JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

Proposed Method to Employ the Robust Correlation in Structural Equations Modeling

Saif Ramzi Ahmed, Postgraduate

Bashar A. Al-Talib Asst. Prof.

saiframz525@gmail.com

bashar.altalib@uomosul.edu.iq

Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

Abstract

The idea of the research is to fortify the path coefficients in the structural equations model through the use of the robust correlation coefficients. Three methods of the robust correlation were compared in addition to the Pearson correlation. A comparison has been made between the proposed method and the traditional method through a causal model proposed by the researchers. Several simulation scenarios have been applied in the experimental aspect. In order to test the efficiency of the generated models, samples of different sizes (n=20,100,1000) were used. Furthermore, the explanatory variables in the causal model were contaminated with different contamination percentages. Path parameters were estimated using the four-fold correlation matrix and the results were compared through some statistical criteria. The study concluded that the pathway coefficients method using the Robust Correlation has given the best results compared to the traditional methods.

Keyword: Robust correlation, outlier, path analysis, structural equation.

Introduction

Structural Equation Modeling (SEM) is sometimes used in pathway analyzes and Confirmatory Factor Analyses (CFA). It investigates causal relationships between one or more independent variables (IVs). It can be intermittent or continuous. One or more dependent variables (DVs), which can also be discontinuous or continuous. According to some studies, the application of SEM effectively contributes to solving complex relationships in variables in studies of social sciences, management, nature and other sciences. It is called modeling and estimating parameter values on the hypotheses of the SEM test by causal inference. In such cases, the results are considered accurate if the verification of the data revealed the problems available in it for the purpose of choosing the appropriate method for estimating the path parameters (Yuan *et al*., 2000).

Structural Equations Modeling

Structural equation modeling is a methodology or method in research used to estimate, analyze and test models that describe and determine the relationships between variables. In other words, it is a comprehensive statistical model for testing hypotheses to identify the relationship between the observed variables and the independent variables. They describe the representation, estimation, and testing of a network of linear relationships between variables to test hypothesis patterns of direct and indirect relationships between variables. Watching and not watching (Suhr, 2006).

The structural equation model is defined as a pattern or hypothetical model of direct and indirect relationships between a set of latent variables. Hence, it can represent a complete path model for a relationship between a set of variables that can be described in a Path Diagram that shows the characteristics of the relationship between this set of variables, which is an extension The General Linear Model (Al-Mahdi, 2013). Furthermore, structural equation modeling is a method for

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

analyzing multiple types of models, such as models of multiple regression analysis, path analysis, as well as confirmatory factor analysis, which are special cases of this modeling.

Stages of Structural Equation

The structural equation modeling technique proposes a methodology that includes several sequential steps. It explains the results for each stage and the next stage that precedes it as clarified by (Shook, et al., 2004). There are several important factors to identify the structural models, which are as follows:

1. Data Characteristics
2. Reliability and Validity
3. Evaluating Model Fit
4. Model Respecification
5. Equivalent Models Reporting

Path Analysis

The path analysis method is one of the efficient statistical methods in analyzing the data. It enables the researcher to analyze and clarify the possible causal relationships for a group of factors and indicate their direct, indirect and total effects on the phenomenon to be studied. Its importance is highlighted by the ability to study the effects of several factors on a particular phenomenon indirectly through several explanatory factors. It is different from the regression analysis method, which is based on direct relationships. Path analysis is a statistical method which is firstly presented by the geneticist Sewall Wright in 1921 through its use in studying the degree of relationship between relatives in studying the genetic behavior of many genetic traits (Olobatuyi, 2006).

This method has spread in most research and in various fields, including social sciences research by Duncan (Vasconcelos *et al*., 1998). He showed in 1966 the relationship between path analysis and Structural Equation Models SEM and set some examples as an aid to an analysis of the path in social research. Moreover, in 1975 the scientist Duncan included all aspects of the structural equations in path analysis and the path analysis method was used extensively in the high environment in 1970 (Olobatuyi, 2006).

Concept of Path Analysis

Path analysis is also known as causal modeling (Jackson *et al.*, 2005). It is also considered a statistical method based on regression and multiple correlation analysis and can be used to establish the probability of the relationship between many variables and examine them in a system of linear equations, whether the variables are continuous or discontinuous (Hadiya, 2011). Moreover, the path analysis method is an extension of multiple regression analysis and is usually used in the study of causal models on the basis that the researcher visualizes the pattern of the relationship between the relevant variables (Davidson, 2012).

