THE USE OF OUTLIERS IN THE DETECTION OF SUSPICIOUS EXAMINATION MALPRACTICES IN THE SCORES OF STUDENTS IN A NIGERIAN UNIVERSITY

Consul, Juliana Iworikumo
(iwori2001@yahoo.com)

and

Ndiwari, Godgift
(gifttola@yahoo.com)
Department of Mathematics and Computer Science, Niger Delta University, Bayelsa State, Nigeria

ABSTRACT
It is very important that we detect outliers as it deviates from the majority of the data. In our study, we have used both graphical and analytical methods to detect outliers which may be mistakes or else accurate but unexpected suspicious examination or test scores. These observations may be due to malpractices. In cases where we might want to perform statistical analysis like simple linear regression models, outliers may lead to erroneous parameter estimate and inferences from the model. We recommend that suspicious data points should be further investigated and in the case of examination malpractices, appropriate measures should be taken. The masking and swamping effect of outliers appeared in this study.
Keywords: Outlier, influential, examination malpractice, residual, masking and swamping effect.

1.0 INTRODUCTION

An outlier in a set of data is defined to be an observation or a subset of observations which appear to be inconsistent with the remainder of the set of data. Outliers may not be genuine members of the main population (Barnett and Lewis (1994) and Cook (1977)). Outliers can also be defined as observations which are found to differ so much from other observations. These observations are often seen as contaminating the data. They either reduce or distort the information about the source of the data. We might want to seek means of interpreting, categorising or handling outliers in other to reduce the effect on statistical analysis if required. Outliers may frustrate attempts to draw inferences about the basic population. Barnett and Lewis (1994) showed that extreme observations and contaminants may or may not be outliers.

Outliers may arise from any of measurement, recording, execution or calculating error in the data. Barnett and Lewis (1994) gave an example of which a recording error may appear as an outlier in a set of data. Outliers can be handled either by being retained or rejected.

A very important part of statistical practice is considering the study of regression models. Suppose that we have a linear regression model

$y = \beta_0 + \beta_1 x + \in$.

We have that the expectation of $\in$, E ($\in$)= 0 and $Var(\in) = \sigma^2$ which is a constant. An observation that seems to deviate so much from the linear model than other observations is said to be an outlier. The outlier breaks the expected pattern of results for the regression model. One very interesting effect of an outlier is its influence on the estimation of the parameters in the regression model. Barnett and Lewis (1994) records that

outlier inflates the estimate of σ^2. They also showed that robust methods of inference can be employed to the data to minimise the influence of outliers.

Onyibe *et al.* (2015) mentioned that Education is a process of teaching and learning which is evaluated through examination at the end of the learning period. It serves as feedback for the examiners as to ascertain the level of knowledge acquisition of the student. This knowledge is meant to be retained by the student. Examination is part of evaluation in an educational system that is aimed at determining a students' level of intellectual competence or understanding after learning. Examination malpractice can therefore be defined as a deliberate wrong doing, misconduct or improper practice during or after examination as contrary to official examination rules in view to obtaining good results by fraudulent means (Onyibe *et al.* (2015)).

The causes of examination malpractice may be due to inadequate teaching or learning facilities, an intension of someone to help a friend, poor parental upbringing, ineffective preparation by students, ill-equipped library facilities or dubious admission policy. Some forms of examination malpractice include impersonation, giraffing (stretching out of neck to copy from a fellow student), inscription, use of phones, bribery et.c.

2.0 Data Collection

In this study, the sample is made up of Year 1 undergraduate students in the Faculty of Science in the Niger Delta University, Bayelsa State of Nigeria. The sample is made up of 177 students who registered for the session. This study seeks to detect strange and inconsistent result as compared to the remainder of a set of data collected from the semester or session result of some particular students. The research is also aimed at identifying offending observation (outliers) prior to any further processing or analysis of data.

In the Niger Delta University, there is a Senate Student Disciplinary Committee which is a Senate Committee that handles all matters related to student discipline in the University. The membership includes the Vice Chancellor, the Registrar, the Dean of Student Affairs, two members representing the Senate, the Deans of Faculties, two student representatives. The functions of the Senate Student Disciplinary Committee include investigating disciplinary cases involving students and imposing appropriate sanctions, reporting concluded matters to the Governing Council and Senate for information, making recommendations to Senate or any matter pertaining to the proper discipline of students and to consider any other disciplinary matter referred to from time to time.

