2 - \tau_{12}\sum_{j=1}^{r}\varepsilon_{1j} = 0 \qquad (18)$$

Equation (18) is multiplied by $\sum_{j=1}^{r}\varepsilon_{1j}$, then it will give the form in equation (19).

$$\sum_{j=1}^{r}\varepsilon_{1j}\sum_{j=1}^{r} y_{2j} - r\mu_2\sum_{j=1}^{r}\varepsilon_{1j} - \tau_{12}\left(\sum_{j=1}^{r}\varepsilon_{1j}\right)^2 = 0 \qquad (19)$$

Equation (17) is multiplied by *r*, then it will give the form in equation (20).

$$r\sum_{j=1}^{r}\varepsilon_{1j}y_{2j} - r\mu_2\sum_{j=1}^{r}\varepsilon_{1j} - r\tau_{12}\sum_{j=1}^{r}\varepsilon_{1j}^2 = 0 \qquad (20)$$

Subtraction equation (19) from equation (20), then it will give the form in equation (21).

$$r\sum_{j=1}^{r}\varepsilon_{1j}y_{2j} - r\tau_{12}\sum_{j=1}^{r}\varepsilon_{1j}^2 - \sum_{j=1}^{r}\varepsilon_{1j}\sum_{j=1}^{r} y_{2j} + \tau_{12}\left(\sum_{j=1}^{r}\varepsilon_{1j}\right)^2 = 0 \qquad (21)$$

Furthermore, the value of τ_{12} that maximize the log-likelihood function is denoted by

$$\tau_{12} = \frac{r\sum_{j=1}^{r}\varepsilon_{1j}y_{2j} - \sum_{j=1}^{r}\varepsilon_{1j}\sum_{j=1}^{r} y_{2j}}{r\sum_{j=1}^{r}\varepsilon_{1j}^2 - \left(\sum_{j=1}^{r}\varepsilon_{1j}\right)^2} \quad \text{or} \quad \tau_{12} = \frac{\sum_{j=1}^{r}\varepsilon_{1j}y_{2j} - r\bar{\varepsilon}_1'\bar{y}_2'}{\sum_{j=1}^{r}\varepsilon_{1j}^2 - r(\bar{\varepsilon}_1')^2}.$$

Therefore, the maximum likelihood estimator of τ_{12} is given by

$$\hat{\tau}_{12} = \frac{\sum_{j=1}^{r} e_{1j}y_{2j} - r\bar{e}_1'\bar{y}_2'}{\sum_{j=1}^{r} e_{1j}^2 - r(\bar{e}_1')^2} \qquad \bar{y}_2' = \frac{1}{r}\sum_{j=1}^{r} y_{2j} \text{ and } \bar{e}_1' = \frac{1}{r}\sum_{j=1}^{r} e_{1j}$$

$$e_{1j} = y_{1j} - \hat{\delta}_0 - \hat{\delta}_1 x_{1j} \qquad j = 1, 2, ..., r\ ;$$

, thTherefore $\mu_2 = \bar{y}_2' - \tau_{12}\bar{\varepsilon}_1'$. or $r\mu_2 = \sum_{j=1}^{r} y_{2j} - \tau_{12}\sum_{j=1}^{r}\varepsilon_{1j}$ maximum likelihood estimator of parameter μ_2 is given by $\hat{\mu}_{2_{\text{Proposed}}} = \bar{y}_2' - \hat{\tau}_{12}\bar{e}_1'$.

