Journal Of Mathematical And Computational Science
Vol.1, Issue 1, June 2016

# Differences Between Pearson's Product Moment Correlation Coefficient and An Absolute Value Correlation Coefficient In The Presence of Outliers

**Norafefah Binti Mohamad Sobri[1]*, Prof Dr. Habshah Midi [2], Nurul Bariyah Ibrahim[3] , Nor Azima Ismail[4], Wan Faizah Wan Yaacob[5] and Mohd Azry Abdul Malik[6]**

[1,3,4,5,6]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kelantan, Bukit Ilmu, Machang, Kelantan, Malaysia
noraf378@kelantan.uitm.edu.my
[2]Faculty of Sciences, University Putra Malaysia
*Corresponding author

**Abstract**: **The correlation coefficient is one of the most commonly used statistical measures in all branches of statistics. The empirical evidence shows that this correlation coefficient is sufficiently non-robust against outliers. The aim of this study is to compare the performance of the estimator of correlation coefficient. In this study, Pilot-plant data was considered at first stage. Second stage of this study, the simulation data were generated based on normal and uniform distribution at its four contaminated form. The methods of analysis used in this study were Pearson's correlation coefficient and An Absolute Value correlation coefficient. It can be conclude that an Absolute Value correlation coefficient performs well and more robust compared to Pearson's correlation coefficient in existence of outliers. Then we investigated the bias, standard error (SE) and root mean square error (RMSE) to judge their performance. The result shows that an Absolute Value performs better than Pearson's correlation coefficient. In general An Absolute Value correlation coefficient appears to be a good estimator because it has the lowest values of bias, standard error and RMSE.**

**Keywords:** ***Absolute Value Correlation Coefficient, contaminated , Pearson's Correlation Coefficient, outliers.***

## 1 Introduction

The correlation coefficient is one of the most commonly used statistical measures in all branches of statistics. The correlation coefficient is a standard tool in applied regression analysis. Although it is not always thoughtfully used, it remains an informative summary measure of predictive of the selected regression model. Karl Pearson [5] was greatly influenced by Sir Francis Galton to systematize the application of correlation and developed the present day version of the Pearson's Product Moment correlation coefficient in 1896. An excellent review of the development of correlation coefficient is available in Shevlyakov and Vilchevski .

During fifties and sixties of the last century, attention was drawn to the fact that inferences which are based on the product moment correlation heavily depend on the assumption of bivariate normality and they are very sensitive to outliers. The Pearson's Product Moment correlation coefficient is based on the sample means $\bar{x}$ and $\bar{y}$ respectively, which are known to be very sensitive to outliers. The statistics which measure the effect of a possible outliers (x,y) at the correlation coefficient is called the influence function. In this respect, Romanazzi [6] illustrated the non-robustness of product moment correlation by showing that its ability to influence.

A number of robust correlation coefficients are suggested over the years. Devlin et al.[2] attempted to identify observations which may unduly distort the correlation coefficient and tried to develop new methods which are not affected by a small fraction of outliers. Rodger and Nicewander [5] provided thirteen ways to look at the product moment correlation coefficient and suggested a new one.

Gideon [3] offered a generalized interpretation of Pearson correlation coefficient and formulated three estimators of correlation coefficient that are An Absolute Value correlation coefficient , An absolute Value from Median correlation coefficient and Median – type correlation coefficient. An Absolute Value correlation coefficient was introduced by Gideon [3]. It is a modification of Pearson's Product

Moment Correlation Coefficient and it is based on absolute values. It is noticeable that the same heuristic motivation for Pearson's holds for absolute value correlation coefficient. This correlation coefficient is new and not much extensive study has yet been made. Although a number of correlation coefficient are available in the literature not much investigation has been done on their performance in the presence of outliers. Therefore the purpose of this study is to compare the effect of Pearson's correlation coefficient and an absolute correlation coefficient in the presence of outliers.

