Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 905-914
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Using Residual Estimators to Detect Outliers and Potential Controlling Observations in Structural Equation Modelling: QQ Plot Approach

A. R. Abdul-Aziz[1], Albert Luguterah[2], Bashiru I. I. Saeed[3]

[1]Department of Mathematics & Statistics, Kumasi Technical University (KsTU), Kumasi, Ghana
[2]Department of Statistics, Faculty of Mathematical Sciences, C.K. Tedam University of Technology and Applied Sciences (CKT-UTAS), Navrongo, Ghana
[3]Department of Statistical Sciences, Tamale Technical University (TaTU), Tamale, Ghana
Email: agbazorlic@gmail.com

How to cite this paper: Abdul-Aziz, A.R., Luguterah, A. and Saeed, B.I.I (2020) Using Residual Estimators to Detect Outliers and Potential Controlling Observations in Structural Equation Modelling: QQ Plot Approach. *Open Journal of Statistics*, **10**, 905-914.
https://doi.org/10.4236/ojs.2020.105053

Received: April 20, 2020
Accepted: October 27, 2020
Published: October 30, 2020

Abstract

The structural equation model (SEM) concept is generally influenced by the presence of outliers and controlling variables. To a very large extent, this could have consequential effects on the parameters and the model fitness. Though previous researches have studied outliers and controlling observations from various perspectives including the use of box plots, normal probability plots, among others, the use of uniform horizontal QQ plot is yet to be explored. This study is, therefore, aimed at applying uniform QQ plots to identifying outliers and possible controlling observations in SEM. The results showed that all the three methods of estimators manifest the ability to identify outliers and possible controlling observations in SEM. It was noted that the Anderson-Rubin estimator of QQ plot showed a more efficient or visual display of spotting outliers and possible controlling observations as compared to the other methods of estimators. Therefore, this paper provides an efficient way identifying outliers as it fragments the data set.

Keywords

Outliers, Controlling Observations, Estimators, QQ Plots, Structural Equation Modelling

1. Introduction

Issues associated with outliers are often looked at in textbooks, whilst in practical sense academics tend to have divergent views on its meaning and how it can rightfully be determined and managed, if possible [1]. Managing outliers of various kinds require different techniques. According to [2] there are 14 varied perspectives about outliers including, but not restricted to, issues of high leverage and the ability to overwhelm parameter estimation and model fitness in SEM.

Outliers are different from controlling observations as was established by [3]. Moreover, [1] noted that outliers often cause dissimilar stir on model adequacy as well as parameter estimation. Some techniques for spotting outliers and possible controlling observations in SEM were the likelihood, Mahalanobis and Cook's distances [2] [3] [4]. Other studies by [5] and [6] and [7] identified a linear notation for modelling outlier residuals in SEM. However, contemporary methods for identifying and controlling outliers and possible controlling observation in SEM require scientists to utilize special programs, which creates more burden for researchers [4] [8].

In a normal SEM model, very little portion of outliers and potential controlling observations can have a huge impact on model fit and parameter estimates. For instance, [9] and [10] demonstrated mathematically that existence of outliers can hugely inflate the Type I error rates of likelihood ratio test (LRT) and associated test statistics balancing for non-normality when using maximum likelihood (ML). The LRT statistic could be exaggerated by, at least, five times in figures as opined by [5]. It was also showed that in confirmatory factor analysis (CFA), by [5] and [11], that about 3% of outliers could necessarily bias the estimates of factor loading by not less than 50% and increase the covariance estimates and the latent factor variance about 3 - 10 times, as opposed to about 3% of bad controlling observations which could yield even higher biases on all parameter estimates. Again, [6] demonstrated mathematically with improved actual data sets that outliers produce worse fit indices; including but not limited to RMSEA and CFI, whereas potential controlling observations can lead to poor RMSEA value but better CFI for some cases.

According to [9] and [6], SEM based on normal-theory is not sturdy to outliers to the extent that a little presence outlier and potential controlling observations could bias both the model fits and parameter. However, robust modelling, which is achieved by substituting the normality assumption with an error term, that follows heavier-tailed t distribution has since been developed in regression models and multilevel models by [12] and [13]. Moreover, using the t-based SEM and other robust SEM methods is preferred, as opposed to deleting outliers and controlling observations directly, since the complex nature of SEM makes it very challenging to apply common methods including Mahalanobis and Cook's distance to identify outliers and potential controlling observations [8] [14]. Outliers and controlling observations are not mutually exclusive, notwithstanding the conceptual disparity, as some outliers can also exert strong influence on research results [2] [3] [4] [15].

