World Applied Sciences Journal 15 (7): 1046-1053, 2011
ISSN 1818-4952

Procedures of Generating a True Clean Data in Simple Mediation Analysis

[1]*Anwar Fitrianto and* [2]*Habshah Midi*

[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia Mathematics Building, 43400 Universiti Putra Malaysia, Serdang Selangor, Malaysia
[2]Laboratory of Computational Statistics and Operation Research, Institute for Mathematical Research Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Malaysia

Abstract: Simulation study is very important in model validation. It is invaluable and versatile tool especially in statistical problems and modeling where analytical technique is inadequate. In fitting to a model, problems will raise when there exists one or more high-leverage points in the data set. Due to the fact that the presence high-leverage points are commonly occurred in models fitting, we propose a new algorithm in mediation analysis which guarantees clean data set without any high-leverage points. The new proposed algorithm employs the newly proposed Modified Diagnostic-Robust Generalized Potentials. By incorporating ModDRGP in the proposed algorithm has rectified the problem of having high leverage points in the generated clean data set, especially for mediation models. We found that in 10000 simulation runs, only about 31.14% of the clean-generated dataset were obtained by direct simulation. The results also reveal that as the sample size increases, the percentage of obtaining direct clean dataset decreases.

Key words: Mediation analysis · Potentials · Monte Carlo

INTRODUCTION

Prior to statistical analysis, it is very crucial to have careful inspection of the research data to avoid making erroneous findings and conclusions. The computer that is used to analyse the data does not know whether the data is accurate or not. However, researcher will not be able to discern the extent to which the results are valid simply by examining the computer output. They will then just 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 alert researcher with potential data problems by identifying data entry errors, missing values, possible outliers, non-normal distributions and other data features.

Due to this reason, before analyzing a data, the entire data set should be checked. Many statistical softwares such as SAS deal with this situation. 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 code errors which might be difficult to detect.

Some Reviews on Data Screening: Tabachnick and Fidell [1] suggested an appropriate sequence for screening a data. The order of the screening is important as decisions at the earlier steps will influence decisions to be taken at later steps. For example, if the data is both non-normal and has outliers, then we are confronted with the decision to delete values or transform the data. If transformation is undertaken first, there is likely to be fewer outliers, yet if the outliers are deleted or modified first, there are likely 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 of the input data will ensure the appropriate of the use of a particular data set. Screening will aid in the isolation of strange data and allow the data to be adjusted in advance for further analysis. The checklist isolates key decision points which need to be assessed to prevent analysis problem induced by a poor data. Consideration and resolution of problems encountered in the screening of a data set is necessary to ensure a robust statistical assessment.

Tabachnick and Fidell [1] also suggested two factors for initial consideration of the data screening. First, data

Corresponding Author: Anwar Fitrianto, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia Mathematics Building, 43400 Universiti Putra Malaysia, Serdang Selangor, Malaysia.

screening techniques relate directly to the choice of statistical method of choice and the assumptions of the method. They used an example in regression analysis. They mentioned that analyzing data using logistic regression (a log-linear) would not require the same screening process as multiple regression (a linear method) due to the different assumptions that are required for each procedure. A second consideration for data screening procedures is data grouping. Group data, such as attitudinal differences based on ethnic classification, may demand different screening procedures as compared to data that does not compare groups, such as relationship between two attitudinal measures in validity study. In other words, Tabachnick and Fidell [1] advised that screening procedures are dependent on the analysis used. The procedures have been developed for screening data for normality prior to hypothesis testing and provide strategies for correcting it.

Outlier Detection as Part of Data Screening: In data analysis, researchers are often dealt with the need to screen for outliers. Statistical tests for outliers are one part of the data validation process where data are screened and examined in various ways before being placed in a data bank and used for estimating population parameters or making decisions. Nelson *et al.* [2] discussed data screening and validation procedures for air quality data.

Several literatures mentioned that outlier identification is part of the data screening process which should be done routinely before any statistical analysis [3, 4, 1]. Iglewicz and Hoaglin [5] stated that, "We recommend that data be routinely inspected for outliers, because outliers can provide useful information about the data." Fink [6] 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 can be used in deciding how to handle the outlier. In recent paper by Banerjee and Boris [7], it is mentioned that screening data for outliers and checking distributional assumptions is an important, but often underused, part of a careful statistical investigation.

