Weighted Maximum Likelihood Correlation Coefficient to Handle Missing Values and Outliers in Dataset

JUTHAPHORN SINSOMBOONTHONG*
Department of Statistics, Faculty of Science
Kasetsart University
Bangkok, 10900
THAILAND

SAICHON SINSOMBOONTHONG
Department of Statistics, School of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520
THAILAND

Abstract: - The proposed estimator, namely weighted maximum likelihood (WML) correlation coefficient, for measuring the relationship between two variables to concern about missing values and outliers in the dataset is presented. This estimator is proven by applying the conditional probability function to take care of some missing values and pay more attention to values near the center. However, outliers in the dataset are assigned a slight weight. These using techniques will give the robust proposed method when the preliminary assumptions are not met data analysis. To inspect about the quality of the proposed estimator, the six methods—WML, Pearson, median, percentage bend, biweight mid, and composite correlation coefficients—are compared the properties in two criteria, i.e. the bias and mean squared error, via the simulation study. The results of generated data are illustrated that the WML estimator seems to have the best performance to withstand the missing values and outliers in dataset, especially for the tiny sample size and large percentage of outliers regardless of missing data levels. However, for the massive sample size, the median correlation coefficient seems to have the good estimator when linear relationship levels between two variables are approximately over 0.4 irrespective of outliers and missing data levels.

Key-Words: - maximum likelihood, correlation coefficient, missing, outliers, bias, mean squared error

Received: August 30, 2021. Revised: August 16, 2021. Accepted: August 26, 2021. Published: September 6, 2021.

1 Introduction

Outliers and incomplete data are common problems in many research studies and often cause wrong conclusions in data analysis when conventional techniques are used to analyse these data. In addition, many fields of researches such as research in biology, medical and engineering often be interested in the relationship analysis of two variables in the dataset. However, the collected data may be sometimes lost and outliers are contaminated in the data. Therefore, the well-designed methods result in accurate analysis and conclusions as close to the actual value as possible. The following researches mention about the estimators for the correlation measure of two variables. Pearson correlation coefficient [1] is a familiar method for estimation the correlation coefficient of two random variables—X and Y—when data are sampled from a bivariate normal distribution [2]. Using a traditional estimator such as Pearson correlation coefficient to analyse incomplete data, it tends to result an inefficient estimator [3, 4]. In addition, the researches of [5-16] were found that analysis of incomplete dataset result in biased estimators and make the wrong conclusions. We know that the major problem in many research studies are outliers occurrence in the dataset and this problem may occur from different reasons [17, 18]. Therefore, the correlation coefficient estimation using Pearson's correlation coefficient can give inaccurate conclusions when the sample data are outlier [19]. These findings are consistent with the researches of [20-24]. In 1997, the median correlation coefficient was proposed [24] which is based on the sample median and it is the robust method when dataset are composed with outliers. Additionally, the research study of [22] is confirmed that this estimator has the good properties for two criteria, i.e., unbiased property and small variance for a huge sample size (*n* = 1,000) under data are sampled from normal distribution. In addition, the type M measure of correlations, e.g., the percentage bend and biweight mid correlation coefficients are suggested by Wilcox [25]. These estimators prevent outliers during the incremental distribution when dealing with outliers. Later, the robust estimator for correlation measure—composite correlation coefficient—was proposed [26] for data are bivariate normally distributed and having outliers. This study was found that the composite correlation coefficient performs well when there are outliers in the dataset. In this study, an estimator of correlation coefficient for incomplete dataset with outliers is proposed. This method is called weighted maximum likelihood (WML) correlation coefficient. The proposed estimator was proved by using the conditional likelihood function [27] when some observations are missing and concern with the Winsorized distribution. That is, the tails of Winsorized distribution can dominate outliers. In addition, the simulation data were conducted to compare the robustness features in two properties—bias and mean squared error (MSE)—of the six methods, namely, WML, Pearson, median, percentage bend, biweight mid, and composite correlation coefficients. This finding will give an advantage to many field of researches because the proposed estimator uses all observations that are sampled. That is, the proposed estimator can capture information from all observations in the sample data regardless outliers that contaminated in the dataset.

2 Materials and Methods

In this section, random samples $\left(X_i, Y_i\right)$; $i=1,2,...,n$ are taken from the bivariate normal distribution with vector of mean $\left(\mu_X,\mu_Y\right)$ and matrix of variance covariance in the form of $\Sigma=\begin{bmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{bmatrix}$, where ρ is the correlation coefficient between *X* and *Y*. Assume observations (x,y) are non-missing value for *r* pairs and the remaining $n-r$ observations $\left(0<r<n\right)$ on *y* are lost. Moreover, data *y* are supposed missing completely at random mechanism [3] and they are composed of mild outliers [28]. In this study, the missing data pattern is represented is Fig. 1.

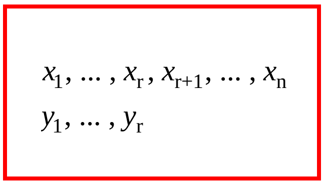

Fig.1: Missing data pattern of variables *X* and *Y*

2.1 Correlation Coefficient

The following five methods of correlation coefficients are studied and calculated from the available-case analysis.

2.1.1 Pearson Correlation Coefficient

Pearson correlation coefficient is a familiar estimator that measure linear relationship of two random variables [1]. Let r_P be the Pearson correlation coefficient that formula is given as follows: $r_P=\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$ where $\bar{x}$ and $\bar{y}$ are the sample means of data *x* and data *y*, respectively.

2.1.2 Median Correlation Coefficient

The median correlation coefficient $\left(r_M\right)$ is proposed by Shevlyakov [29] in 1997. This method is derived by using the sample medians and this estimator is calculated as $r_M=\dfrac{M_R^2-M_S^2}{M_R^2+M_S^2}$ where M_S and M_R are the sample medians of data *s* and *r*, respectively. Let data *s* and *r* be in the forms of

$$s_i=\left|\frac{x_i-M_X}{MAD_X}-\frac{y_i-M_Y}{MAD_Y}\right| \quad \text{and}$$

$$r_i=\left|\frac{x_i-M_X}{MAD_X}+\frac{y_i-M_Y}{MAD_Y}\right| \ ; \ i=1,2,...,n,$$

where the sample median absolute deviations for data *x* and *y* are denoted by symbols MAD_X and MAD_Y, respectively.

2.1.3 Percentage Bend Correlation Coefficient

When data compose of outliers, one of the robust estimators [30] is the percentage bend correlation coefficient which is denoted by symbol r_{PB}. The formula of this method is given as follows:

$$r_{PB} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}$$

where $A_i = \max\left(-1, \min\left(1, U_i^*\right)\right) = \psi\left(U_i^*\right)$, $U_i^* = \frac{1}{\hat{\omega}_X}\left(x_i - \hat{\phi}_X\right)$ for $\hat{\phi}_X = \frac{S_X + \hat{\omega}_X\left(i_2 - i_1\right)}{n - i_1 - i_2}$; i_1 and i_2 are the number of observations x_i which corresponding to inequalities of $\frac{1}{\hat{\omega}_X}\left(x_i - M_X\right) < -1$ and $\frac{1}{\hat{\omega}_X}\left(x_i - M_X\right) > 1$ respectively, M_X is the sample median of observation *x*, $\hat{\omega}_X = W_{(m)}$ is the statistic of order m for variables $W_i = \left|x_i - M_X\right|$, $m = \lfloor n(1-\beta) \rfloor$ and the value of $n(1-\beta)$ is rounded down to the nearest whole integer. Moreover, the formula of B_i is denoted by $B_i = \max\left(-1, \min\left(1, V_i^*\right)\right) = \psi\left(V_i^*\right)$ and variable V_i^* is the counterpart form of variable U_i^*. However, $V_i^* = \frac{1}{\hat{\omega}_Y}\left(y_i - \hat{\phi}_Y\right)$ which is calculated by using the observation y_i.

2.1.4 Biweight Mid Correlation Coefficient

A robust estimator [30] when data are contained with outliers was proposed. This is called a biweight mid correlation coefficient $\left(r_B\right)$ and its formula can be calculated as follows:

$$r_B = \frac{S_{bxy}}{\sqrt{S_{bxx} S_{byy}}} \text{ where}$$

$$S_{bxy} = \frac{n\sum_{i=1}^{n} a_i\left(1-U_i^2\right)^2\left(x_i - M_X\right) b_i\left(1-V_i^2\right)^2\left(y_i - M_Y\right)}{\left[\sum_{i=1}^{n} a_i\left(1-5U_i^2\right)\left(1-U_i^2\right)\right]\left[\sum_{i=1}^{n} b_i\left(1-5V_i^2\right)\left(1-V_i^2\right)\right]},$$

$$S_{bxx} = \frac{n\sum_{i=1}^{n} a_i\left(1-U_i^2\right)^4\left(x_i - M_X\right)^2}{\left[\sum_{i=1}^{n} a_i\left(1-5U_i^2\right)\left(1-U_i^2\right)\right]^2},$$

$$S_{byy} = \frac{n\sum_{i=1}^{n} b_i\left(1-V_i^2\right)^4\left(y_i - M_Y\right)^2}{\left[\sum_{i=1}^{n} b_i\left(1-5V_i^2\right)\left(1-V_i^2\right)\right]^2},$$

$$a_i = \begin{cases} 1 & for \;\; -1 \le U_i \le 1 \\ 0 & for \;\; U_i < -1 \;\; or \;\; U_i > 1, \end{cases}$$

$$b_i = \begin{cases} 1 & for \;\; -1 \le V_i \le 1 \\ 0 & for \;\; V_i < -1 \;\; or \;\; V_i > 1, \end{cases}$$

$$U_i = \frac{x_i - M_X}{9 \times MAD_X}, \; V_i = \frac{y_i - M_Y}{9 \times MAD_Y}.$$

The symbols M_X and M_Y are the sample median of data *x* and data *y*, respectively. Additionally, MAD_X and MAD_Y are the sample median absolute deviation of data *x* and data *y*, respectively.

