Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Use the robust RFCH method with a polychoric correlation matrix in structural equation modeling When you are ordinal data

Omar.S.Ibraheem and Mohammad J. Mohammad

Department of Informatics & Statistic, College of Computer & Mathematical Science, University of Mosul, Mosul, Iraq

Article information

Article history:
Received May 1, 2022
Accepted June 5, 2022
Available online December 1, 2022

Keywords: a polychoric correlation matrix, outlier, robust RFCH, SEM, fit indexes, WLSMV,ULSMV

Correspondence:
Omar.S.Ibraheem
Omarsalim85@uomosul.edu.iq

Abstract

Structural Equation Modeling is a statistical methodology commonly used in the social and administrative sciences and all other. In this research, the researcher made a comparison between methods of estimation Unweighted Least Squares with Mean and Variance Adjusted(ULSMV) and weighted Least Squares with Mean and Variance Adjusted (WLSMV). When we have a five-way Likert scale, the data is treated as ordinal using the polychoric matrix as inputs for the weighted methods with robust corrections. With robust standard errors ULSMV and WLSMV .No study compared these methods and the impact of outliers on them. where a robust algorithm is proposed to clean the data from the outlier, as this proposed algorithm calculates the robust correlation matrix Reweighted Fast Consistent and High Breakdown (RFCH), which consists of several steps and has been modified by taking the clean data before calculating the RFCH correlation matrix, where these data are data clean from outlier to add in the methods and to calculate a correlation matrix for each method where the purpose is to keep the ordinal data to calculate the polychoric matrix, which is robust to the violation of the assumption of normal distribution .By conducting a simulation experiment on different sample sizes and the degree of distribution to observe the accuracy of the proposed method for obtaining clean data. On methods ULSMV and WLSMV before and after the treatment process by calculating the absolute bias rate For the standard errors and the estimated parameters, in addition to studying the extent of their effect on the quality of fit indicators for each of the chi-square index, Comparative fit index (CFI), Tucker-Lewis Index (TLI), and Root-Mean-Squared-Error-of Approximation(RMSEA), Standardized Root Mean square Residual (SRMR), , with the robust corrections in the chi-square index for each of the methods WLSMV and ULSMV the accuracy of the proposed.

DOI: 10.33899/IQIOSS.2022.176201,

1. Introduction

Researchers and specialists have addressed various estimating approaches for structural equations. Modeling components, measurement errors, and correlation among the various factors are estimated, and the independent variables with the direct and indirect relationships connect the various independent variables. Social and behavioral research researchers use SEM, which has gained widespread appeal in the previous decades, to solve big problems. With the wide range of statistical analytic features that SEM

offers, researchers may build models that account for latent variables and measurement errors. Using ML as well as other techniques, such as to estimate methods for (GLS). When certain conditions are met, possess desirable asymptotic distribution, such as unbiased, consistency, and efficiency(Gregory R. Hancock and Ralph O. Mueller 2013) Therefore, the researchers recommend addressing the problem of outlier data before using estimation methods. For this reason, a robust method has been proposed to address the problem of outlier data through the use of a proposed RFCH robust algorithm to trim the data from outlier values and the use of both methods WLSMV and ULSMV with robust corrections in standard errors and fit indexes where These robust correction methods work with data that has non normal distribution but is also sensitive to outliers The proposed algorithm for cleaning the data from the outlier and calculates a robust RFCH(Reweighted Fast Consistent and High Breakdown)matrix of an outlier, where the researcher made a simple modification to the algorithm by taking the final data he reached by going through several estimators before calculating the matrix to be hired these robust data in all methods and to calculate a polychoric correlation matrix When we deal with data ordinal.

2- Objective

The researcher aims to address the problem of outliers when we have a Likert scale questionnaire form, so there are responses of individuals on a paragraph more than others, in addition to errors in data entry because the modeling requires a large sample size and the entry error is very likely. Studying the effect of an outlier on estimation methods and using the same estimation methods before and after treatment using robust RFCH aims to study the effect of the sample size and the degree of distribution on the estimator bias and standard error bias and use the same estimation methods before and after treatment using robust RFCH.It aims to study the effect of the sample size and the degree of distribution on the model's overall fit indexes.

3- The problem

Researchers in psychological and administrative sciences often use the ML and GLS estimation method without resorting to any test because the technique requires the assumption of a normal distribution. Thus other estimation methods deal with the non normal distribution, especially when the data are ordinal, and these methods are WLSMV, ULSMV. The problem of an outlier, as the outlier values affect the estimation of parameters, standard errors, and the fit indexes, although there are methods that deal with no normal distribution, the methods are not Robust for outlier values, so they require treatment before using the method of estimation by using robust method RFCH. When we have a Likert scale of five categories. We use new methods and corrections when we treat the data as ordinal using the polychoric matrix.

4- Structural equation models (SEM)

An important two-part of models employed in SEM includes measurement models and structure models. . CFA is used to correct for indicator measurement error, shaping the latent variables (factors). A model in which the exogenous variable x and the endogenous variables y are being measured is defined as

$$\mathbf{x} = \mathbf{\Lambda}_x \xi + \delta$$
$$y = \Lambda_y \eta + \varepsilon \quad (1)$$

The full structural Equation model is defined as

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \quad (2)$$

The covariance matrix is obtained as follows by

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(\mathrm{I}-\mathrm{B})^{-1}[\Gamma\Phi\Gamma' + \Psi](\mathrm{I}-\mathrm{B})^{-1}\Lambda_y' + \Theta_\varepsilon & \Lambda_y(\mathrm{I}-\mathrm{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'(\mathrm{I}-\mathrm{B})^{-1}\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{bmatrix} \quad (3)$$

Therefore the matrix of covariance was proven. (Timm 2002) (Byrne 2013)(Bollen 1989).