## 2 Outliers

Outliers can arise from several different mechanisms or causes. Anscombe (1960) sorted outliers into two categories: those arising from errors in the data and those arising from the inherent variability of the data. Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening.

Occurrence of an outlier has a low probability that is originated from the same statistical distribution as the other observation in the data set. On the other hand, an outlier is also an extreme value , an observation that might have a low probability of occurrence but cannot be statistically shown to originate from a different distribution than the rest of the data.

However, outliers can provide useful information about a process. An outlier can be created by a shift in the location (mean) or in the scale (variability) of the process. Though an observation in a particular sample might be a candidate as an outlier, the process might have shifted. Hence, outliers lower the significance of the fit of a statistical model because they do not coincide with the model's prediction. Thus, outliers also have adverse effects on the correlation coefficients.

## 3 Data and methodology

### *A. Data*

In this study, a Pilot-plant chemical data from Daniel and Wood [1] will be analyzed. The response variable (Y) correspond to acid content which is determined by titration and the explanatory variable (X) is the organic acid content determined by extraction and weighting. This data consist of 20 observations. The data will be analyzed using Pearson Moment Correlation (Pearson, 1896) for non-robust technique and for robust technique, an Absolute Value correlation coefficient will be utilized . The comparison between both methods will be done. Simulation study was conducted in this study using normal and uniform distribution. The simulation study was carried out using S-Plus software.

### *B. Pearson Product Moment Correlation Coefficient*

In statistics, the Pearson Product Moment Correlation Coefficient (referred as PMCC and denote by ( $r_p$ ) is a measure of the correlation between two variables predictor (X) and response (Y) giving a value between +1 and -1 inclusive. It was developed by Karl Pearson (1856) from a similar but slightly different idea introduced by Francis Galton in the 1880s. Correlation methods for determining the strength of the linear relationship between two of more variables are among the range of wide applied statistical techniques. Theoretically, the concept of correlation is considered a starting point of a building block in the development of many areas of statistical research.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho = corr(XY) = \frac{Cov(X,Y)}{[\text{var}(X)\,\text{var}(Y)]^{1/2}} \tag{1}$$

Where $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$ (2)

The above formula defines the population correlation coefficient, commonly represented by the Greek letter ρ (rho). Substituting estimates of the covariance and variances based on a sample gives the sample correlation coefficient, commonly denoted as $r_p$ .

Now let $(x_1, y_1), (x_2, y_2), ....(x_n, y_n)$ be a bivariate data set. The usual mean notation will be used (are the centered data):

$$x_i^* = x_i - \bar{x}, y_i^* = y_i - \bar{y}, i = 1,2,..n \quad (3)$$

The sample covariance is proportional to $\sum x_i^* y_i^*$

This covariance is rewritten,

$$\sum x_i^* y_i^* = (\sum (x_i^* + y_i^*)^2 - \sum (x_i^* - y_i^*)^2)/4 \quad (4)$$

In the uncentered notation, this can be written as;

$$\sum (x_i - \bar{x} + y_i - \bar{y})^2 - \sum (x_i - \bar{x} - y_i + \bar{y})^2)/4 \quad (5)$$

This form of the covariance function appears as an interpretation of Pearson's in Rodgers and Nicewater [5], when their rescaled variance interpretation is added together.

Therefore the Pearson's Product Moment Correlation Coefficient can be defined as:

$$r_p = \left[ \sum \left( \frac{x_i^*}{\sqrt{SS_x}} + \frac{y_i^*}{\sqrt{SS_y}} \right)^2 - \sum \left( \frac{x_i^*}{\sqrt{SS_x}} - \frac{y_i^*}{\sqrt{SS_y}} \right)^2 \right] / 4 \quad (6)$$

where

$$SS_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \; ; \qquad SS_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \quad (7)$$

The Pearson is definition of (4) is equal to {(standardized distance from perfect negative correlation) - (standardized distance from perfect positive correlation)} divided by a constant, that puts the value between -1 and +1 .

