Journal of Theoretical Educational Science, 15(1), 90-103, January 2022
Kuramsal Eğitimbilim Dergisi, 15(1), 90-103, Ocak 2022
[Online]: https://dergipark.org.tr/akukeg
DOI number: http://doi.org/10.30831/akukeg.917194

Effect of Influential Cases on Factor Analysis Results

Etkili Vakaların Faktör Analizi Sonuçları Üzerindeki Etkisi

Akif AVCU[*]

Received: 15 April 2021 **Research Article** **Accepted:** 01 November 2021

ABSTRACT: When performing regression analysis, one way to examine the normality of data is to screen outliers. Outliers, on the other hand, do not always have an effect on regression results. In reality, cases with a large amount of residuals that affect regression analysis results are referred to as influential cases. It is important to detect them in the dataset because they can lead to erroneous conclusions. The influence of influential cases has already gotten a lot of attention in the regression literature, while it has gotten a lot less attention in factor analysis. The aim of this paper is to show how influential cases affect factor analysis results when they are detected using the Forward Search algorithm. The data was collected from 686 university students ranging in age from 17 to 30. The data was gathered using the Self-Regulation Scale (SRS). The results revealed that the removal of influential cases had an effect on the observed correlation matrice for the SRS items, the factorability results, the number of dimensions extracted, CFA fit indices, and the amount of factor loadings and associated errors. Later, in light of related literature, these results were discussed and the researchers were recommended to consider the effect of influential when applying factor analysis.

Keywords: Influential cases, factor analysis, forward search.

ÖZ: Regresyon analizi gerçekleştirirken verilerin normalliğini araştırmak için uç değerlerin incelenmesi kullanılan yaklaşımlardan bir tanesidir. Gerçekte ise, uç değerlerin regresyon sonuçları üzerinde etkili olması bir gereklilik değildir. Aslında, regresyon analizi sonuçlarını etkileyen gözlemler, büyük miktarda artık barındıran etkili vakalar olarak adlandırılır. Veri setinde yanıltıcı sonuçlara yol açabilecek yüksek artık içeren gözlemleri tespit etmek önemlidir. Etkili vakaların etkisi, regresyon alan yazını hâlihazırda dikkat çekmişken faktör analizinde ise daha az vurgulanmıştır. Gerçekleştirilen bu çalışmanın amacı, ileri arama algoritması kullanılarak etkili durumlar belirlendiğinde bu vakaların faktör analizi sonuçları üzerindeki etkilerini göstermektir. Veriler, yaşları 17 ile 30 arasında değişen 686 üniversite öğrencisinden toplanmıştır. Çalışmada veri toplama aracı olarak Öz Düzenleme Ölçeği (ÖDÖ) kullanılmıştır. Sonuçlar, etkili vakaların veri setinden kaldırılmasının ÖDÖ maddeleri için gözlemlenen korelasyon matrisini, faktörlenebilirlik sonuçlarını, çıkarılan boyutların sayısını, doğrulayıcı faktör analizi uyum indekslerini ve faktör yüklerinin miktarını ve yüklere ait hataları etkilediğini ortaya koymuştur. Bu bulgular daha sonra ilgili alanyazın kapsamında tartışılmış ve araştırmacılar faktör analizi gerçekleştirirken etkili vakaların etkilerini dikkate almaları önerilmiştir.

Anahtar kelimeler: Etkili vakalar, faktör analizi, ileri arama.

[*] Dr., Marmara University, İstanbul, Turkey, avcuakif@gmail.com, https://orcid.org/0000-0003-1977-7592

Citation Information
Avcu, A. (2022). Effect of influential cases on factor analysis results. *Kuramsal Eğitimbilim Dergisi [Journal of Theoretical Educational Science], 15*(1), 90-103.

ISSN: 1308-1659

The existence of outliers is a frequently encountered problem in studies involving data collection. Although various definitions have been proposed so far, outliers can simply be defined as a data point that is far outside the norm of observations of a universe and as a value with extreme relationships (Rasmussen, 1988). They are the observations that seem inconsistent with the rest of the data (Barnett & Lewis, 1994). Researchers must carefully consider the issue of outliers since any statistical test based on sample means and variance can provide biased estimations in the presence of outliers. As in other analyses techniques, regression modeling is not exempt from the effect of extreme values on the estimates. In the related literature, this effect has been well established (Dan & Ijeoma, 2013; Hadi & Simonoff, 1993; Liu et al., 2004). Outliers in the dataset lead to model misspecification, the inflated sum of squares, distorted p values, biased parameter estimated, and drawing wrong conclusions. In the context of regression analysis, outliers are values that cause a large number of residuals when the dependent variable is predicted or modeled (Fox, 2008).