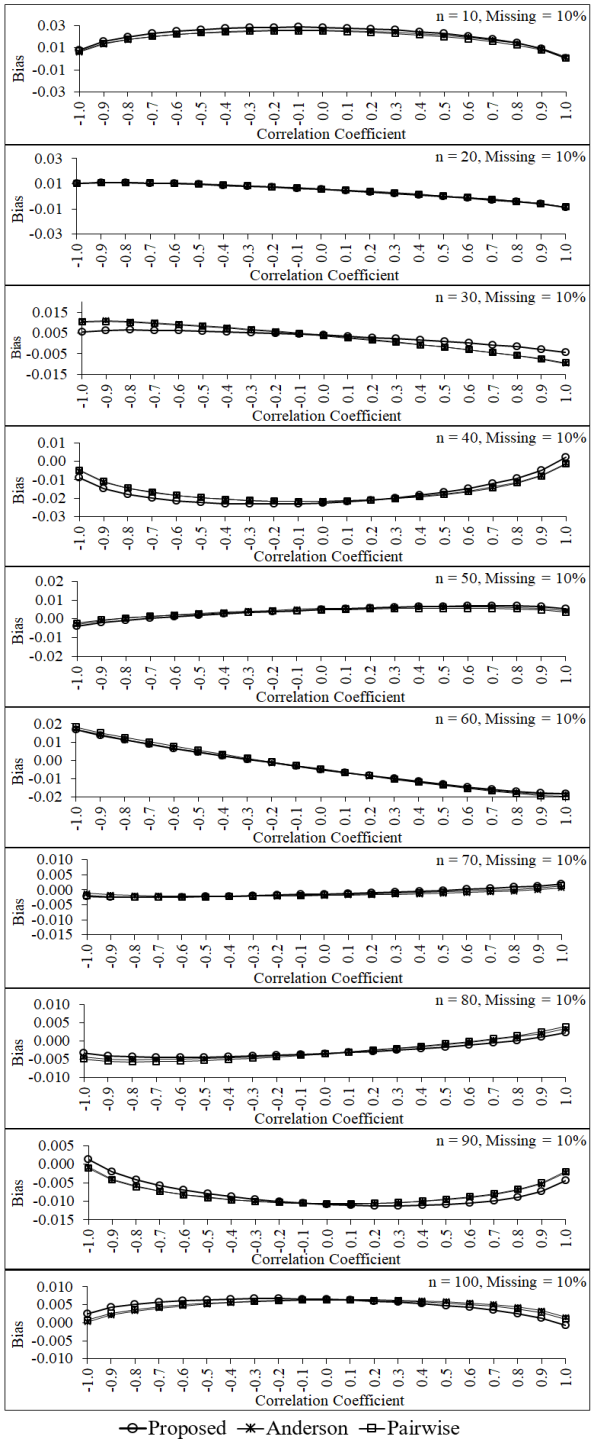

Figure 4: Biases of the three estimators for percentage of missing data equals 10 of each sample size

3. Results of a Simulation Study

The efficiency investigation of the proposed estimator and comparison of its efficiency with the two estimators—Anderson and pairwise deletion estimators—are studied via the simulation data. Moreover, these data are generated 630 situations and repeated 50,000 times for each situation. In this section, the criteria in terms of bias and mean square error are used for efficiency comparison. The population data of random variables Y_1 and Y_2 are generated in the form of bivariate normal distribution with mean vector $\underline{\mu} = (\delta_0 - \delta_1 X_1,\ \mu_2)$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.