### *C. An Absolute Value Correlation Coefficient*

An absolute Value Correlation Coefficient $r_{av}$ was introduced by Gideon in 1998. Gideon [3] offered a generalized interpretation of Pearson correlation coefficient and formulated three new estimators of correlation coefficient. An Absolute Value correlation coefficient uses original data directly. Let us consider $(x_i^*, y_i^*), i = 1,2,...,n$ be the n pair of observations deviation from their mean i.e $x_i^* and\ y_i^*$. The sum of absolute values about the mean is denoted by $SA_x\ and\ SA_y$ for x and y variables respectively. Thus the Absolute Value correlation coefficient as $r_{av}$ is defined as:

$$r_{av} = \frac{1}{2}\left\{\Sigma\left(\frac{x_i^*}{SA_x} + \frac{y_i^*}{SA_y}\right)^2 - \Sigma\left(\frac{x_i^*}{SA_x} - \frac{y_i^*}{SA_y}\right)^2\right\} \tag{8}$$

the denominator is 2 since

$$\Sigma\left|\frac{x_i^*}{SA_x}\right| + \Sigma\left|\frac{y_i^*}{SA_y}\right| = 2 \tag{9}$$

where

$$x_i^* = x_i - \bar{x} \;; \qquad y_i^* = y_i - \bar{y} \tag{10}$$

and

$$SA_x = \Sigma|x_i - \bar{x}| \;; \qquad SA_y = \Sigma|y_i - \bar{y}| \tag{11}$$

It is noticeable that the same heuristic motivation for Pearson's holds for Absolute Value correlation coefficient.

### *D. Simulation Study*

In this section, we carried out a simulation study to look at the result of the correlation coefficients $r_p$ and $r_{av}$ as defined by Eq.6 and Eq.8. For 'good' data, we simulated 1000 times according to the linear relation $\mathrm{y_i} = 2 + 1.0\mathrm{x_i} + \varepsilon_i$ where $x_i$ is normally distributed with mean 5 and variance 1, $\varepsilon_i$ drawn from $\mathrm{N}(0, \sigma^2), \sigma = 0.2$ Then we obtained results for $r_p$ and $r_{av}$ . Because the data were uncontaminated, we suspect that all the correlation coefficients yielded values which are close to the original $\rho = 1$ .

Then, we started to contaminate the data at 5%, 10%, 15% and 20%. We deleted one 'good' observation and replace it with outliers. The contaminated data were generated according to the linear relation $\mathrm{y_i} = 2 + 1.0\mathrm{x_i} + \varepsilon_i$ where $x_i$ is uniformly distributed on interval (5, 10) and, $\varepsilon_i$ is normally distributed with mean 2 and standard deviation 0.2. A sample of size 20, 60 and 100 respectively were generated according to the above relation. Then we obtained results the for $r_p$ and $r_{av}$ for uncontaminated data and contaminated data.

### *E. Measuring Estimate of Correlation Coefficient Performance*

*Bias*

The bias of an estimator provides a measure of the average error in the estimator $\hat{\rho}$ of a parameter $\rho$ that is the error which arises when estimating a quantity. The bias of an estimator is defined as the difference between the expected value of the estimator and the actual value

$$bias(\hat{\rho}) = E[\hat{\rho}] - \rho \tag{12}$$

An estimator is unbiased if the expected value of the estimator equals the true parameter value, i.e. $E[\hat{\rho}] = \rho$ . Otherwise, the estimator is biased.

This study provides us the distribution is known so we used the bias to look at the performance between the Pearson's Correlation Coefficient and An Absolute Value Correlation Coefficient.