Path Analysis Assumption

The path analysis method can be used if some of the following hypotheses are fulfilled (Nair, 2007):

1. The relationship between the variables is linear.
2. The relationship between the variables is an association relationship (there is no interaction between the variables).
3. The relationship between variables is causal.
4. There should not be a correlation between the residual variables with each other.
5. There should not be a correlation between the residual variables and the other variables, i.e., the independent and dependent variables.
6. The relationship between the variables is one-way causation and there is an inverse causal relationship in the model.
7. The measured independent variables are free from any measurement errors.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

8. Assuming that the model is free from assignment errors, that is, the model contains the theoretically possible variables to explain all the variance of the dependent variables.

Outliers

They are those observations that are illogical in the data set. They show a large and clear deviation from the rest of the data in the selected sample group in which that observation was found, and it seems illogical when compared to the rest of the data set. The stray values were known by some researchers by the same method. It seems inconsistent observations with the rest of the observations of the sample under study (Keller & Brian, 2000).
The first steps that the researcher takes in a particular study is the process of examining the data for the phenomenon under study in order to notice the presence of inconsistent values with the rest of the observations that affect the realization of the imposition of a normal distribution. The next step is the process of purifying the data from these values or observations that were called stray values before entering into any statistical analysis depends on the normal distribution. The statistical analysis depends mainly on the data and its purification from any abnormal, stray or inconsistent observations. It constitutes a clear deviation from the rest of the observations and thus distorts the estimated model towards it (Freeman, 1980).

Outliers Detection

The detection of outlier values is applied to regression models and experiments designed; they were tested in multivariate data and by different methods, either through partial detection or total detection of observations. There are two methods used for detecting and identifying outlier values including univariate methods. These methods test each variable separately. The second is multivariate methods that take into account the correlations between the variables in the same data set and highlight the outlier values in both methods because they are far from the observation values of the problem under study (Dan & Ijeoma, 2013). There are many methods for detecting outlier values, including the standard deviation method, the standard score, the modified-standard score, the Tukey's Method (Boxplot), the adjusted boxplot, the median method and others (Al-Talib & Shaker, 2018).

Robust Correlation

The correlation coefficient is one of the most important statistical measures that shows the degree of relationship between two random variables. Its value ranges between (+1, -1). Likely, it is not consistent with the rest of the data in any phenomenon.
Devlin et al. (1975) suggested some graphical methods for diagnosing the observations that could affect the value of the correlation coefficient. More studies about the correlation coefficient were followed. Shevlyakov (1997) relied on simulation experiments to compare the known estimates of the correlation coefficient and some of the fortified estimates that depend on dealing with the polluted binary normal distribution, which was called the Median Correlation Coefficient instead of using the arithmetic mean. It can be calculated using the following formula:

$$r_{\alpha}(\psi) = \frac{\sum_{\alpha} \psi(x_i - \hat{x})\psi(y_i - \hat{y})}{(\sum_{\alpha} \psi^2(x_i - \hat{x})\psi^2(y_i - \hat{y}))^{1/2}} \qquad \dots(1)$$

Since: x̂, ŷ represent the median magnitude of each of x, y respectively, while ψ represents the monotonic function of Huber.

Methods for Estimating Path Coefficients Using Correlation Matrix

There are many ways to estimate the path parameters, including:

1. Estimation of the path analysis coefficient using the simple correlation coefficient. The linear relationship between two random variables is described by the so-called simple correlation coefficient; it is calculated according to Rodgers and Nicewander (1988) using the following formula:

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{E(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} \qquad \dots(2)$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

So, the $Cov(x,y)$: represent the covariance between the two variables x,y whereas the $\sigma_x\sigma_y$ symbolize the standard deviation of the variable y and the variable x, respectively.

So, the simple correlation coefficient, or the so-called Pearson correlation coefficient is affected by outliers. Several alternative impregnable methods have been proposed, including:

1. Robust Correlation: Biweight midcorrelation

The Biweight midcorrelation is determined by the vectors $x_i, y_i \quad , i = 1,2,3, \dots, m$ for each vector according to Langfelder and Horvath (2012) as follows:

$$x = [x_1 \quad x_2 \dots \quad x_m]$$
$$y = [y_1 \quad y_2 \dots \quad y_m]$$

The components of these vectors are defined as follows:

$$u_i = \frac{x_i - Med(x)}{9Mad(x)} \qquad \dots (3)$$

$$v_i = \frac{y_i - med(y)}{9mad(y)} \qquad \dots (4)$$

$Med(x)$ represents the magnitude of the median vector x, $Med(y)$ stands for the magnitude of the median vector $y, Mad(x)$:represents the mean absolute deviation of vector x, $Mad(y)$ stands for the mean absolute deviation of vector y.