When conducting regression analysis, one method for checking the normality of data is to screen outliers. The Mahalanobis distance (MD) statistic has long been used in regression analysis to detect outliers. (Fox, 2008). The MD statistics help researchers to detect the cases away from the centroid (which could be regarded as the overall mean of multivariate data) and determine whether an observation is a multivariate outlier with respect to a set of explanatory variables. The Cook Distance (CD) and the generalized distance of Cook (gCD) are the two other well-known outlier case detection statistics used in the regression analysis (Pek & MacCallum, 2011). These approaches are also known as deletion statistics since the impact of cases on the model is measured over the entire dataset, and outlier cases are removed sequentially, resulting in a "*clean*" dataset (Cook & Weisberg, 1982).

One of the drawbacks of deletion statistics is that if a dataset includes a large of outliers, one of the two very similar cases can mask each other, resulting in only one of the two cases being deleted (Bendre & Kale, 1987). Although these statistics have gained popularity in the identification of outlier cases, their efficacy against masking effects is limited. On the other hand, the forward search method (Poon & Wong, 2004), which was originally designed to detect aberrant cases in multivariate methods such as cluster and discriminant analysis (Atkinson, 1994) and was modified to be used in regression models, does not have this limitation (Atkinson et al., 2004). Residual-based approaches and Cook's statistics are known as backward methods, which begin by fitting all the data to the model and deleting one observation at a time until the model fit remains at an acceptable level. On the other hand, forward search begins by fitting the model to a subset of the total data and then iteratively adding all the observations maintaining the fit.

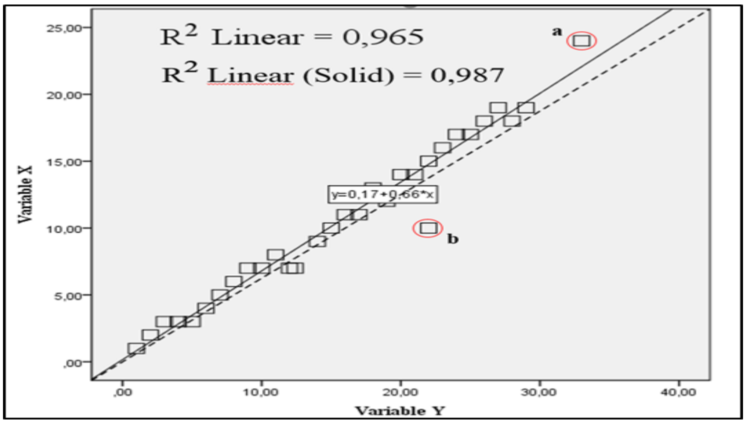
Briefly introducing, the forward search process continues by adding the observations that maintain the fit of the model by sorting the data units according to their contribution to the fitted model. In this process, the unit is selected primarily among the possible sub-units that are free of extreme values. These observations form the starting unit and are also referred to as the "*basic*" unit. The remaining units form "*non-basic*" units. Then, the process continues by adding observations from these "*non-basic*" units to the "basic unit" set one by one. Basic statistics such as model parameter estimates and goodness of fit statistics can be kept preserved in this method by adding cases that maintain the model's fit (Mavridis & Moustaki, 2008). According to a recent study, the forward search method outperforms deletion statistics when it comes to preventing the masking effect (Riani et al., 2008).

Outliers, contrary to popular opinion, do not always have an effect on regression outcomes. In fact, it is expected that a regression outlier has an important effect on regression results depending on its influence. At this point, an important distinction has to be made between outliers and influential cases. Although some overlap is expected between outliers and influential cases, they are not exactly the same. In fact, outliers don't always have an effect on regression analysis' outcomes. Indeed, the cases that affect the results of regression analysis are called influential cases, which have a large number of residuals. As a result, some outliers are referred to as "*good*" because they have a minor influence on the outcome, whereas others are referred to as "*bad*" because of their excessive influences. Researchers should realize the importance of looking into the influence as well as the extremeness of the cases. Truncating or censoring outliers prior to model analysis is strongly discouraged because removing influential outliers improves model fit while removing non-influential cases reduces the data size and statistical power (Tabachnick & Fidell, 2001).