Outliers can exist in both univariate and multivariate situations, among dichotomous and continuous variables and among independent variables as well as dependent variables [1]. The simplest and the most researched case is the identification of univariate outliers, where the distribution of a single variable is examined [7, 8]. Extreme data values are obvious outlier candidates. When the distribution is symmetric, we suspect that candidate outliers are the extremes of the left or right tail. In a skewed distribution, the suspect outliers are likely to be the extremes of the longer tail. Multivariate outlier detection is more difficult, because the multivariate distribution has no tails [9, 10]. According to Tabachnick and Fidell, [1], univariate outliers are cases with extreme scores on a single variable. In order to determine whether univariate outliers were present, a standardized residual was calculated for each case. Meanwhile, with regard to multivariate outliers, Mahalanobis distance refers to the degree to which a case differs from the centroid created as a function of means for the combination of the variables across multidimensional space. Many methods for detecting outliers are discussed by Beckman and Cook [3], Hawkins [11] and Barnett and Lewis [9]. Burr [12] provided many references on control chart techniques. Kinnison [13] discussed extreme value statistics, which are closely connected with the ideas of outlier detection.

With data sets consisting of a small number of variables, detection of univariate outliers can be relatively simple. This can be accomplished by visually inspecting the data, either by examining a frequency distribution or by obtaining a histogram and looking for unusual values. Meanwhile, multivariate outliers consist of unusual combinations of scores on two or more variables. Individual z-scores may not indicate that the case is a univariate outlier (i.e. for each variable, the value is within the expected range), but the combination of variables clearly separates the particular case from the rest of the distribution. Multivariate outliers are more subtle and, therefore, more difficult to identify, especially by using any of the previously mentioned techniques. Fortunately, a statistical procedure (known as *Mahalanobis distance)* exists which can be used to identify outliers of any type, [14]. Mahalanobis distance was first introduced by Mahalanobis [15]. It is defined as the distance of a case from the centroid of the remaining cases where the centroid is the point created by the means of all the variables [1].

The Proposed Method: 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, [16]. It implies a causal hypothesis whereby an independent variable causes a mediator which causes a dependent variable [17, 18]. Mediation is typically assessed by using a sequence of independent regression equations to measure the various paths in a complex

model, as initially suggested by Judd and Kenny, [19]. Detailed regression equation needed to establish a mediation analysis is as follows:

$$Y = i_1 + cX + \varepsilon_1 \quad (1)$$
$$M = i_2 + aX + \varepsilon_2 \quad (2)$$
$$Y = i_3 + c'X + bM + \varepsilon_3 \quad (3)$$

where Y is the dependent variable, X is the independent variable, M is the mediating variable or mediator. Coefficient c represents the relation between the independent variable to the dependent variable in the first equation, c' is the parameter relating the independent variable to the dependent variable adjusted for the effects of mediator, a is coefficient of the relationship between X and M and b is the parameter relating the mediator to the dependent variable adjusted for the effects of the independents variable to the mediating variable. Note that the ε_1, ε_2 and ε_3 represent error variability and i_1, i_2 and i_3 are the intercepts. The intercepts are not involved in the estimation of mediated effects and could be left out of the equations, [16]. Note that, both c and c' are parameters relating the independent variable to the dependent variable, but c' is a partial effect, adjusted the effects of mediator.

A simple mediation analysis involves simple and multiple linear regressions. Whenever a simple linear regression analysis, such as Eq 1 and Eq. 2 are performed, addressing issues of outliers and high influence points is not a very frightening task. A simple scatter plot of the response variable versus the regressor variable would illustrate graphically the nature of the data. The reliance on having a regression estimator that is able to sift out and detect potentially damaging observations is not at all crucial in this scenario. How one deals with these special observations is, of course, important, but their mere existence can easily be determined. When the regression analysis involves multiple independent variables, such as Eq. 3 in simple mediation model, 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 the method of analysis needs to be addressed carefully in order to avoid poor regression analysis.