5- Polychoric Correlations

Polychoric correlation, explained by (Olsson 1979) can be calculated when ordinal data is involved. Ordinal variabley_1 and ordinal variabley_2 have distinct r and s class categories.

Usually, using the two-stage method, polychoric correlations computed by Olsson (1979) defined. The proportions of data for the category of an ordinal univariate variable are utilized independently in the first phase to approximate each latent univariate response variable's threshold values. gives both variables ordinal y1, with a_i, denotes, $a_i, i = 0, \dots, s$ and ordinal y2, with b i,j=0,…,r The first step is to set the thresholds at the estimated value of r and s.

$$a_i = \Phi_1^{-1}(Pi\cdot) \quad (4)$$

And $b_j = \Phi_1^{-1}(P{\cdot}j)$ (5)

The univariate standard normal for cumulative distribution function is denoted as Φ1, and *P ij* denotes in the proportion cell (*i, j*), *Pi*·and *P*·*j* denote the proportions cumulative marginal. (Flora and Curran 2004) (Yang-Wallentin, Jöreskog, and Luo 2010)

6- Estimation of Model Parameters

a -Weighted Least Squares (WLS)

When data are non normal, the most generally advocated estimate strategy is the asymptotically distribution-free (ADF) system (Browne, 1984).

When continuous and ordinal data stray greatly from normality, the use of this method is allowed. ' In the general situation, θ is the ADF estimator under the following GLS method: the vector that minimizes this function is

$F(\theta) = 2^{-1}[\text{vecs}\{\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\}]^T \text{V}^{-1}[\text{vecs}\{\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\}]$ (6)

the stochastic weight matrix V has a positive definite vector structure. can be written WLS minimizes the fit function . (Muthén and Asparouhov 2002) (DiStefano 2002)

b- Diagonally weighted squares and Robust DWLS with Corrections to Robust Standard Errors and Robust Test Statistics

The estimate of Diagonally WLS (DWLS) was developed to address the limitations of the full estimate of the WLS. Specifically, by decreasing the statistical sensitivity associated with the complete WLS estimator, DWLS eliminates the need for a large sample size DWLS may also incorporate scaling similar to the SB scaling approach, resulting in robust DWLS estimation(Gregory R. Hancock and Ralph O. Mueller 2013). The general form of the RWLS fit function is:

$\text{F}_{\text{DWLS}} = (S - \sigma(\theta))^T (W_D)^{-1} (S' - \sigma'(\theta))$ (7)

In ordinary data, one technique fits the SEM model with the polychoric correlation matrix rather than the sample covariance matrix called cat-DWLS. $W_D = \hat{\Gamma}_{\text{D}}^{*} = \text{diag}(\Gamma_c^{*})$ includes only diagonal elements of a polychoric corelation , and threshold projections approximate asymptotic covariance matrix. (Bollen, 1989; Muthén & Muthén, 2010).

However, the typical test statistics TWLS are not sufficient for model fit evaluation because the test statistics provided by cat-DWLS are no longer distributed asymptotically chi-square. This robust correction requires both corrections. The mean-adjusted chi-square statistic can also be implemented in the cat-DWLS estimator (Asparouhov and Muth 2010) proposed a new way to compute the mean- and variance-adjusted χ^2 (denoted as TDWLS-MV). The method of estimating this correction is called WLSMV: developed ways to compute the robust χ^2 test

$T_{\text{WLSMV}} = aT_{\text{DWLS}} - b$ (8)

Where $a = \sqrt{\frac{df}{\text{tr}[(\text{U}\hat{\Gamma}_{\text{c}}^{*})^2]}}$, and $b = df - \sqrt{\frac{df[\text{tr}(\text{U}\hat{\Gamma}_{\text{c}}^{*})]^2}{\text{tr}[\text{U}\hat{\Gamma}_{\text{c}}^{*}\text{U}\hat{\Gamma}_{\text{c}}^{*})\,]}}$

$df = s - t$, and $\hat{\text{U}} = \text{W}_{\text{D}}^{-1} - \text{W}_{\text{D}}^{-1}\hat{\Delta}'(\hat{\Delta}'\,\text{W}_{\text{D}}^{-1}\hat{\Delta}')^{-1}\boldsymbol{\Delta}'\,\text{W}_{\text{D}}^{-1}$

$\hat{\Gamma}_{\text{c}}^{*}$ Is the estimated asymptotic covariance matrix of s, s = the number of unique elements in s, and t = the number of independent model parameters. The method of estimating this correction is called WLSMV(Weighted Least Squares with Mean and Variance Adjusted).(Jia 2016)(Muthén 2002)

c- unweighted Least-squares and Robust ULS Robust Corrections to Standard Errors and Test Statistics

The ULS approach is simply a type of OLS estimation that minimizes the total squared differences between the sample and the covariance's expected by the model. This can obtain unbiased estimates through random samples. A downside of the ULS approach is the necessity that all variables observed be on the same scale. One benefit is that the ULS approach does not need a positive-definite covariance matrix, including ML(Kline 2016) estimation method does not require distributional assumption(Nalbantoğlu Yılmaz 2019)

The cat-ULS parameter estimates $\hat{\theta}_{\text{LS}}^{-}$ a saturated threshold structure by minimizing the fit can be represented as follows

$F_{\text{UIS}} = (r - \rho(\theta))'(r - \rho(\theta))$ (9)

Where r represent polycoric corelation matrix. . (Savalei and Rhemtulla 2013)