2.1.5 Composite Correlation Coefficient

In 2016, the composite correlation coefficient $\left(r_C\right)$ was proposed by Sinsomboonthong [26]. This estimator was proved by using two combination of estimator—adaptive Blest and Blest correlation coefficients—with equal weights of them are assigned. This estimator can be computed as $r_C = n\hat{\delta} - \left(1 - \frac{1}{n}\right)\sum_{i=1}^{n}\hat{\delta}_{(-i)}$ where $\hat{\delta}$ and $\hat{\delta}_{(-i)}$ are defined as follows:

$$\hat{\delta} = \frac{2n+1}{n-1} - \frac{6}{n(n-1)(n+1)^2}\sum_{i=1}^{n}\left[\left(n+1-q_i\right)^2 p_i + \left(n+1-p_i\right)^2 q_i\right]$$

$$\hat{\delta}_{(-i)} = \frac{2n-1}{n-2} - \frac{6}{(n-1)(n-2)n^2}\sum_{j \ne i}^{n}\left[\left(n-q_{ij}\right)^2 p_{ij} + \left(n-p_{ij}\right)^2 q_{ij}\right]$$

p_{ij} and q_{ij} are the ranks of x_j and y_j, respectively, of the i^th^ jackknife sample $S_{(-\mathrm{i})}$; $i=1,2,...,n$.

2.2 Proposed Correlation Coefficient

When some observations are lost and data with outliers, the estimator of the correlation measure (ρ) for the bivariate normal distribution is proposed in this section. The important problem for the average of observation y is the tail of distribution contains of outliers which can dominate its average. Hence, one technique for solving this problem, the small weights for observations in the tail of distribution are assigned and set more important weights to those near the central of distribution. This procedure dealing with the Winsorized distribution [30]. For $\mathrm{Q}_1-1.5(\mathrm{IQR}) \le y_i \le \mathrm{Q}_3+1.5(\mathrm{IQR})$, let function of Winsorized mean be given by equation (1).

$$\mathrm{F}(y_i) \;=\; y_i \tag{1}$$

For $y_i < \mathrm{Q}_1-1.5(\mathrm{IQR})$ or $y_i > \mathrm{Q}_3+1.5(\mathrm{IQR})$, let the function of the Winsorized mean be computed by equation (2).

$$\mathrm{F}(y_i) \;=\; \frac{\sum_{i=1}^{r} w_i^* y_i}{\sum_{i=1}^{r} w_i^*} \tag{2}$$

where IQR is the interquartile range of y values, Q_1 and Q_3 are the 1st and 3rd quartiles of data y respectively, and w_i^* is the weight function of y which is denoted by equation (3).

$$w^*(\mathrm{y}_i) = \begin{cases} 0 \text{ for } y_i < \mathrm{Q}_1 - 1.5\,\mathrm{IQR} \text{ or} \\ \qquad\quad y_i > \mathrm{Q}_3 + 1.5\,\mathrm{IQR} \\ 1 \text{ for } \mathrm{Q}_1 - 1.5\,\mathrm{IQR} \le y_i \le \mathrm{Q}_3 + 1.5\,\mathrm{IQR} \end{cases} \tag{3}$$

or $w^*(y_i)$ can be symbolized with w_i^* as illustrated in equation (3); $\mathrm{i} = 1, 2, ..., r$. Let W_i be the random sample after make a Winsorization and distribution of these as same as distribution of Y_i. The values for W_i can be given in the form of equation (4).

$$w_i \;=\; \mathrm{F}(y_i) \tag{4}$$

The independent and identically distributed of all data pairs (x_i, w_i) are considered. In addition, the joint distribution of random variables X and W corresponds to the joint distribution of random variables X and Y [30]. The proposed estimator of ρ is called the weighted maximum likelihood (WML) correlation. This estimator is proved by using the likelihood function approach as given in Theorem 1.

Lemma 1 The bivariate random vector (X, W) is taken from a bivariate normal distribution with vector of mean $\mu = \begin{bmatrix} \mu_X \\ \mu_W \end{bmatrix}$, matrix of variance covariance $\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XW} \\ \sigma_{XW} & \sigma_W^2 \end{bmatrix}$, and ρ is correlation coefficient between X and W. The joint distribution of random variables X and W is assigned in equation (5).

$$f(x,\mathrm{w}) = \frac{1}{2\pi\sigma_X\sigma_W\sqrt{1-\rho^2}}E$$

$$;E = e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_W}{\sigma_X}\right)\left(\frac{w-\mu_W}{\sigma_W}\right) + \left(\frac{w-\mu_W}{\sigma_W}\right)^2\right\}} \tag{5}$$

for $x \in (-\infty, \infty)$ and $w \in (-\infty, \infty)$. There exists the marginal distribution of random variable X is given by equation (6).

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \quad \text{for } -\infty < x < \infty \tag{6}$$

Proof Assume (X, W) is the bivariate random vector with joint distribution in the form of equation (5). Let $A = \left(\dfrac{x-\mu_X}{\sigma_X}\right)$ and $B = \left(\dfrac{w-\mu_W}{\sigma_W}\right)$. Then, the marginal distribution of X can be derived as follows:

$$f(x) \;=\; \int_{-\infty}^{\infty} f(x,\mathrm{w})\,dw$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_W\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{A^2 - 2\rho AB + B^2\right\}} dw$$

$$=\frac{1}{\sqrt{2\pi}\sigma_X}e^{-\frac{1}{2}A^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_W\sqrt{1-\rho^2}}e^{-\frac{1}{2\left(1-\rho^2\right)}(B-\rho A)^2}dw \qquad (7)$$

Consider the term in equation (7) as follows:

$$\int_{-\infty}^{\infty}\frac{1}{\sigma_W\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{1}{2\left(1-\rho^2\right)}(B-\rho A)^2}dw.$$

This term can be written in the other form as shown in equation (8).

$$\int_{-\infty}^{\infty}\frac{1}{\sigma_W\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{1}{2\left(1-\rho^2\right)}(B-\rho A)^2}dw$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sigma_W\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{1}{2\sigma_W^2\left(1-\rho^2\right)}\left[w-\left\{\mu_W+\frac{\rho\sigma_W}{\sigma_X}\left(x-\mu_X\right)\right\}\right]^2}dw \qquad (8)$$

Since the formula

$$\frac{1}{\sigma_W\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{1}{2\sigma_W^2\left(1-\rho^2\right)}\left[w-\left\{\mu_W+\frac{\rho\sigma_W}{\sigma_X}(x-\mu_X)\right\}\right]^2}$$

is the function of probability density of the normal distribution that has mean $\mu_W+\frac{\rho\sigma_W}{\sigma_X}\left(x-\mu_X\right)$ and variance $\sigma_W^2\left(1-\rho^2\right)$. Therefore, the integration of this normal function over w from $-\infty$ to ∞, it's equal to one and equation (8) can be written as equation (9).

$$\int_{-\infty}^{\infty}\frac{1}{\sigma_W\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{1}{2\left(1-\rho^2\right)}(B-\rho A)^2}dw=1 \qquad (9)$$

By substituting equation (9) in equation (7), the marginal probability density function of X is obtained. That is, $f\left(x\right)=\frac{1}{\sqrt{2\pi}\sigma_X}e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}$ for $-\infty<x<\infty$. Hence, X is normally distributed with the average and variance are equal to μ_X and σ_X^2 respectively.

Lemma 2 Assume $\left(X,W\right)$ is the vector of random bivariate with distribution having vector of mean as $\mu=\begin{bmatrix}\mu_X\\ \mu_W\end{bmatrix}$, matrix of variance covariance is given as $\Sigma=\begin{bmatrix}\sigma_X^2 & \sigma_{XW}\\ \sigma_{XW} & \sigma_W^2\end{bmatrix}$, and ρ is the correlation coefficient of *X* and *W*. A joint distribution of *X* and *W* is shown in equation (5). For some *x* such that $f\left(x\right)$ is greater than zero, the conditional distribution of random variable *W* given *X*=*x* is shown as equation (10).

$$f(w\,|\,x)=\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{W|X}}}e^{-\frac{1}{2}\left(\frac{w-\mu_{\mathrm{W|X}}}{\sigma_{\mathrm{W|X}}}\right)^2}\text{ for }-\infty<w<\infty$$

where $\mu_{\mathrm{W|X}}=\beta_0+\beta_1x,\ \beta_0=\mu_{\mathrm{W}}-\beta_1\mu_X,$

$\beta_1=\rho\frac{\sigma_W}{\sigma_X}$ and $\sigma_{\mathrm{W|X}}=\sigma_{\mathrm{W}}\sqrt{1-\rho^2}$ (10)

Proof Assume $\left(X,W\right)$ is the bivariate random vector with the joint distribution $f\left(x,\mathrm{w}\right)$ as illustrate in equation (5). The marginal distribution of *X* is illustrated in equation (6). For some *x* such that $f\left(x\right)$ is greater than zero, the conditional distribution of random variable *W* given *X*=*x* is written in the form of equation (11).

$$f(\mathrm{w}\,|\,x)=\frac{f(x,w)}{f(x)} \qquad (11)$$

From Lemma 1, substitute equations (5) and (6) into equation (11), this conditional density function is given as follows:

$$f(\mathrm{w}\,|\,x)=\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{W|X}}}e^{-\frac{1}{2\sigma_{\mathrm{W|X}}^2}\left[w-\mu_{\mathrm{W|X}}\right]^2}$$

for $\mu_{\mathrm{W|X}}=\beta_0+\beta_1x$, $\beta_1=\rho\frac{\sigma_W}{\sigma_X}$, $\beta_0=\mu_{\mathrm{W}}-\beta_1\mu_X$ and $\sigma_{\mathrm{W|X}}=\sigma_{\mathrm{W}}\sqrt{1-\rho^2}$. Hence, the variable *W* given $X=x$ is distributed of normal with average equals $\mu_{\mathrm{W|X}}=\beta_0+\beta_1x$ and variance equals $\sigma_{\mathrm{W|X}}^2=\sigma_{\mathrm{W}}^2\left(1-\rho^2\right)$.