*Mean Square Error*

The mean square error (MSE) is the expected value of the square error. Let be $\hat{\rho}$ a parameter and be $\rho$ an estimator of the parameter, the mean squared error of the estimator is defined as:

$$MSE(\hat{\rho}) = E[(\hat{\rho} - \rho)^2] \tag{13}$$

It is sometimes more useful to rewrite the MSE equation in terms of the bias and variance. The first step of the rewriting is to expand the expected value on the right-hand side of Eq.13 to get:

$$MSE(\hat{\rho}) = E[(\hat{\rho}^2 - 2\hat{\rho}\rho + \rho^2)] = E[\rho^2] - 2\rho E[\hat{\rho}] + \rho^2 \tag{14}$$

The next step of the rewriting is to add to and subtract $(E[\hat{\rho}])^2$ from the right-hand side of Eq. 14 so that

$$MSE(\hat{\rho}) = E[\hat{\rho}^2] - (E[\hat{\rho}])^2 - 2\rho E[\hat{\rho}] + \rho^2 \tag{15}$$

By simplifying Eq.15, the mean squared error can be written as

$$\begin{aligned} MSE(\hat{\rho}) &= E[\hat{\rho}^2] - (E[\hat{\rho}])^2 + (E[\hat{\rho}] - \rho)^2 \\ &= V(\hat{\rho}) + [bias(\hat{\rho})]^2 \end{aligned} \tag{16}$$

*Standard Error*

The standard error gives a measure of the precision of the estimators. The standard error of an estimator $\hat{\rho}$ is defined as the standard deviation of its sampling distribution

$$SE(\hat{\rho}) = \sqrt{V(\hat{\rho})} = \sigma_{\hat{\rho}} \tag{17}$$

*Root Mean Square Error (RMSE)*

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the value actually observed from the thing being modeled or estimated. RMSE is a good measure of precision and these individual differences are also called residuals, and the. RMSE serves to aggregate them into a single measure of predictive power.

The RMSE of an estimator with respect to the estimated parameter is defined as the square root of the mean square error:

$$RMSE(\hat{\rho}) = \sqrt{MSE(\hat{\rho})} = \sqrt{E((\hat{\rho} - \rho)^2)} \tag{18}$$

For an unbiased estimator, the RMSE is the square root of the variance, known as the standard error.

## 4 Result and discussion

The results that are obtained by running the data using S-Plus®. Analyses on modified Pilot-plant data will discuss. They are inclusive of the results of Pearson's correlation coefficient and an Absolute Value correlation coefficient at 0%, 5%, 10%, 15% and 20% contamination. After that the analysis and results based on simulation studies are discussed.

### *F. Result on contaminated data using correlation coefficient.*

The results of Pearson's correlation coefficient and Absolute Value correlation coefficient based on the modified data at 5%, 10%, 15% and 20% contamination are presented in Table 1.

TABLE 1 : THE VALUES OF $r_p$ AND $r_{av}$ FOR THE MODIFIED PILOT-PLANT DATA.

| Contamination | Pearson's CC | Absolute Value CC |
|---|---|---|
| **0%** | 0.9973 | 0.9583 |
| **5%** | 0.3656 | 0.5833 |
| **10%** | 0.1760 | 0.3426 |
| **15%** | -0.0277 | 0.1021 |
| **20%** | -0.0252 | 0.0134 |

The result shows that at 0% contamination, the value of Pearson's Moment correlation coefficient is reasonably closed to the Absolute Value correlation coefficient. The Pearson's correlation coefficient value is 0.9973 in which it is slightly higher than an Absolute Value correlation which is only 0.9583. In the presence of outliers, at 5% contamination (one outlier exists in the data) the result shows that the Pearson's correlation coefficient was immediately affected. The value of Pearson's decreases from 0.9973 to 0.3656 which is from strong positive correlation to weak positive correlation. The value of An Absolute Value correlation coefficient also changes when outlier exists in the data, but the change is not drastic as compared to Pearson's correlation.

The value of Pearson's correlation coefficient moves away from the true value as the percentage of outliers increases. The increase in the percentage of outliers at 15% and 20% contamination has changed the sign of Pearson's correlation coefficient from the positive relationship to the negative relationship. This gives the big impact to the value of Pearson correlation coefficient. The Absolute Value Correlation coefficient seems to be slightly affected by the outliers because the value of correlation still in the positive relationship at 15% and 20% contamination.