In this study, an artificial dataset with two variables (named variable X and variable Y) was simulated to better understand the difference between the outlier and influential cases. The relationship between them is shown with a scatter plot in Figure 1 below. The case "*a*" in the figure is most likely an extreme value for the X and Y variables, as well as their linear combination, but it has little effect on the relationship between the variables since it is close to the solid line representing the model. Case "*b*", on the other hand, is closer to the data center and would most likely not be recognized as an extreme value as strongly as case "*a*", but it does have an effect on modeling the relationship between the two variables. The dashed line shows the model calculated by including the case "b". As shown, adding it to the equation changes the line's slope. The exclusion of the case "*a*" has little influence on the model results, while dropping case "*b*" improves the model's fit. All in all, despite the fact that both cases "*a*" and "*b*" are outliers, only case "*b*" is influential and bad.

Figure 1

Scatter Plot of an Artificial Data Showing the Difference Between Influential and Non-Influential Case.



Note: Solid Line Represents the Model Fitted without Case b, Intermittent Line Represents the Model Fitted with All Cases.

A data analysis could result with higher or lower correlations and biased regression coefficients when influential cases are ignored (Dan & Ijeoma, 2013). Obviously, such influential observations should be identified and a decision needs to be given about their inclusion in statistical analysis. As previously mentioned, the effect of influential cases has received attention in regression literature, but it is less so in factor analysis. On the other hand, the basic principles of regression analysis can be applied to common factor analysis since the common factor model is a version of the linear regression model (Chalmers & Flora, 2015). Independent variable(s) are used to explain a dependent variable in regression models, while factors are used to explain the relationships between observed variables in factor analytic models. Hence, the regression diagnostics, for example, which is used to analyze problematic data for the model, is also applicable to Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) techniques (Flora et al., 2012). Although they are viewed as techniques involving very different processes, both techniques regard factors as independent variables and observed variables as dependent variables. Since the usage of Pearson correlations and covariances is still the main approach, normality and linearity are among the basic assumptions to conduct factor analysis. For this reason, strong relationships may not be captured when a dataset contains influential cases.

Influential cases are often ignored in factor analysis because researchers using factor analysis and structural equation models are more concerned with covariance or correlation matrices while residual values are at the heart of regression analysis. Since the researchers are primarily interested in these matrices, they pay no attention to the individual cases that generate them. The key explanation for this ignorance is that because some researchers mistakenly believe that the effect of assumption violations and the presence of influential cases in datasets is minimal when conducting factor analysis since factor analysis is performed with a large number of cases and is a robust technique to such cases (Flora et al., 2012). As a result, the application of influential case detection techniques to factor analytic models is relatively new (e.g., Lee & Wang, 1996; Tanaka et al., 1991). As a result, only a few studies investigating the impact of influential cases on factor analysis results have been published. For instance, Yuan and Zhong (2008) compared three robust MD-based statistics with the simulated datasets. Similarly, Liu et al. (2012) use simulated data and investigated how outliers affect the decisions about the number of factors in exploratory factor analysis.

Further, Koran and Jaffari (2020) studied the accuracy of conventional deletion statistics when they are applied to confirmatory factor models. Finally, Mavridis and Moustaki (2008) investigated the efficacy of the Forward Search approach in detecting influential cases when factor analysis was performed. According to that study, the forward search method is more efficient than deletion statistics in detecting influential events. This study was also conducted with simulated data. In summary, previous studies on this subject use simulated data and focus on methodological comparisons. I provided extensive technical information. Furthermore, only one study used the forward search algorithm that is more powerful technique for finding influential cases than conventional deletion statistics.

From this perspective, it was thought that an illustrative study, which uses real data and includes less technical terminology, would be useful for researchers who use factor analysis in social science studies. In this way, interested readers will be able to see firsthand how powerful influential cases can be in determining construct dimensionality and model fit.

As a result, the aim of this paper is to show how influential cases affect factor analysis results when using the Forward Search algorithm. The Forward Search was used in this study because of its applicability for multivariate data and its superiority over deletion methods for a potential masking effect (Flora et al., 2012) and efficiency in finding influential cases. This study is expected to help researchers pay more attention to influential cases in their datasets and incorporate the forward search approach into their standard test validation processes.

Method

Participants

The participants of the current study were 686 university students ranging in age from 17 to 30 (*mean*=22.13; *ss*=2.84). Three hundred and seventy-six (54.8%) were enrolled in a public university located in a large metropolis, while 310 (45.2%) were enrolled in a private university located in a mid-sized metropolis. The convenient sampling was used to select the universities and the students. Furthermore, 559 (81.5%) of the students are female, while 127 (18.5%) are male. Participants were included in the research on a voluntary basis and were told that the information they provided would be kept confidential. The data collection process was done by an online data collection platform due to the 2020 pandemic outbreak.