Theorem 1 The random vector (X,Y) is taken from bivariate normal distribution with vector of mean $\mu=\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix}$, matrix of variance covariance equals $\Sigma=\begin{bmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{bmatrix}$, and ρ is the correlation coefficient of *X* and *Y*. Suppose that data (x,y) are non-missing observed for *r* pairs and the rest $n-r$ observations (for r between zero and n) on *y* are lost. Furthermore, data *y* are supposed missing completely at random mechanism and they are composed of mild outliers. Let W be a random variable with the same distribution as the random variable Y after Winsorization of the data. The values for W_i are written by $w_i=\mathrm{F}(y_i)$; $i=1, 2,\ldots,r$ where the function $\mathrm{F}(\mathrm{y}_i)$ are shown in equation (1) or (2). Then, the weighted maximum likelihood correlation coefficient estimator of ρ is computed in equation (12). This estimator is called the weighted maximum likelihood (WML) correlation coefficient and symbolized by r_{WML}.

$$r_{WML}=\frac{\hat{\beta}_1\hat{\sigma}_X}{\hat{\sigma}_W} \qquad (12)$$

where $\hat{\beta}_1=\dfrac{\sum_{i=1}^{r}x_iw_i-r\bar{x}'\bar{w}'}{\sum_{i=1}^{r}x_i^2-r\bar{x}'^2}$, $\hat{\sigma}_X=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$,

$$\hat{\sigma}^2_{\mathrm{W|X}}=\frac{1}{r}\sum_{i=1}^{r}\left\{(w_i-\bar{w}')-\hat{\beta}_1(x_i-\bar{x}')\right\}^2,$$

$$\hat{\sigma}_W=\sqrt{\hat{\sigma}^2_{\mathrm{W|X}}+\hat{\sigma}^2_X\hat{\beta}_1^2},\quad \bar{w}'=\frac{\sum_{i=1}^{r}w_i}{r},\quad \bar{x}'=\frac{\sum_{i=1}^{r}x_i}{r}$$

and $\bar{x}=\dfrac{\sum_{i=1}^{n}x_i}{n}$.

Proof Suppose the random vector (X,Y) is taken from bivariate normal distribution. The value of random variable *W* is symbolized as $w_i=\mathrm{F}(y_i)$; $i=1, 2, \ldots, r$ and $\mathrm{F}(y_i)$ are illustrated in equation (1) or (2). Hence, the random vector (X,W) is bivariate normally distributed and the joint distribution of variables *X* and *W* is explained in equation (4). The likelihood function of vector parameter $\underline{\theta}=(\mu_X, \mu_W, \rho, \sigma_X^2, \sigma_W^2)$ can be demonstrated as equation (13).

$$L(\underline{\theta}\,|\,x_1,x_2,\ldots,x_n,\mathrm{w}_1,\mathrm{w}_2,\ldots,\mathrm{w}_r)=\prod_{i=1}^{r}f(\mathrm{x}_i,\mathrm{w}_i)\prod_{i=r+1}^{n}f(\mathrm{x}_i)$$

$$=\prod_{i=1}^{n}f(\mathrm{x}_i)\prod_{i=1}^{r}f(\mathrm{w}_i\,|\,x_i) \qquad (13)$$

From Lemmas 1 and 2, equations (6) and (10) are substituted into equation (13), then equation (14) is obtained.

$$L(\underline{\theta}\,|\,x_1,x_2,\ldots,x_n,\mathrm{w}_1,\mathrm{w}_2,\ldots,\mathrm{w}_r)=C\times D$$
(14)

where

$$C=\left(2\pi\sigma_X^2\right)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n}(x_i-\mu_X)^2},$$

$$D=\left(2\pi\sigma^2_{\mathrm{W|X}}\right)^{-\frac{r}{2}}e^{-\frac{1}{2\sigma^2_{\mathrm{W|X}}}\sum_{i=1}^{r}\left[w_i-(\beta_0+\beta_1x_i)\right]^2},$$

$\beta_0=\mu_{\mathrm{W}}-\beta_1\mu_X$, $\beta_1=\rho\dfrac{\sigma_W}{\sigma_X}$ and $\sigma^2_{\mathrm{W|X}}=\sigma_W^2(1-\rho^2)$. The log likelihood function of equation (14) can be proved in equation (15).

$$\ln L(\underline{\theta})=\ln L(\underline{\theta}\,|\,x_1,x_2,\ldots,x_n,\mathrm{w}_1,\mathrm{w}_2,\ldots,\mathrm{w}_r)$$

$$=\frac{-n}{2}\ln\left(2\pi\sigma_X^2\right)+\frac{-1}{2\sigma_X^2}\sum_{i=1}^{n}(x_i-\mu_X)^2$$

$$+\frac{-r}{2}\ln\left(2\pi\sigma^2_{\mathrm{W|X}}\right)+\frac{-1}{2\sigma^2_{\mathrm{W|X}}}\sum_{i=1}^{r}\left[w_i-(\beta_0+\beta_1x_i)\right]^2$$

(15)

The maximum likelihood estimators of μ_X, σ_X^2, β_0, β_1 and $\sigma^2_{\mathrm{W|X}}$ will be proved on taking the partial derivative of the log-likelihood function as shown in equation (15) with respect to each of parameters μ_X, σ_X^2, β_0, β_1 and $\sigma^2_{\mathrm{W|X}}$, and assigning to zero for each. These can be derived as follows:

$$\frac{\partial}{\partial\mu_X}\ln L(\underline{\theta})=\frac{\partial}{\partial\mu_X}\left(\frac{-1}{2\sigma_X^2}\sum_{i=1}^{n}(x_i-\mu_X)^2\right)=0$$

$$=\frac{-1}{2\sigma_X^2}\left(-2\sum_{i=1}^{n}(x_i-\mu_X)\right)=0$$

So that $\mu_X = \dfrac{\sum_{i=1}^{n} x_i}{n}$ (16)

Additionally, the maximum likelihood estimator of parameter μ_X is symbolized by $\hat{\mu}_X = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$.

Next, considering the partial derivative of the log-likelihood function with respect to parameter σ_X^2.

$$\frac{\partial}{\partial \sigma_X^2} \ln L(\underline{\theta}) = \frac{\partial}{\partial \sigma_X^2}\left(-\frac{n}{2}\ln\left(2\pi\sigma_X^2\right) + \frac{-1}{2\sigma_X^2}\sum_{i=1}^{n}\left(x_i - \mu_X\right)^2 \right) = 0$$

$$= -\frac{n}{2}\left(\frac{2\pi}{2\pi\sigma_X^2}\right) + \frac{1}{2\left(\sigma_X^2\right)^2}\sum_{i=1}^{n}\left(x_i - \mu_X\right)^2 = 0$$

$= \dfrac{1}{2\sigma_X^2}\left(-n + \dfrac{1}{\sigma_X^2}\sum_{i=1}^{n}\left(x_i - \mu_X\right)^2 \right) = 0.$ After rearrange this formula, the variance of X is computed as $\sigma_X^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \mu_X\right)^2}{n}$ and μ_X is replaced with equation (16). Hence, the maximum likelihood estimator of parameter σ_X^2 is denoted by symbol $\hat{\sigma}_X^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$. Moreover, the estimators of parameters β_0 and β_1 can be derived by the same way as previous proof. That is,

$$\frac{\partial}{\partial \beta_0} \ln L(\underline{\theta}) = \frac{\partial}{\partial \beta_0}\left(\frac{-1}{2\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left[w_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2 \right) = 0$$

$$= \frac{-2}{2\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left[w_i - \beta_0 - \beta_1 x_i\right](-1) = 0$$

$$\sum_{i=1}^{r} w_i - r\beta_0 - \beta_1 \sum_{i=1}^{r} x_i = 0$$
(17)

$$\frac{\partial}{\partial \beta_1} \ln L(\underline{\theta}) = \frac{\partial}{\partial \beta_1}\left(\frac{-1}{2\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left[w_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2 \right) = 0$$

$$= \frac{-2}{2\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left[w_i - \beta_0 - \beta_1 x_i\right](-x_i) = 0$$

$$\sum_{i=1}^{r} x_i w_i - \beta_0 \sum_{i=1}^{r} x_i - \beta_1 \sum_{i=1}^{r} x_i^2 = 0$$
(18)

Equation (17) × $\sum_{i=1}^{r} x_i$, it is written in equation (19).

$$\sum_{i=1}^{r} x_i \sum_{i=1}^{r} w_i - r\beta_0 \sum_{i=1}^{r} x_i - \beta_1\left(\sum_{i=1}^{r} x_i\right)^2 = 0 \qquad (19)$$

Equation (18) × r, it is written in equation (20).

$$r\sum_{i=1}^{r} x_i w_i - r\beta_0 \sum_{i=1}^{r} x_i - r\beta_1 \sum_{i=1}^{r} x_i^2 = 0$$
(20)

Solving equation (19) and equation (20), then the maximum likelihood estimators of parameters β_0 and β_1 can be symbolized as $\hat{\beta}_0 = \bar{w}' - \hat{\beta}_1 \bar{x}'$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{r} x_i w_i - r\bar{x}'\bar{w}'}{\sum_{i=1}^{r} x_i^2 - r\bar{x}'^2}$, respectively. Further, a partial derivative of the log-likelihood function with respect to parameter $\sigma_{\mathrm{W|X}}^2$ and setting it to zero that can be proved as

$$\frac{\partial}{\partial \sigma_{\mathrm{W|X}}^2} \ln L(\underline{\theta})$$

$$= \frac{\partial}{\partial \sigma_{\mathrm{W|X}}^2}\left(-\frac{r}{2}\ln\left(2\pi\sigma_{\mathrm{W|X}}^2\right) + \frac{-1}{2\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left[w_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2 \right)$$

$$= 0$$

$$= -\frac{r}{2}\left(\frac{2\pi}{2\pi\sigma_{\mathrm{W|X}}^2}\right) + \frac{1}{2\left(\sigma_{\mathrm{W|X}}^2\right)^2}\sum_{i=1}^{r}\left(w_i - \beta_0 - \beta_1 x_i\right)^2 = 0$$

$$= -\frac{1}{2\sigma_{\mathrm{W|X}}^2}\left\{ r - \frac{1}{\sigma_{\mathrm{W|X}}^2}\sum_{i=1}^{r}\left(w_i - \beta_0 - \beta_1 x_i\right)^2 \right\} = 0$$

$$\sigma_{\mathrm{W|X}}^2 = \frac{1}{r}\sum_{i=1}^{r}\left(w_i - \beta_0 - \beta_1 x_i\right)^2 \qquad (21)$$