We use a diagnostic technique of identifying multiple high-leverage points which was newly proposed by Fitrianto and Habshah [20], Modified DRGP, as the main part of the screening procedures. Fitrianto and Habshah's Modified DRGP is mainly based on the DRGP which was developed by Midi *et al.* [21]. It employs the Q_n estimator instead of MAD. The Q_n scale estimate is motivated by the Hodges-Lehmann estimate of location, [21]:

$$\hat{\mu} = \text{median}\left(\frac{x_i + x_j}{2}\right), 1 < i \le j < n$$

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

Croux and Rousseeuw [23] 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 has inspired us to incorporate the Q_n instead of MAD in the formulation of the proposed method. We expect to obtain more powerful scheme in the detection of outliers in mediation analysis which involves several regression equations, by using the Q_n instead of the MAD in the formulation of cut-off point in the DRGP. The procedure of Fitrianto and Habshah, [20] is summarized as follows:

Step 1 : For each i point on (x_i , m_i) pair, calculate the RMD_i

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

Step 3 : Based on the above D and R sets, compute the p_{ii}^* using the following formula,

$$p_{ii}^* = \begin{cases} \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}}; & \text{for } i \in R \\ w_{ii}^{(-D)}; & \text{for } i \in D \end{cases}$$

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

Fitrianto and Habshah [20] refers the above method as ModDRGP1 where the MAD is incorporated in the second step of the ModDRGP1 algorithm. Another identification procedure which Fitrianto and Habshah [20] called ModDRGP2, will also be used in the screening procedures of this paper.

The Proposed Algorithm of Generating Clean Data (GCD) in Simple Mediation Analysis: This section presents a study from simulated data that violate the assumption underlying standard approach of mediation analysis. Simulated data are presented because the underlying structure of this data is known with certainty. This, in turn, allows accurate assessment of the difference in results between the standard and the suggested approach. Specifically in mediation analysis, Shaver [24] carried out a study when data in management research would violate assumptions underlying tests of mediating variables. He also mentioned that violating underlying assumptions in mediation analysis could have severe consequences. The resulting coefficient estimates could be biased and inconsistent which may lead to incorrect conclusion.

We call the procedure of generating data in simple mediation model which contain outlier screening as GCD. It involves data collection activities with an initial screening of available data sources and detail iterative processes were outlined. This kind of generating and screening processes may be time-consuming. The proposed GCD algorithm to generate a set of data in a Simple Mediation Analysis is based on ModDRGP1 or ModDRGP2 Fitrianto and Habshah [20] and has the following steps:

- 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.
- Define a condition of outlier detection method for the screening stage. In this study we use ModDRGP2 as the outlier detection method and ModDRGP1 may be used as an alternative.
- Conduct outlier detection procedure in each of the generated data using an outlier detection method in step 2. Assign '0' if an observation is an outlier (or high leverage point) and '1' if it is not. From this step, we define a variable, let say *D*, as a binary variable containing 0 and 1 values only.
- Count the number of '1' value in variable *D* and if the number of '1' equals to n, then stop, otherwise repeat steps 1-3.

The data structure for the simulation study is 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 easily be understood. The simulated data are such that *X* affects *Y* and *X* affects *M*. Meanwhile, *M* also affects *Y*.

Table 1: Data structure for the simulations study

Variable	Definition
X	Random variable drawn from a uniform distribution *U* [0,20]
M	1.75*X* + random error drawn from a normal distribution with mean=0 and standard deviation=1
Y	-0.1*X* + 0.85 *M* + random error drawn from a normal distribution with mean=0 and standard deviation=1

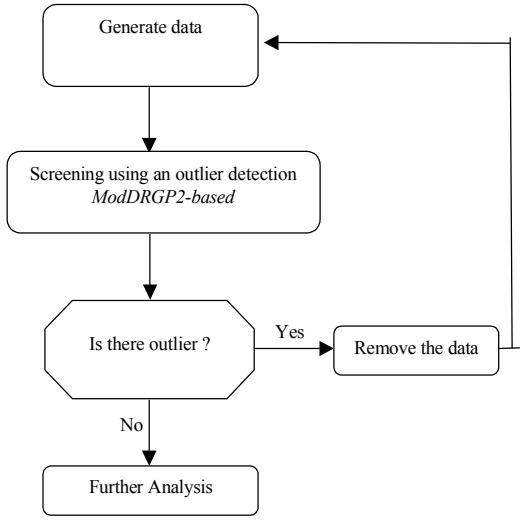

Fig. 1: Flow chart of the GCD algorithm in simple mediation analysis

Simulation Designs: In statistics, simulation study is very important in model validation. Simulation is invaluable and versatile tool especially in statistical problems and modeling where analytical technique is inadequate. It is a numerical technique for conducting experiment on a computer. Computer simulation is used in engineering and the sciences to rerun physical experiments with selected changes in parameters or operating conditions. In statistics, simulation experiments are almost often used to study properties of statistical methods. It is very useful in presenting key statistical concepts. For example, the Central Limit Theorem, probability density functions, asymptotic consistency and even confidence limits all become far more intelligible when a researcher is shown how these results actually operate in practice. Many real statistics processes can be simulated on computer with the aid of random number. Here, random generating process takes important rule in simulation as displayed in Figure 1.