Measurement Tool

A real dataset was used in this study to show the impact of influential cases in a more concrete way. As a result, the Self-regulation Scale was used to collect data (SRS). The scale was developed by Schwarzer et al. (1999). The scale consisted of seven items which are scored by rating each item on a 4-point Likert type (1=Completely wrong, 4=Completely correct) scale. A person’s high score implies s/he has more ability to control and collect attention. The English version of the scale was adapted by Diehl et al. (2006) and the Turkish version was adapted by Demiraslan et al. (2015). For the adaptation study, the Exploratory and Confirmatory Factor Analysis methods were used to examine the construct validity of the Turkish version. Additionally, the internal consistency of the Demir Aslan SRS was evaluated by the Cronbach Alpha coefficient and the test-retest correlation coefficient was also computed to evaluate the stability of SRS scores. Furthermore, The General Self-Efficacy Scale and Academic Self-Efficacy Scale were utilized for investigating criterion-related validity. As a result of factor analysis, it was observed that the Turkish version of the SRS was unidimensional. Cronbach Alpha internal consistency coefficient of the scale was found to be .84, test-retest reliability coefficient was .67. It was found that the SRS scores showed a significant and positive relationship with the Demir Aslan scores obtained from the General Self-Efficacy Scale and Academic Self-Efficacy Scale. The analyzes showed that the Turkish form of the Self-Regulation Scale is a valid and reliable measurement tool for university students.

Statistical Analysis

The current study carried out the whole statistical procedures on the R statistical program (R Core Team, 2020). The analyses were repeated three times, and different sample sizes were used for each repetition. Each analysis depicts a variety of testing scenarios that are experienced in validation studies. The entire dataset was used for the first scenario that specifies a situation in which a researcher uses a factor analytic approach without removing any outliers. At the second step, the analysis was repeated by removing outlier cases based on MDs of the cases. This scenario describes a situation in which a researcher is aware of the impact of outlier cases. Nevertheless, it uses a traditional and less effective deletion statistic that is susceptible to masking effect. For this analysis, the R function of "*mahalanobis*" was used. At this stage, case removal was conducted by considering *p<.001* significance level. As a result, influential cases still remain in the dataset. In the final stage, the forward search algorithm was used to remove influential cases that could be omitted in the previous step. This final step describes a scenario in which a researcher is aware of the impact of influential cases and employs the forward search approach to efficiently eliminate them. For this analysis, the "*forward.search*" command was used, which is available in the "*faoutlier*" package (Chalmers & Flora, 2015).

At each stage, both the EFA and CFA were conducted and the results were presented. The EFA was carried out with the R command the "*fa*," which is available in the "*psych*" package developed by (Revelle, 2015). The CFA was performed using the command "*cfa*" which is available in the "lavaan" package (Rosseel, 2012). The results of each step were compared in order to demonstrate how different scenarios produce different dimensional structures and estimates.

Ethical Procedures

Ethical approval and written permission were obtained from Marmara University Institute of Educational Sciences Research and Publication Ethics Committee with the decision dated 19.11.2020 and numbered 2020/85. The research was carried out in accordance with ethical rules at every stage. Participation in the research took place on a voluntary basis.

Results

The "*complete*" dataset, including 686 cases, was used in the first step to conduct both the EFA and the CFA. The MDs were measured later, and 17 cases were identified as outliers at *p*<.001 significance level, and they were excluded from the dataset for further study. Following the removal of these cases, the EFA and CFA analyses were repeated with "*outlier case-free*" dataset composed of 669 cases. Finally, the Forward Search algorithm was used to detect the cases with the highest influence statistics and the first fifty cases were removed. This "*influential case-free*" dataset was used to repeat the EFA and CFA analyzes with the remaining 619 cases.

The Results of the Analysis for the Complete Dataset

The factorability of the complete dataset was firstly examined before the EFA was performed. Accordingly, KMO sample adequacy value of the scale was found to be .807. The Bartlett's Test of Sphericity result is statistically significant at *p<.001* level.

Anti-image correlations were in the range of .860 to .731, and communalities were found to be varied between .506 to .688. All these results supported the factorability of the SRS. After the investigation of the factorability results, the EFA was performed and the results provided by the EFA suggested a two-factor solution. The two-factor solution was found to contribute to 58.49 percent of the total variance: the first factor with eigenvalue of 2.95 explains 42.17 percent of the total variance, while the second dimension with an eigenvalue of 1.14 explains 16.31 percent of variance. Based on the EFA results, the 4th, the Ross eel 5th, and the 6th items of the SRS were found to gather on a second dimension while the rest of the items were placed in the first dimension. Additionally, the CFA results showed that the single factor solution was not confirmed for the complete dataset [χ^2=155.88, df=14, χ^2/df=11.13, CFI=.87, TLI=.81, NFI=.86, GFI=.93, AGFI=.86, RMSEA=.12, SRMR=.07]. Subsequently, 17 cases were eliminated from the dataset based on the MDs and the second phase of analysis was conducted. In the rest of this manuscript, the information regarding the results provided by factor analyzes carried out at the second and third phases was presented together for the convenience of readers.