The maximum likelihood estimator of parameter $\sigma_{\mathrm{W|X}}^2$ can be calculated by replacing β_0 with its estimator as $\hat{\beta}_0 = \bar{w}' - \hat{\beta}_1 \bar{x}'$ in equation (21) and rearrange the formula. Therefore, the maximum likelihood estimator of parameter $\sigma_{\mathrm{W|X}}^2$ is denoted as $\hat{\sigma}^2_{\mathrm{W|X}} = \dfrac{1}{r}\sum_{i=1}^{r}\left(w_i - \left(\bar{w}' - \hat{\beta}_1\bar{x}'\right) - \hat{\beta}_1 x_i\right)^2$ or

$$\hat{\sigma}^2_{\mathrm{W|X}} = \frac{1}{r}\sum_{i=1}^{r}\left\{\left(w_i - \bar{w}'\right) - \hat{\beta}_1\left(x_i - \bar{x}'\right)\right\}^2.$$

In equation (10), parameters β_1 and $\sigma^2_{\mathrm{W|X}}$ can be written in the formulas of equation (22) and (23).

$$\beta_1 = \rho\frac{\sigma_W}{\sigma_X} \quad \text{or} \quad \rho = \beta_1\frac{\sigma_X}{\sigma_W} \qquad (22)$$

$$\sigma^2_{\mathrm{W|X}} = \left(1-\rho^2\right)\sigma^2_W \qquad (23)$$

Equation (22) is substituted into equation (23), then equation (24) is obtained.

$$\sigma^2_{\mathrm{W|X}} = \sigma^2_W\left[1-\left(\beta_1\frac{\sigma_X}{\sigma_W}\right)^2\right] = \sigma^2_W\left(1-\beta_1^2\frac{\sigma^2_X}{\sigma^2_W}\right)$$

$$= \sigma^2_W - \beta_1^2\sigma^2_X \qquad (24)$$

The variance of random variable *W* can be written in the formula $\sigma^2_W = \sigma^2_{\mathrm{W|X}} + \beta_1^2\sigma^2_X$. In equation (22), we know that $\rho = \beta_1\dfrac{\sigma_X}{\sigma_W}$. Due to a result of invariance feature for the maximum likelihood estimator, then the WML correlation coefficient of ρ is symbolized by r_{WML} and this estimator can be written in the formula $r_{WML} = \dfrac{\hat{\beta}_1\hat{\sigma}_X}{\hat{\sigma}_W}$. Similarly, the maximum likelihood estimator of parameter σ_W can be proved by using its invariance feature, and the WML estimator of parameter σ_W is given in the formula $\hat{\sigma}_W = \sqrt{\hat{\sigma}^2_{\mathrm{W|X}} + \hat{\beta}_1^2\hat{\sigma}^2_X}$.

3 Results of the Simulation Study

A performance validity of the proposed estimator is performed to compare the reliability in term of robustness properties through the simulated data. The comparison of six estimators, namely r_{WML}, r_P, r_M, r_{PB}, r_B and r_C, are investigated through the simulated data across 240 situations. In this study, the bivariate normal distributed data are generated by Monte Carlo technique. The simulated data are composed of mild outliers [28] and missing observations in variable Y for 10% and 20% of the total cases. The performance comparisons of the six estimators are considered from two features: bias and mean squared error. The random sample (x,y) of six sizes as 10, 20, 30, 40, 50 and 100 are taken from the standard bivariate normal distribution with the correlation measure coefficient ρ equals 0, 0.1, 0.2, …, 0.9. The results of this study are illustrated in Fig. 2 to Fig. 13. Consider Fig.2, it is found that bias of the proposed correlation measure—r_{WML}— tends to be smaller than those of four methods—r_P, r_M, r_{PB} and r_B—for sample size (n) equals 10 and outliers equal 20% regardless of missing data levels and levels of ρ. Moreover, for sample size equals 10 and outliers in dataset equals 20%, it is found that the mean squared error of WML estimator gives the smallest values for all levels of correlation coefficients ρ that show in Fig. 3. That is, the r_{WML} seems to be the best method or it provides both of the smallest bias and mean squared error for a sample size equals 10 when outliers equals 20% regardless levels of ρ and percentages of missing data. For a sample size equals 20 as shows in Fig. 4, the composite and median correlation coefficients tend to have the smallest biases. However, mean squared error of the WML estimator tends to be less than those of the composite and median correlation coefficients for ρ approximates not greater than 0.5 regardless levels of missing data and outliers in the dataset as shows in Fig. 5. For larger sample sizes, i.e., n = 30, 40, 50, 100, the results are shown in Fig. 6, 8, 10 and 12, these are found that bias of the median correlation coefficients tend to be lower than those of five estimators—r_P, r_{WML}, r_{PB}, r_B and r_C— irrespective of outliers and missing data levels. However, the mean squared error of WML estimator seems to have the lowest value for the small correlation coefficients ρ, e.g., ρ = 0.1, 0.2, 0.3, 0.4, irrespectively of percentages of outliers and missing data which these results are shown in Fig. 7, 9, 11 and 13. For the largest sample size $(n=100)$, Fig. 12 to Fig. 13 indicate that r_M seems to have a good properties for both criteria, namely, bias and mean squared error. That is, it tends to give the small bias and mean squared error for all the levels of outliers and missing whatever the level of ρ will be.

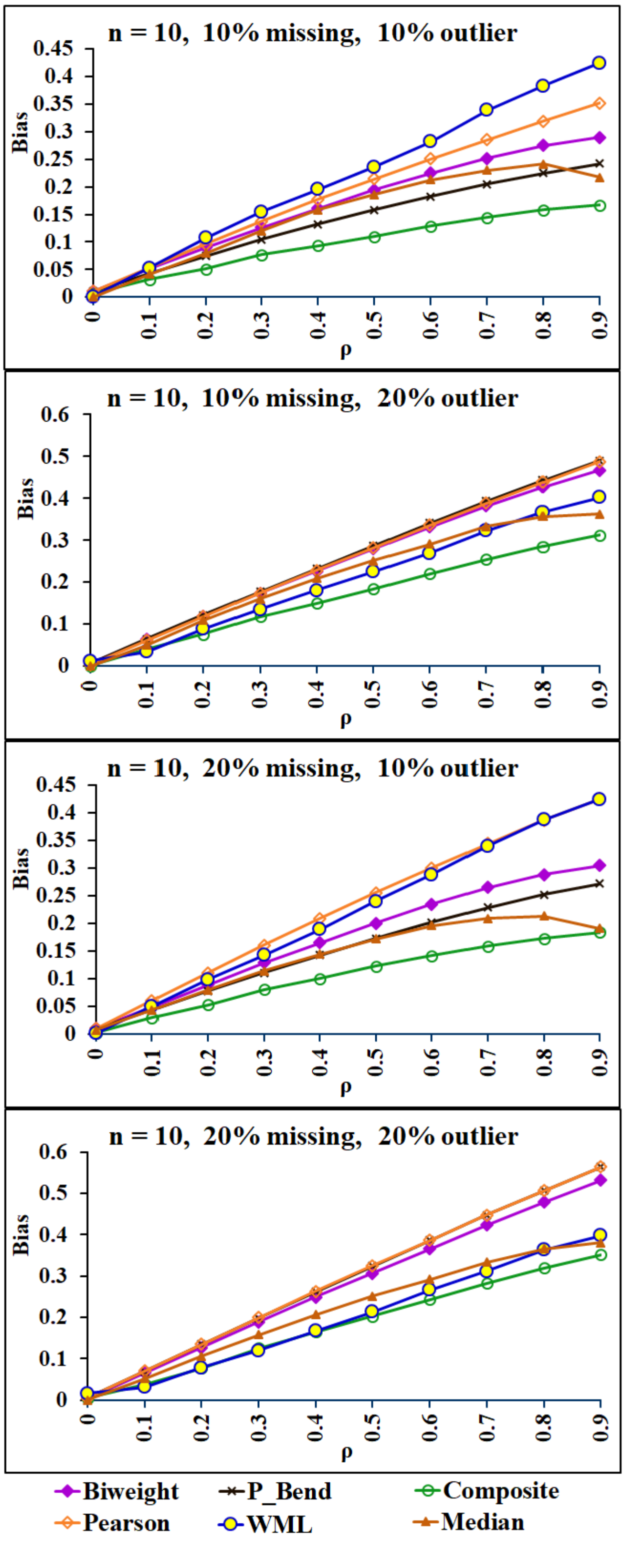

Fig.2: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when n=10.

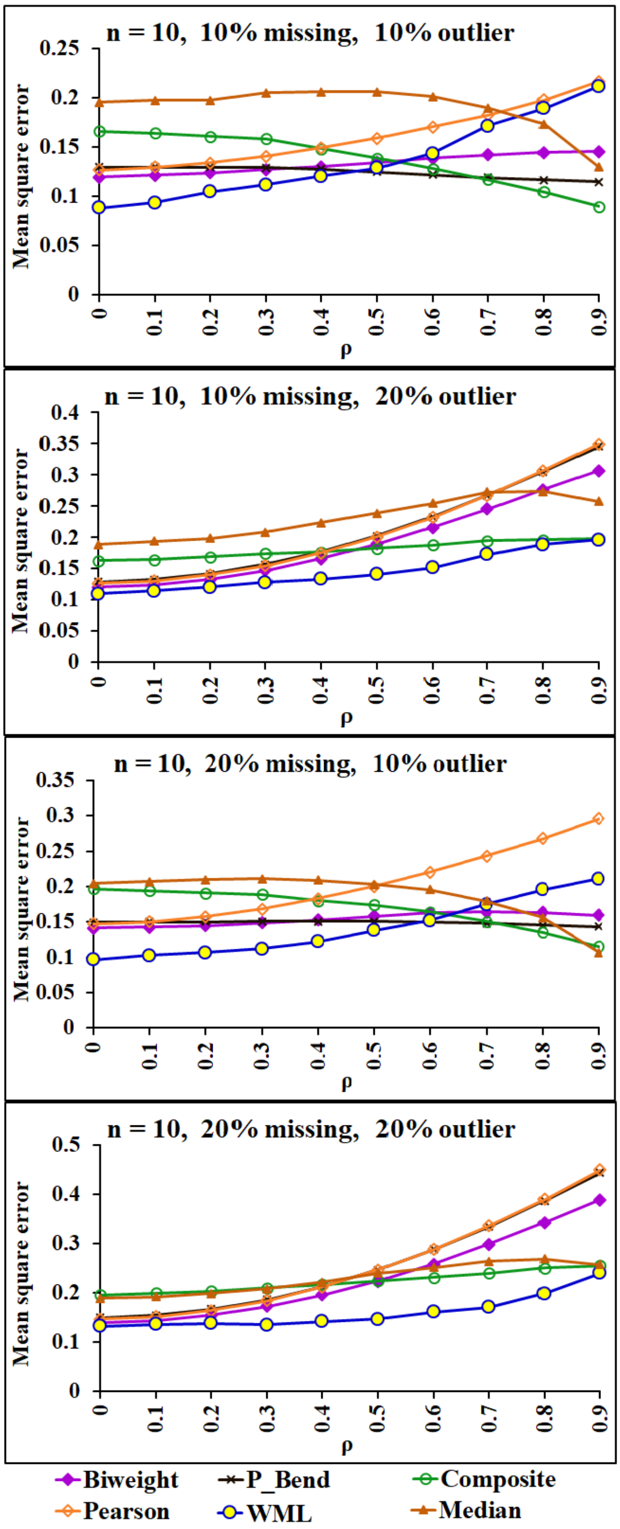

Fig.3: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=10.