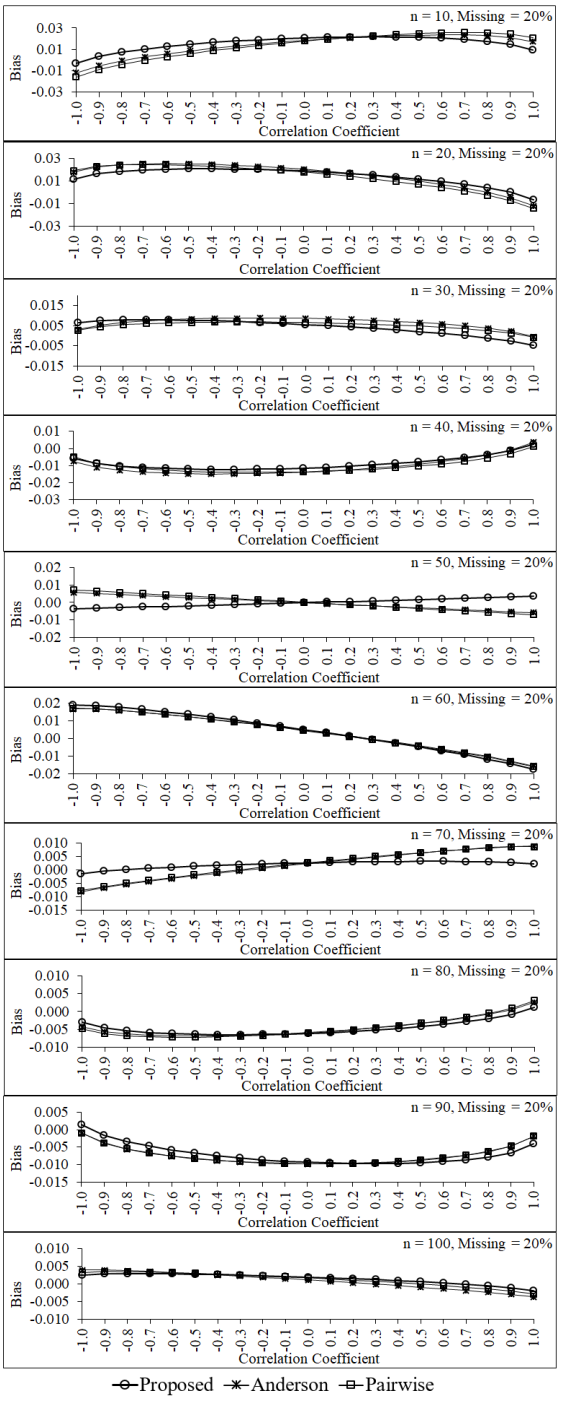

Figure 5: Biases of the three estimators for percentage of missing data equals 20 of each sample size

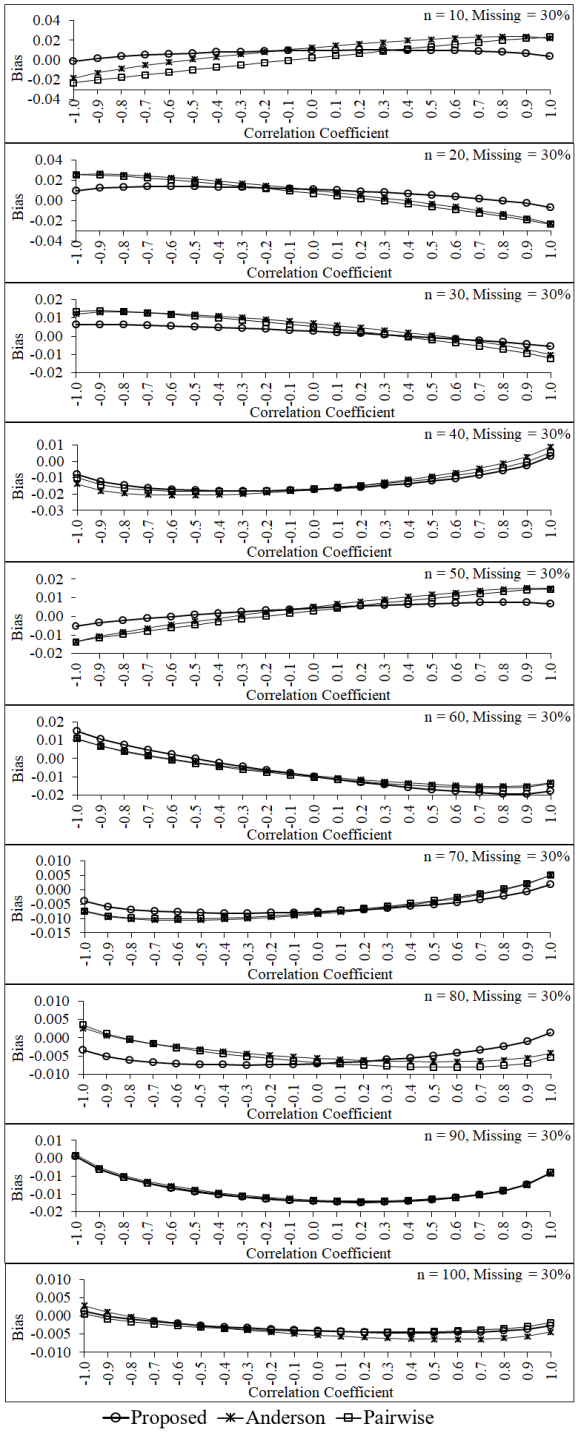

Figure 6: Biases of the three estimators for percentage of missing data equals 30 of each sample size