A recent proposal by (Asparouhov and Muth 2010) to implement an amendment in the second-order that does not modify the degree of the freedom. The Cat-ULS estimator determines the next method for the new mean and variance-adjusted statistics: ULSMV

$$T_{\text{ULSMV}} = aT_{\text{ULS}} - b \quad (10)$$

Where $a_{\text{ULS}} = \sqrt{\frac{df}{\text{tr}(\hat{U}_{\text{UIS}}\hat{\Gamma}^{*}\hat{U}_{\text{UIS}}\hat{\Gamma}^{*})}}$, $b_{\text{ULS}} = df - a_{\text{UIS}}\text{tr}(\hat{U}_{\text{UIS}}\hat{\Gamma}^{*})$, Represent $\text{df} = {}^{1}\!/_{2}\, k(k+1) - t$, $\hat{\Delta}_{\text{ULS}} = \frac{\partial p(\theta)}{\partial \theta'}|_{\text{LS}}$,

It is a standard matrix of ${}^{1}\!/_{2}\, k(k+1) * t$. 'These statistics are similar to the chi-square scaled by the so-called Satorra – Bentler, famous for continuous results. This applies to a chi-square distribution of df degrees of freedom, but that is just the approximate asymptomatic distribution. (Savalei and Rhemtulla 2013) (Xia and Yang 2018)(Asparouhov and Muth 2010)

7- The proposed method for processing data from outlier values represented by estimation Reweighted Fast Consistent and High Breakdown (RFCH)

Olive and Hawkins (2010) developed Reweighted Fast Consistent and High breakdown (RFCH) estimators of location and scatter, which was faster than the fast MCD developed by Rousseeuw and Driessen (1999). The attractive feature of the RFCH technique is that not only its computation is very fast, which is even faster than Fast MCD (Zhang et al., 2012), but it is $\sqrt{n}$ Consistent estimators. The RFCH utilizes the $\sqrt{n}$ Consistent DGK (Devlin et al., 1981) estimator and high breakdown Median Ball (MB) (Olive & Hawkins, 2008) estimators as attractors.

Mahalanobis (1936) defined Mahalanobis Distance (MD) to measure the deviation of a data point from its center. Let us write the i^{th} vector of predictor variables as:

$X_i' = (1, X_1, X_2, \dots, X_P) = (1, x_i)$

where x_i Is a p -dimensional row vector. The mean vector and the variance-covariance matrix are calculated as:

$\bar{x} = 1/n \sum_{i=1}^{n} x_i$ and $C = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$, respectively.

Subsequently, the (MD) for each observation is written as Equation:

$$MD_i = \sqrt{(x_i - T(X))'C(X)^{-1}(x_i - T(X))}\; i = 1,2, \dots, n \quad (11)$$

where $T(X)$ is the mean vector $(\bar{x})$ and $C(X)$ is the variance-covariance matrix (C).

8- The RFCH algorithms can be summarized as follows:

The RFCH consists of three steps where; in the first step, the Fast Consistent and High breakdown (FCH) attractors of Olive and Hawkins (2010) is determined based on the final attractors of DGK and MB estimators that adhere to the following rules:

The T_{FCH} and C_{FCH} are determined as:

$$T_{FCH} = \begin{cases} T_{K,DGK} \text{ if } & \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ T_{K,MB} & \text{Otherwise} \end{cases} \quad (12)$$

And Equation (12)

$$C_{FCH} = \begin{cases} \frac{\text{Med}\left(MD_i(T_{K,DCK}, C_{K,DGK})\right)}{x^2(p,0.5)} \times C_{K,DGK}, & \text{if} \quad \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ \frac{Med\left(MD_i(T_{K,MB}, C_{K,MB})\right)}{x^2(p,0.5)} \times C_{K,MB}, & \text{otherwise} \end{cases} \quad (13)$$

$$C^{*}_{2,RFCH} = \frac{\text{Med}\left(MD_i(T_{2,RFCH}, C_{2,RFCH})\right)}{x^2(p,0.5)} * C_{2,RFCH} \quad (14)$$

with the new cut-off point until convergence to get the final attractors $(T_{\text{RFCH}}, C_{\text{RFCH}})$ and $\tilde{X}_{\text{RFCH}}$, Subsequently, the Mahalanobis Distance based on is computed, and a new set of data is constructed using the following Equation (15) ;

$$clean\ data \quad = \tilde{X}_{3,RFCH} = \{X_{jl}: MD_i(T_{2,RFCH}, C^*_{2,RFCH}) \le x^2(p, 1-\alpha)\}, \quad j = 1,2,\dots,k, l = 1,2,\dots,m \tag{15}$$

(Olive and Hawkins 2010) (Uraibi and Midi 2019) (Zhang 2011)(Rousseeuw and Van Driessen 1999) (Olive and Hawkins 2008)(D. J. Olive, 2017)

9- Model evaluation

1- Robust Model-fit Indexes with methods robust estimation

The robust chi-square statistic, model degrees of freedom, scale factor, and shift factor for WLSMV and ULSMV is denoted as T, d, a, and b, respectively. PR model-fit indexes are determined for a sample size of n.

$$\text{RMSEA}_{.n} = \sqrt{max(0, \frac{a_M(n-1)\hat{F}_M + b_M - d_M}{(n-1)d_M})} \tag{16}$$

$$\text{CFI}_{.n} = 1 - \frac{a_M(n-1)F_M + b_M - d_M}{a_B(n-1)F_B + b_B - d_B} \tag{17}$$

$$\text{TLI}_{.n} = 1 - \frac{a_M(n-1)F_M + b_M - d_M}{a_B(n-1)F_B + b_B - d_B} \cdot \frac{d_M}{d_n} \tag{18}$$

(Xia and Yang 2019)(Savalei 2018)(Asparouhov and Muth 2010)

2- Residual-Based Fit Indices

a- Residual Matrix.