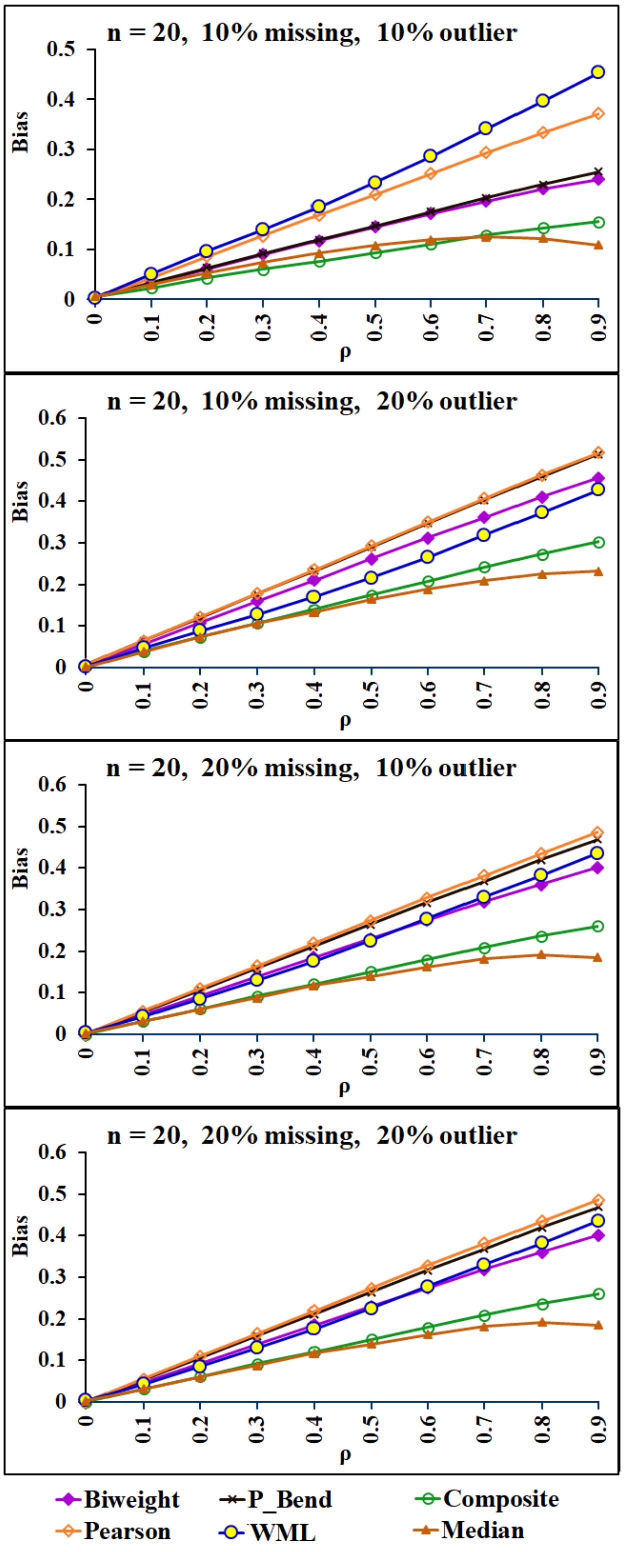

Fig.4: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when n=20.

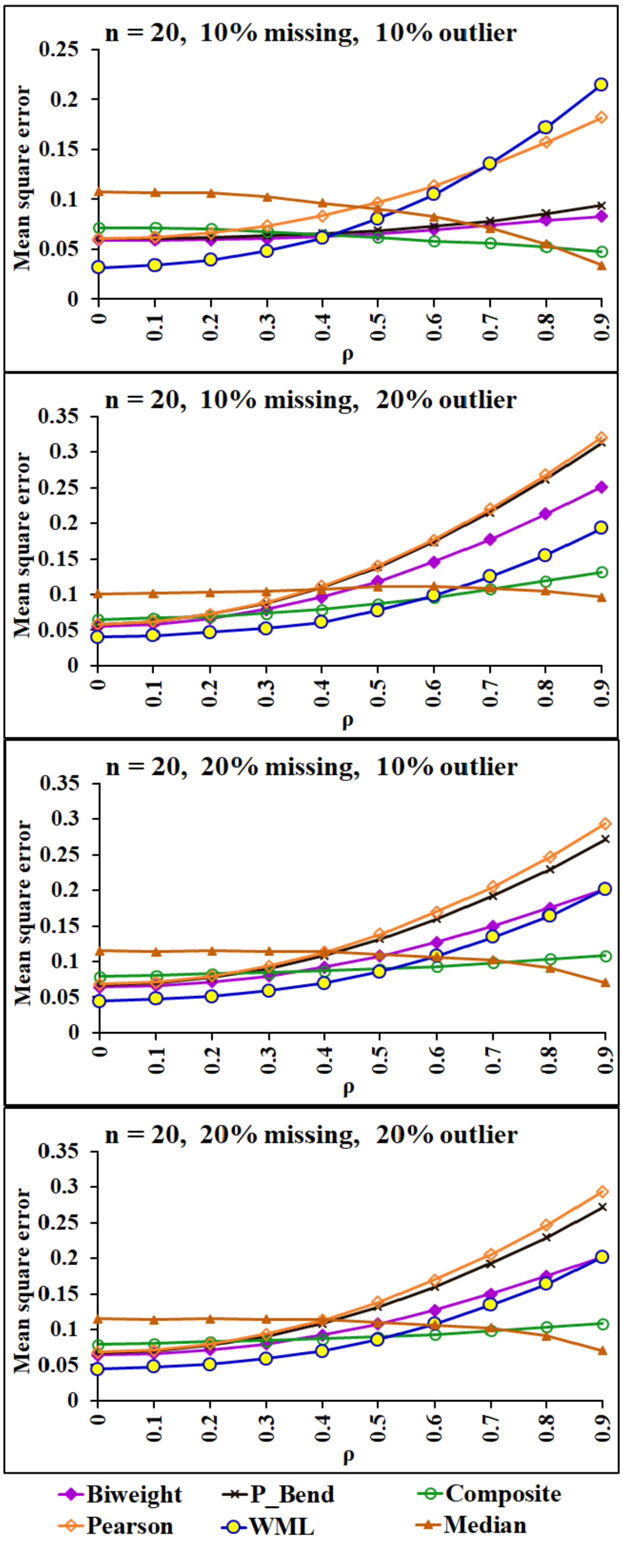

Fig.5: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=20.

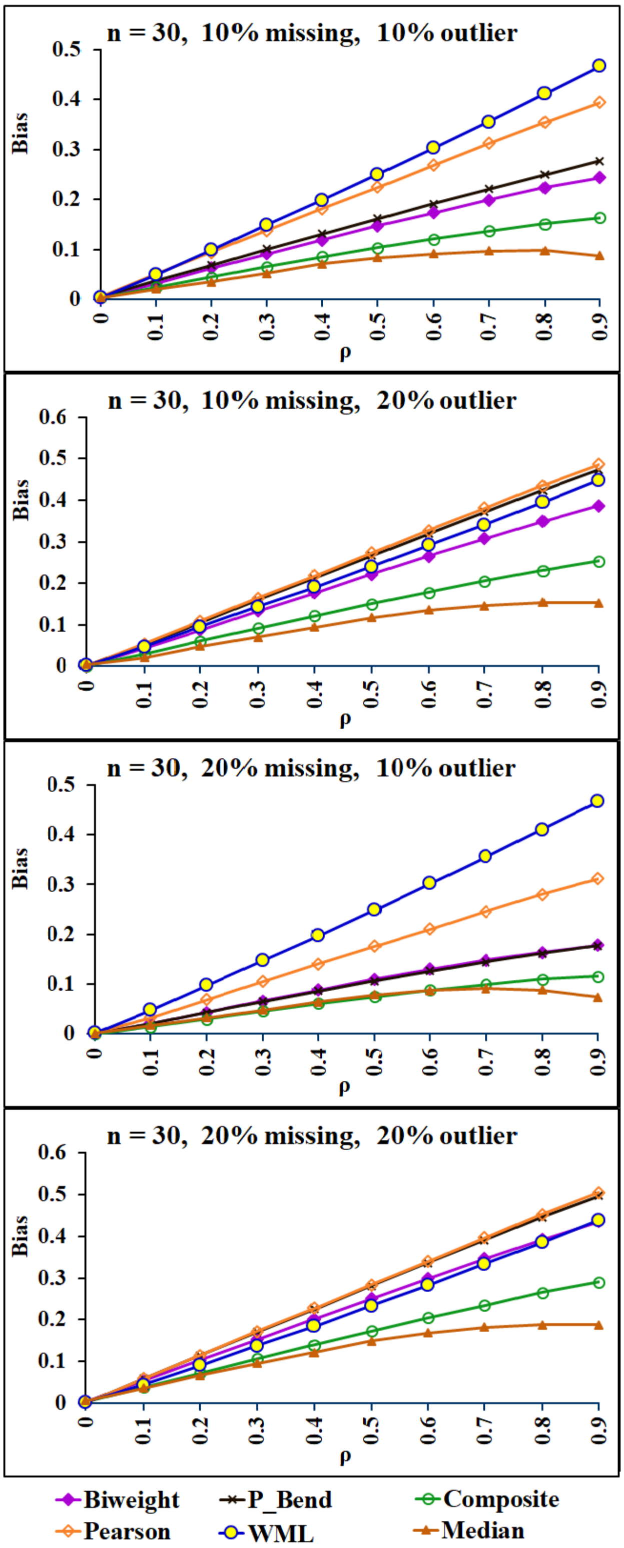

Fig.6: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when n=30.