In this study, the values of parameters are defined as follows: $\delta_0 = 2$, $\delta_1 = 3$, $\mu_2 = 5$, $\sigma_2^2 = 9$ and the correlations between Y_1 and Y_2 are given by $\rho = -1.0, -0.9, ..., 0, ..., 0.9, 1.0$. Then, the samples of size n = 10, 20, 30, ..., 100 are randomly taken from these populations. Missing data mechanism in the form of MCAR [5] for three levels—10%, 20% and 30%—are constructed from each sample. The simulation results are shown in Figure 4 to Figure 9. Figure 4 to Figure 6 show that when percentages of missing data equal 10, 20 and 30 of each sample size, bias of the proposed estimator tends to be no difference from those of pairwise deletion and Anderson estimators for almost all sample sizes and all levels of the correlation between two variables in the data set. Moreover, some situations (e.g., n = 20, 30 and percentage of missing data in the data set equals 30) and negative high correlation between two variables, its bias tends to be smaller than the bias of pairwise deletion and Anderson estimators.

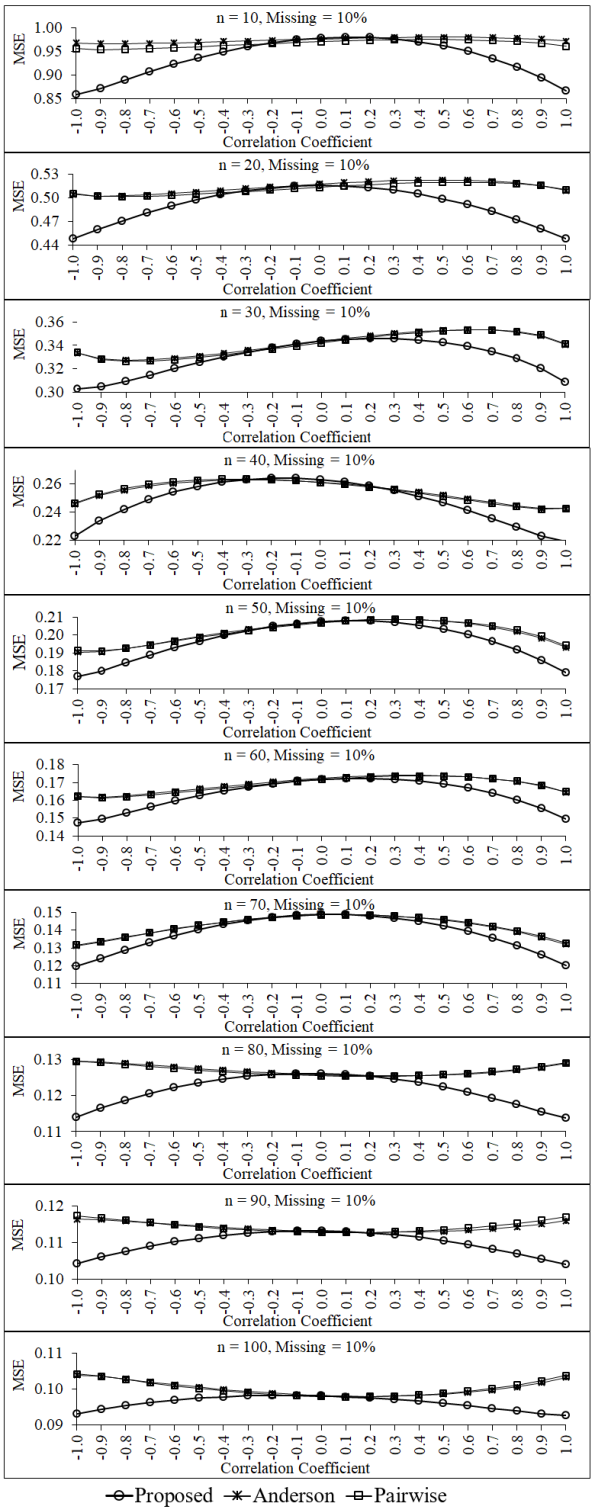

Figure 7: Mean square errors of the three estimators for percentage of missing data equals 10 of each sample size