The defined weights will be used as follows

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \qquad \dots (5)$$

$$w_i^{(y)} = (1 - v_i^2)^2 I(1 - |v_i|) \qquad \dots (6)$$

Since: I represents a unary function and defined in the following form:

$$I(x) = \begin{cases} 1 & , if \quad x > 0 \\ 0 & , \quad otherwise \end{cases}$$

That is, this function $I(1 - |u_i|)$ is equal to 1 if it is $(1 - |u_i|) > 0$, and is equal to zero otherwise. Also, the weights $w_i^{(x)}$ whose value approaches one if the vector x_i is equal to the median value of the vector x, and approaches zero if the magnitude of the median for the vector $x_i$approaches $9mad(x)$.

Therefore, we can define the Biweight midcorrelation for each of the vectors x and y, which is symbolized by $bicor(x,y)$ as follows:

$$\overline{x}_i = \frac{(x_i - Med(x))w_i^{(x)}}{\sqrt{\sum_{j=1}^{m}\left[(x_j - Med(x))w_j^{(x)}\right]^2}} \qquad \dots (7)$$

$$\overline{y}_i = \frac{\left(y_i - Med(y)\right)w_i^{(y)}}{\sqrt{\sum_{j=1}^{m}\left[\left(y_j - Med(y)\right)w_j^{(y)}\right]^2}} \qquad \dots (8)$$

$$bicor(x,y) = \sum_{i=1}^{m} \overline{x}_i \overline{y}_i \qquad \dots (9)$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

It can be formulated as follows:

$$bicor(x,y) = \frac{\sum_{i=1}^{m}\left(x_i - Med(x)\right)w_j^{(x)}\left(y_i - Med(y)\right)w_j^{(y)}}{\sqrt{\sum_{j=1}^{m}\left[\left(x_j - Med(x)\right)w_j^{(x)}\right]^2 \sum_{j=1}^{m}\left[\left(y_j - Med(y)\right)w_j^{(y)}\right]^2}} (10)$$

2. Percentage Bend Correlation

This type of correlation, which is called the Percentage Bend Correlation, has proven relatively successful in controlling the probability of error of the first type of independent hypothesis test, which was named P.B.Corr. It is estimated through what is shown according to Wilcox (2011) as below:

If we have a random sample consisting of two pairs of defined variables $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ and according to the observations $X_1, X_2, \dots, X_n$, and assuming that M_x represents the sample median, and the selection of of β is a defined quantity of $0 \leq \beta \leq 0.5$ to calculate:

$$W_i = |X_i - M_x|$$
$$m = [(1-\beta)n]$$

We note that the amount $(1-\beta)n$ is close to the integer values.

Assuming that $W_{(1)} \leq \cdots \leq W_{(n)}$ so that the values of W_i are taken in descending order, that is:

$$\widehat{\omega}_x = W_{(m)}$$

Assuming that we have i_1 of the values of X_i such that $(X_i - M_x)/\widehat{\omega}_x < -1$, and we have i_2 as another number of values of $X_i$such that $(X_i - M_x)/\widehat{\omega}_x > 1$, then it will be calculated as:

$$S_x = \sum_{i=i_1+1}^{n-i_2} X_{(i)} \quad \dots (11)$$

and:

$$\widehat{\emptyset}_x = \frac{\widehat{\omega}_x(i_2 - i_1) + S_x}{n - i_1 - i_2} \quad \dots (12)$$

Let:

$$U_i = \frac{\left(X_i - \widehat{\emptyset}_x\right)}{\widehat{\omega}_x} \quad \dots (13)$$

By repeating a set of calculations for the values of Y_i, we get:

$$V_i = \frac{\left(Y_i - \widehat{\emptyset}_y\right)}{y_x} \quad \dots (14)$$

Thus, we get:

$$(15) \dots \Psi(x) = Max[-1, \text{Min}(1, x)]$$

Putting:

$$A_i = \Psi(U_i) \quad ، \quad B_i = \Psi(V_i)$$

We get the percentage of the twisted correlation between the variables X, Y estimated as follows:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

$$r_{pb} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2 \sum B_i^2}} \quad \dots (16)$$

