A Screening Algorithm in Simulation of Mediation Models

Anwar Fitrianto and Habshah Midi

Abstract—Several literatures mentioned that outlier identification is a part of the data screening process which should be done routinely before statistical analyses. Specifically in mediation analysis, the clean data generating process in a simulation is needed before continuing to further analysis so that allows accurate result of an analysis can be obtained. We proposed a screening algorithm in mediation analysis, named Clean-Assured Data Generating Procedures (CADGP). The screening method uses ModDRGP, a newly outlier detection method which provided a procedure for obtaining clean datasets. The generated dataset through CADGP will be free of high leverage points. The CADGP is needed especially in simple mediation analysis which is usually used in social sciences.

Keywords—Mahalanobis distance, Mediation analysis, outlier, outlier detection, simulation

I. INTRODUCTION

Prior to statistical analysis, it is very crucial to have careful inspection of the research data to avoid from making erroneous findings and conclusions. The computer that uses to analyse the data does not know whether the data is accurate or not. However, the researcher will not be able to discern the extent to which the results are valid simply by examining the computer output. The researchers will then proceed to interpret the results and draw conclusions accordingly. They are not aware of the erroneous conclusions due to the analysis of inaccurate data. Thus, it is imperative to employ a data screening methodology to provide researcher with a means to detect potential data problems by identifying data entry errors, missing values, possible outliers, non-normal distributions, and other data features.

Due to that reason, before analyzing data, the entire dataset should be reviewed. Many statistical programs deal with this situation. SAS has procedures by which we calculate several outlier statistics. They can produce a report and output data set which requires further manipulation to use. This data manipulation opens the door of opportunity for program errors which might be difficult to detect. The more statisticians can rely on known code, the more comfortable they can be with the outcome.

II. SOME REVIEWS ON DATA SCREENING

[1] suggested an appropriate sequence for screening the proposed data. The order of the screening is important as decisions at the earlier steps which will influence decisions to be taken at later steps. For example, if the data are both non-normal and have outliers, the decision to delete values or transform the data is confronted. If transformation is undertaken first, there is likely to be fewer outliers, yet if the outliers are deleted or modified first, there are likely of be fewer variables with non-normality. Transformation of the variable is usually preferred as it typically reduces the number of outliers, and is more likely to produce normality, linearity, and homoscedasticity. Screening the input data will help assessing the appropriateness of the use of a particular data set. Screening will aid in the isolation of data strange and allow the data to be adjusted in advance of further analysis. The checklist isolates key decision points which need to be assessed to prevent poor data induced analysis problems. Consideration and resolution of problems encountered in the screening of a data set is necessary to ensure a robust statistical assessment.

A. Outlier Detection as a Part of Data Screening

At various times, many researchers are faced with the need to screen for outliers. Statistical tests for outliers are one part of the data validation process wherein data are screened an examined in various ways before being placed in a data bank and used for estimating population parameters or making decisions. [2] discussed data screening and validation procedures for air quality data. It includes several procedures: 1) routine check during the data processing, 2) tests for the internal consistency of a data set, 3) comparing the current data with historical data to check for consistency over time, and 4) tests to check for consistency with parallel data sets, that is, data sets obtained presumably from the same population.

Several literatures mentioned that outlier identification is a part of the data screening process which should be done routinely before statistical analyses ([3]; [1]). [4] stated that, “We recommend that data be routinely inspected for outliers, because outliers can provide useful information about the data.” [5] also suggested that screening for outliers should be the first step to screen a dataset. He also suggested that if we...
find outliers, the analysis should be done twice: with and without outlier. By doing this way, the effects of outliers can be determined and the results used in deciding how to handle the outlier. At recent paper by [6], it is mentioned that screening data for outliers and checking distributional assumptions is an important, but often underused, part of a careful statistical investigation.

B. The Proposed Screening Method

We propose a diagnostic technique to identify multiple high-leverage points in mediation analysis. To the best of my knowledge, no robust methods and diagnostic techniques have been done on mediation analysis. The proposed approach of a diagnostic technique will be based on simple mediation analysis for its simplicity. We will describe the proposed approach in a procedure that can be easily understood. It is mainly based on the DRGP that has been proposed by [7]. The proposed method employs the $Q_n$ estimator instead of MAD. The $Q_n$ scale estimate is motivated by the Hodges-Lehmann estimator of location [8]:

$$\hat{\mu} = \text{median} \left( \frac{x_i + x_j}{2} \right); \quad 1 < i \leq j < n.$$ 

The Hodges-Lehmann estimator might be viewed as a "smooth version" of the median since it possesses a smooth influence function [8].

[9] verified that both MAD and $Q_n$ have the same breakdown point that is 50\%. Nonetheless, the efficiency of the $Q_n$ is higher (86\%) than the MAD (37\%). This work inspire us to incorporate the $Q_n$ instead of MAD. By using the $Q_n$ rather than the MAD in the equation of cut-off point in the DRGP, we hope a more powerful scheme that can detect more outliers in mediation analysis which involves several regression equations.