With the simulation structure as described above, the values of the inlaying or "clean" observations' regressor were selected at random from a uniform distribution with range of [0, 20]. The error is normally distributed with mean of 0 and variance of 1, $N(0,1)$ in both Equations 2 and 3. In the study, we performed simulation study based on various sample sizes, n. In particular, the sample sizes considered were 20, 50 and 100.

A SAS Macro is then constructed for the GCD algorithm. The SAS macro makes use of Proc IML and Call MVE facility in SAS. The MVE subroutine computes the minimum volume ellipsoid estimator. These robust locations and covariance matrices can be used to detect multivariate outliers and leverage points. For this purpose, the MVE subroutine provides a table of robust distances. Beside using MVE subroutine, the important part of the SAS macro is the screening code.

- The pseudo code of the screening in SAS language is constructed as follows:

```
do j = 1 to n;
xi=matXM[j,];
r =loc(RD<CutRMD);
r2 =loc(RD<CutRMD2);
xibrack= matXM[r,];
xibrack2= matXM[r2,];
wiiD=xi*inv((xibrack`*xibrack))*xi`;
wiiD2=xi*inv((xibrack2`*xibrack2))*xi`;
piistar[j,1]=choose(RD[j,1]>CutRMD,wiiD,wiiD/(1-wiiD));
piistar2[j,1]=choose(RD[j,1]>CutRMD2,wiiD2,wiiD2/(1-
wiiD2));
end;

do j=1 to n;
Medpiis=Median(piistar);
MADss=(Median(abs(piistar-Medpiis)))/0.6745;
MADss2=Mad(piistar,"qn");
CutDRGP=Medpiis+(3*MADss);
CutModDRGP2=Medpiis+(3*MADss2);
DRGPInd[j,1]=choose(piistar[j,]<CutDRGP, 1, 0);
ModDRGP2Ind[j,1]=choose(piistar2[j,]<CutModDRGP2,
1, 0);
sDRGP=DRGPInd[+,];
sModDRGP2=ModDRGP2Ind[+,];
end;
```

RESULTS

The simulation study is carried out to explore the behaviour of the proposed procedure. Table 2 exhibits the simulation results of the GCD using ModDRGP2.

Table 2: Simulation result of the GCD using ModDRGP2 with 10,000 simulation runs

	Sample size		
Measurement	$n = 20$	$n = 50$	$n = 100$
HLP mean	4.167	10.167	4.692
HLP median	1.0	2.5	4.0
NoDGP mean	3.226	13.322	133.752
% First run	31.14%	7.75%	0.69 %

Note:

NoDGP: Number of data generating process in each simulation run

% FR: Percentage number of simulation runs to get clean data in the first run

It displays and highlights the importance of the screening step using the proposed procedures in a model simulation. As has been described in the previous section, we use sample of size $n = 20$, $n = 50$ and $n = 100$. Due to time constraint, the simulation was not done for n more than 100. Simulation has been done on a 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 $n = 100$. In each simulation runs there were 10,000 replications. We believe that the time needed in this simulation study will increase geometrically rather than arithmetically if the sample size are increased.

It can be observed from Table 2 that on the average, about 4 high leverage points (HLP) were detected for $n = 20$. The NoDGP equals to 3.226 for $n = 20$ in the Table 2, implies that, on the average, it needs about 3 data generating processes in each simulation run to get the first clean dataset. The results of Table 2 also signify that as the sample size increases, the mean of NoDGP also increases.

In this study, we also interested to know valuable information regarding the number of clean data set obtained directly from a loop of simulation run. A direct simulation implies that a clean dataset is successfully obtained in the first generating process in each simulation run. To simplify the presentation of results, Table 2 also presents the percentage of clean dataset obtained by direct simulation in 10000 simulation runs. It can be seen from Table 2 that about 31.14% of the clean-generated dataset were obtained by direct simulation. It is worth mentioning that a smaller sample size will result in a higher possibility of getting direct clean dataset in a simulation run.