***3.* Robust Correlation via Robust Principle Variables**

Suppose we have the correlation coefficients defined as follows:

$$\rho = \frac{var(U) - var(V)}{var(U) + var(V)} \quad \dots (17)$$

Such that, $V = \left(\frac{X}{\sigma_1} + \frac{Y}{\sigma_2}\right)/\sqrt{2}$ ، $U = \left(\frac{X}{\sigma_1} + \frac{Y}{\sigma_2}\right)/\sqrt{2}$

They represent basic variables so that there is no common variance between them, i.e., $Cov(U,V) = 0$

And: $Var(U) = 1 + \rho$ ، $Var(V) = 1 - \rho$

Presenting the standard function $S(X)$ as follows*:*

$$S(aX + b) = |a|S(X) \quad \dots (18)$$

Therefore, we can write the Robust Principle Variables of $S^2(.)$corresponding to what is fixed in equation (15) to get the correlation as follows:

$$\rho^*(X,Y) = \frac{S^2(U) - S^2(V)}{S^2(U) + S^2(V)} \quad \dots (19)$$

By substituting the Robust Principle Variables of the sample in equation (17), the Robust Principle Variables of the correlation become as follows:

$$\hat{\rho} = \frac{\hat{S}^2(U) - \hat{S}^2(V)}{\hat{S}^2(U) + \hat{S}^2(V)} \quad \dots (20)$$

Replacing the estimator of the standard deviation with the median absolute deviation, so that: $\hat{S} = MAD(x)$ *that is* defined in equation (18) to get a strong estimator known as the correlation coefficient of the median absolute deviation. It is known by Shevlyakov and Smirnov (2011) through the following equation:

$$r_{MAD} = \frac{MAD^2(u) - MAD^2(v)}{MAD^2(u) + MAD^2(v)} \quad \dots (21)$$

As each of u, v represent the basic immune variables; they are calculated as follows:

$$u = \frac{x - Med(x)}{\sqrt{2}MAD(x)} + \frac{y - Med(y)}{\sqrt{2}MAD(y)} \quad \dots (22)$$

$$v = \frac{x - Med(x)}{\sqrt{2}MAD(x)} + \frac{y - Med(y)}{\sqrt{2}MAD(y)} \quad \dots (23)$$

7. Criteria to Choose Best Model

There is a set of statistical criteria that will be addressed in this research to select the best models:

1. Akaike's Information Criteria

It is expressed mathematically according to Wei (2006) as follows:

$$AIC = 2k - 2\ln L \quad \dots (24)$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

Since AIC expresses the Akaike's criterion of information, L: expresses the function of greatest possibility.

2. Bayesian Information Criteria

Its formula can be explained according to Akaike, (1981) as follows:

$$BIC(k) = n \log \sigma_a^2 - (n-k) \ln\left(1 - \frac{k}{n}\right) + k \ln n + k \ln\left[\left(\frac{\sigma_Y^2}{\sigma_a^2} - 1\right)/k\right] \quad \ldots(25)$$

Since: σ_Y^2: represents the amount of variance; the model will be selected that corresponds to the lowest value of this criterion.

1. Chi- Square Criterion (χ^2)

The model becomes more suitable for the data depending on the value of the chi-square criterion. Whenever the ratio of the calculated chi-square value to the tabular value of chi-square is small, that is, the ratio is greater than (2). This indicates that the inappropriateness of the model for the data used. However, if this ratio is Less than (2), the model becomes suitable for the data (Abu Zaid & Bassiouni, 2021).

2. Absolute Fit Criterion

According to Taghza (2012), the Absolute Fit Criterion includes:

a) Goodness Fit Criterion is symbolized by (GFC), and its value ranges between (0-1). It is very similar to the coefficient of determination R^2. So, if the value of the goodness fit criterion is greater than 0.9, then this means that the model exists, and if its value is equal to integer values, it indicates that the proposed model matches the assumed model.
b) If Root Mean Square Error of Approximation (RMSER) value is less than or equal to 0.05, this means that the model matches the data. Likely, if RMSER value ranges between (0.08-0.05), it means that the model has been specifically successful. Furthermore, if the RMSER value is greater than 0.08, it means that there is a defect and the model is rejected.
c) The value of Root Mean Square Residual Standard (RMSR) ranges between (0-0.1), and its low value indicates the extent to which the model is consistent with the data used.