The Results of the Analysis for Outlier Case-free and Influential Case-free Datasets

Before presenting the EFA and CFA results, the inter-item correlations and item statistics that were calculated for both "outlier-free" and "influential case-free" datasets were presented in Table 1 below. The coefficients in the upper diagonal were calculated with the outlier the case-free dataset, while the coefficients in the lower diagonal were calculated with the influential case-free dataset. The findings revealed that small increases were observed for correlation values after removing influential cases from the dataset.

Table 1

Pearson Correlations of Items for Both the Outlier Case-Free and the Influential Case-Free Datasets

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Item 1	-	.44	.48	.46	.30	.30	.22
Item 2	.46	-	.57	.46	.27	.26	.15
Item 3	.50	.56	-	.53	.30	.32	.24
Item 4	.50	.51	.58	-	.47	.43	.18
Item 5	.36	.35	.36	.48	-	.46	.00
Item 6	.35	.34	.39	.46	.43	-	.07
Item 7	.24	.17	.28	.24	.09	.14	-

The statistics obtained for item statistics and internal consistency coefficients for both datasets were given in Table 2 below. As presented in the table, the corrected item-total correlations calculated for the influential case-free dataset are a little higher. Similarly, the value of the Cronbach α calculated for this dataset showed an increase as compared to the outlier-free dataset. That is, influential cases have an impact on both item-total correlations and internal consistency of the SRS items. On the other hand, when the effect of inclusion of each item to the test on the reliability level was investigated, it was seen that the removal of each item has the same impact on test reliability for both datasets. For example, for both datasets, removing the 7th item increases the level of reliability.

Table 2

Item Statistics and Reliabilities of Datasets with and without Influential Cases

	Complete with Influential Cases		Data without Influential Cases	
Items	Corrected Item Total Cor.	Alpha if Item Deleted	Corrected Item Total Cor.	Alpha if Item Deleted
Item 1	.55	.72	.58	.76
Item 2	.53	.72	.57	.76
Item 3	.61	.70	.65	.74
Item 4	.65	.69	.69	.73
Item 5	.45	.73	.49	.77
Item 6	.46	.73	.51	.77
Item 7	.19	.80	.26	.82
Cronbach α		.76		.79

When the EFA results were investigated, it was seen that a two-factor solution was suggested as the first factor has the eigenvalue of 3.10 and explains 44.31% of the total variance. In comparison, the second dimension has the eigenvalue of 1.12 and explains 15.96% of the total variance for outlier case-free dataset. These two factors contributed to 60.27% of the total variance. Another result showed that data removal based on the MD did not affect the factor structure.

Then, with the removal of the fifty most influential cases, the analyzes were repeated. As a result, in line with the original structure of the SRS, one-dimensional structure was supported. This dimension had an eigenvalue of 3.33 and explained 47.54% of the total variance. As it can be inferred, removal of influential cases yields a factor structure compatible with the original factorial structure of the SRS.

Table 3

The EFA Results with Complete Data and the Data without İnfluential Cases

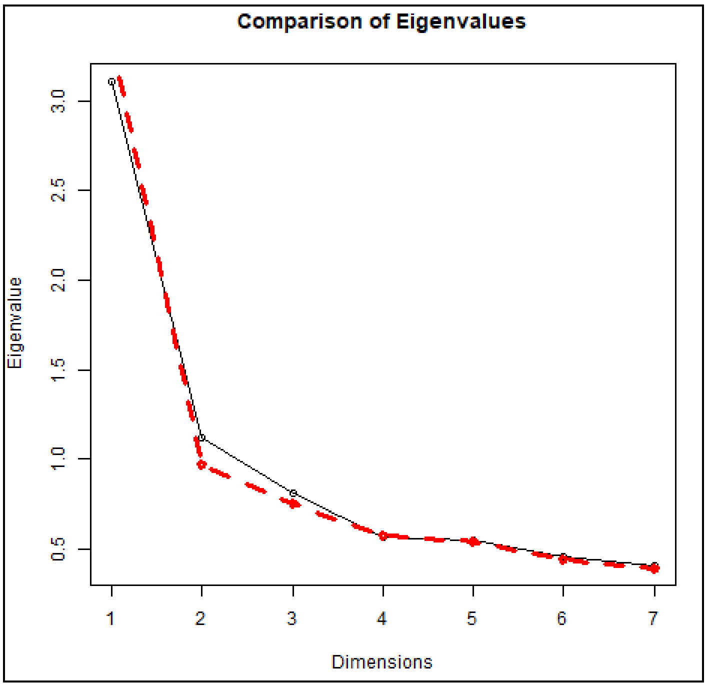
	Data with Influential Cases			Data without Influential Cases	
Items	Dim. 1	Dim. 2	Communality	Dim. 1	Communality
Item 1	.65		.54	.73	.53
Item 2	.66		.55	.73	.54
Item 3	.73		.65	.79	.62
Item 4	.53	.61	.65	.82	.67