Therefore it can be concluded that an Absolute Value of correlation coefficient gives a better estimate of correlation coefficient than the Pearson's Product Moment correlation coefficient in the presence of outlier at 5%, 10%, 15% and 20%. It can also be conclude that an Absolute Value correlation coefficient performs well and more robust compared to Pearson's correlation coefficient in the existence of outliers.

We continue our study by looking at graphical method. Scatter plots of pilot-plant data was plotted. *Figure 1* apparently display the scatter plot between acid content (Y) and organic acid content (X).

In the Figure 1, it can be concluded that the existence strong positive linear relationship between organic acid content (X) and acid content (Y) when there is no presence of outlier(s).

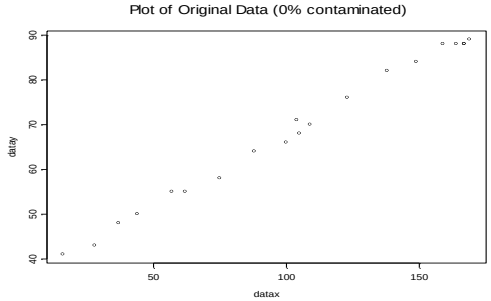


FIGURE 1: SCATTER PLOT OF PILOT-PLANT DATA WITH NO PRRSENCE OF OUTLIERS

The data were modified and it has considered outliers. The *Figure 2* below shows the scatter plot of pilot-plant data.

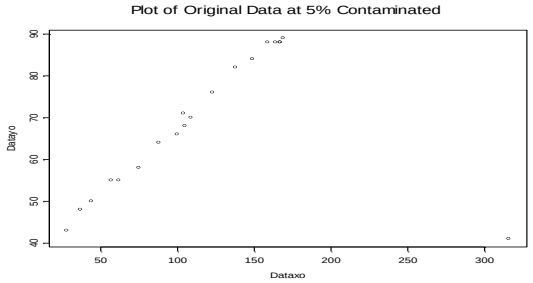


FIGURE 2: SCATTER PLOT OF PILOT-PLANT DATA WITH 5% CONTAMINATION.

The data was modified and it considered having 5% outliers i.e. the x-value of the ninth observation has been modified as 16 to 316. Then by using graphical method we look the effect of presence outlier(s) in the x-direction has been identified. *Figure 2* above shows the scatter plot of pilot-plant data at 5% contamination. The plot apparently displays the magnitude of correlation in which it has become weak due to the existence of one outlier in the data set.

Now the data were modified so that it contained 10% outliers. *Figure 3* below shows the scatter plot of pilot-plant data at 10% contamination.

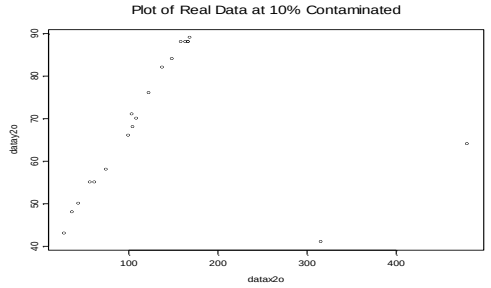


FIGURE 3: SCATTER PLOT OF PILOT-PLANT DATA WITH 10% CONTAMINATION.

The data were changed i.e. the x-value of the ninth and 15$^{th}$ observations has been modified as 16 to 316 and 88 to 480. *Figure 3* above, shows the scatter plot of pilot-plant data at 10% contamination. The plot apparently displays the magnitude of correlation becomes weaker as two outliers exist in the data set.

The study continues with the modification of the data to ensure it contain 15% outliers. *Figure 4* below shows the scatter plot of pilot-plant data at 15% contamination.