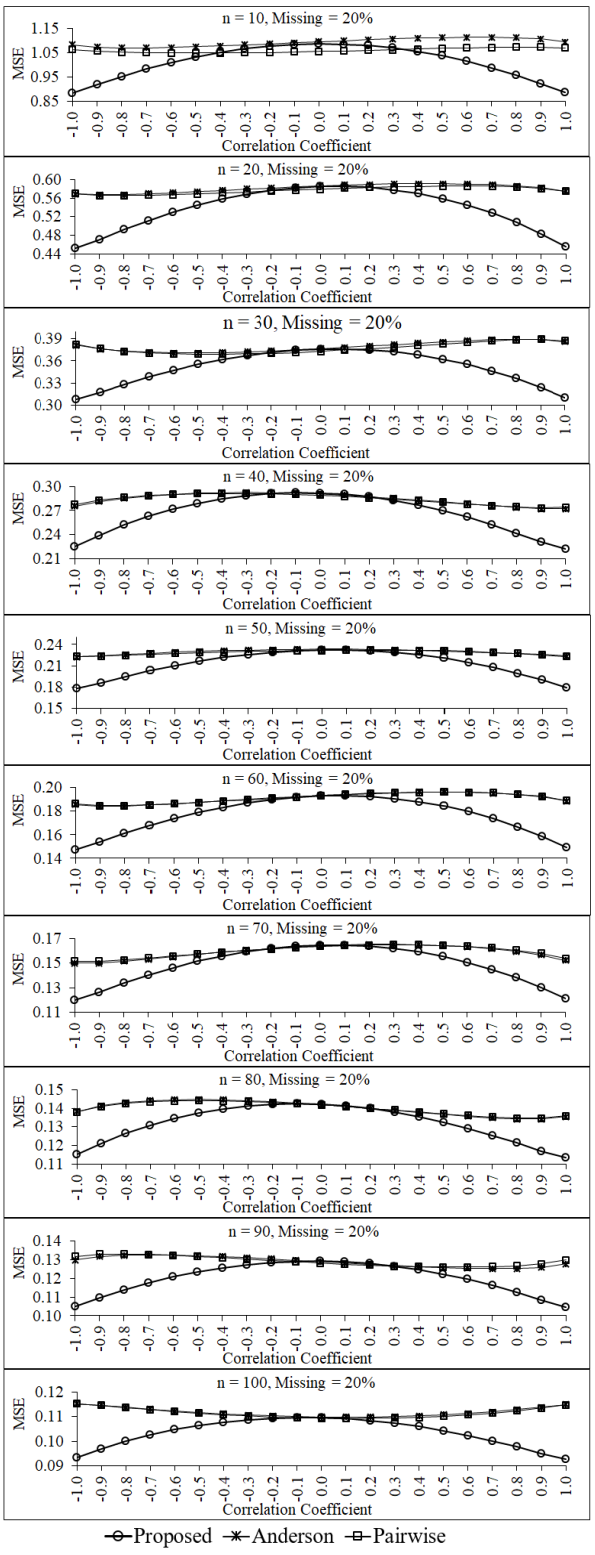

Figure 8: Mean square errors of the three estimators for percentage of missing data equals 20 of each sample size

When considering the performance of the proposed estimator in term of mean square error in Figure 7, it is found that the mean square error of the proposed estimator tends to be lower than those of pairwise deletion and Anderson estimators for the large correlation levels between two variables in the data set and all sample sizes when the data have 10 % of missing data.