Item 5		.81	.67	.64	.41
Item 6		.73	.57	.65	.43
Item 7	.71		.60	.37	.14
Eigenvalues	3.10	1.12		3.33	
Var. explained (%)	44.31	15.96		47.54	

Figure 2 shows how the eigenvalues of extracted factors are affected as a result of removing influential cases from the dataset. As can be shown, the scatter plot’s slope became steeper as the influential cases were removed, and the unidimensional form became more interpretable based on the plot.

Figure 2

Effect of Influential Cases on the Eigenvalues. Red Line Pertains to the Data Without Influential Cases



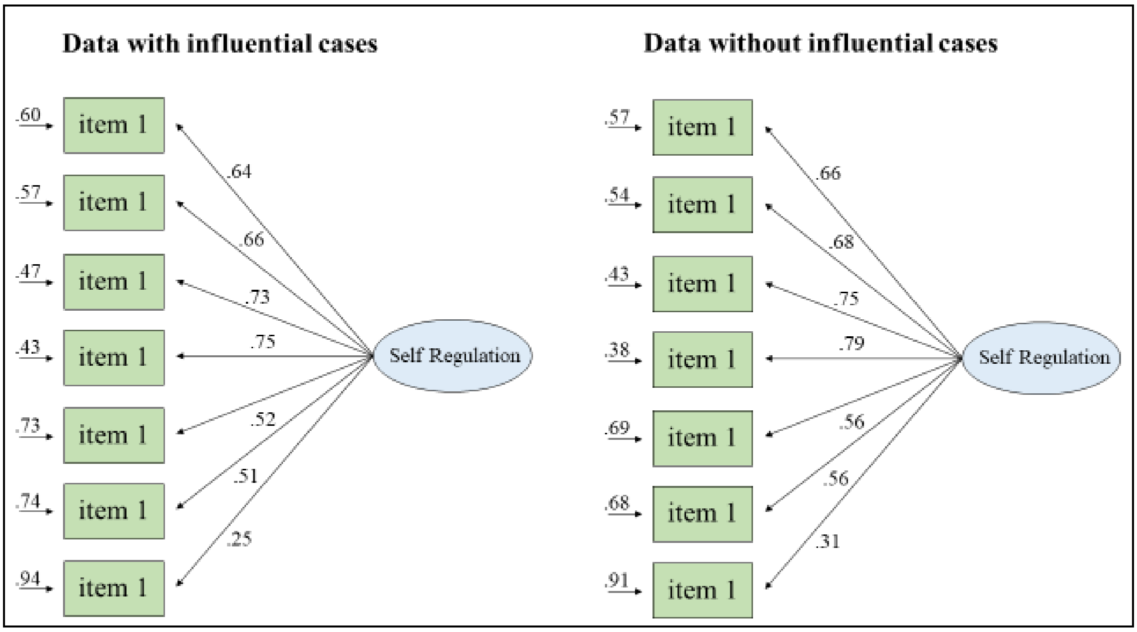
Not: Red line pertains to the results obtained from the dataset without influential cases.

After removing the influential cases, the CFA findings also confirmed the SRS's unidimensionality. The CFA findings showed that unidimensional structure cannot be verified by using the outlier cases-free dataset [χ^2=137.94, df=14, χ^2/df 9.86, CFI=.90, TLI=.85, NFI=.85, GFI=.94, AGFI=.87, RMSEA=.12, SRMR=.06]. The unidimensional structure was largely supported when the study was repeated after removing the most influential fifty cases from the dataset [χ^2=58.03, df=14, χ^2/df=4.15, CFI=.96, TLI=.95, NFI=.95, GFI=.97, AGFI=.94, RMSEA=.07, SRMR=.04]. This finding demonstrates that existence of the influential cases may have an effect on the SRS's dimensionality, not only for the CFA but also for the EFA.