3.0 DETECTION OUTLIERS IN SIMPLE LINEAR REGRESSION

It will be unrealistic to declare that an observation is an outlier based on just one variable but rather it is better to collectively consider all variables of the observation. In this study, we collectively use regression model. Regression is a tool which is used to establish a relationship between a dependent variable, y called the response or output variable and one or more independent variable(s) $x_1, x_2, x_3 \ldots\ldots x_p$ called predictors or explanatory variables. When $p = 1$ it is called simple linear regression but when $p > 1$ it is called multiple linear regressions. Our interest in this research will be on simple linear regression.

We consider a set of n observations such that $(y_1, x_1), (y_2, x_2), \ldots\ldots (y_n, x_n)$ are related by $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. We have that E $(\in) = 0$ and $Var(\in) = \sigma^2$ for $i = 1,2, \ldots\ldots n$. We will also note that β_0 and β_1 are unknown but estimable parameters. The estimable parameters of β_0 and β_1 $(\hat{\beta}_0, \hat{\beta}_1)$, can be obtained using the least squares method as follows:

$$\hat{\beta}_1 = \frac{n\ \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\ \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where the regression line passes through the means $(\bar{x}, \bar{y})$, that is $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. We can estimate ϵ_i by $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. It will be reasonable to examine the relative sizes of the ϵ_i for us to seek for outliers.

We might confront a problem while dealing with regression in the presence of outliers in the data. The detection of outliers and influential point's is an important step of the regression analysis. It is usual that we detect outlier either by using graphical or analytical methods. In graphical method like scatterplots, normal probability plot, boxplot, residual e.t.c, the presence of outlier is identified by the shape of the plot or graph. In this research, we use the scatterplot where our set of data $(x_1, y_1), (x_2, y_2), \ldots \ldots (x_n, y_n)$ are plotted. The scatterplot of the observed data points with points that are standing apart from the majority of the set of data indicate the presence of outliers. On the other hand, some statistical values of the data are computed in the analytical methods and these values are used to identify the presence of outliers. Some statistical values that can be used to identify the presence of outliers include studentized residual, standard residual, PRESS residual, HAT matrix, Cook's square distance e.t.c.

One possible approach to the detection of a discordant outlier is to test for discordancy by examining the studentized residuals

$$\epsilon_i = \frac{\hat{\epsilon}_i}{s_i} = \hat{\epsilon}_i / \{ S \sqrt{(1 - 1/n) - (x_i - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2} \}$$

where $S^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} / (n-2)$ is an unbiased estimate of σ^2. Thus, we have that ${S_i}^2$ is an unbiased estimate of $Var(\hat{\epsilon}_i)$. We then detect and test for discordancy of a single outlier through the maximum absolute studentized residual which is expressed in terms of statistics as $t = max \left| \frac{\hat{\epsilon}_i}{s_i} \right|$. If t is sufficiently large then the observation yielding $t = max \left| \frac{\hat{\epsilon}_i}{s_i} \right|$ is said to be a discordant outlier.

We will note that the simple linear regression can also be represented in a matrix form as follows:

$\underline{Y} = X\underline{\beta} + \underline{\in}$

where $\underline{X}$ is a full rank matrix known to be $n \times p$, $\underline{Y}$ is an $n \times 1$ vector, $\underline{\beta}$ is an unknown $p \times 1$ vector and $\underline{\in}$ is an $n \times 1$ error vector.

We will look for observations that are surprisingly far away from the main group. It is not so reasonable to sought for outliers merely only at extremes. An observation might disturb the general pattern to a degree which it is discomforting. This observation may lie outside the typical relationship between the dependent and independent variables as revealed by the remaining data. It might be obvious by visual inspection but a bit complicated and not straight forward in regression models with many independent variables. It is very possible that the outlier might influence or distort the estimates of the parameter and hence inflate the standard errors. Barnett and Lewis (1994) also records that outliers are often influential but not all influential observations need to be outliers. Therefore, there is the need to distinguish carefully between outliers which are influential and influential observations which may not be outlying.

Another statistical value of the data that can be used to identify outlier is the standardized residual. This is computed as

$$S_{di} = \frac{\in_i}{\sqrt{MSE}}$$

where MSE is the mean squared error. A large value of S_{di} (S_{di}>3) indicates that the i^{th} observation is an outlier following (Montgomery et al. 2003).

The PRESS residual can also be used to identify an outlier. This is computed as

$$P_i = \frac{\in_i}{(1 - h_{ii})}$$

where h_{ii} is the i^{th} diagonal element of the matrix $H = X(X'X)^{-1}X'$. Again, we follow Montgomery et al. (2003) that a large value of P_i (P_i>3) indicates that the i^{th} observation is an outlier.