3. Incremental Fit Criterion (IFC), according to Al-Areeqi (2015), it includes:

a) The Standard Fit Criterion (NFC) has a value that ranges between (0-1). The high value of NFC indicates the extent to which the model matches the data.
b) Non-Normal Fit Criterion (NNFC) has a value which ranges between (0-1) and the model is consistent with the data when its value is greater than or equal to 0.95.
c) The value of Comparative Fit Criterion (CFC) ranges between (0-1). The value of 0.95 or more indicates a better fit for the model.

4. Root Mean Square Error (RMSE OR RSE)

RMSE represents the mean of the squares of the error, because the root is measured in the original units of the same values of the variables; it is calculated through the following equation:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \quad \ldots(26)$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

Simulation experiments

In this research, simulation experiments were performed for the problem under study. The researcher assumed a causal model that contains a several types of variables, i.e., external, internal and median random ones. The former includes random variables (x_1, x_2) and the second including (x_5), and finally i.e., median random variables include (x_3, x_4). The details and divisions of this causal model can be seen through the following figure:

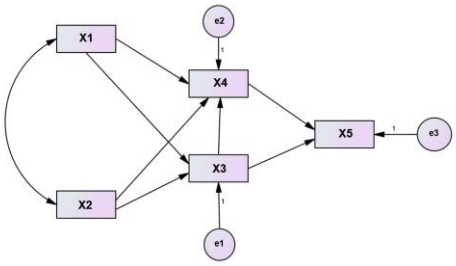

Figure 1. the causal model that includes the extrinsic, intrinsic, and median variables

Figure 1 shows that many relationship scenarios were applied using the R program to illustrate the simulation experiments. Many data were generated for samples of different sizes, $(n = 20,100,1000)$ and polluting the explanatory variables in the causal model with different contamination rates for each sample size, i.e., $(5\%, 20\%, 35\%)$. This was applied in generating data for random variables in order to form different sets of data with different problems.

The random vector $\underline{x}$, which includes the five variables of all kinds (external, median, and internal). They follow a multiple normal distribution with P variables, has been generated with mean μ, covariance and covariance matrix Σ. These variables can be clarified according to Chalmers and Adkins (2020) as:

$$\underline{x} \sim N_P\left(\underline{\mu}, \Sigma\right) \quad \ldots(27)$$

Such that: the random vector $\underline{x}$ includes the variables shown in the causal model in Figure 1. The covariance matrix Σ includes the variances of the variables in the causal model as well.

Assuming that there is a relationship between the variables in the causal model shown in Figure 1. Figure 1 also shows that there is a correlation between the variables through a correlation matrix that the researcher developed. Moreover, the mean score vector was determined, the covariance matrix was found among the five random variables. Then, the polluting process was applied, as the researcher relied on some of the above-mentioned percentages, which were chosen randomly, after which the abnormal values of the univariate data were generated using the box graph were revealed.

Furthermore, the process of estimating the path coefficients was carried out based on the aforementioned estimation methods. Because the correlation is a binary relationship between two variables, the anomalous values will be found through the bivariate boxplot that depends on the ranks as the box moves from the observations of rank $\frac{n}{4}$ to the observations of the order $\frac{3n}{4}$. The central bar of the box is drawn in the middle, and in general, the concept of a bivariate box graph expresses what the scientist Tuky (1975) described is called the biplot. Figure 2 demonstrates:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

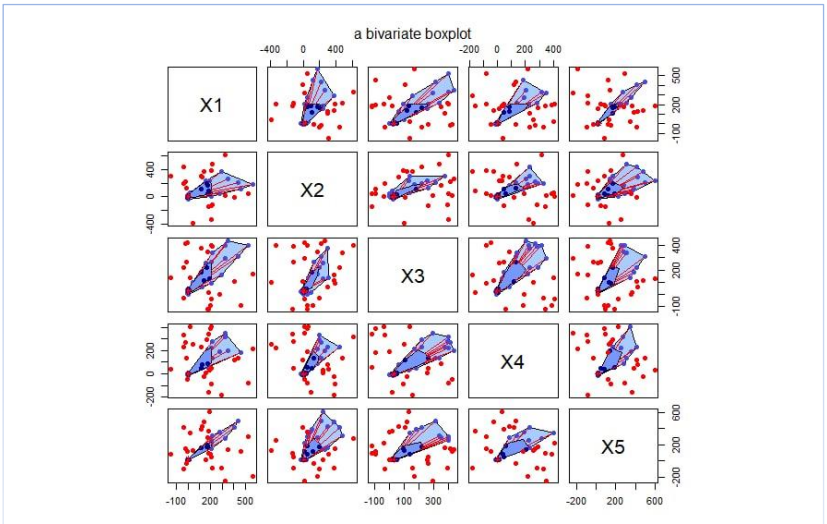

Figure 2. The bivariate box diagram.