In addition, Figure 3 shows the estimated factor loadings based on CFA and the number of corresponding errors. As can be seen in the graph, removing the influential cases from the dataset increased factor loadings and reduced the amount of error. For example, the factor loading increased from .25 to .31 for the seventh item, while the amount of error decreased from .94 to .91.

Figure 3

Effect of Influential Cases on Eigenvalues



Discussion and Conclusion

The aim of this analysis was to illustrate the impact of influential cases on the EFA and the CFA outcomes. Real polytomous data were collected for this purpose. There were 686 cases in the initial dataset. The following datasets were used for the EFA and the CFA: (a) complete dataset (b) the outlier case-free dataset that was consisted of 669 cases after the removal of 17 cases based on MDs, and (c) the influential case-free dataset, which composed of 619 cases after fifty of the most influential cases were removed. As suggested by Chalmers and Flora (2015), a forward search algorithm was used to evaluate influential cases. Correlation matrices for the SRS items, the factorability results, the number of dimensions extracted, the CFA-based fit indices, and the amount of factor loadings and errors were compared between these datasets after the analyses were completed.

As a result of removing the influential cases from the dataset, the inter-correlation values obtained for the SRS items and the internal consistency level of the scale both increased. This result implies that the existence of influential cases affects the homogeneity of the items. These findings are consistent with the existing literature (i.e., Liu & Zumbo, 2007).

According to the EFA results obtained for the entire dataset, the SRS has a two-dimensional structure, which contrasts with the SRS's original factorial structure. The number of factors extracted did not alter after removing the extreme cases based on MDs, but when the influential cases were removed further from the dataset, the factor analysis results for the SRS revealed a unidimensional structure, as originally proposed by Demiraslan et al. (2015). This finding showed that before starting the analysis, MDs and influential cases should be examined carefully because after removing the influential cases, the fit statistics of CFA analysis mostly supported unidimensional structure. Otherwise, researchers can come to incorrect conclusions about the scale's factor structure because this study concretely showed that existence of influential cases in dataset adversely affected the factorability results, communality values, factor loadings and, percentage of variance explained by the first factor. These findings were in line with the existing literature (i.e., Liu et al., 2012). At the same time, it was found that the factor loadings had increased and that the amount of error for items had decreased. These results are consistent with Bollen and Arminger's findings (1991).

Implications

The impact of influential cases on regression analysis is well-documented. On the other hand, factor analytic approaches have only recently begun to be studied in terms of the effect of influential cases on them as a variant of regression-based modeling. All previous studies (Koran & Jaffari, 2020; Liu et al., 2012; Mavridis & Moustaki, 2008; Yuan & Zhong, 2008) used simulated datasets, compared different approaches, tested their efficiency across different regulated conditions and used highly technical terminology. Furthermore, except for Mavridis and Moustaki (2008), none of these studies used Forward Search. In this way, this research is unique. It uses a real dataset and shows how false conclusions can be drawn when influential cases are ignored using a factor analytic approach.

The findings obtained in this study should be interpreted with caution for some reasons. Firstly, in this study, the most effective 50 individuals were removed from the datasets without closer investigation on them. However, in a real study, the final decision to delete these cases will have to be made after a qualitative examination of these cases: investigation of particular response vectors and questionnaire forms (Bollen, 1987; Kleinbaum et al., 1988). None of the influential case detection methods, including the forward search algorithm, are adequate to explain a case's deletion from the dataset. This fact was not taken into account in this study. Second, in this study, the most influential fifty cases were all eliminated at the same time. The number of cases that were deleted was decided arbitrarily because there was no cut-off threshold value for deciding how many cases should be removed. It should be noted that this figure cannot be used as a guideline for future research.

Since this is a demonstration report, the results can only be considered in light of the current dataset. Analyses conducted with a different dataset could not yield similar results. The factor structure of the dataset was not monitored or manipulated in this research. Future research should examine whether the current study's findings are generalizable to other scales using datasets with a different number of variables and different psychometric characteristics.

Conflicts of Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Bio

Akif Avcu is research assistant in Educational Sciences Department, Marmara University. His research interests include, test validation, item response theory, network psychometrics and behavioral addiction. He is actively supervising the academic studies conducted by Green Crescent Foundation. In addition, he teaches different courses like Introductory Statistics, Advanced Statistics, Measurement & Evaluation and Research Methodology.