Residual matrix To examine the hypothesis that Σ = Σ(θ) you must calculate Σ−Σ(θ). A nonzero member in a null matrix indicates model definition error. To find S, you would use Σ(θ) as a substitution for Σ, and then you would use S − Σ(θ) to form S-Σ(θ) has elements, where each element is calculated as S − Σ(θ). Each parameter determines whether the model predicts covariance levels between observed variables *i* and *j* in the negative or positive definite. the correlation residuals(Hildreth 2013) (Ibrahim and Mohammed 2021)

$$r_{ij} - \hat{\rho}_{ij} = \frac{s_{ij}}{(s_{ii}s_{jj})^{1/2}} - \frac{\hat{\sigma}_{ij}}{(\hat{\sigma}_{ii}\hat{\sigma}_{jj})^{1/2}},\ (i,j = 1,\dots,p) \tag{19}$$

b- Standardized Root Mean square Residual (SRMR)

This formula is known as the "Root Mean Square Residual" (SRMR). Dr. Stephen Bentler created SRMR in 1995.SRMR is calculated the sample estimate and population is follows:

$$SRMR = \sqrt{\frac{1}{s}\sum_{i=1}^{p}\ \sum_{j=1}^{i}\ \frac{\left(\sigma^*_{ij} - \sigma_{0,ij}\right)^2}{\sigma^*_{ii}\sigma^*_{\hat{y}}}} \tag{20}$$

Where s = k(k + 1)/2). And $\sigma^*_{ij}, \sigma_{0,ij}, s_{ij}, \hat{\sigma}_{ij}$ are elements of $\mathbf{\Sigma}^*, \mathbf{\Sigma}_0, \mathbf{S}$, and $\hat{\mathbf{\Sigma}}$ Respectively. Represent s_{ij} is the sample covariances, $\hat{\sigma}_{ij}$ Is the model implied covariances, and *sii* and *s jj* are observed standard deviations. SRMR value has a value of 0 or 1, with 0 being the optimum fit and 1 representing the worse fit. (Kline 2016) (Schermelleh-Engel, Moosbrugger, and Müller 2003)

10-Simulation Design

The simulation was conducted to answer the research objectives and problems of the research. The simulation design, data generation and analysis procedures and evaluation of the results will be described. Continuous data were generated using the R program according to the method of(Vale and Maurelli 1983) and(Rhemtulla, Brosseau-Liard, and Savalei 2012) for a multivariate normal distribution with skewness and kurtosis of 0 and 0 and a distribution of moderate normal with skewness and kurtosis 2 and 7, and the number of variables required for the variance-covariance matrix as defined in the model, and then a set of thresholds are determined to convert each continuous variable into an ordered categorical variable, as the number of categories is equal to 5, and this is common in research. It is Generating data with different sample sizes and 500 replicates for each group with 20% contamination average for each sample size, randomly, where the proposed modified robust system is applied to clean the data from an outlier. The following Table shows the design of the simulation experiment for the model, sample sizes, and distributions.

11- Simulation population parameter models

The first model consists of four factors and 12 variables; each factor has three variables. We have three exogenous factors and one endogenous factor, and the indicators are loaded on the first three factors at 0.70. with making the indicators for one factor, they are generated random normality, with a mean equal to 0.5 and standard deviation 0.05, the scheme The following describes the design of a simulation experiment for a model

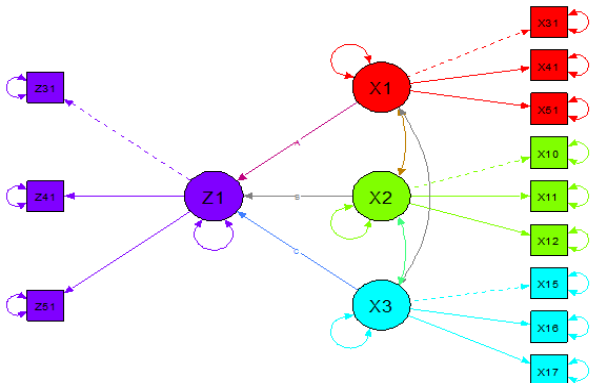

Diagram (1) the design of a model of the hierarchical model paths for the estimated parameters

As for the simulation model, it was designed as follows

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.7 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & rnom(1,0.5.0.05) \\ 0 & 0 & 0 & rnom(1,0.5.0.05) \\ 0 & 0 & 0 & rnom(1,0.5.0.05) \end{bmatrix}$$

where **Λ** load factors for X and Z, respectively

$$\Phi = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

As for Φ, it represents the correlation between the exogenous latent variables, as the correlation with the value 0.2 is shown in the matrix below

$$\Phi = \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}$$

$$dig(\boldsymbol{\Theta}) = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]$$

Also, the covariance matrix **Θ** represents the measurement error, or the variance of the residuals on the independent and dependent variables (indicators), which equals 1. In contrast, the covariance matrix of ψ reflects the correlations or variances of the factors located on the latent variables.

$$dig(\psi) = [1 \quad 1 \quad 1 \quad 1 \quad]$$

Whereas the matrix Γ represents the paths between the exogenous and endogenous latent variables so that these paths were generated with a multivariate normal distribution with a mean equal to 0.3 and standard deviation of 0. 5

$$\Gamma = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ rnom(1,0.3.0.5) & rnom(1,,0.3.0.5) & rnom(1,0.3.0.5) & \mathbf{0} \end{bmatrix}$$

