(Cheung, S. F., & Pesigan, I. J. A. (2023). FINDOUT: Using either SPSS commands or graphical user interface to identify influential cases in structural equation modeling in AMOS. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2022.2148089. This is the version accepted, not the final published version.

OSF project page: https://osf.io/dbm5c/)

FINDOUT: Using Either SPSS Commands or Graphical User Interface to Identify Influential Cases in Structural Equation Modeling in AMOS

Shu Fai Cheung[1] and Ivan Jacob Agaloos Pesigan[1]

University of Macau

[1] Department of Psychology, University of Macau, Macao SAR, China

Author Note

This study was supported by the Multi-Year Research Grant (MYRG2019-00072FSS) from University of Macau, Macao SAR, China. Correspondence concerning this article should be addressed to Shu Fai Cheung, Department of Psychology, Faculty of Social Sciences, University of Macau, Avenida da Universidade, Taipa, Macao SAR, China, or by email (sfcheung@um.edu.mo).

Abstract

The results in a structural equation modeling (SEM) analysis can be influenced by just a few observations, called *influential cases*. Tools have been developed for users of R to identify them. However, similar tools are not available for AMOS, which is also a popular SEM software package. We introduce the FINDOUT toolset, a group of SPSS extension commands, and an AMOS plugin, to identify influential cases and examine how these cases influence the results. The SPSS commands can be used either as syntax commands or as custom dialogs from pull-down menus, and the AMOS plugin can be run from AMOS pull-down menu. We believe these tools can help researchers to examine the robustness of their findings to influential cases.

Keywords: influential cases, outliers, structural equation modeling, AMOS, sensitivity analysis, SPSS

FINDOUT: Using Either SPSS Commands or Graphical User Interface to Identify Influential Cases in Structural Equation Modeling in AMOS

Structural equation modeling (SEM) is now a popular technique in behavioral research to model the multivariate relations among variables. However, there is one aspect of modeling that is well-developed in multiple regression but did not receive sufficient attention in SEM: Identifying influential cases, one form of sensitivity of analysis that we focus on in this manuscript (other forms exist, such as parameter influence, Lee & MacCallum, 2015). Methods have been proposed and published for a long time (e.g., Pek & MacCallum, 2011, in *Multivariate Behavioral Research*). However, these methods were rarely adopted, probably due to the lack of tools. There are tools in R (e.g., *faoutlier* by Chalmers & Flora, 2015; *influence.SEM* by Pastore & Altoe, 2018). However, to our knowledge, there is no tool for identifying influential cases in AMOS (Arbuckle, 2021), which is also popular among applied researchers. A crude search using *AMOS* and *structural equation* identified about 18,600 entries published in 2021 or later in Google Scholar (April 7, 2022). To facilitate researchers who use AMOS to identify influential cases in SEM, we developed FINDOUT, which is composed of three SPSS extension commands and an AMOS plugin, to implement some of the methods. We first present the basic techniques for identifying influential cases in SEM, and then we illustrate how researchers can use the commands and plugin to implement these techniques in AMOS.

A Brief Overview

We define an influential case as a case that, if removed from an SEM analysis, will result in a substantial difference in one or more aspects of the results (Pek & MacCallum, 2011). What is considered substantial need to be judged with respect to the influence of all cases using tools like index plots (illustrated later) and conclusion to be drawn (e.g., when an effect is nonzero or trivial). As shown by other authors, an influential case is not necessarily

an outlier, and an outlier may not be influential (e.g., Flora, LaBrish, Chalmers, 2012). Moreover, outliers can be "good cases" that will worsen model fit if removed, while influential cases may not be extreme enough to be identified as outliers (Pek & MacCallum, 2011). Therefore, outlier screening cannot replace sensitivity analysis, and doing outlier screening and not sensitivity analysis may give ungrounded confidence in the robustness of results to influential cases.