References

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*, 1329-1339. https://doi.org/10.1080/01621459.1994.10476872

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. Springer-Verlag.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Wiley.

Bendre, S. M., & Kale, B. K. (1987). Masking effect on tests for outliers in normal samples. *Biometrika*, *74*(4), 891-896. https://doi.org/10.1093/biomet/74.4.891

Bollen, K. A. (1987). Outliers and improper solutions: a confirmatory factor analysis example. *Sociological Methods & Research*, *15*(4), 375-384. https://doi.org/10.1177/0049124187015004002

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In Mardsen, P. V. (Eds.), *Sociological methodology* (Vol. 21, pp. 235-262). Blackwell Publishing.

Chalmers, R. P., & Flora, D. B. (2015). Faoutlier: An R package for detecting influential cases in exploratory and confirmatory factor analysis. *Applied Psychological Measurement*, *39*(7), 573-374. https://doi.org/10.1177/0146621615597894

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman & Hall.

Dan, E., & Ijeoma, O. A. (2013). Statistical analysis/methods of detecting outliers in a univariate data in a regression analysis model. *International journal of education and research*, *1*(5), 1-24.

Demiraslan, Y. C., Haşlaman, T., Filiz, K., Mumcu, F., & Gökçearslan, Ş. (2015). Özdüzenlemenin dikkat kontrolü boyutu: bir ölçek uyarlama çalışması-Control dimension of self-regulation: a scale adaptation study. *Başkent University Journal of Education*, *2*, 229-238.

Diehl, M., Semegon, A. B., & Schwarzer, R. (2006). Assessing attention control in goal pursuit: A component of dispositional self-regulation. *Journal of Personality Assessment*, *86*(3), 306-317. https://doi.org/10.1207/s15327752jpa8603_06

Flora, D., LaBrish, C., & Chalmers, P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, *3*, 1-21. https://doi.org/doi:10.3389/fpsyg.2012.00055

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Sage Publications, Inc.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American statistical association, 88*(424), 1264-1272. https://doi.org/10.1080/01621459.1993.10476407

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Variable reduction and factor analysis. Applied regression analysis and other multivariable methods*. PWS Kent Publishing Co.

Koran, J., & Jaffari, F. (2020). Deletion statistic accuracy in confirmatory factor models. *Methodological Innovations, 13*(2), 1-10. https://doi.org/10.1177/2059799120918349

Lee, S. Y., & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika*, *61*, 93-108. https://doi.org/10.1007/BF02296960

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering*, *28*(9), 1635-1647. https://doi.org/10.1016/j.compchemeng.2004.01.009

Liu, Y., & Zumbo, B. D. (2007). The impact of outliers on Cronbach’s coefficient alpha estimate of reliability: Visual analogue scales. *Educational and Psychological Measurement*, *67*, 620-634. https://doi.org/10.1177/0013164406296976

Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A Demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. *Educational and Psychological Measurement*, *72*(2), 181-199. https://doi.org/10.1177/0013164411410878

Mavridis, D., & Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research*, *43*(3), 453-475, https://doi.org/DOI: 10.1080/00273170802285909

Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research*, *46*(2), 202-228. https://doi.org/10.1080/00273171.2011.561068

Poon, W.-Y., & Wong, Y.-K. (2004). A forward search procedure for identifying influential observations in the estimation of a covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 357-374, https://doi.org/DOI: 10.1207/s15328007sem1103_4

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D squared and Comrey Dk. *Multivariate behavioral research, 23*(2), 189-202. https://doi.org/10.1207/s15327906mbr2302_4

R Core Team. (2020). R*: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2015). psych: Procedures for personality and psychological research [Computer software manual]. http://cran.r-project.org/web/packages/psych/ (R package version 1.9.12)

Riani, M., Cerioli, A., Atkinson, A. C., Perrotta, D., & Torti, F. (2008). Fitting mixtures of regression lines with the forward search. *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security, 19*, 271.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Schwarzer, R., Diehl, M., & Schmitz, G. S. (1999). Self-Regulation scale. http://userpage.fu-berlin.de/~health/selfreg_e.htm

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed). Needham Heights, Allyn and Bacon.

Tanaka, Y., Watadani, S., & Ho Moon, S. (1991). Influence in covariance structure analysis: With an application to confirmatory factor analysis. *Communications in Statistics-Theory and Methods, 20*(12), 3805-3821. https://doi.org/10.1080/03610929108830742

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: using robust procedures to minimize their effect. *Sociological Methodology*, *38*(1), 329-368. https://doi.org/10.1111/j.1467-9531.2008.00198.x