Moreover, [2] reviewed 232 varied methods on organizational science matters about outliers and controlling observations, and just five of them were related to SEM, in spite of the popularity of SEM in the recent times. The main possible reason is that practical guidelines on managing outliers and potential controlling observations were evolved just recently [2] [3]. Notwithstanding the reported essence of outliers and controlling observations, detection and diagnostics of such observations were rarely performed and practice in real research, and in particular the use of SEM methods that are robust has been very scarce though SEM is notably made up of a measurement model(s) and a structural model [16] [17]. To this end, this study will apply a class of residual estimators, via simulation, to detect outliers and potential controlling observations in SEM using QQ plots.

2. Methodology

Seven different plots in residuals have been utilized for purposes of identifying outliers [18] but the uniform horizontal QQ plots are yet to be explored.

Therefore, the current method adopts a different approach which is the uniform horizontal QQ plot. Now, take a given linear regression equation:

$$y_i = x_i'\beta + \varepsilon_i$$

For y_i represents the outcome for observation *i*, x_i represents the predictor vector of size $p \times 1$ for observation *i*, β represents a vector of unestimated parameters of size $p \times 1$ and ϵ_i the random error term (0, σ^2). The predicted values, y_i, could then be plotted by a QQ residual plot as defined by

$$\hat{y}_i = x_i'\hat{\beta}$$

observed in the x-axis versus the residuals, e_i, defined as

$$\begin{aligned} e &= y_i - \hat{y}_i \\ &= y_i - x_i'\hat{\beta}_i \end{aligned}$$

observed in the y-axis. This could be extended in SEM by constructing the residuals $\hat{v}_i\left(\hat{\theta}\right)$ and $\hat{\zeta}_i\left(\hat{\theta}\right)$ versus its predicted counterparts in $\hat{v}_i\left(\hat{\theta}\right)$ and $\hat{\zeta}_i\left(\hat{\theta}\right)$ respectively [19]. By using residual estimators $\hat{v}\left(\hat{\theta}\right) = \left(I - \hat{\Lambda}\hat{W}\right)z$ and $\hat{\zeta}\left(\hat{\theta}\right) = \hat{M}\hat{W}z$ for the i^{th} observation, then

$$\hat{v}_i\left(\hat{\theta}\right) = \left(I - \hat{\Lambda}\hat{W}\right)z_i \tag{1}$$

$$\hat{\zeta}_i\left(\hat{\theta}\right) = \hat{M}\hat{W}z_i \tag{2}$$

Given that $\hat{\theta} = \hat{\theta}_{MLE}$.

Again, we obtain the predicted observations $\hat{z}_i\left(\hat{\theta}\right)$ and $\hat{\eta}_i\left(\hat{\theta}\right)$ which are linked to

$$\hat{z}_i\left(\theta\right) = \Lambda L_i + v_i$$

$$\hat{\eta}_i\left(\theta\right) = B\eta_i + \Gamma\xi_i + \zeta_i$$

The factor scores then replaces the observations in L_i to give $\hat{L}_i = Wz_i$ which then provides predicted observations with estimators [19];

$$\hat{z}_i(\theta) = \Lambda W z_i \tag{3}$$

$$\hat{\eta}(\theta) = [B\Gamma] W z_i \tag{4}$$

For practical implementation purposes estimators for predicted observations utilized, the vector θ with their sample counterparts $\hat{\theta}$, were

$$\hat{z}(\hat{\theta}) = \hat{\Lambda}\hat{W}z$$

$$\hat{\eta}(\hat{\theta}) = [\hat{B}\hat{\Gamma}]\hat{W}z$$

which could be predicted for the i^{th} value

$$\hat{Z}_i(\hat{\theta}) = \hat{\Lambda}\hat{W}z_i \tag{5}$$

$$\hat{\eta}_i(\hat{\theta}) = [\hat{B}\hat{\Gamma}]\hat{W}z_i \tag{6}$$

As QQ plots in a given linear equation, the predictors in (5) and (6) were plotted with their counterpart residuals $\hat{v}(\hat{\theta}) = (I - \hat{\Lambda}\hat{W})z$ and $\hat{\zeta}(\hat{\theta}) = \hat{M}\hat{W}z$ respectively.