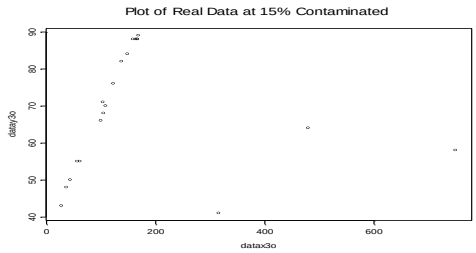


FIGURE 4: SCATTER PLOT OF PILOT-PLANT DATA WITH 15% CONTAMINATION.

The data are modified i.e. the x-value of the $9^{th}$ , $14^{th}$ and $15^{th}$ observations has been changed as 16 to 316 , 75 to 750 and 88 to 480. *Figure 4* above shows the scatter plot of pilot-plant data with 15% contamination. The plot apparently shows the result as the percentage of outliers' increase, the correlation now changes to weak negative relationship.

Now we modified the pilot-plant data so that it containing 20% outliers. *Figure 5* below, showed the scatter plot of pilot-plant data at 20% contamination.

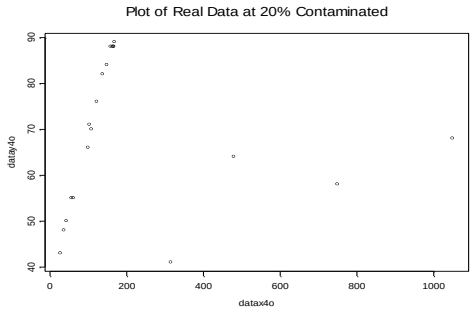


FIGURE 5: SCATTER PLOT OF PILOT-PLANT DATA WITH 20% CONTAMINATION.

The data were modified i.e. the x-value of the $9^{th}$ , $12^{th}$, $14^{th}$ and $15^{th}$ observations has been changed as 16 to 316 , 105 to 1050, 75 to 750 and 88 to 480 . The plot apparently displays that as the percentage of outliers increase at 20% contaminated data, the values of correlation coefficient turn to weak negative relationship.

### *G. Analysis and Results Based on Simulation*

A series of simulations were employed in this study to further assess the performance of the correlation coefficient; Pearson's correlation coefficient and Absolute Value correlation coefficient as were defined by Eq.6 and Eq.8. We begin by generating 100 'good' observations according to the linear relation: $y_i = 2 + 1.0x_i + \varepsilon_i$ where $x_i$ is normally distributed with mean 5 and variance 1, drawn from $N(0, \sigma^2), \sigma = 0.2$ and the true value of $\rho = 1$ .The normal variates were generated by the S-Plus program. Using these data, we applied Equation 6 and 8. The obtained results for $r_p = 0.9991$ and $r_{av} = 0.9781$ . The simulation result shows that the Pearson's correlation coefficient is fairly closed to the Absolute Value correlation when there is no outlier in the data. Because the data were uncontaminated, it was expected that all the correlation coefficients yielded values which are close to the original $\rho = 1$

The second stage would be to contaminate the data. At each step, we deleted one 'good' observation and replaced it with a 'bad' data point. The contaminated point was generated according to the linear

relation $y_i = 2 + 1.0x_i + \varepsilon_i$ where $x_i$ is uniformly distributed on interval (5, 10) and, $\varepsilon_i$ is normally distributed with mean 2 and standard deviation 0.2. This is repeated up to 20% contamination.

Table 2 presents the values of $r_p$ and $r_{av}$ when 'good' observations were replaced by a certain percentage of outliers i.e. 5%, 10%, 15% and 20%.

TABLE 2: THE VALUES OF $r_p$ AND $r_{av}$ FOR SIMULATION DATA WITH n = 100.

| **Contamination** | $r_p$ | $r_{av}$ |
|---|---|---|
| 0 | 0.9991 | 0.9781 |
| 5 | 0.1857 | 0.4827 |
| 10 | -0.0756 | 0.1556 |
| 15 | -0.0165 | 0.1101 |
| 20 | -0.2853 | -0.1643 |

From the results of Table 2, it can be observed that at 5% contamination, the Pearson's correlation coefficient $(r_p)$ equal to 0.1857 compared to an Absolute Value correlation coefficient $(r_{av})$ 0.4827. This shows that Pearson's correlation coefficient was immediately affected by outliers and its values moved away from the true value as the percentage of outliers' increases.