There are no simple closed-form formulas in SEM to compute the changes in results after removing a case. Therefore, deletion statistics are often used to assess case influence. Readers can refer to the review by Flora et al. (2012) for the methods proposed. FINDOUT focuses on the generalized Cook's distance (GCD, Cook, 1977) and deletion statistics. GCD of Case *i* is a measure of the overall difference in parameter estimates if this case is removed (Pek & MacCallum, 2011):

$$\left(\hat{\theta} - \hat{\theta}_{(-i)}\right)' V\left(\hat{\theta}_{(-i)}\right)^{-1} \left(\hat{\theta} - \hat{\theta}_{(-i)}\right)$$

where $\hat{\theta}$ is the vector of parameter estimates in the full sample, $\hat{\theta}_{(-i)}$ is the vector of parameter estimates with Case *i* deleted, and $V\left(\hat{\theta}_{(-i)}\right)^{-1}$ is the inverse of the variance-covariance matrix of $\hat{\theta}_{(-i)}$. The larger the GCD, the larger the overall differences in estimates with and without this case. GCD can also be computed for a subset of parameters, such as all path coefficients. A deletion statistic is the raw difference on a measure between the full sample and a sample with a case removed: $a - a_{(-i)}$, where *a* can be a parameter estimate or a goodness of fit measure such as model χ^2, or derived parameters such as the standardized factor loading.

Installing the FINDOUT Toolset

To install the SPSS extension commands and the AMOS plugin in FINDOUT, first download these files: *FINDOUT_GENERATE_DATA.spe*, *FINDOUT_SAVE_RESULTS.spe*, *FINDOUT_MERGE_RESULTS.spe*, and *FINDOUT_RUN_LEAVE_ONE_OUT.vb* (available

from the OSF page for this manuscript: https://osf.io/dbm5c). In SPSS, select *Install Local Extension Bundle* from the *Extensions* menu, and then select each of the files and install it (video demonstration: https://osf.io/wg2da). Custom dialogs will also be installed. The AMOS plugin, `FINDOUT Run Leave-One-Out`, can be installed by copying the file *FINDOUT_RUN_LEAVE_ONE_OUT.vb* to the plugin folder (video demonstration: https://osf.io/4bdgs)

A Demonstration on Using FINDOUT

We developed the SPSS extension commands because we believe users with access to AMOS usually also have access to SPSS and the skills to use it. All files used in the demonstration can be downloaded from the OSF page of this manuscript (video demonstration at https://osf.io/wxnz5/). Researchers who prefer syntax commands can type the commands directly (sample syntax file at https://osf.io/mw53f). Researchers who prefer using the graphical user interface can run the commands directly from the custom dialogs (skipping the "Paste" steps in the video). We recommend users to use the dialogs to generate the syntax commands, such that the analysis can be repeated or reproduced if necessary.

Sample Dataset and the Model Being Fitted

The classic dataset from Holzinger and Swineford (1939, cited in Rosseel, 2012) is used, with nine variables hypothesized to load on three factors: x1, x2, and x3 on the visual factor, x4, x5, and x6 on the textual factor, and x7, x8, and x9 on the speed factor. For illustration, we adapted the procedure by Pek and MacCallum (2011) to generate a subset of 100 cases from the original dataset, with five cases highest on Mahalanobis distance (MD) and 95 cases not high on MD. A 3-factor model is fitted in AMOS using maximum likelihood. The covariances of estimates were requested for computing GCD.

Step 1: Generate *n* Leave-One-Out Samples By `FINDOUT GENERATE DATA`

First, run `FINDOUT GENERATE DATA` to generate *n* SPSS datasets from the dataset

used in the AMOS model, each with one case removed, *n* being the number of cases in the SPSS dataset. Users can use the custom dialog (*Analyze* > *FINDOUT extension commands* > *Generate N Leave-One-Out Data Files*) or type it as below:

```
FINDOUT GENERATE DATA
  OUTFOLDER = "E:\demo\one_out"
  USER_CASE_ID = case_id
 /OPTIONS OVERWRITE_DIAG_CASE_ID_DATA_FILE = YES.
```

`OUTFOLDER` is the folder to store the generated datasets; `USER_CASE_ID` is the case identification variable in the original dataset. The commands will use an internally generated case ID, and an SPSS dataset will be generated to match the internally generated case ID to the user case ID. The leave-one-out files are named `diag_case_id_x.sav`, `x` being the internal ID. The file matching the user case ID to the auto-generated ID is under *Data file matching diag_case_id* in the output window. Users can use this file to find out which case an internal ID refers to.