The result are not encouraging for larger sample size. For instances, only 7.75% and 0.69% of the clean-generated dataset obtained from direct simulation,

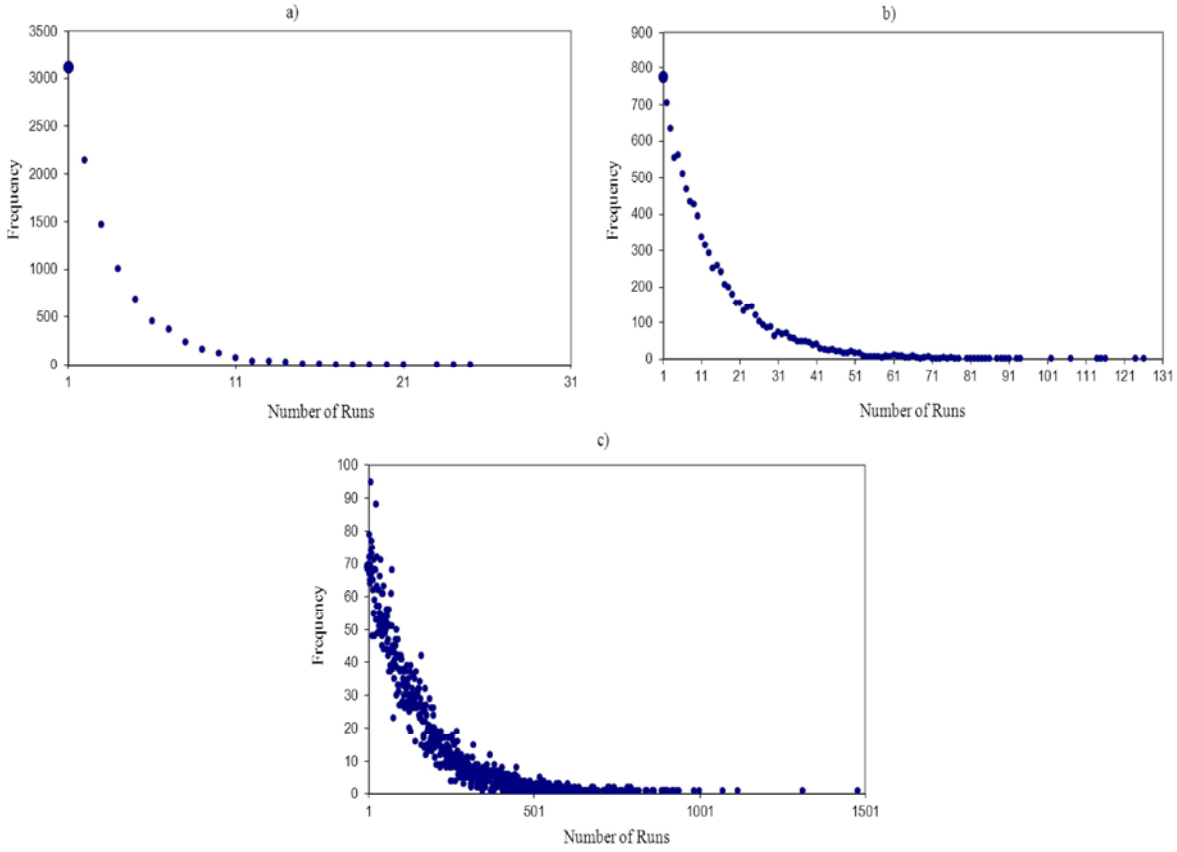

Fig. 2: Scatter plot of frequency distribution of the number of simulation required to obtain clean data after 10000 simulation runs, panel a) n = 20, panel b) n = 50, panel c) n = 100

for n equals 50 and 100, respectively. Figure 2 of panel a, panel b and panel c display a clear description of the result. In the panel a, the scatter plot demonstrates that 2886 clean datasets are successfully generated in the first simulation run (symbolized with bigger dot in the figure). The frequency of simulation runs gradually decreases as the sample size increases.