The increase in the percentage of outliers from 0% (no contamination) up to 10% contamination has changed not only the values but also the signs of Pearson's correlation coefficient i.e. from the positive to negative values of the correlation coefficient. It appears that an Absolute Value correlation coefficient holds in positive sign until 15% contamination before it changes the sign to negative correlation at 20% contamination. Therefore, the result shows an Absolute Value correlation coefficient is more robust than the Pearson's correlation coefficient in simulated data.

### *H. Simulation 1000 times correlation coefficient based on n = 20,60 and 100.*

The properties of these sample correlation coefficients were investigated further by looking at five summary statistics, namely the bias, variance, standard error (SE), mean square error (MSE) and root mean square error (RMSE) in 1000 trials. In each trial t (t = 1,2, ….. 1000) a sample of size 20, 60 and 100 respectively, was generated according to the sampling situations described earlier. The average of the sample correlation coefficient $\hat{\rho}$ is $\bar{\rho} = \sum \rho / 1000$ which yield the bias $\bar{\rho} - \rho$.The variance is given by $v(\hat{\rho}) = \sum (\hat{\rho}_t - \bar{\rho})^2 / 1000$ which can be used to compute the MSE as: $MSE(\hat{\rho}) = [bias]^2 + v(\hat{\rho})$. Accordingly, the standard error (SE) is given by $\sqrt{v(\hat{\rho})}$ and the RMSE by $\sqrt{MSE(\hat{\rho})}$ . These summary statistics are presented in Table 3 for n = 20, 60 and 100.

TABLE 3: SUMMARY STATISTICS FOR $r_p$ AND $r_{av}$ FOR N = 20, 60, 100 AND $\rho = 1$ .

| Contamination (%) | Correlation Coefficient | n = 20 | | | n = 60 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Bias** | **SE** | **RMSE** | **Bias** | **SE** | **RMSE** | **Bias** | **SE** | **RMSE** |
| **0** | $r_p$ | -0.00084 | 0.00042 | 0.00094 | -0.00823 | 0.00023 | 0.00085 | -0.00081 | 0.00017 | 0.00083 |
| | $r_{av}$ | -0.02041 | 0.00539 | 0.02111 | -0.02039 | 0.00301 | 0.02061 | -0.02029 | 0.00228 | 0.02042 |
| **5** | $r_p$ | -0.81011 | 0.27571 | 0.85575 | -0.80051 | 0.15600 | 0.81557 | -0.80234 | 0.12436 | 0.81192 |
| | $r_{av}$ | -0.53518 | 0.15305 | 0.55663 | -0.52224 | 0.08755 | 0.52953 | -0.52218 | 0.00692 | 0.52674 |
| **10** | $r_p$ | -1.11281 | 0.24543 | 1.13956 | -1.10190 | 0.14296 | 1.11114 | -1.10066 | 0.11209 | 1.10636 |
| | $r_{av}$ | -0.87418 | 0.17584 | 0.89169 | -0.85895 | 0.10431 | 0.86526 | -0.85456 | 0.08282 | 0.85856 |
| **15** | $r_p$ | -1.29289 | 0.21147 | 1.31007 | -1.26671 | 0.12085 | 1.27246 | -1.26393 | 0.09556 | 1.26754 |
| | $r_{av}$ | -1.11023 | 0.17500 | 1.12394 | -1.08226 | 0.10140 | 1.08700 | -1.07958 | 0.07958 | 1.08251 |
| **20** | $r_p$ | -1.38426 | 0.19044 | 1.39729 | -1.37471 | 0.10705 | 1.37887 | -1.37193 | 0.08217 | 1.37439 |
| | $r_{av}$ | -1.24923 | 0.16739 | 1.26039 | -1.23713 | 0.09459 | 1.24074 | -1.23493 | 0.07327 | 1.23710 |