We use the following procedure to identify potential outliers in mediation analysis as follows:

Step 1 : For each $i$ point on $(x_i, m_i)$ pair, calculate the $\text{RMD}_i$.

Step 2 : An $i^{th}$ point with $\text{RMD}_i$ exceeds cut-off point of $\text{Median}(\text{RMD}) + 3 \text{MAD}(\text{RMD})$ is suspected a high-leverage point and included in the deleted set $D$. The rest of the points are put into the $R$ set.

Step 3 : Based on the above $D$ and $R$ sets, compute the $p_i^*$ using the formula written in the equation 18, that is

$$p_i^* = \begin{cases} 
\frac{w_i^{(-D)}}{1 - w_i^{(-D)}}; & \text{for } i \in R \\
 w_i^{(-D)}; & \text{for } i \in D 
\end{cases}$$

Step 4 : Any deleted point having $p_i^*$ exceeds cut-off point of $\text{Median}(p_i^*) + c Q_n(p_i^*)$ is finalized and declared as the high-leverage points, where $c = 3$.

For convenience, we refer the above new method of identifying potential outliers in mediation analysis as ModDRGP1 where the MAD is incorporated in the second step of the ModDRGP1 algorithm. In this paper we also propose another DRGP, which has the following procedures for identifying potential outliers in mediation analysis. We called the second proposed method as ModDRGP2. The ModDRGP2 modifies the criteria of determining cut-off point in the step 2. Instead of using $\text{Median}(\text{RMD}) + 3 \text{MAD}(\text{RMD})$, we employed the $\text{Median}(\text{RMD}) + 3 Q_n(\text{RMD})$.

Then the proposed ModDRGP2 procedures for identifying potential outliers in mediation analysis are:

Step 1 : For each $i$ point on $(x_i, m_i)$ pair, calculate the $\text{RMD}_i$.

Step 2 : An $i^{th}$ point with $\text{RMD}_i$ exceeds cut-off point of $\text{Median}(\text{RMD}) + 3 Q_n(\text{RMD})$ is suspected a high-leverage point and included in the deleted set $D$. The rest of the points are put into the $R$ set.

Step 3 : Based on the above $D$ and $R$ set, compute the $p_i^*$ using the formula written in the equation 18, that is

$$p_i^* = \begin{cases} 
\frac{w_i^{(-D)}}{1 - w_i^{(-D)}}; & \text{for } i \in R \\
 w_i^{(-D)}; & \text{for } i \in D 
\end{cases}$$

Step 4 : Any deleted point having $p_i^*$ exceeds cut-off point of $\text{Median}(p_i^*) + c Q_n(p_i^*)$ is finalized and declared as the high-leverage points, where $c = 3$.

III. THE PROPOSED CLEAN ASSURED DATA GENERATING PROCEDURES (CADGP) ALGORITHM

The basic mediation model is a causal sequence in which the independent variable ($X$) causes the mediator ($M$) which in turn causes the dependent variable ($Y$), therefore explaining how $X$ had its effect on $Y$ ([10]; [11]). It implies a causal hypothesis whereby an independent variable causes a mediator which causes a dependent variable ([12]; [13]). Mediation is typically assessed by using a sequence of independent regression equations to measure the various paths in a complex...
A simple mediation analysis involves simple and multiple linear regressions. Whenever a simple linear regression analysis, such as Eq. 1 and Eq. 2 of simple mediation analysis, is performed, addressing issues of outliers and high influence points is not a very frightening task. The reliance on having a regression estimator that is able to separate out and detect potentially damaging observations is not at all crucial in this scenario. How one deal with these special observations is, of course, important, but their mere existence can be easily determined. When the regression analysis involves multiple regressor variables, such as Eq. 3 in simple mediation analysis, the existence of outliers and high leverage points is generally less evident (or even not evident at all) by a casual viewing of the data. The level of sophistication of an analysis method needs to be raised considerably in order to avoid an inferior regression analysis.

We name the procedure of generating data in simple mediation model in which contain outlier screening as Clean-Assured Data Generation Procedures (CADGP). It involves data collection activities with an initial screening of available data sources. This will be an iterative process where details of data that are available are built up. This generating and screening process may be slow and require longer time as the cost of the usefulness of the clean-resulted data set. The CADGP algorithm which we propose to generate set of data necessary in a Simple Mediation Analysis has the following steps:

1. Generate a dataset of size \( n \) which have a certain design structure \((X, M, Y)\). The design structure is developed based on the three linear regression models required in a simple mediation analysis.

2. Define a condition of outlier detection method for the screening stage. In this study we use ModDRGP2 as the outlier detection method.

3. Conduct outlier detection on each of the generated data using an outlier detection method on step 2. Assign ‘0’ if an observation is an outlier (or high leverage point) and ‘1’ if it is not. From this step, we can make a variable, let say \( D \), is a binary variable containing 0 and 1 values only.

4. Count the number of ‘1’ value in variable \( D \), if the number of ‘1’ equals to \( n \), then stop, otherwise repeat steps 1-3.