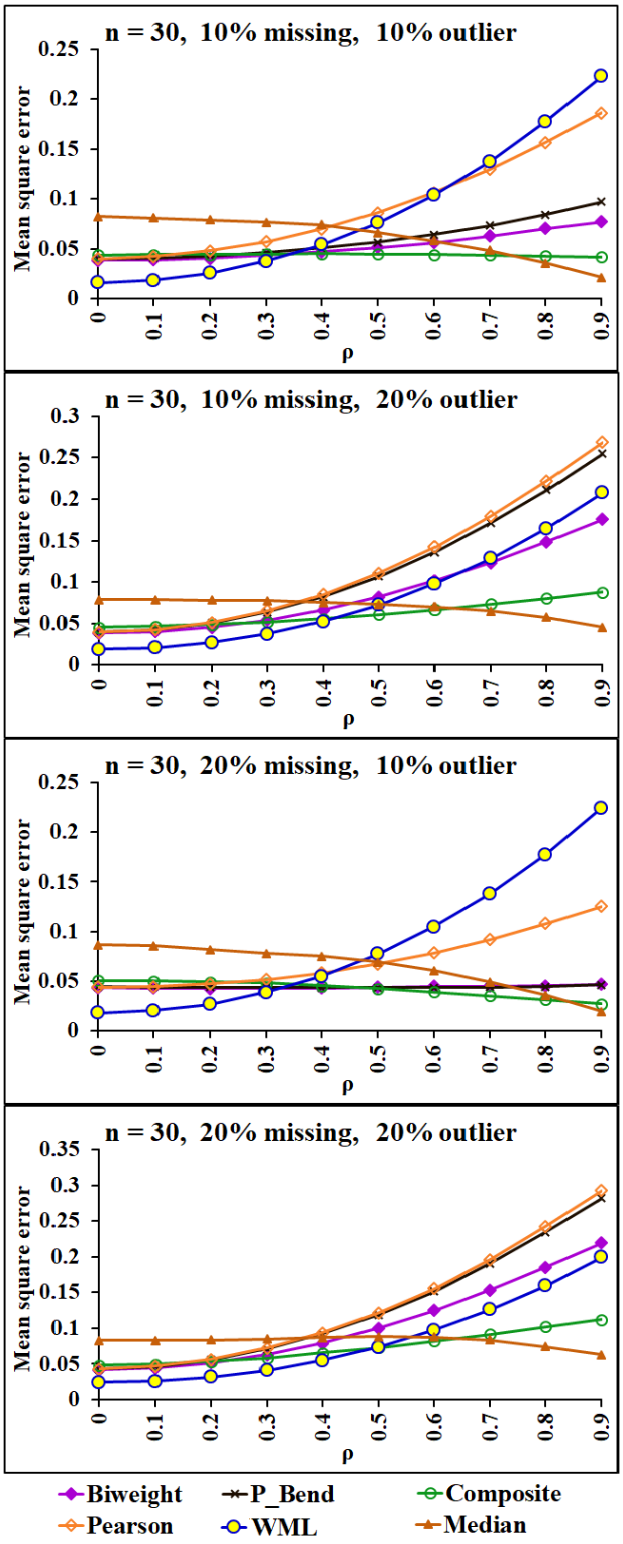

Fig.7: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=30.

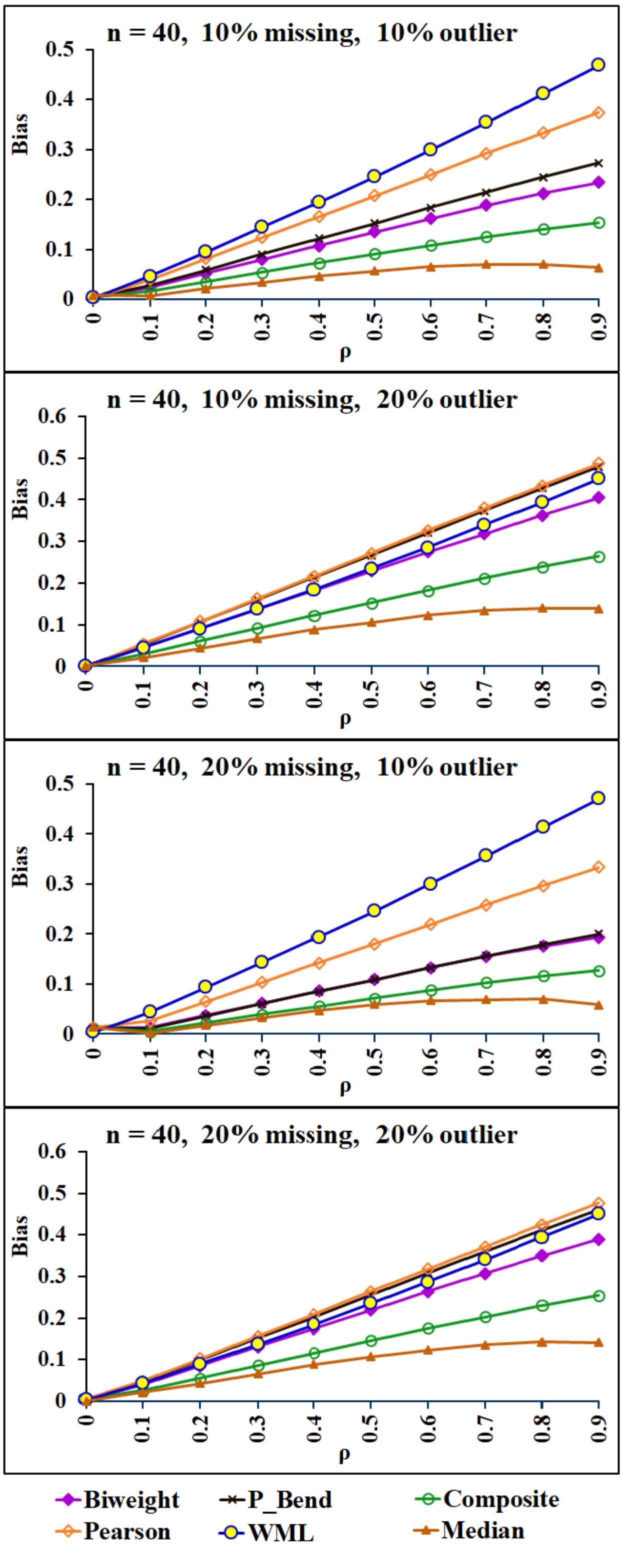

Fig.8: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when when n=40.

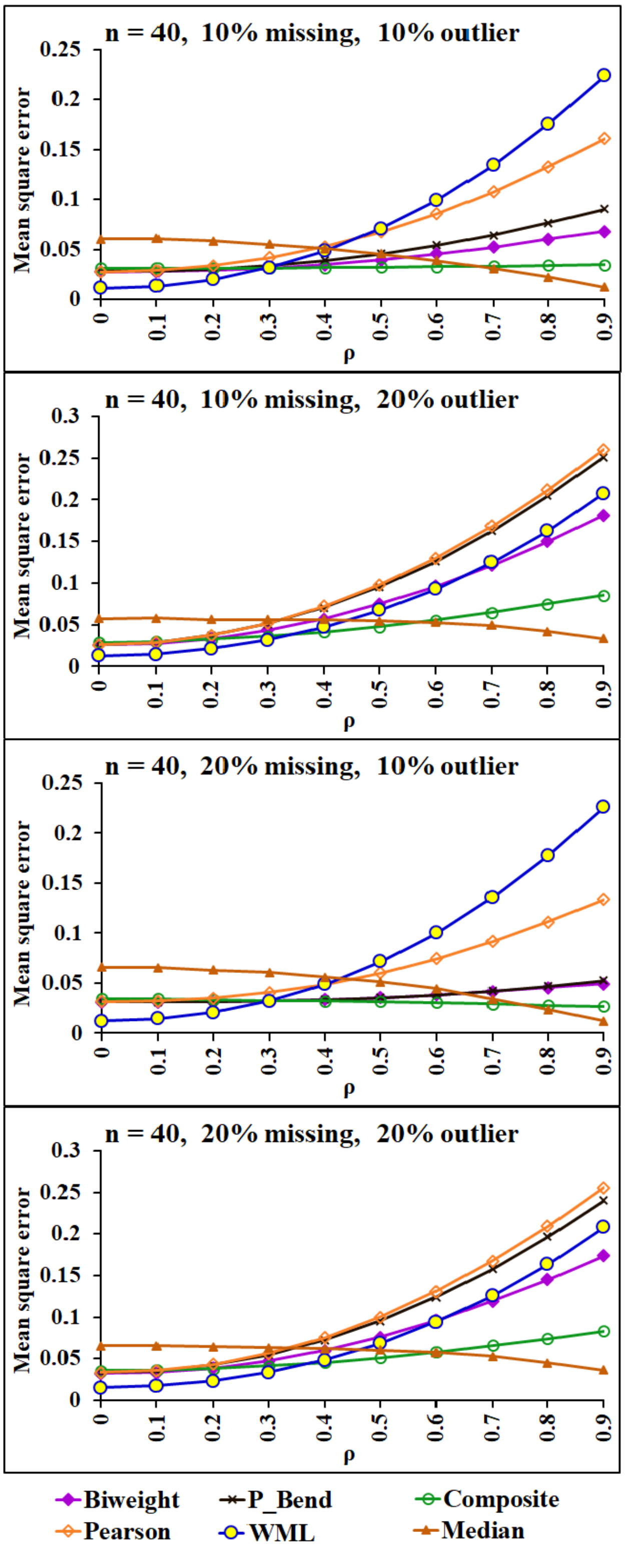

Fig.9: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=40.