The Cook's distance (Cook, 1977) can also be used to detect outlier. The cook's distance is a measure which is computed with respect to a given regression model. It computes the influence exerted by each data point on the predicted outcome. The Cook's distance is defined as

$$C_{Di} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{p\, s^2}$$

where $\hat{\beta}_{(i)}$ is the estimator of $\beta_{(i)}$ calculated without the i^{th} case, s^2 is the mean square error given as

$$s^2 = \frac{\in'\in}{n - p}$$

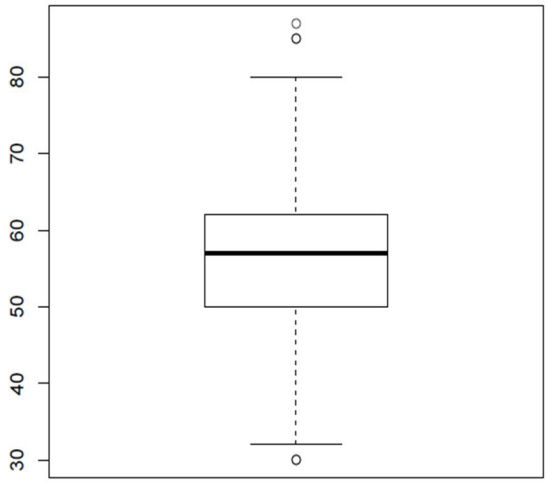
where $\in = Y - \hat{Y}$.

Stephen and Senthamarai (2017) detected the presence of outliers in the simple linear regression models for medical data set where he used the standardized scores for detecting outliers without the use of predicted values. They also compared some outlier detection procedures in multiple linear regression. These methods include Mahalanobis distance, Cook's square distance and DEFFIT distance. Cook (1977) investigated the performance of Mahalanobis distance, Cook's square distance and DEFFIT distance at different proportions of outliers for various sample sizes. Ka – Veng Yuen and He-Qing Mu (1985) discussed the detection of outliers using the outlier probability for robust linear regression by considering a set of regular points. Farid and Swallow used sequential testing of the maximum studentized residual, Marasinghe (1985) multistage procedure and the recursive method to detect outliers. Adikaram *et al*. (2014) introduced a non-parametric outlier detection method for a linear series based on sum of arithmetic progression.

4.0 APPLICATION TO DATA SET, RESULTS AND DISCUSSIONS

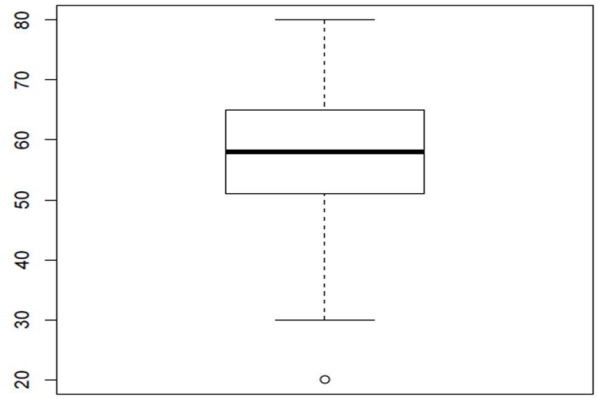
We may wish to identify outliers in the data set which may represent cases of examination malpractices by the respective students. We will use the Boxplot as a graphical approach to the detection of outliers. Boxplots will give an insight into whether outliers exist or not in our data set. We will note that all plots and results were done using R software (R Development Core Team (2008)). Figure 1 shows the Boxplot of the dependent variable (y) and Figure 2 shows the Boxplot of the independent variable (x).

Figure 1: Boxplot of examination scores



We can see few outliers in the boxplot. In Figure 1, we have that the outliers in the dependent variable are the extreme values (likely scores 85, 87 and 30).