The structure which we implement is a simulation as presented in Table 1. Following the standard approach of testing for mediating variables, it is needed to establish a relationship between \( X \) and \( M \), and also a relationship between \( Y, X, \) and \( M \).

The choice of structuring the simulation is in such way that the description of the simple mediation analysis with some contamination can be easily understood. The simulated data are such that \( X \) affects \( Y \) and \( X \) affects \( M \). Meanwhile, \( M \) also affects \( Y \).

IV. SIMULATION DESIGNS

Based on the structure of the simulation we have illustrated, the values of the inlaying or "clean" observations' regressor were selected at random from a uniform distribution with range of \( U[0;20] \). The error is normally distributed with mean of 0 and variance of 1, \( N(0;1) \) in the set of equations. In the study, we performed simulation study based on various sample size \( n \). In particular, sample sizes considered were 20, 50, and 100.

![Fig. 2 scatter plot of number of simulation needed to get clean data after 10000 simulation runs, \( n=20 \)](image)

A simulation study is carried out to explore the behaviour of the proposed CADGP procedure. We start to reveal the result by displaying Table 2 below. It displays and highlights the importance of the screening step using CADGP in a model simulation. As has been described previously, we use sample size of \( n=50 \) and \( n=100 \). Due to time constraint, the simulation was not done at \( n \) more than 100. Simulation with the CADGP we did were using personal computer with specification of Intel Dual Core E2200 2.2GHz processor and 2 Gigabytes RAM, takes about 5 minutes for \( n=20 \), 3 hours for
n = 50, and 9 hours for n = 100. All of them were done at 10000 simulation runs. We believe that the time needed in this of simulation will increase geometrically instead of arithmetically if the sample size is increased.

From the Table 2, we can see that at n = 20, we can get 4.167 and 1 for mean and median of number of high leverage points (HLP), respectively with the range of simulation runs is between 1 and 25. Mean and median for the number of data generating process in each simulation which is needed before obtaining first clean dataset in each simulation runs is represented in the third row of the table. NoDGP mean of n = 20 is equal to 3.226, means that, on average, it needs about 3 data generating process in each simulation runs before getting a clean dataset. If the size of sample is increased, that is n = 50 for this study, NoDGP is greatly increased become 13.322. The result is even extremely high at n = 100, which is 133.752. For NoDGP median, the result also displays obvious trend of increasing its value when the sample increases.

One more valuable information in this study is regarding how many or percentage of simulation runs which can generate a clean dataset directly without necessary any number of loops in a simulation run. It is reasonable that smaller sample size we want to generate, higher possibility to get direct clean dataset in a simulation run. This study has proven the statement. In the Table 2, for sample size of 20, after 10000 simulation runs, about 31.14% of the clean-generated dataset were obtained by direct simulation. In other words, percentage of obtaining directly a clean dataset is about 31.14%. A clean-generated dataset which are obtained by direct simulation means that the clean dataset is obtained after one simulation only.

The result even worst if we want to generate larger sample size. In our study, we obtained only 7.75% and 0.63% of the simulation runs which can get clean dataset directly. Figure 2, 3, 4 and 5 display a clear description about that result. In the Figure 2, the scatter plot demonstrates that at the number of simulation runs equals to 1, the frequency is equal to 2886 (symbolized with bigger dot in the figure). The frequency of simulation runs gradually decreases when the simulation runs increases.

VI. CONCLUSION

Generating data is a common step needed in statistics to provide and justify evidence of mathematical models. Unfortunately, many people do not realize that generated data need to follow necessary assumptions of a model they need to prove. A simulation study is required in model validation. But as before doing further steps of an analysis of the simulated data, researchers must make sure that the initiated data is clean so that an examination of a method can be valid.

In social sciences in which simple mediation analysis are commonly being used, researcher used to use large sample size in their study. The CADGP procedure using ModDRGP1 and ModDRGP2 has great performance in large sample. CADGP incorporating ModDRGP2 that we proposed provides a procedure for obtaining clean datasets, especially in mediation analysis and multiple linear regression models. The generated dataset through CADGP will be free of high leverage points. The CADGP is needed especially in simple mediation analysis.

### Table II

<table>
<thead>
<tr>
<th>Measurement</th>
<th>n = 20</th>
<th>n = 50</th>
<th>n = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLP mean</td>
<td>4.167</td>
<td>10.167</td>
<td>4.692</td>
</tr>
<tr>
<td>HLP median</td>
<td>1</td>
<td>2.5</td>
<td>4</td>
</tr>
<tr>
<td>NoDGP mean</td>
<td>3.226</td>
<td>13.322</td>
<td>133.752</td>
</tr>
<tr>
<td>NoDGP median</td>
<td>2</td>
<td>9</td>
<td>94</td>
</tr>
<tr>
<td>% First run</td>
<td>31.14%</td>
<td>7.75%</td>
<td>0.69%</td>
</tr>
</tbody>
</table>

Note:
- NoDGP is the number of data generating process each simulation run, while the % FR is percentage number of simulation runs to get clean data in the first run.
which is usually used in social sciences. In social sciences, researchers are commonly needs larger sample size.

REFERENCES