The model consists of two parts, measurement the model, which is represented by the following mathematical equations

$$\left\{\begin{array}{ll} X_{31}=\lambda x_{31}X_1+\delta_3 & X_{153}=\lambda x_{153}X_3+\delta_{15} \\ X_{41}=\lambda x_{41}X_1+\delta_4 & X_{163}=\lambda x_{163}X_3+\delta_{16} \\ X_{51}=\lambda x_{51}X_1+\delta_5 & X_{173}=\lambda x_{173}X_3+\delta_{17} \\ X_{102}=\lambda x_{102}X_2+\delta_{10} & Z_{31}=\lambda z_{31}Z_1+\delta_3 \\ X_{112}=\lambda x_{112}X_2+\delta_{11} & Z_{41}=\lambda z_{41}Z_1+\delta_4 \\ X_{122}=\lambda x_{122}X_2+\delta_{12} & Z_{51}=\lambda z_{51}Z_1+\delta_5 \end{array}\right\} \quad (21)$$

As for the structural model, it is written in the following format

$$Z_1=\gamma_{11}X_1+\gamma_{12}X_2+\gamma_{13}X_3+\zeta_1 \quad (22)$$

The parameters λx_{11} ... λy_{51} , γ_{11} …. $\gamma_{13}$are unknown, and their estimation is required. The factor loads of the standard model, the measurement errors on the measured variable, and the structural model parameters represent a path analysis between the underlying variables.

12- The absolute bias average for standard errors

To determine the overall fit of the standard errors of the parameters, the total absolute bias average of the standard errors was calculated as shown in tables (1), which represents the bias for both factor loading, structural coefficients, correlations, influence by two methods estimation with the presence of outlier values and using the proposed method RFCH and according to the distribution normality and moderate distribution non-normality, as it was noted that the relative bias of errors decreased in all sample sizes and all methods, which indicates the quality of the proposed method to clean the data from an outlier in addition to the effect of an outlier on standard errors.

Table (1) represents the absolute bias average of the standard errors of the small model

	Sample size	200	400	600	800	1000
Dist.	ULSMV	0.4341	0.4116	0.4462	0.4227	0.4371
	CULSMV	0.16983	0.16893	0.2016	0.1704	0.18444
	WLSMV	0.337	0.30244	0.3402	0.3199	0.3571
	CWLSMV	0.16932	0.16792	0.20101	0.1784	0.18386
skew =2, kurtosis = 7	ULSMV	0.5096	0.4939	0.4156	0.4083	0.443
	CULSMV	0.1866	0.1705	0.1854	0.1805	0.19634
	WLSMV	0.4176	0.3498	0.361	0.36039	0.3612
	CWLSMV	0.1818	0.17297	0.18712	0.1884	0.19545

The (C) symbol is represented in front of each method using the clean data of the proposed method RFCH represent CULSMV and CWLSMV.As for the methods that deal with the data as ordinal by calculating the polychoric matrix in addition to using the robust corrections in the standard errors and the robust corrections in the chi-square, the values of the absolute bias average for the method of ULSMV before cleaning ranged between 0.4462- 0.4227, while CULSMV after using the method The proposed ranged between 0.16983 - 0.12016, and from this result. It is clear from this result that there is a clear difference using the RFCH method, as the errors were very small and less than using the WLSMV method directly with contaminated data.

By comparing the two methods, it is clear that both methods are ideal in terms of the relative bias of the standard errors of the clean data, and they give close results. And in some sample sizes, the WLSMV method is superior, and in other sample sizes, the ULSMV method is superior.

13- Bais parameter estimation

The total quality of the estimated parameters was calculated by calculating the absolute bias average for the parameters before and after cleaning with the presence of outlier using the proposed method, as it was noted through Table (2) that all the parameters estimated using the robust RFCH were very small compared to the contaminated data and for all methods.

Table (2) the absolute bias average for the parameters of the small model

Dist.	Sample size	200	400	600	800	1000
	ULSMV	333.559	300.937	49.9154	18.053	32.395
	CULSMV	1.394	0.227	0.18716	0.19073	0.19064
	WLSMV	91.9735	19.9744	9.17086	8.1207	31.702
	CWLSMV	1.42	0.2262	0.18677	0.1909	0.19071
skew =2, kurtosis = 7	ULSMV	660.797	471.36	60.455	98.805	32.1393
	CULSMV	2.9102	0.2823	0.2561	0.2792	0.245
	WLSMV	238.545	6.5361	15.7378	2.28352	4.9672
	CWLSMV	2.3659	0.28103	0.2557	0.2792	0.24503

It was noted through the data that was normality generated with outlier and cleaned using the proposed method that the overall bias average of the parameters is much smaller than the data that were assumed by the nonnormal distribution so that the performance of the robust and weighted methods without outlier is better for both two distributions, in addition to the evaluation of the model through the relative bias average For standard errors and estimated parameters, the quality of the proposed method is evaluated after cleaning from the outlier through the residual matrix, which represents the difference between the real parameter and the estimated parameter

14- fit indexes for small model

Through the simulation results of the previous model, the data follow the two distributions of first: skewness 2, kurtosis 7, and second: skewness 0, kurtosis 0, in the presence of an outlier. They are cleaned by the proposed method RFCH from outlier and use five sample sizes: 200, 400, 600, 800, and 1000. as well as it was noted that the fit indexes of differing according to the estimated method Because some methods use the correction robust chi-square, in addition, some fit indicators are based on the chi-square correction robust.

1-chi-Square fit Index

In comparison, its value was less when the nonnormal data is distributed, which indicates the robustness of the correction to deal with nonnormal data. And all chi values decreased after using the proposed method RFCH as shown in Table (3).