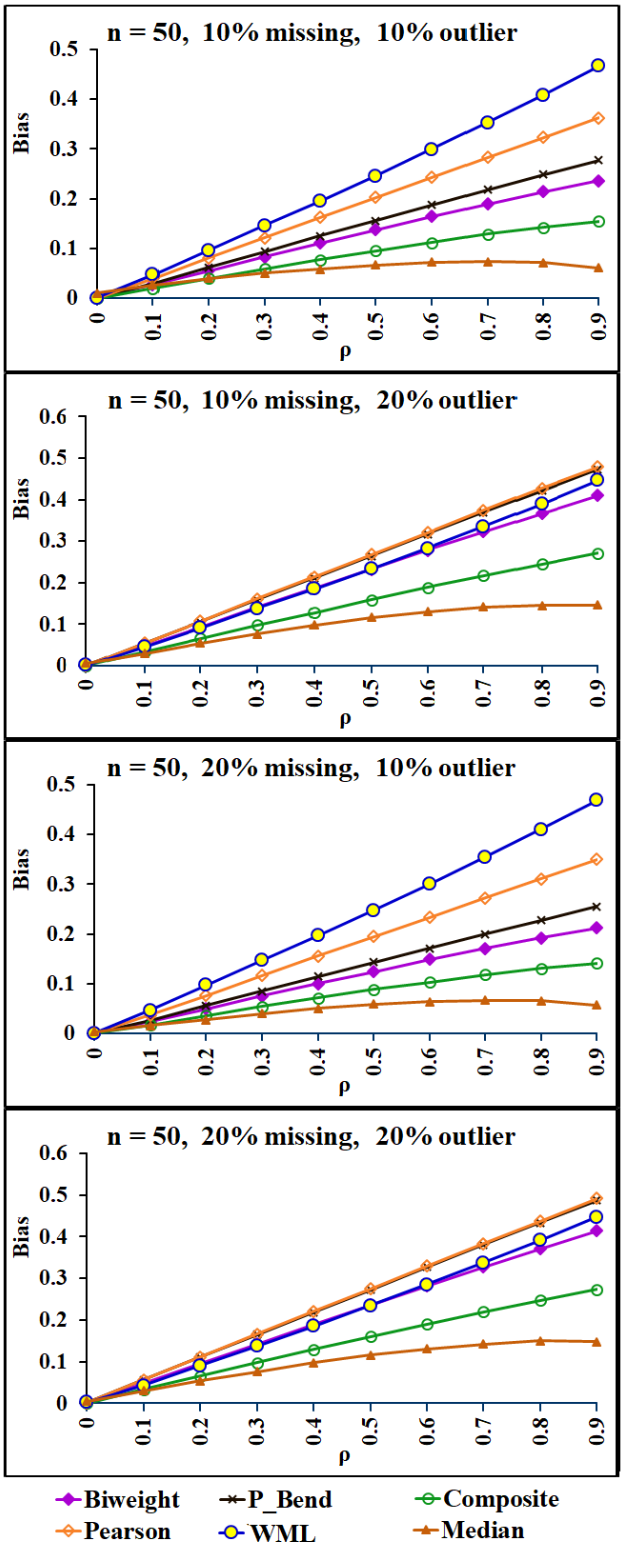

Fig.10: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when n=50.

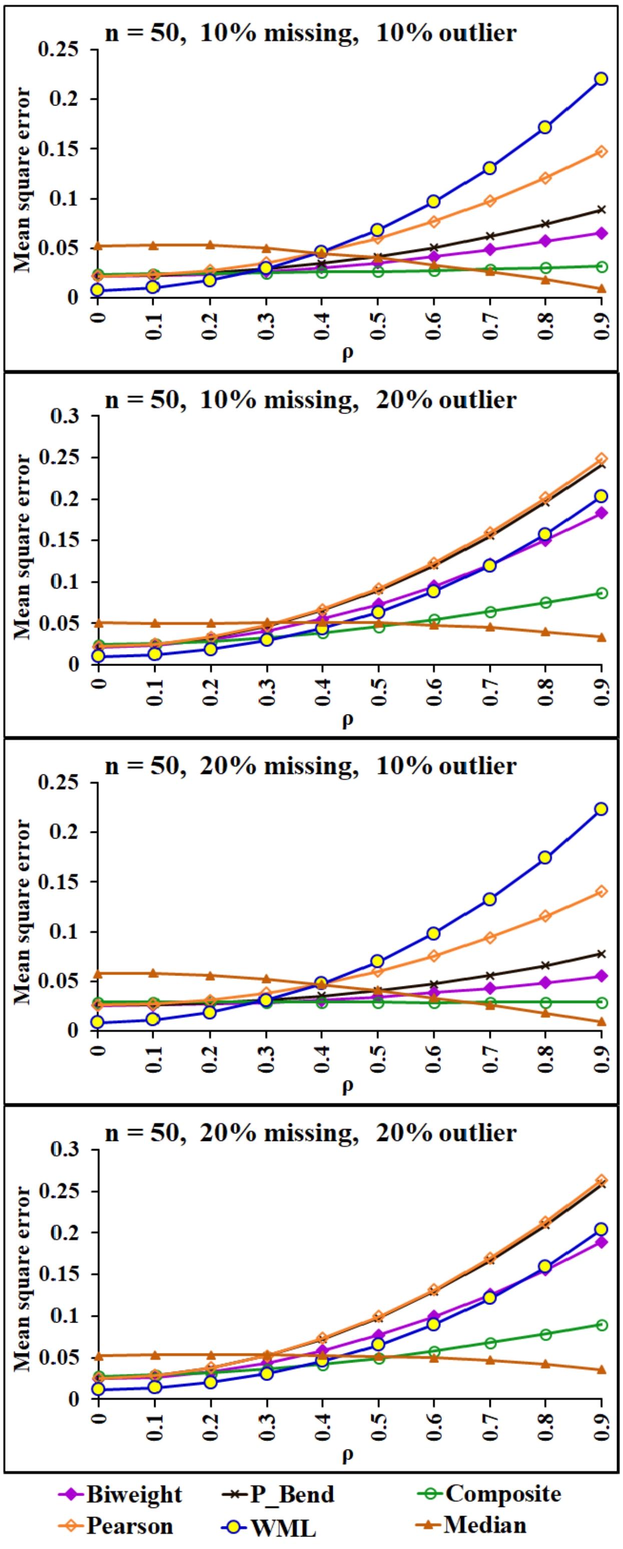

Fig.11: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=50.

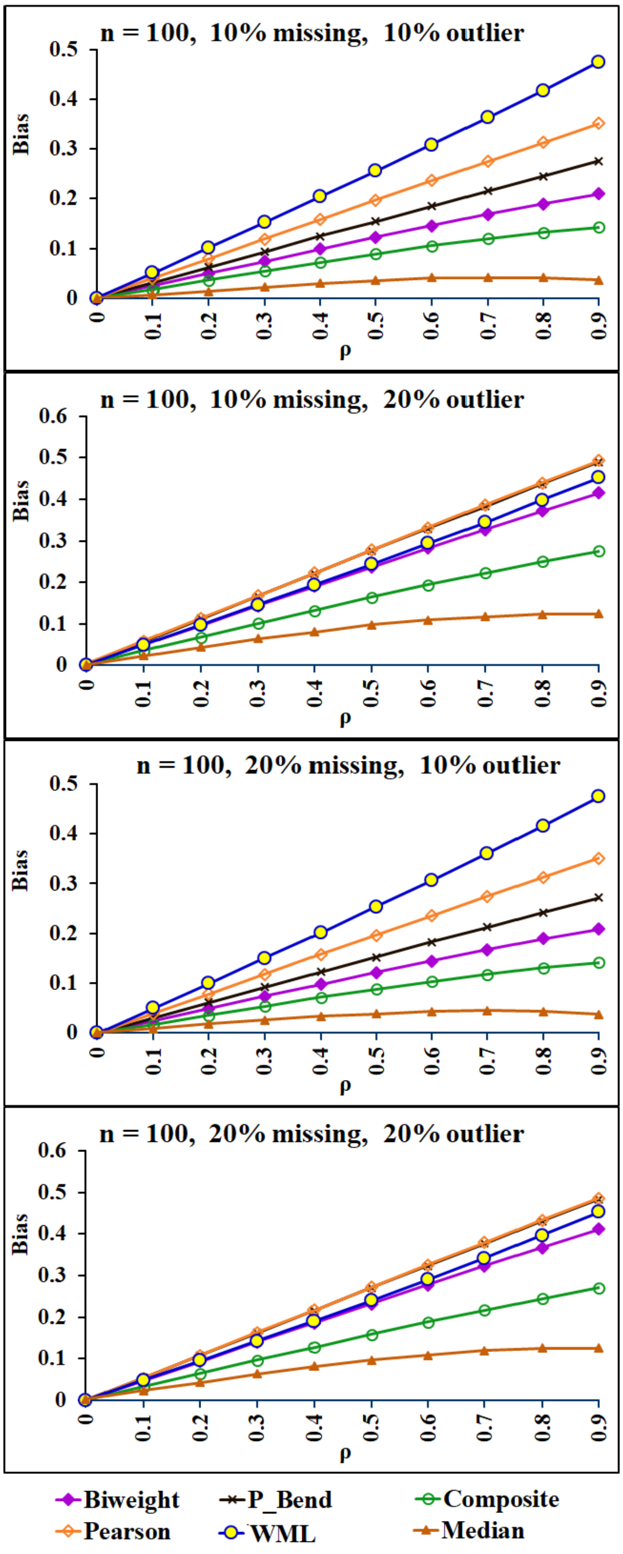

Fig.12: Biases comparison of the six estimators for different levels of missing and outliers in the sample data when n=100.