Another graphical display is presented to show the importance of using the ModDRGP2 procedures in generating data, especially in Simple Mediation Analysis. Figure 3 presents the percentage of obtaining direct clean data set versus sample size. The graph suggests the necessity of screening procedures in generating data of simple mediation analysis. This is evident by looking at the plot shows the probability of obtaining a clean dataset without screening step can be considered low (30%) for n = 20. The results of the simulation study reveals that as the sample size increases, the percentage of obtaining direct clean dataset decreases.

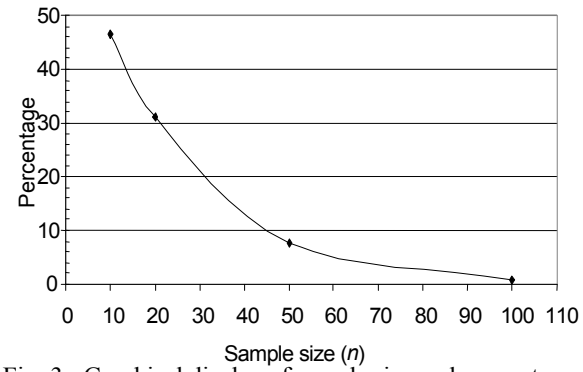

Fig. 3: Graphical display of sample size and percentage of obtaining direct clean data in generating data using GCD of ModDRGP2

We also would like to compare the GCP based on the Modified Diagnostic Robust Generalized Potentials 1 (ModDRGP1) and ModDRGP2. The performances of the procedures are evaluated based on the average number of data generating process to obtain a correct clean dataset and percentage of the number of simulation runs for

Table 3: Comparisons of GCD between ModDRGP1 and ModDRGP2 with 10,000 simulation runs

		Diagnostic Methods	
Sample size	Statistics	ModDRGP1	ModDRGP2
n=10	NoDGP mean	2.182	2.129
	% First run	46.06	46.50
n=20	NoDGP mean	3.475	3.226
	% First run	28.93	31.14
n=50	NoDGP mean	14.129	13.322
	% First run	7.18	7.75
n=100	NoDGP mean	139.263	133.752
	% First run	0.71	0.69

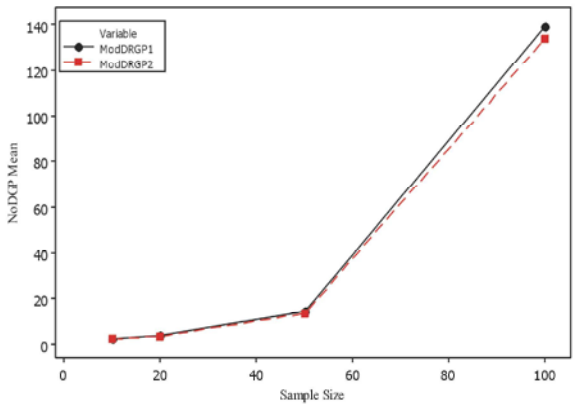

Fig. 4: Comparisons of NoDGP mean of GCD between ModDRGP1 and ModDRGP2 after 10000 simulation runs

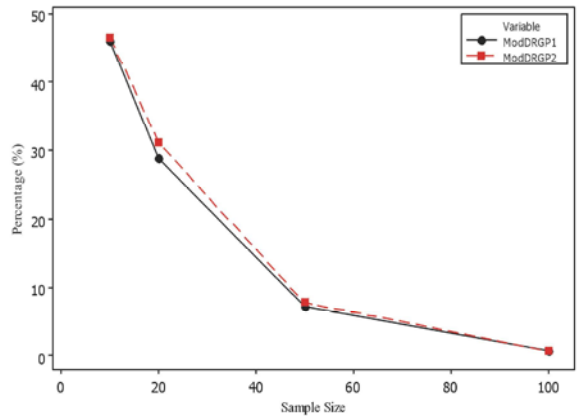

Fig. 5: Comparisons of first run percentage of GCD between ModDRGP1 and ModDRGP2 after 10000 simulation runs

obtaining a clean dataset in the first run. A good procedure is the one which has smaller value of NoDGP and higher percentage value of simulation runs to get the correct clean dataset. The results are presented in the Table 3, Figure 4 and Figure 5.

It can be seen from the Table 3 that in almost all sample sizes considered, the NoDGP and the percentage of first run of the GCP ModDRGP2-based is smaller and higher, respectively than the GCP ModDRGP1-based procedure. Hence, the GCP ModDRGP2-based is slightly better than the GCP ModDRGP1-based.