Table (3) the chi-square fit index for the small model

Dist.	Sample size	200	400	600	800	1000
	method	Chisq	chisq	chisq	chisq	chisq
skew 0=, kurtosis = 0	ULSMV	113.8416	170.9354	231.5337	297.2292	346.9776
	C.ULSMV	46.45603	45.58651	45.83298	45.64137	45.59435
	WLSMV	91.59424	133.8841	176.1701	222.8137	273.8744
	C.WLSMV	46.23925	45.6918	45.98901	45.70388	45.65046
skew =2, kurtosis = 7	ULSMV	110.4685	164.4508	210.8787	276.7401	330.7945
	C.ULSMV	47.19695	46.52312	46.1589	46.16313	46.52194
	WLSMV	97.19711	154.4712	204.8167	268.7435	320.2169
	C.WLSMV	47.70195	46.64953	46.24243	46.2121	46.65148

For the ULSMV and WLSMV methods, the use of (Asparouhov, & Muthén, 2010) correction robust method for mean and variance, especially when the distribution assumption is violated and in the presence of outliers It gave better results and noted that the chi-square index is biased for the sample size, the model size and affected by the degree of distribution, so other matching indicators have been developed based on

the chi -Square and the immune-corrected chi-Square robust, even though the process of cleaning from the outlier made all the values of chi-Square and for all methods close.

2- RMSEA fit Index

This is the most fitting indicator based on the estimation technique; It was noticed through Table (4) when using the proposed method and for all sample sizes that the RMSEA values had decreased and became within the ideal limits close to zero, and it was also noted that the value of RMSEA with the increase in the sample size approached to zero using the RFCH method and that the use of robust corrections for chi Square in the RMSEA index gave better results

Table (4) the RMSEA fit index values for the small model

Dist.	Sample size	200	400	600	800	1000
	method	RMSEA	RMSEA	RMSEA	RMSEA	RMSEA
skew 0=, kurtosis = 0	ULSMV	0.07729	0.07638	0.076292	0.077398	0.07632
	C.ULSMV	0.00996	0.006982	0.005776	0.005002	0.004306
	WLSMV	0.064078	0.064761	0.065101	0.0659	0.067281
	C.WLSMV	0.009658	0.007118	0.005906	0.005064	0.004389
skew =2, kurtosis = 7	ULSMV	0.07618	0.075328	0.072846	0.074898	0.074478
	C.ULSMV	0.01084	0.00759	0.006024	0.005238	0.00481
	WLSMV	0.06855	0.07249	0.071814	0.074092	0.07363
	C.WLSMV	0.01147	0.007781	0.006045	0.00536	0.004942

For the ULSMV and WLSMV estimation methods, we note that the fit index values are very small before and after cleaning and that they are smaller than the fit indicators for other methods and all sample sizes. We also note that the ULSMV method is superior to the methods by giving it a relatively lower value than the WLSMV method. This is if the data does not follow a normal distribution. But if the data follow a normal distribution, then through the table results, it was noted that the values of ULSMV for the clean data ranged between 0.004306-0.00996, while the WLSMV method ranged between 0.009658-0.004389.

3- SRMR fit index

This fit index is less affected by the chi-square determinants, which is an index of the covariance matrix of the residuals, and the closer to zero indicates that there is no error and that the recommended minimum is 0.08.

Table (5) the SRMR fit index values for the small model

Dist.	Sample size	200	400	600	800	1000
	method	SRMR	SRMR	SRMR	SRMR	SRMR
skew 0=, kurtosis = 0	ULSMV	0.0787	0.070284	0.06711	0.066576	0.066152
	C.ULSMV	0.0521	0.03536	0.028706	0.024638	0.022024
	WLSMV	0.07347	0.064963	0.061885	0.060871	0.060586
	C.WLSMV	0.052101	0.035412	0.028711	0.024651	0.02202
skew =2, kurtosis = 7	ULSMV	0.08379	0.076872	0.072374	0.07307	0.072196
	C.ULSMV	0.04853	0.034448	0.0282	0.024356	0.021946
	WLSMV	0.079676	0.073774	0.069234	0.069611	0.067858
	C.WLSMV	0.04853	0.034476	0.028205	0.024378	0.021973

The results are shown in Table (5) for all methods, whether normal or nonnormal distribution, the value of SRMR falls within the ideal limits. However, some methods such as WLSMV and ULSMV before cleaning also fall within the acceptable limits for the use of robust corrections in errors, as noted through The Table shows that these methods have the lowest SRMR compared to other methods when we treat the data as ordinal, where the ULSMV values for the nonnormal distribution ranged between 0.04853-0.021946 for the clean data, which indicates a perfect fit for the residuals of standard errors, which represent the difference between the sample matrix for the real data and the estimated matrix from the model, while the values of the WLSMV method ranged between 0.04853-0.021973, as it was noted that

with the increase in the sample size and for all methods after cleaning, it approaches more than zero and the least error for the residuals.

4- fit indexes TLI and CFI

These fit indicators the high value indicates a perfect fit through the results in the tables (6) (7) for all methods and all sample sizes and the two distributions, and the values of the two fit indicators lead to the rejection of the model when the data contains outlier values for most methods. At the same time, the values after using the proposed method RFCH obtained an ideal fit quality and were close to one. However, most methods after cleaning give very close results, especially when the data is normally distributed. We conclude from That is, with the increase in the sample size, increase the accuracy and robustness of fit indexes, as shown in the tables.