Now, take a general set of sample quantiles to be sorted as

$$\mu_{(1)} < \mu_{(2)} < \mu_{(3)} < \cdots < \mu_{(n-1)} < \mu_{(n)},$$

The subscripts in the parentheses show an ordered data. The first ordered observation will lie in the horizontally in the middle of $(0, 1/n)$, the next in the middle of $(1/n, 2/n)$ and the last to be in the middle of interval $\left(\left(\frac{n-1}{n}\right), 1\right)$. Thus, we take as the theoretical quantile value

$$\xi_q = q = \frac{1-0.5}{n} \tag{7}$$

For q corresponding i^{th} ordered sample value. The quantity 0.5 is subtracted such that the data is exactly in the middle of the interval $\left(\left(\frac{i-1}{n}\right), i/n\right)$.

Now the QQ plot can precisely be defined. First, we compute through simulation the n expected values of the data, which we pair with the n data points sorted in ascending order. For the uniform density, the QQ plot is composed of the n ordered pairs

$$\left(\frac{1-0.5}{n}, u_{(i)}\right), \text{ for } i = 1, 2, \cdots, n$$

Deviations from the horizontal pattern allow for the spotting of possible issues outliers and/or controlling observations.

3. Results

Generally, it can be seen, from the QQ plots below, that at any percentile, the observations lie within a uniform horizontal scale of observations. Slight deviations from the horizontal scale show evidence of an outlier. However, observations that depart farther away from the uniform horizontal scale indicates evidence of potential controlling observation. The QQ plots proposed in this study differs from other methods and from one another based on the estimated residuals of the measurement errors using either the Anderson-Rubin or Bartlett's or Regression based methods for the simulated data using the EM method in detecting outliers and potential controlling observations for the SEM model.

It can be noticed from **Figure 1**, based on the Anderson-Rubin method, that there was evidence of an controlling observation within the first quartile (25th). Again, in the second quartile (50th) there was evidence of outliers which are observations deemed to lie close, about 0.5 cm, to the horizontal plane whereas the observation which lies farther away from the horizontal plane was identified as controlling observation within the median. Also, there were evidence of both outliers and controlling observations in the third quartile (75th). In the last quartile, observations can be seen lying almost on the horizontal plane and others lying within the 1.5 cm distance which were all deemed to be outliers with some few observation found to lie outside the reference distance or father away from the horizontal plane and a such were deemed to be controlling observations.

From **Figure 2**, based on the Bartlett's method, that there was no evidence of both outliers and controlling observations within the first quartile (25th). Meanwhile, the second quartile (50th) showed evidence of outliers which are observations deemed to lie close, about 1.5 cm, to the quantile horizontal plane whereas the observation which lies farther away from the quantile horizontal plane were represents the controlling observations within the median. Also, as can be seen

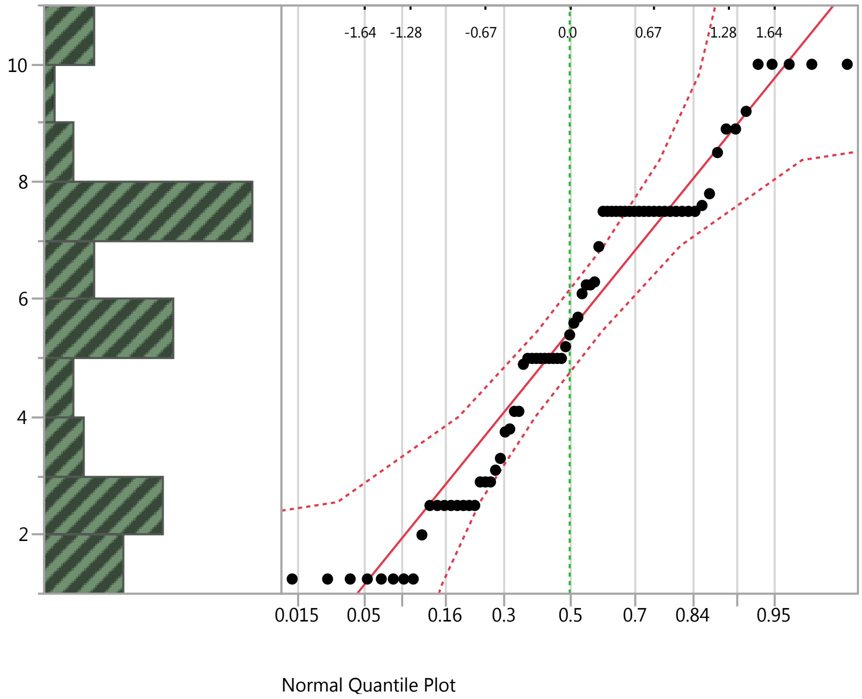

Figure 1. QQ plot for Anderson-Rubin based method.