After that, the path coefficients of the correlation matrices were found in the four ways for different sample sizes and contamination percentages that the researcher set. The results can be clarified through the Table 1.

Table 1. The contamination rates and criteria for the four methods with a sample size (n = 20)

	$(n = 20)$				
5%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	278.909	264.662	249.324	151.847
	BIC	291.853	277.607	262.268	164.791
	R$RMSE$	0.303	0.268	0.109	0.000
	$RMSR$	0.038	0.079	0.037	0.007
	$Chi - Square$	0.965	0.059	0.290	0.087
	CFI	0.975	1.000	0.990	0.860
20%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	271.916	274.218	263.503	229.186
	BIC	284.861	287.163	276.448	242.131
	R$RMSE$	1.318	0.000	0.281	0.406
	$RMSR$	0.218	0.023	0.049	0.119
	$Chi - Square$	0.775	0.000	0.076	0.014
	CFI	0.493	1.000	0.913	0.789
35%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	258.456	249.324	151.847	274.694
	BIC	271.400	262.268	164.791	287.642
	R$RMSE$	0.303	0.109	2.750	0.274
	$RMSR$	0.038	0.037	0.115	0.080
	$Chi - Square$	0.059	0.290	0.000	0.082
	CFI	0.088	0.990	0.112	0.975

Table 1 includes the estimation path of coefficients at different contamination rates, and at a sample size of n = 20 for the four methods. We note through the statistical criteria and at the 5% contamination rate, the method of the relative skewness correlation, symbolized by r_{pb}, is superior by obtaining the lowest values of the statistical criteria, which is a standard Akaike and the Bees criterion for information, as well as the root mean squares of residuals criterion (RMSER)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

and the root mean squares of residuals criterion as shown in the yellow color in Table 1. The goodness of fit criterion, in which the value of this criterion is close to the correct one, which indicates that the proposed model matches with the assumed model, followed by the Robust correlation method, symbolized by r_{MAD}, as shown in the green color in Table 1.

At 20% contamination rate, Table 1 shows that the relative torsion correlation method, symbolized by r_{pb}, has the lowest values of information standards, which are the Bees and Akaike standards, as shown in yellow in the Table 1. It was followed by the superiority of the binary correlation weighted method (Robust correlation) r_{bic} which got the lowest values of the statistical criteria, i.e., the root mean squares of error and the root of the mean of the squares of the residuals, as well as the chi-square criterion and the goodness of fit criteria, also shown in the yellow color in the table, followed by the robust correlation method r_{MAD} by obtaining the lowest values of four statistical criteria, which are the Akaike criterion and the Biz criterion information and the RMS error and RMS and the root mean square residual.

As for the percentage of contamination of 35%, we note the superiority of the robust correlation method symbolized by r_{MAD}. It obtains the lowest values of information criteria, followed by the second weighted correlation method (robust correlation) symbolized by r_{bic}. Likely, we also note the superiority of the weighted binary correlation method in other criteria, which are two root mean standards, the error squares and the root mean squares of the residuals, followed by the simple correlation method which symbolized by r. Generally, we notice from Table 1 a discrepancy in the values of the statistical criteria for the aforementioned four methods.

But when the sample size (n = 100), the path coefficients for the correlation matrices of the four methods and the contamination rates proposed by the researcher were found, and the results are shown in Table 1:

Table 2. The contamination rates and criteria for the four methods (n = 100)