Table (6) the CFI fit index values of the small model

Dist.	Sample size	200	400	600	800	1000
	method	CFI	CFI	CFI	CFI	CFI
skew 0=, kurtosis = 0	ULSMV	0.80913	0.82075	0.82066	0.820054	0.82902
	C.ULSMV	0.98448	0.994214	0.996556	0.997476	0.998014
	WLSMV	0.878486	0.883508	0.884433	0.883871	0.880352
	C.WLSMV	0.986554	0.993926	0.996297	0.997353	0.997891
skew =2, kurtosis = 7	ULSMV	0.798505	0.820174	0.835068	0.828842	0.83122
	C.ULSMV	0.987735	0.9947	0.996604	0.997592	0.998048
	WLSMV	0.851528	0.844848	0.850562	0.844627	0.848025
	C.WLSMV	0.985737	0.993998	0.996187	0.997245	0.997745

Table (7) the TLI fit index values of the small model

Dist.	Sample size	200	400	600	800	1000
	method	TLI	TLI	TLI	TLI	TLI
skew 0=, kurtosis = 0	ULSMV	0.73755	0.753544	0.753398	0.752518	0.764868
	C.ULSMV	1.06101	1.00882	1.004278	1.003484	1.002636
	WLSMV	0.83364	0.839841	0.841095	0.840322	0.83548
	C.WLSMV	1.034713	1.008622	1.004102	1.00351	1.002683
Skew =2, kurtosis = 7	ULSMV	0.723005	0.752768	0.773196	0.764664	0.767928
	C.ULSMV	1.00595	1.004124	1.003232	1.00244	1.001554
	WLSMV	0.796838	0.786672	0.794524	0.786364	0.791032
	C.WLSMV	1.003515	1.004126	1.003423	1.002665	1.001587

In addition, the TLI and CFI fit indicators for the normal distribution, whether for contaminated data and clean data, after using the proposed method give greater results than if the data distribution is no normal.

15- Conclusions

We conclude from the simulation results that all methods with robust corrections in the weighted standard errors affected by the outlier. Using the proposed method RFCH, the absolute bias rate for standard errors and parameters and all models decreases significantly, indicating the algorithm's quality to get clean of outliers and improve the quality of parameters and reduce errors. We conclude that the absolute bias rate for parameters and standard errors is affected by the degree of distribution. It is less accurate when the data is not distributed normally. Through the simulation results after using the proposed method and for the clean data, we conclude through the comparison between the methods that the best methods are the ULSMV weighted and WLSMV; when we deal with the data, it is ordinal by calculating the polychoric matrix as input, In addition to the strong corrections in the standard errors because it has the least bias rate in standard errors and the least bias in the estimate parameters. By simulating different sample sizes and with an increase in the sample size, at a contamination rate of 20%, the absolute bias rate of errors

increases due to the percentage of contamination, but with the use of the proposed method RFCH, we conclude that the standard errors after cleaning and with the same sample size obtain stability, which indicates the quality of the method. Through the total quality based on the fit indexes, we conclude that all fit indexes decrease after using the proposed method and are within the limits of the ideal cut-off after cleaning. We conclude that the chi-square value is biased the sample size, as it rises with the increase in the sample size and the degree of distribution, so it is not recommended to rely on it. Through the simulation results, all the fit indexes are affected by the sample size, so we notice an increase in the accuracy of the quality of the fit indexes after using the proposed method for clean data as the sample size increases. Whereas TLI and CFI are close to one, so modeling requires a large sample size. Through the results, we conclude that the quality of fit indexes is affected by the degree of distribution. When the data are distributed in a normal distribution and free of an outlier, the fit indexes are more ideal than no normal distribution. By drawing the residual matrix for all methods, we conclude that the residuals approach zero and the normal distribution after cleaning using the proposed method. The use of the robust corrections of (Asparouhov, & Muthén,2010) in the estimation methods ULS and DWLS gave results and quality of fit greater by using correlation polychoric, especially when the data is distributed nonnormal, because of the robustness of this Correction on data that are not normally distributed.

References

Asparouhov, Tihomir, and Bengt Muth. 2010. “Simple Second Order Chi-Square Correction.” 1–8.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. John Wiley \& Sons.

Byrne, Barbara M. 2013. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. New York london: routledge.

DiStefano, Christine. 2002. “The Impact of Categorization With Confirmatory Factor Analysis.” *Structural Equation Modeling* 9(3):327–46.

Flora, David B., and Patrick J. Curran. 2004. “An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data.” *Psychological Methods* 9(4):466–91.

Gregory R. Hancock and Ralph O. Mueller. 2013. *Structural Equation Modeling: A Second Course (2nd Ed.)*. United States of America: Iap.

Hildreth, Laura. 2013. “Residual Analysis for Structural Equation Modeling.” *Statistics* PhD.

Ibrahim, Omar Salim, and Mohammed Jasim Mohammed. 2021. “A Proposed Method for Cleaning Data from Outlier Values Using the Robust Rfch Method in Structural Equation Modeling.” *International Journal of Nonlinear Analysis and Applications* 12(2):2269–93.

Jia, Fan. 2016. “Methods for Handling Missing Non-Normal Data in Structural Equation Modeling.” *Doctoral Dissertation, University of Kansas* 111.

Kline, R. B. 2016. *Principles and Practices of Structural Equation Modelling*. 4th ed. The Guilford Press.

Muthén, Bengt O. 2002. *Mplus Technical Appendices*. Retrieved from http://www.statmodel.com/download/techappen. pdf.

Muthén, Bengt O., and Tihomir Asparouhov. 2002. “Latent Variable Analysis With Categorical Outcomes: Multiple-Group And Growth Modeling In Mplus.” *Mplus Web Notes:* 5(4):1–22.

Nalbantoğlu Yılmaz, Funda. 2019. “Comparison of Different Estimation Methods Used in Confirmatory Factor Analyses in Non-Normal Data: A Monte Carlo Study.” *International Online Journal of Educational Sciences* 11(4):131–40.

Olive, David J. 2017. *Robust Multivariate Analysis*. USA: Springer.

Olive, David J., and Douglas M. Hawkins. 2008. “High Breakdown Multivariate Estimators.” 1–29.