Step 2: Run AMOS *n* Times Using `FINDOUT Run Leave-One-Out`

In this step, users run the `FINDOUT Run Leave-One-Out` AMOS plugin to fit a model *n* times, once on each leave-one-out dataset from Step 1. As shown in the video, the plugin will first ask two questions: the folder to find the leave-one-out datasets, and the folder to save the AMOS output files (`E:\demo\amos_out` in this example). If some output files already exist in the output folder, the plugin will ask whether to overwrite them. If not, then only leave-one-out datasets without corresponding output files will be processed. This process is fast because it is conducted inside AMOS. The demonstration took about 1.5 minute for the 100 files when ran on a Windows computer (Intel Core i7-7700, 16G RAM). It will take longer when the sample size is large, but the duration should still be acceptable.[1]

Step 3: Save AMOS Results to SPSS Data Files By `FINDOUT SAVE RESULTS`

[1] We did a stress test on a 500-case sample with missing data on the same computer. It took about less than 5 minutes to fit the model 500 times.

In this step, run `FINDOUT SAVE RESULTS` to extract statistics from the AMOS output files and save them to SPSS data files Users can use the custom dialog box (*Analyze* > *FINDOUT extension commands* > *Save SEM Results to Data Files*) or type the command as below:

```
FINDOUT SAVE RESULTS
  AMOS_OUT_FOLDER = "E:\demo\amos_out"
  FULL_SAMPLE_AMOSOUTPUT =
"E:\demo\HolzingerSwineford1939.AmosOutput"
  RESULTS_OUT_FOLDER = "E:\demo\amos_to_spss"
 /OPTIONS OVERWRITE_RESULTS = NO
          ALL_DELETION_STATISTICS = YES.
```

`AMOS_OUT_FOLDER` is the folder with the *n* AMOS output files, `RESULTS_OUT_FOLDER` is the folder to store the results, `FULL_SAMPLE_AMOSOUTPUT` is the AMOS output file on the full sample, which is required if deletion statistics are to be computed. Setting `ALL_DELETION_STATISTICS` to `YES` requests GCD and raw changes of fit measures and all point estimates. If `OVERWRITE_RESULTS` is `YES`, existing result files will be overwritten.

Step 4: Merge the AMOS Results to One File By `FINDOUT MERGE RESULTS`

Last, users use `FINDOUT MERGE RESULTS` to merge all *n* leave-one-out SPSS datasets to one dataset by the custom dialog (*Analyze* > *FINDOUT extension commands* > *Merge Extracted SEM Results*) or the following code:

```
FINDOUT MERGE RESULTS
 RESULTS_OUT_FOLDER = "E:\demo\amos_to_spss".
```

`RESULTS_OUT_FOLDER` is the folder in which the SPSS datasets with extracted AMOS results are stored. After running this command, the merged dataset will become the active dataset, but this command will not save it. We recommend the researchers to examine the merged dataset first and then save it manually or use the `SAVE` syntax command of SPSS to save the file automatically.

Identify Influential Cases

The merged dataset can then be used to identify influential cases on all common SEM results supported by `FINDOUT SAVE RESULTS`. The deletion statistics for model χ^2 will be computed if requested: DIFF_GOF_CMIN ($\Delta\chi^2$ in Pek & MacCallum, 2011). If the value for a case is positive, it means that the full sample value is greater than the value with this case removed (improved model fit if this case is removed). Three GCDs will be computed if requested: GCD (for all free parameters), GCD_B (for all regression coefficients, including loadings), and GCD_VarCov (for all variances and covariances). Deletion statistics for estimates available in the full sample output will be computed if requested, all prefixed with DIFF_EST_. Therefore, if the standardized solution, R^2, and indirect effects are requested in AMOS, raw changes on these estimates for removing each case will also be computed.