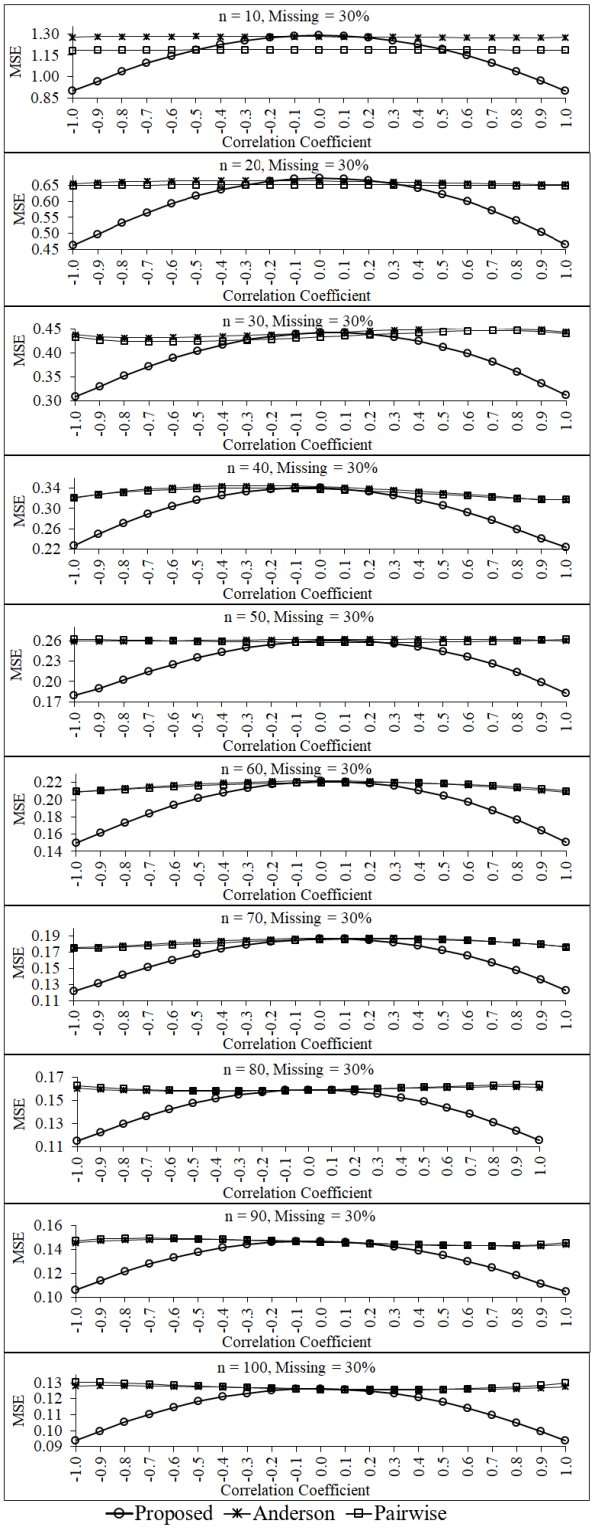

Figure 9: Mean square errors of the three estimators for percentage of missing data equals 30 of each sample size

For higher percentages of missing data of each sample sizes as show in Figure 8 and Figure 9, the performance of the proposed estimator in term of mean square error are similar to the case of the small percentages of missing data as mention above. Additionally, the mean square error of the proposed estimator tends to be obviously lower than those of pairwise deletion and Anderson estimators for the large correlation levels between two variables in

the data set whatever the sample sizes will be. However, for the small correlation levels between two variables in the data set, the three estimators tend to have the same performances in terms of both two criteria—bias and mean square error—for all sample sizes and all percentage levels of missing data. This simulation study is found that the mean square errors of three estimators tend to be decrease when the sample size increases for all levels of the correlations between two variables in the data set and all levels of the percentages of missing data. In addition, the mean square error of the proposed estimator tends to be lower than those of the two estimators—pairwise deletion and Anderson estimators—for the small sample sizes (e.g., n = 10, 20, 30) and high correlations (e.g., ρ = -0.1, -0.9, -0.8, 0.8, 0.9, 1.0) between two variables in the data set, especially the percentage of missing data is equal to 30. However, the mean square errors of three estimators tend to have a similar performances for the low correlations between two variables in the data set and all levels of the percentages of missing data.

4. Discussion

In this study, the simulation results show that pairwise deletion estimator tends to be a biased estimator for the small sample sizes as mention by [5,9]. Moreover, the maximum likelihood estimator of the population average for incomplete data set is derived by using factorization of the likelihood function approach [14] tends to have a good performance for the large correlation levels between two variables in the data set and small sample sizes. This conforms to the studies of [14,16]. In addition, the maximum likelihood estimation of the population mean for incomplete data set tends to have a good efficiency for small sample sizes as the study of [7]. This discovery of the proposed estimator will benefit for some applications in the real life data, especially nowadays it is the era of big data analysis which has the large number of variables in data set. Therefore, we should find the relationships of some attributes in data set before estimating the average of the interested variables for incomplete data analysis. Further, this proposed estimator will lead to correct estimate as possible.