Olive, DJ, and DM Hawkins. 2010. “Robust Multivariate Location and Dispersion.” *Unpublished Manuscript Available From (Http://Www. Math. Siu. Edu/Olive/ Pphbmld.Pdf)* 1–30.

Olsson, Ulf. 1979. “Maximum Likelihood Estimation of the Polychoric Correlation Coefficient.” *Psychometrika* 44(4):443–60.

Rhemtulla, Mijke, Patricia É. Brosseau-Liard, and Victoria Savalei. 2012. “When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods under Suboptimal Conditions.” *Psychological Methods* 17(3):354–73.

Rousseeuw, Peter J., and Katrien Van Driessen. 1999. “A Fast Algorithm for the Minimum Covariance Determinant Estimator.” *Technometrics* 41(3):212–23.

Savalei, Victoria. 2018. “On the Computation of the RMSEA and CFI from the Mean-And-Variance

Corrected Test Statistic with Nonnormal Data in SEM." *Multivariate Behavioral Research* 53(3):419–29.

Savalei, Victoria, and Mijke Rhemtulla. 2013. "The Performance of Robust Test Statistics with Categorical Data." *British Journal of Mathematical and Statistical Psychology* 66(2):201–23.

Schermelleh-Engel, Karin, Helfried Moosbrugger, and Hans Müller. 2003. "Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures." *Methods of Psychological Research Online* 8(2):23–74.

Timm, Neil H. 2002. *Applied Multivariate Analysis*. Verlag New York, Inc: Springer,New York Berlin Heidelberg.

Uraibi, Hassan S., and Habshah Midi. 2019. "On Robust Bivariate and Multivariate Correlation Coefficient." *Economic Computation and Economic Cybernetics Studies and Research* 53(2):221–39.

Vale, C. David, and Vincent A. Maurelli. 1983. "Simulating Multivariate Nonnormal Distributions." *Psychometrika* 48(3):465–71.

Xia, Yan, and Yanyun Yang. 2018. "The Influence of Number of Categories and Threshold Values on Fit Indices in Structural Equation Modeling with Ordered Categorical Data." *Multivariate Behavioral Research* 53(5):731–55.

Xia, Yan, and Yanyun Yang. 2019. "RMSEA, CFI, and TLI in Structural Equation Modeling with Ordered Categorical Data: The Story They Tell Depends on the Estimation Methods." *Behavior Research Methods* 51(1):409–28.

Yang-Wallentin, Fan, Karl G. Jöreskog, and Hao Luo. 2010. "Confirmatory Factor Analysis of Ordinal Variables with Misspecified Models." *Structural Equation Modeling* 17(3):392–423.

استخدم طريقة RFCH القوية مع مصفوفة الارتباط متعدد الألوان في نمذجة المعادلة الهيكلية عندما تكون بيانات ترتيبية

عمر سالم ابراهيم و محمد جاسم محمد

قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل ، العراق

قسم الاحصاء/كلية الادارة والاقتصاد/جامعة بغداد/بغداد/العراق

الخلاصة

نمذجة المعادلات الهيكلية هي منهجية إحصائية شائعة الاستخدام في العلوم الاجتماعية والإدارية وجميع المجالات الأخرى. أجرى الباحث في هذا البحث مقارنة بين طرق تقدير المربعات الصغرى غير الموزونة ذات المعدل المتوسط والمتغير المعدل (ULSMV) والمربعات الصغرى الموزونة ذات المعدل المتوسط والمتغير المعدل (WLSMV). عندما يكون لدينا مقياس ليكرت خماسي الاتجاهات ، يتم التعامل مع البيانات على أنها ترتيبية باستخدام مصفوفة متعددة الألوان كمدخلات للطرق الموزونة مع تصحيحات قوية. مع وجود أخطاء قياسية قوية ULSMV و WLSMV. لم تقارن أي دراسة بين هذه الأساليب وتأثير القيم المتطرفة عليها. حيث يتم اقتراح خوارزمية قوية لتنظيف البيانات من الخارج ، حيث تحسب هذه الخوارزمية المقترحة مصفوفة الارتباط القوية المعاد قياسها سريع الاتساق وعالي التفصيل (RFCH) ، والتي تتكون من عدة خطوات وتم تعديلها عن طريق أخذ البيانات النظيفة قبل حساب مصفوفة ارتباط RFCH ، حيث تكون هذه البيانات نظيفة من الخارج لإضافتها إلى الأساليب ولحساب مصفوفة الارتباط لكل طريقة حيث يكون الغرض هو الاحتفاظ بالبيانات الترتيبية لحساب المصفوفة متعددة الألوان ، والتي تعتبر قوية لانتهاك الافتراض من خلال إجراء تجربة محاكاة على أحجام عينات مختلفة ودرجة التوزيع لمراقبة دقة الطريقة المقترحة للحصول على بيانات نظيفة. حول طرق ULSMV و WLSMV قبل وبعد عملية المعالجة عن طريق حساب معدل التحيز المطلق للأخطاء المعيارية والمعلمات المقدرة ، بالإضافة إلى دراسة مدى تأثيرها على جودة مؤشرات الملاءمة لكل من مؤشر مربع كاي ، مؤشر التوافق المقارن (CFI) ، ومؤشر تاكر لويس (TLI) ، ومتوسط الجذر التربيعي للخطأ التقريبي (RMSEA) ، ومتوسط الجذر القياسي المتبقي (SRMR) ، مع التصحيحات القوية في مؤشر مربع كاي لكل من طرق WLSMV و ULSMV دقة المقترح.

الكلمات المفتاحية: مصفوفة ارتباط متعددة الألوان ، خارجية ، قوية RFCH، SEM ، فهارس مناسبة ، WLSMV، ULSMV.