	$(n = 100)$				
5%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	1290.003	1291.998	1194.939	1252.694
	BIC	1323.870	1325.865	1228.806	1286.562
	$RMSEA$	0.361	0.000	0.253	0.000
	$RMSR$	0.120	0.014	0.032	0.013
	$Chi-Square$	0.661	0.000	0.001	0.868
	CFI	0.845	1.000	0.949	1.000
20%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	1236.407	1188.275	1207.899	1248.599
	BIC	1270.274	1222.142	1241.766	1282.466
	$RMSEA$	0.500	0.000	0.000	0.000
	$RMSR$	0.087	0.010	0.019	0.013
	$Chi-Square$	52.030	1.022	1.789	1.100
	CFI	0.830	1.000	1.000	1.000
35%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	1367.881	1302.107	1166.004	1336.694
	BIC	1401.749	1335.975	1199.871	1370.561
	$RMSEA$	0.273	0.104	0.084	0.307
	$RMSR$	0.093	0.035	0.019	0.090
	$Chi-Square$	0.181	0.000	0.126	0.000
	CFI	0.811	0.984	0.995	0.835

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

Table 2 includes the estimation of the path coefficients for the correlation matrices of the four methods, the use of statistical criteria for comparison, and the statement of the preference of the methods used at sample size (n = 100) and at different contamination rates. As we note through the statistical criteria at 5% contamination that there is a discrepancy between the three methods, which is the robust correlation method denoted by r_{MAD} and the relative torsion correlation r_{pb} method, as well as the weighted binary correlation method r_{bic} or through the values of the statistical criteria. We were unable to indicate any of the methods are better in determining the path parameters. At 20% contamination, we clearly notice the superiority of the weighted binary correlation method r_{bic}, as shown in yellow in Table 2; it varies in the rest of the methods used.

While at 35% contamination, the superiority of the robust correlation method, symbolized byr_{MAD}, is noticeable. it obtained the lowest values of the statistical criteria, as shown in yellow in Table 2, followed by the r_{bic} weighted binary correlation method shown in green in Table 2.

In general, we notice a discrepancy between the two methods of r_{bic} and the robust correlation, symbolized by r_{MAD}. They obtain the lowest values of the statistical criteria.
But when the sample size is (n = 1000), the path coefficients of the correlation matrices of the methods used can be estimated and the results compared through the statistical criteria, as the results are shown in Table 2:

Table 3. the contamination rates and criteria for the four methods (n = 1000)

	$(n = 1000)$				
5%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	11941.689	11992.342	5105.224	6481.167
	BIC	12005.490	12056.143	5169.025	7112.235
	$RMSEA$	0.562	0.000	0.060	0.020
	$RMSR$	0.010	0.004	0.011	0.006
	$Chi - Square$	0.609	0.000	0.010	0.247
	CFI	0.935	1.000	0.997	1.000
20%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	11916.612	2310.372	11476.444	6099.820
	BIC	11980.413	2374.173	11540.245	6163.620
	$RMSEA$	0.421	0.000	0.724	0.052
	$RMSR$	0.006	0.006	0.070	0.003
	$Chi - Square$	0.382	0.026	0.000	0.000
	CFI	0.971	1.000	0.722	0.999
35%		r	r_{bic}	r_{MAD}	r_{pb}
	AIC	10444.363	2037.352	10957.779	2085.164
	BIC	10508.164	1973.551	11021.580	2021.363
	$RMSEA$	0.277	0.050	2.686	0.027
	$RMSR$	0.008	0.002	0.002	0.002
	$Chi - Square$	0.030	0.000	0.000	0.000
	CFI	0.991	0.999	0.184	0.991

We note from the above Table 3 which includes the estimation of the path coefficients for the correlation matrices by the four methods, at a sample size (n = 1000) and at different contamination rates. Furthermore, we note through the statistical criteria at 5% contamination, we note the superiority of the weighted binary correlation method r_{bic}shown in yellow in Table 3 for four statistical criteria followed by the r_{bic} relative skew correlation method, which are shown in green in Table 3.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1724-1736
https://publishoa.com
ISSN: 1309-3452

At 20% contamination rate, we notice clearly the superiority of the r_{bic} binary correlation method, which is shown in yellow, followed by the relative torsion correlation r_{pb} method, which is shown in green in Table 3.
While at 35% contamination, we also notice the superiority of the weighted binary correlation r_{bic}, which is shown in yellow color in Table 3. It obtains the lowest values of the statistical criteria, followed by the method of the relative torsion correlation r_{pb}. Furthermore, it is also followed by the robust correlationr_{MAD}, as shown in the yellow color in Table 3.
General, we notice that the larger the sample size is, the higher the contamination rate, the weighted binary correlation method and the hippocampal correlation excel in obtaining the lowest values of the statistical criteria, and its preference in estimating the path coefficients of the correlation matrix in this method.
The results obtained for the methods, sample sizes and the used contamination percentages can be summarized in Table 4.