From Table 3, it is apparent that Pearson's $(r_p)$ correlation coefficient provides the best results when no contamination occurs in the model. As a result, an Absolute Value correlation

From Table 3, it is apparent that Pearson's $(r_p)$ correlation coefficient provides the best results when no contamination occurs in the model. As a result, an Absolute Value correlation coefficient performs somewhat less than the based counterparts in the normal situation. It can be noted that the bias is negligible in this situation and the variance makes up most of the MSE. In addition, as percentage of the outliers increases in the data, the, the Pearson correlation coefficient decreases systematically at these contaminated samples and they have very high MSE values. For this correlation coefficient, the bias makes up most of the MSE.

In terms of standard error, the performance of Pearson's correlation coefficient is very poor in the presence of outliers. The value of standard error for Pearson's correlation coefficient is higher than an Absolute Value correlation coefficient. On the other hand, the Absolute Value correlation coefficient performs reasonably well as it give minimum bias, standard error and root mean square error compared to Pearson's Correlation Coefficient. The results seem to be consistent in all 1000 trial and for each sample n = 20, n = 60 and n = 100. Therefore, it can be summed up that the Absolute Value correlation coefficient has better performance than the Pearson's correlation coefficient in the presence of outliers.

## Conclusion

In this study, the performance of the estimated Pearson's Product Moment Correlation Coefficient and An Absolute Correlation Coefficient in the presence of outliers has been studied. Both methods have been employed on an illustrative example with no contaminant data, which resulted in a conclusion that the Pearson's Product Moment Correlation Coefficient performed equally good as the Absolute Value correlation coefficient. However, the results of both estimators changed when the example data has been modified to contain several outliers at 5%, 10%, 15% and 20. The Pearson's correlation

coefficient was affected immediately by outliers and its values moved away from the true value as the percentage of outliers increased. The increase in the percentage of outliers at 10% up to 20% contamination has changed not only the values but also the sign of Pearson's correlation coefficient i.e. from the positive to negative values of the correlation coefficient. The Absolute Value correlation coefficients seem not much affected by outliers. The increase in outlier up to 20% contamination has not changed the sign of an Absolute Value correlation coefficient. Thus, an Absolute Value correlation coefficient performed better than Pearson's correlation coefficient in the presence of outliers.

A series of simulation studies were then conducted in order to have more general picture on the performance of both methods under different conditions. The simulation studies yielded the following conclusion (i) The Pearson's Correlation Coefficient has lesser bias and more efficient than an Absolute Value correlation coefficient when there is no problem of outliers in a dataset and (ii) In the presence of outliers in a dataset, an Absolute Value correlation coefficient performed better than the Pearson's correlation coefficient. The result showed that the values of bias, standard error and root mean square give minimum value on an Absolute Value correlation coefficient compared to Pearson's correlation coefficient.

**Acknowledgements**

The authors would like to express the greatest acknowledgement to Prof. Dr. Habshah Midi for her valuable guidance and comment pertaining to this study.

**References**

[1] C. Daniel, and F.s. Wood, "Fitting Equations to Data," New York : John Wiley, 1980.
[2] J.S. Devlin, R. Gnanadesikan and J.R. Kettenring, "Robust Estimation and Outlier Detection With Correlation Coefficient," Biometrica 62, vol. 3, 1975, pp. 531-559.
[3] R.A Gideon and R.A. Hollister, "A Rank Correlation Coefficient Resistant to Outliers," J. Amer. Stat. Assoc. 82, 1987, pp. 656-666.
[4] R.A. Gideon, "A Generalized Interpretation of Pearson's r," http://www.math.umt.cdu/gideon.
[5] J.L Rodger, and W.A. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," The American Statistician, 42, 1988, pp 59-66.
[6] M. Romanazzi, "Influence in Canerical Correlation Coefficient Analysis," Biometrica, 57, 1992, pp. 237-259.