5. Conclusion

The proposed estimator of the population mean for incomplete dataset was derived by using the linear relationship between some variables in the data set and the factorization of likelihood function [14] was created to derive the proposed maximum likelihood estimator. Additionally, the investigation of this proposed estimator was studied via the simulation data for 630 situations to compare the efficiency in terms of bias and mean square error with two estimators, namely pairwise deletion and Anderson estimators. It is found that the efficiency of the proposed estimator tends to be better than those of two above mention estimators, especially for case of the high percentages of missing data and the strong linear correlation between two variables (e.g., the degree of ρ close to -1 or 1) whatever the sample size will be. However, for the small correlation between two variables (e.g., the degree of ρ close to zero), the three estimators tend to have the similar efficiencies for all sample sizes and all percentage levels of missing data.

Acknowledgment

The authors would like to express our special thanks of gratitude to head of Kasetsart University Research and Development Institute (KURDI) for financial support of this research.

References

[1] S. Gaucher, O. Klopp, G. Robin, "Outlier detection in networks with missing links," Computational Statistics & Data Analysis, **164**, 107308, 2021, doi:10.1016/j.csda.2021.107308.

[2] L.A. Vale-Silva, K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," Scientific Reports, **11**(1), 1–12, 2021, doi:10.1038/s41598-021-92799-4.

[3] J.A. Smith, J.H. Morgan, J. Moody, "Network sampling coverage III: Imputation of missing network data under different network and missing data conditions," Social Networks, **68**(June 2021), 148–178, 2022, doi:10.1016/j.socnet.2021.05.002.

[4] N. Kumar, M.A. Hoque, M. Sugimoto, "Kernel weighted least square approach for imputing missing values of metabolomics data," Scientific Reports, **11**(1), 1–12, 2021, doi:10.1038/s41598-021-90654-0.

[5] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley&Son, 2002.

[6] M.N. Norazian, Y.A. Shukri, R.N. Azam, A.M.M. Al Bakri, "Estimation of missing values in air pollution data using single imputation techniques," ScienceAsia, **34**(3), 341–345, 2008, doi:10.2306/scienceasia1513-1874.2008.34.341.

[7] P.T. Von Hippel, "The Bias and Efficiency of Incomplete-Data Estimators in Small Univariate Normal Samples," Sociological Methods and Research, **42**(4), 531–558, 2013, doi:10.1177/0049124113494582.

[8] A.F. C.R. Rao, H. Toutenburg, Linear models: least squares and alternatives, 2nd ed., Springer Verlag, 1999.

[9] A.C. Acock, "Working With Missing Values," Journal of Marriage and Family, **67**(November), 1012–1028, 2005.

[10] D.W. A. Rotnitzky, "A Note on the biased of estimators with missing data," Biometrics, **50**, 1163–1170, 1994.

[11] M.H. Gorelick, "Bias arising from missing data in predictive models," Journal of Clinical Epidemiology, **59**(10), 1115–1123, 2006, doi:10.1016/j.jclinepi.2004.11.029.

[12] P.L. Roth, J.E. Campion, S.D. Jones, "The impact of four missing data techniques on validity estimates in Human Resource Management," Journal of Business and Psychology, **11**(1), 101–112, 1996, doi:10.1007/BF02278259.

[13] G. Fitzmaurice, "Missing data: implications for analysis," Nutrition, **24**, 200–202, 2008.

[14] T.W. Anderson, "Maximum likelihood estimates for the multivariate normal distribution when some observations are missing," Journal of the American Statistical Association, **52**, 200–203, 1957.

[15] A.M. C. Gourieroux, "On the problem of missing data in linear models," Review of Economic Studies, **48**(4), 579–586, 1981.

[16] J. Sinsomboonthong, "Jackknife maximum likelihood estimates for a bivariate normal distribution with missing data," Journal of Thai Statistical Association, **9**(2), 151–169, 2011, doi:10.1214/aos/1176345020.