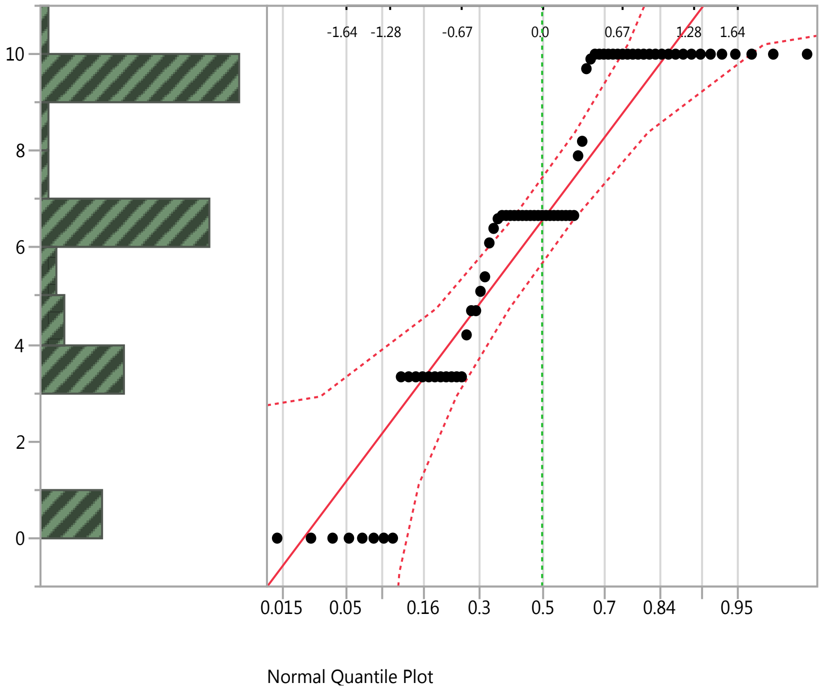

Figure 2. QQ plot for Bartlett's based method.

from **Figure 2**, there was evidence of both outliers and controlling observations in the third quartile (75th). In the last quartile, observations can be seen lying almost on the horizontal plane and others lying within the 0.5 cm distance which were all deemed to be outliers with some few observation found to lie outside the reference distance or father away from the horizontal plane and a such were deemed to be controlling observations.

It can be noticed from **Figure 3**, based on the regression method, that there was evidence of about three outliers and two controlling observations within the first quartile (25th). Also, the second quartile (50th) showed evidence of two outliers which are observations deemed to lie close, about 0.5 cm, to the quantile horizontal plane whereas the two observations which lie farther away from the quantile horizontal plane were identified as controlling observations within the median. Also, there was evidence of about four outliers and two controlling observations in the third quartile (75th).

The fitting indices, as indicated in **Figure 4**, for the QQ plot of the various estimation methods clearly support the earlier view, based on the QQ plots, that Anderson-Rubin based method provides a better visual display for the detection of outliers and influential observations. As seen from **Figure 4**, the first plot from left, which represents the Anderson-Rubin based method plot shows a smaller AIC, BIC and SABIC as compared to Bartlett's and the Regression based methods which are second and third from left.

4. Discussions

The paper applied a group of residual estimators to spot outliers and possible

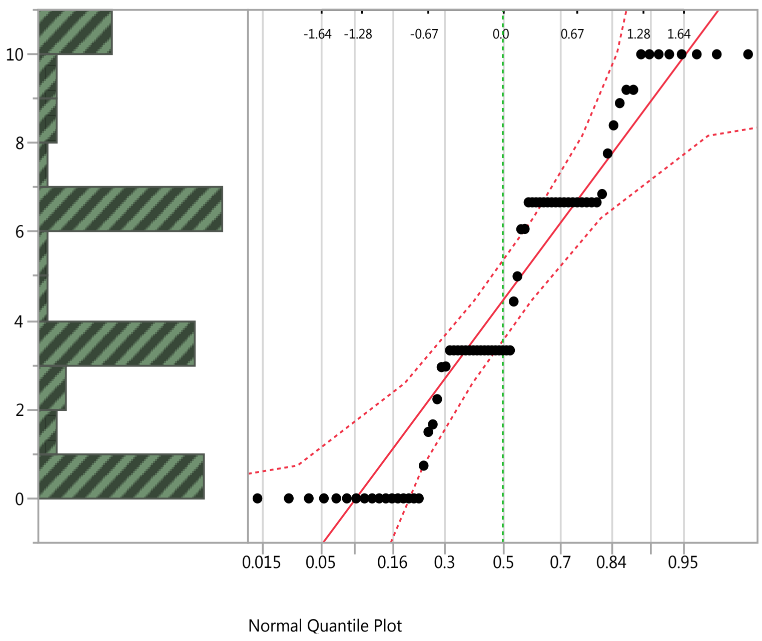

Figure 3. QQ plot for Regression based method.