Figure 4 and Figure 5 clearly show the performance of the GCP ModDRGP2-based compared to the GCP ModDRGP2-based. It is important to point out that the procedures performances decrease as the sample size increases.

CONCLUSIONS

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. However, prior to further steps of an analysis of the simulated data, researchers must make sure that the initial data is clean so that an examination of a method can be valid.

We have witnessed the discouraging outcome of generating clean data especially for large *n*. The claimed clean generated data set is not always true, particularly for larger sample size which usually be encountered in mediation research. In this respect, we proposed a procedure for obtaining clean datasets, especially in mediation analysis and multiple linear regression models. The merit of our proposed procedure is that the generated dataset is free from the presence of high leverage points, even for larger sample size which is commonly used in social science research.

REFERENCES

1. Tabachnick, B. and L. Fidell. 2001. Using Multivariate Statistics, *4nd ed*. New York: Boston and Bacon.
2. Nelson, A.C., D.W. Armentrout and T.R. Johnson, 1980. Validation of Air Monitoring Data, EPA-600/4-80-030, U.S. Environmental Protection Agency, Research Triangle Park, N.C.
3. Beckman. R.J. and R.D. Cook, 1983. Outlier….s, Technometrics, 25: 119-149.
4. Ahmad, S., M. Habshah and M.R. Norazan, 2011. Diagnostics for Residual Outliers Using Deviance Component in Binary Logistic Regression, World Appl. Sci. J., 14(8): 1125-1130.

5. Iglewicz, B. and D.C. Hoaglin, 1993. How to Detect and Handle Outliers. Milwaukee,Wisconsin: Quality Press, American Society for Quality.
6. Fink, A., 1995. How to analyze survey data, 3rd ed. Thousand Oaks, CA: Sage Publications.
7. Banerjee, S. and Boris Iglewicz, 2007. A simple univariate outlier identification procedure designed for large samples, Commun. Stat. Simulation Comput., 36: 249-263.
8. Barnett, V., 1978. The study of outliers: Purpose and Mode, Applied Statistics, 27(3): 242-250.
9. Barnett, V. and T. Lewis. 1994. Outliers in Statistical Data, 3rd ed. New York: Wiley.
10. Gnanadesikan, R. and J.R. Kettenring. 1972. Robust estimates, residuals and outlier detection with multiresponse data, Biometrics, 28: 81-124.
11. Hawkins, D.M., 1980. Identification of Outliers. New York: Chapman Hall.
12. Burr, I.W., 1976. Statistical Quality Control Methods, New York: Marcel Dekker, Inc.
13. Kinnison, R.R., 1985. Applied extreme value statistics. New York: Macmillan.
14. Stevens, R., 1996. Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Erlbaum.
15. Mahalanobis, P.C., 1936. On the generalized distance in statistics, Proceedings National Institute of Science, India, 12: 49-55.
16. MacKinnon, D.P., 2008. Introduction to Statistical Mediation Analysis. New York: Taylor and Francis.
17. Holland, P.W., 1988. Causal inference, path analysis and recursive structural equation models. Sociological Methodol., 18: 449-484.
18. Sobel, M.E., 1990. Effect analysis and causation in linear structural equation models, Psychometrika, 55: 495-515.
19. Judd, C.M. and D.A. Kenny, 1981. Process analysis: Estimating direction in evaluation research, Evaluation Res., 9: 602-618.
20. Fitrianto, A. and Habshah Midi, 2010. Diagnostic-Robust Generalized Potentials for Identifying High Leverage Points in Mediation Analysis. World Appl. Sci. J., 11(8): 979-987.
21. Midi, H., M.R. Norazan and A.H.M. Rahmatullah Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression, J. Applied Statistics. 36(5): 507-520.
22. Hodges, J.L. and E.L. Lehmann, 1963. Estimates of location based on rank tests, Ann. Math. Statist., 34: 598-611.
23. Croux, C. and P.J. Rousseeuw, 1993. Time-efficient algorithms for two highly robust estimators of scale. In Y. Dodge and J.C. Whittaker, (eds*), Computational Statistics*, 1: 411-428. Heidelberg. Physica-Verlag.
24. Shaver, J.M., 2005. Testing for Mediating Variables in Management Research: Concerns, Implications and Alternative Strategies, J. Manage., 31(3): 330-353.