To our knowledge, there is no consensus on cutoff values for deletion statistics in SEM. Therefore, as proposed by Aguinis, Gottfredson, and Joo (2014), we recommend using index plots to examine the influence of each case relative to other cases in the same sample, as in Pek and MacCallum (2011). For example, as shown in Figure 1, the plot of DIFF_GOF_CMIN shows that, even for cases that are not outliers based on Mahalanobis distance (Case 6 to Case 100, available in the AMOS output), some of them can still be moderately influential on model fit (Pek & MacCallum, 2011, labelled a case that resulted in a change of 1.33 in χ^2 as "moderately bad", p. 214). The plot of GCD shows that Case 2 has the largest overall influence on parameter estimates. The plot of DIFF_EST_beta.2@Estimate (standardized loading of x2) shows that the standardized loading of x2 is strongly influenced by Case 2, with an estimate about .08 higher in the full sample than that with Case 2 removed. Last, the plot of DIFF_EST_Corr.3@Estimate (the correlation between visual and textual) shows that Case 5 also has nonnegligible influence on the estimate of this correlation (nearly .10 lower in the full sample).

Insert Figure 1 About Here

Users can use the index plots of other deletion statistics to have a more comprehensive sensitivity analysis. Although there is no consensus on any cutoff values, the index plot can be used to identify cases that are relatively different from others in influence. For example, even if the five cases were positioned randomly in the dataset, they can still be identified easily in the plot of GCD. For cases identified as influential, we recommend checking whether they are due to data error first (Cohen, Cohen, West, Aiken, 2003, pp. 411-412). If not due to data error, researchers can then adopt the approaches recommended by Aguinis et al. (2014), reporting results with and without the influential case. For example, Case 2 can be removed and see whether the other cases are still influential on the results.

Conclusion

FINDOUT has several advantages. It has no known restriction on the model and the estimation method and can handle missing data by FIML (full information maximum likelihood). It can compute changes for all reported estimates, including standardized estimates, R^2, and indirect effects. If the sample size is extremely large, users can use `FINDOUT RUN Leave-One-Out` to fit the models only on selected cases by keeping only files of target cases in the leave-one-out folder, allowing researchers to do targeted sensitivity analysis. Although FINDOUT cannot do a forward search (Mavridis & Moustaki, 2008) to identify clusters of influential cases, we hope it can still help AMOS users to examine potential influential cases to assess the degree of robustness of their results to case influence. We also encourage researchers to report sensitivity analysis as supplementary material to demonstrate to what extent the results are robust to the exclusion or inclusion of a few cases. For example, index plots of GCD and $\Delta\chi^2$ can be provided for the models presented in a manuscript to demonstrate the robustness of the results in model fit and overall parameter estimates.

References

Arbuckle, J. L. (2021). *IBM® SPSS® Amos™ 28 User's Guide*.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Cadigan, N. G. (1995). Local influence in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(1), 13–30. https://doi.org/10.1080/10705519509539992

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), 573–574. https://doi.org/10.1177/0146621615597894

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd Ed.). Lawrence Erlbaum Associates Publishers.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

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

Lee, T., & MacCallum, R. C. (2015). Parameter influence in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 22(*1), 102–114. https://doi.org/10.1080/10705511.2014.935255

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/10.1080/00273170802285909

Pastore, M., & Altoe, G. (2018). Package *influence.SEM* (Version 2.2).

Pek, J., & MacCallum, R. (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

Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). http://www.jstatsoft.org/v48/i02/paper

Figure 1. Deletion Statistics and Generalized Cook's Distance