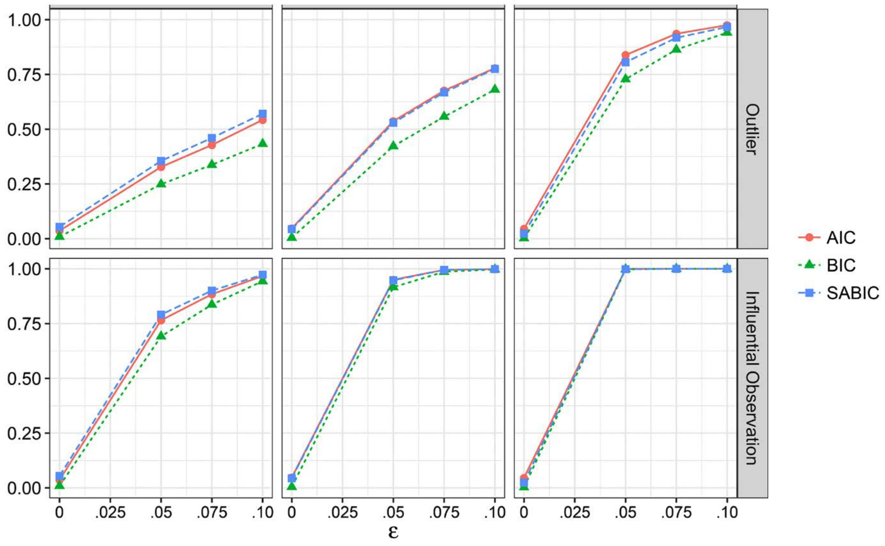

Figure 4. Information criteria for QQ plots.

controlling observations via uniform horizontal QQ plot. The study implemented the QQ plots in SEM using JMP software. The implantation experience supports [8] and [4] who opined that identifying outliers in SEM was rarely accessible due to the complexity of modelling, unlike traditional modelling in statistics including linear regression models.

Our results showed that the presence of outliers and possible controlling observation which affirms the assertion made by [6] who spotted outliers and controlling observations through boxplots. Moreover, it was revealing to note that the ease with which outliers and possible controlling observations could be spotted in this study, particularly with the Anderson-Rubin technique which contradicts

[2] who indicated that only outliers could be identified easily but noting the challenge in spotting controlling observations using boxplot under SEM framework.

Also, the present study found Aderson-Rubin technique the most efficient method of identifying outliers and possible controlling observations under SEM which corroborates the previous studies that utilized general techniques such as Mahalanobis and Cook's distances [8] [14]. These residual estimators provided parameters and fitness that are insensitive to the influence of outliers and possible controlling observations as they were achieved through a robust procedure of estimation by assigning weights. Further, the paper affirms the views noted in other studies that outliers and possible controlling observations need be of interest in their own right and therefore can lead to crucial scientific findings [1] [2] [15].

Again, the present study provides a different perspective to spotting outliers and possible controlling observations through a uniform horizontal QQ plots approach as was opined in earlier methodological works which provided accessible tools to identify outliers and possible controlling observations in SEM [1] [3] [8].

It is worth noting that despite the significant contributions of the study, there were some limitations that call further studies. To begin with, it should be emphasized that, in the current study, we only focused on situations where a small proportion of data is partitioned in each quantile, based on the moderate sample sized used. Also, the data used in the QQ plot was found to be normal and for that matter further studies could ascertain new way(s) of detecting outliers and controlling observations for a non-normal data with the same or similar concept of residual estimators. Corrections for non-normality such as the Satorra-Bentler procedure which relies on sandwich estimator and higher-order moments of the sample data could be adopted as data used under the SEM concept often had skewness and kurtosis deviated from those of a normal distribution [1].