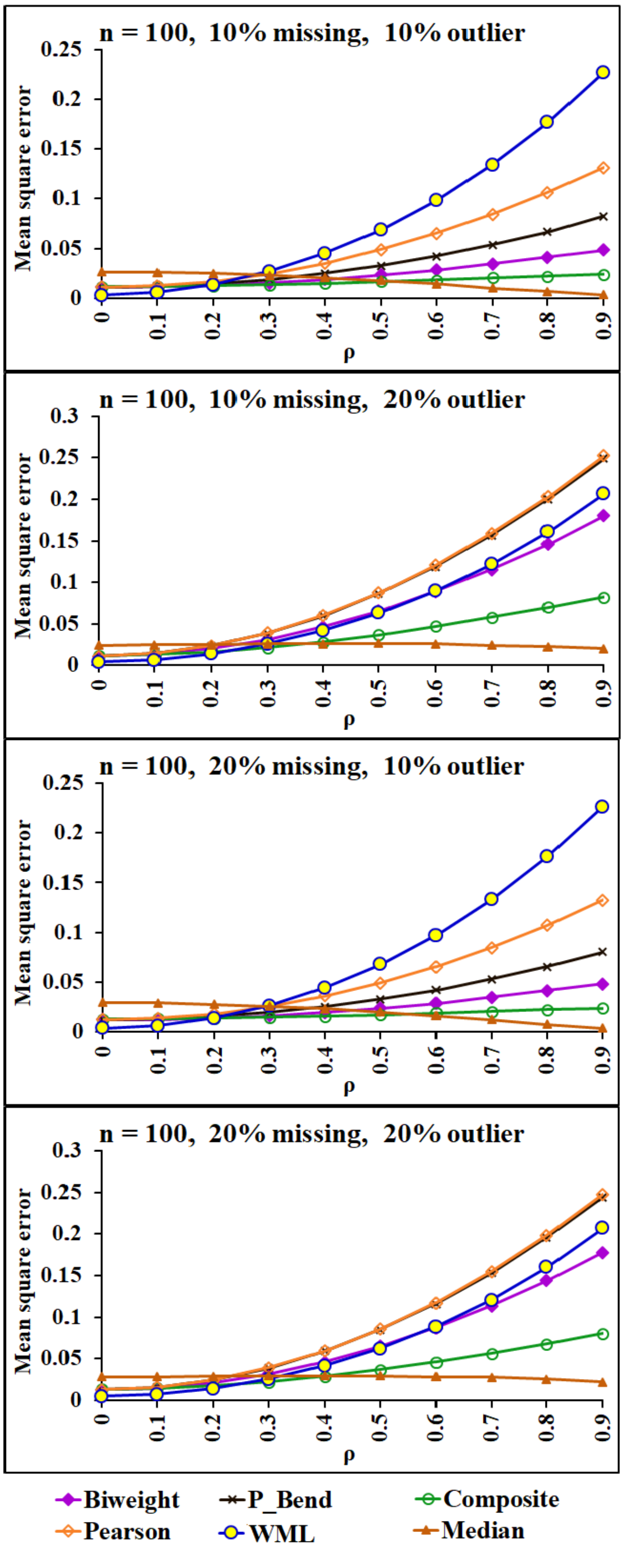

Fig.13: Dipersion comparison of six estimators for different levels of outliers and missing in the sample data when n=100.

4 Discussion

From the results of this research, the median correlation coefficient tends to provide the best performance even if some observations are lost and outliers occur in the dataset. As a results of this method using the robust estimator, i.e., sample median and sample median absolute deviations, then outliers are not effect in the accuracy of estimation. In other words, this method give the lowest bias and mean squared error when the sizes of sample are large by comparing with the other five methods. This study result conforms to the research of [22]. Additionally, the Pearson correlation coefficient is the most ineffective estimator for estimating the relationship between two random variables for the bivariate normal distribution when some observations are lost and outliers occur in the dataset. The result of this research is consistent with the studies of [20-23, 25, 31-32]. Because Pearson estimator based on the sample mean of two variables and it is a sensitive estimator for outliers, hence the estimated of this estimator is not close to the true parameter.

5 Conclusion

The relationship estimator—WML correlation coefficient—is presented when data are sampled from population which has the bivariate normal distribution. This proposed estimator was developed from the conditional probability function principle and concern about the Winsorized distribution concept when the tails of it contains anomalous information. A performance verification of the presented estimator is performed through the simulated data across 240 scenarios to compare the robust property of this estimator when analysis of data dealing with outliers and missing observations occur. The comparison results of six method—WML, Pearson, median, percentage bend, biweight mid, and composite correlation coefficients—are as follow: for almost all levels of ρ, the WML correlation coefficient tends to be the robust estimator although the data in the analysis contain outliers and missing values. Especially, when the sample size is as small as 10 and outliers are presented in the dataset as 20%, regardless of the percentage of data loss. For size of sample equals 100, the median correlation coefficient seems to provide the best performance when ρ approximately over 0.4 at all levels of percentages of missing and outliers in the dataset. Although, the bias of the proposed estimator is greater than this of median method when sample sizes are larger, but its mean squared error tends to be lower than another methods, regardless of the degree of data loss and outliers. This effect is particularly observed when ρ approximately less than 0.5.

References:

[1] Kutner, M.H., Nachtsheim C.J., Neter, J., Li, W., *Applied Linear Statistical Models*, ed. 5, Irwin, 2005.

[2] Cheng, Y.T., Yang, C.C., An approach of stocks substitution strategy using fuzzy interval correlation coefficient, *Communications in Statistics – Simulation and Computation*, Vol.45, No. 4, 2016, pp. 1187–1196.

[3] Little, R.J.A., Rubin, D.B., *Statistical Analysis with Missing Data*, ed. 3, John Wiley & Son, 2019.

[4] Rao, C.R., Toutenburg, H., Fieger, A., *Linear Models and Generalizations: Least Squares and Alternatives*, ed. 3, Springer Verlag, 2007.

[5] Acock, A.C., Working with missing values, *Journal of Marriage and Family*, Vol.67, 2005, pp. 1012–1028.

[6] Rotnitzky, A., Wypij, D., A note on the biased of estimators with missing data, *Biometrics*, Vol.50, 1994, pp. 1163–1170.

[7] Roth, P.L., Campion, J.E., Jones, S.D., The impact of four missing data techniques on validity estimates in human resource management, *Journal of Business and Psychology*, Vol.11, 1996, pp. 101–112.

[8] Gorelick, M.H., Bias arising from missing data in predictive models, *Journal of Clinical Epidemiology*, Vol.59, 2006, pp. 1115–1123.

[9] Fitzmaurice, G., Missing data: Implications for analysis, *Nutrition*, Vol.24, 2008, pp. 200–202.

[10] Sinsomboonthong, J., Estimation of the correlation coefficient for a bivariate normal distribution with missing data, *Kasetsart Journal (Natural Science)*, Vol.45, No.4, 2011, pp. 736–742.

[11] Azimi, I., Pahikkala, T., Rahmani, A.M., Niela-Vilén, H., Axelin, A., Liljeberg, P., Missing data resilient decision-making for healthcare IoT through personalization: a case study on maternal health, *Future Generation Computer Systems*, Vol.96, 2019, pp. 297–308.

[12] Sinsomboonthong J., Sinsomboonthong S., Estimation of the population mean for incomplete data by using information of

simple linear relationship model in data set, *Advances in Science, Technology and Engineering Systems*, Vol.6, No.4, 2021, 161–169.

[13] Choi, J., Dekkers, O.M., le Cessie, S., A comparison of different methods to handle missing data in the context of propensity score analysis, *European Journal of Epidemiology*, Vol.34, No.1, 2019, pp. 23-36.

[14] White, I.R., Carlin, J.B., Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, *Statistics in Medicine*, Vol. 29, No.28, 2010, pp. 2920-2931.

[15] Nagashima, H., Kato, Y., Method for selecting a data imputation model based on programming by example for data analysts, *Proceedings - 2020 IEEE International Conference on Big Data*, Big Data 20209377818, 2020, pp. 4147-4156.

[16] Kumar N., Hoque M.A., Sugimoto M., Kernel weighted least square approach for imputing missing values of metabolomics data, *Scientific Reports*, Vol.11, No.1, 2021, 11108.

[17] Aggarwal, C.C., *Outlier Analysis*, Springer, 2013.

[18] Ibrahim, E., Shouman, M.A., Torkey, H., El-Sayed, A., Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system, *Multimedia Tools and Applications*, Vol.80, No.13, 2021, pp. 20125-20147.

[19] Stockburger, D.W., *Introductory Statistics: Concepts, Models, and Applications*, ed. 3, Missouri State University, 2013.

[20] Evandt, O., Coleman, S., Ramalhoto, M.F., Lottum, C.V., A little-known robust estimator of the correlation coefficient and its use in a robust graphical test for bivariate normality with applications in the aluminium industry, *Quality and Reliability Engineering International*, Vol.20, 2004, pp. 433–456.

[21] Maturi T.A., Elsayigh A., A comparison of correlation coefficients via a three-step bootstrap approach, *Journal of Mathematics Research*, Vol.2, No.2, 2010, pp. 3–10.

[22] Shevlyakov, G., Smirnov, A., Robust estimation of the correlation coefficient: An attempt of survey, *Austrian Journal of Statistics*, Vol.40, No.1&2, 2011, pp. 147–156.

[23] Mukaka, M.M., Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, Vol.24, No.3, 2012, pp. 69–71.

[24] Shevlyakov, G., On robust estimation of a correlation coefficient, *Journal of Mathematical Sciences*, Vol.83, No.3, 1997, pp. 434–438.

[25] Wilcox, R., Inferences based on a skipped correlation coefficient, *Journal of Applied Statistic*, Vol.31, No.2, 2004, pp. 131–143.

[26] Sinsomboonthong J., Robust Estimators for the correlation measure to resist outliers in data, *Journal of Mathematical and Fundamental Sciences*, Vol.48, No.3, 2016, pp.263–275.

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

[28] Barnett, V., Lewis, T., *Outliers in Statistical Data*, ed. 3, John Wiley, 1995.

[29] Shevlyakov, G., On robust estimation of a correlation coefficient, *Journal of Mathematical Sciences*, Vol.83, No.3, 1997, pp. 434–438.

[30] Wilcox, R., *Introduction to Robust Estimation and Hypothesis Testing*, ed. 4, Academic Press, 2017.

[31] Armstrong, R.A., Should Pearson's correlation coefficient be avoided?, *Ophthalmic and Physiological Optics*, Vol.39, No.5, 2019, pp.316-327.

[32] Olivoto, T. et al., Confidence interval width for pearson’s correlation coefficient: A gaussian-independent estimator based on sample size and strength of association, *Agronomy Journal*, Vol.110, No.2, 2018, pp. 503-510, doi: 10.2134/agronj2017.09.0566.

Sources of funding for research presented in a scientific article or scientific article itself

This research was financial supported by Kasetsart University Research and Development Institute (KURDI).

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US