Table 4. The contamination rates and sample sizes for the best methods

	$n = 20$	$n = 100$	$n = 1000$
5%	r_{pb}	r_{MAD}	r_{bic}
20%	r_{pb}	r_{bic}	r_{bic}
35%	r_{MAD}	r_{MAD}	r_{bic}

Table 4 shows according to the statistical criteria used that the larger the sample size and the higher the percentage of contamination of the data, the greater the robust correlation method or the so-called weighted binary correlation in estimating the path coefficients. However, there is a discrepancy in the preference of methods according to different sample sizes and contamination rates.

Conclusion

The research concluded that the robust methods are superior to the classical methods in estimating the coefficients path by using the correlation matrices of the mentioned methods and for all sample sizes and with any percentage of contamination, in addition, the larger the sample size and the percentage of contamination are, the better the results become.

References

1. Abu Zaid, N. R, & Bassiouni, A. A. A. (2021). Using the path analysis method in determining the factors affecting the rate of inflation in Egypt. *Journal of Financial and Commercial Research, 22* (3), 474-485 https://jsst.journals.ekb.eg/
2. Akaike, H. (1981). Likelihood of a model and Information Criteria. *Journal of Econometrics, 16*, 14-30.
3. Al-Areeqi, N. (2015). *The Integrated Amos Model*. http://www.youtube.com/watch?v=Duen58Ipfk
4. Chalmers, R. P., & Adkins, M. C. (2020). Writing Effective and Reliable Monte Carlo Simulations with the SimDesign Package. *The Quantitative Methods for Psychology,16*(4), 248-280.
5. Dan, E. & Ijeoma, O. (2013). Statistical analysis method of detecting outliers in aunivariate data in a regression analysis model. *International Journal of Education and Research, 1* (5), 1-24.
6. Davidson, E. S. (2012). *Predictors of sleep quantity and quality in college students.* (Unpublished Doctoral dissertation), Southern Illinois University Carbondale.
7. Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). Robust Estimation and Outlier Detection with Correlation Coefficients. *Biometrika, 62*, 531-545.
8. Freeman, P. (1980). On the number of outliers in data from a linear model. In Bayesian Statistics, Ed. JM.
9. Hadiya, W. A. (2011). A comparative study using path analysis and regression analysis in the blood pressure model. *Kirkuk University Journal of Administrative and Economic Sciences, 1*(1),154-175

10. Keller, G. & Brian W. (2000). *Statistic for Management and Economics*, (5th Edition). Duxbury, Thomson Learning U.S.A.
11. Langfelder, P., & Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software, 46*(11) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3465711/
12. Nair, P. K. (2007). *A path analysis of relationships among job stress, job satisfaction, motivation to transfer, and transfer of learning: perceptions of occupational safety and health administration outreach trainers*. (Unpublished Doctoral dissertation), Texas A & M University.
13. Olobatuyi, Moses E. (2006). *A user's guide to path analysis*. Maryland University pressof America, America.
14. Rodgers, J. L. & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66. https://doi.org/10.2307/2685263
15. Al-Talib, B.A, &, Shaker, S. M. (2018). Using Weighted Ridge Regression as A Proposed Approach to Remedy Outliyngness and MultiCollinearity in Mediation Analysis. *Kirkuk University Journal of Administrative and Economic Sciences,8* (2), 545-572.
16. Shevlyakov, G. L. (1997). Robust Estimator of the Scaling Parameter of Exponential Distribution in Fault Models. *Automation and Remote Control, 58*(2) 273–277
17. Shevlyakov, G., & Smirnov, P. (2011). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics, 40*(1&2), 147-156.
18. Shook, C. L., Ketchen Jr, D. J., Hult, G. T. M., & Kacmar, K. M. (2004). An assessment of the use of structural equation modeling in strategic management research. *Strategic management journal, 25*(4), 397-404.
19. Suhr, D. (2006). *The basics of structural equation modeling*. *Presented: Irvine, CA, SAS User Group of the Western Region of the United States (WUSS)*. http://www.lexjansen.com/wuss/2006/tutorials/TUT-Suhr.pdf
20. Taghza, M. B. (2012). *Exploratory and confirmatory factor analysis, its concepts and methodology using the SPSS package and LISREL*, (1st edition). Al Masirah for Publishing, Distribution and Printing.
21. Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.
22. Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.
23. Yuan, K. H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *53*(1),31–50. https://doi.org/10.1348/000711000159169