5. Conclusion

It can be deduced from the results on the various simulations of QQ plots that all these methods demonstrate the ability to detect outliers and potential controlling observations in an SEM framework. It is worth noting that the Anderson-Rubin method of QQ plot provided a more efficient and visual display of detecting outliers and potential controlling observations as compared to the other classes of residual estimators. This, therefore, provides an efficient way of expanding the cook's method of detecting outliers and controlling observations with the QQ plot under the SEM framework.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lai, M.H.C. and Zhang, J.Q. (2017) Evaluating Fit Indices for Multivariate t-Based Structural Equation Modeling with Data Contamination. *Frontiers in Psychology*, **8**, 1286. https://doi.org/10.3389/fpsyg.2017.01286

[2] Aguinis, H., Gottfredson, R.K. and Joo, H. (2013) Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, **16**, 270-301. https://doi.org/10.3389/fpsyg.2017.01286

[3] Pek, J. and MacCallum, R.C. (2011) Sensitivity Analysis in Structural Equation Models: Cases and Their Influence. *Multivariate Behavioral Research*, **46**, 202-228. https://doi.org/10.1080/00273171.2011.561068

[4] Yuan, K.H. and Zhang, Z.Y. (2012) Structural Equation Modeling Diagnostics Using R Package Semdiag and EQS. *Structural Equation Modeling*: *A Multidisciplinary Journal*, **19**, 683-702. https://doi.org/10.1080/10705511.2012.713282

[5] Yuan, K.H. and Zhong, X.L. (2008) Outliers, Leverage Observations, and Controlling Cases in Factor Analysis: Using Robust Procedures to Minimize Their Effect. *Sociological Methodology*, **38**, 329-368. https://doi.org/10.1111/j.1467-9531.2008.00198.x

[6] Yuan, K.H. and Zhong, X. (2013) Robustness of Fit Indices to Outliers and Leverage Observations in Structural Equation Modeling. *Psychological Methods*, **18**, 121-136. https://doi.org/10.1037/a0031604

[7] Asparouhov, T. and Muthén, B. (2015) Structural Equation Models and Mixture Models with Continuous Nonnormal Skewed Distributions. *Structural Equation Modeling*: *A Multidisciplinary Journal*, **23**, 1-19. https://doi.org/10.1080/10705511.2014.947375

[8] Sterba, S.K. and Pek, J. (2012) Individual Influence on Model Selection. *Psychological Methods*, **17**, 582-599. https://doi.org/10.1037/a0029253

[9] Yuan, K.H. and Bentler, P.M. (2001) Effect of Outliers on Estimators and tests in Covariance Structure Analysis. *British Journal of Mathematical and Statistical Psychology*, **54**, 161-175. https://doi.org/10.1348/000711001159366

[10] Yuan, K.H. and Zhang, Z. (2015) rsem: Robust Structural Equation Modeling with Missing Data and Auxiliary Variables. R package Version 0.4.6. https://CRAN.R-project.org/package=rsem

[11] Yuan, K.H. and Hayashi, K. (2010) Fitting Data to Model: Structural Equation Modeling Diagnosis Using Two Scatter Plots. *Psychological Methods*, **15**, 335-351. https://doi.org/10.1037/a0020140

[12] Pinheiro, J.C., Liu, C.H. and Wu, Y.N. (2001) Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate *t* Distribution. *Journal of Computational and Graphical Statistics*, **10**, 249-276. https://doi.org/10.1198/10618600152628059

[13] Gelman, A. and Hill, J. (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511790942

[14] Flora, D.B., LaBrish, C. and Chalmers, R.P. (2012) Old and New Ideas for Data Screening and Assumption testing for Exploratory and Confirmatory Factor Analysis. *Frontiers in Psychology*, **3**, 55. https://doi.org/10.3389/fpsyg.2012.00055

[15] O'Connell, A.A., Yeomans-Moldanado, G. and McCoach, D.B. (2015) Residual Diagnostics and Model Assessment in a Multilevel Framework: Recommendations toward Best Practice. In: Harring, J.R., Stapleton, L.M. and Beretvas, S.N., Eds., *Ad-*

vances in Multilevel Modeling for Educational Research: *Addressing Practical Issues Found in Real-World Applications*, Information Age, Charlotte, NC, 97-135.

[16] Bollen, K.A. (1989) Structural Equations with Latent Variables, Wiley, New York. https://doi.org/10.1002/9781118619179

[17] Kline, R.B. (2011) Principles and Practice of Structural Equation Modeling. 3rd Edition, The Guilford Press, New York.

[18] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2004) Applied Linear Statistical Models. 5th Edition, McGraw-Hill/Irwin, New York

[19] Hildreth, L. (2013) Residual Analysis for Structural Equations Modeling. Iowa State University, Ames, IA.