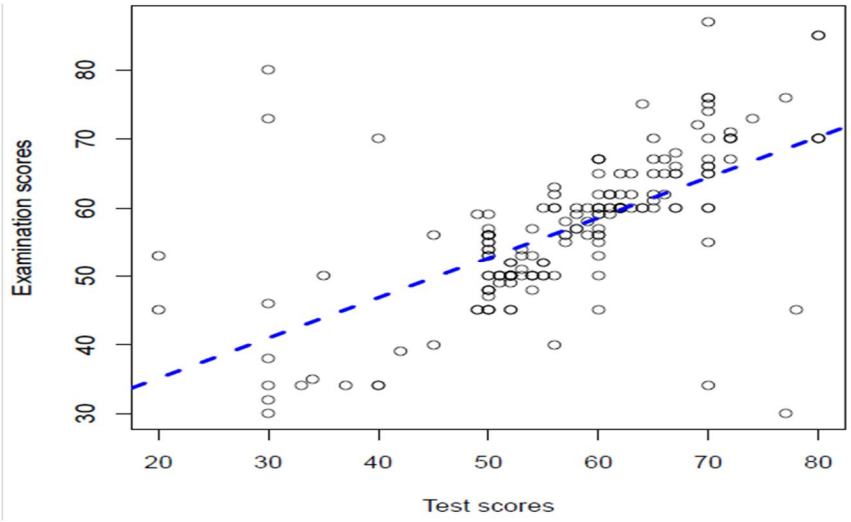
Figure 2: Boxplot of test scores



In Figure 2, we have that the outlier in the independent variable is also the extreme values (likely 20). Outliers could help us choose between the lowest or highest extreme values since outliers can lie to both sides of the data set. In each of the cases, the outliers in the dependent and independent variables are declared based on just one feature which leads to unrealistic inferences. It is better to collectively consider a row of observations. We will fit a simple linear regression model on the complete data set including outliers.

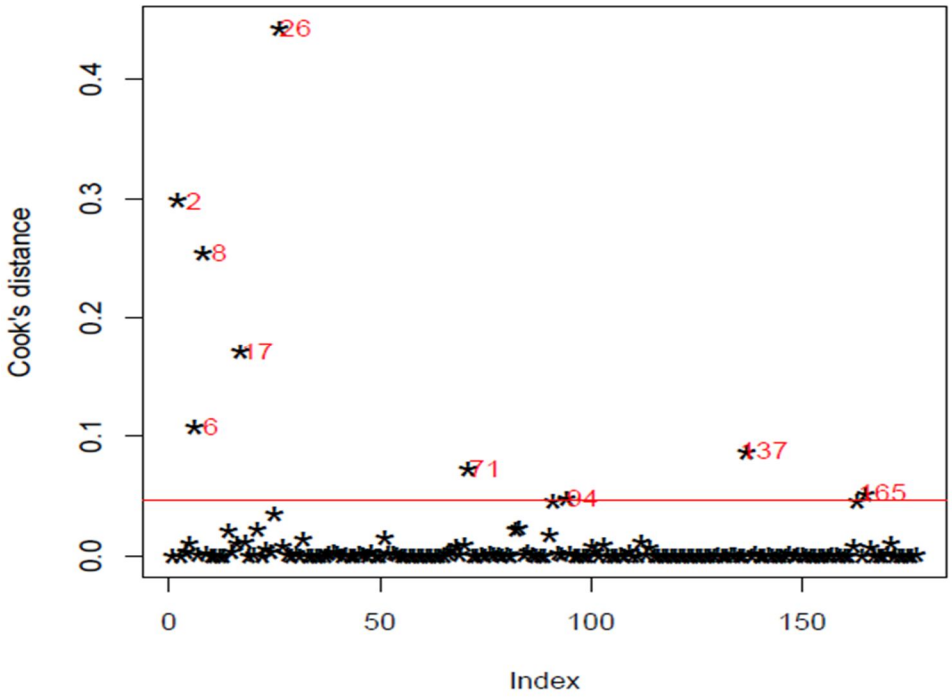
We use R software to fit a simple linear regression model. The regression parameters are given as $\hat{\beta}_0 = 23.41$ and $\hat{\beta}_1 = 0.59$. Hence, the regression line for the i^{th} observation is given as $\hat{y}_i = 23.41 + 0.59\, x_i$. The coefficient of determination R^2 can be used to judge the adequacy of the regression model to the observed sample values of y and x. It measures the amount of variation in the data explained or accounted for by the regression model. The value of $R^2 = 0.39$. This implies that 39% of the total variation is accounted for by the regression model and the remaining 61% is due to error. We will note that the higher the value of R^2, the better the fit. The fit in this case is very poor. Figure 3 gives a plot of the fit of the simple linear regression model.

Figure 3: Plot of the fit of the simple linear regression model of examination scores against test scores.



Now, we might want to decide that if a row or observation is an extreme value or not by collectively considering the variables. We choose to use the Cook's distance. The Cook's distance with respect to the regression is computed using the R software package. The observations that have a Cook's distance greater than 4 times the mean is classified as influential. Figure 4 shows the influential observation by Cook's distance.

Figure 4: A plot of influential observation by Cook's distance.



From Figure 4, the 2nd, 6th, 8th, 17th, 26th, 71st, 94th, 137th and 165th observations are tagged as influential. The observations are as follows:

Observation number	y	x
2	73	30
6	45	78
8	30	77
17	53	20
26	80	30
71	70	40
94	87	70
137	34	70
165	45	20

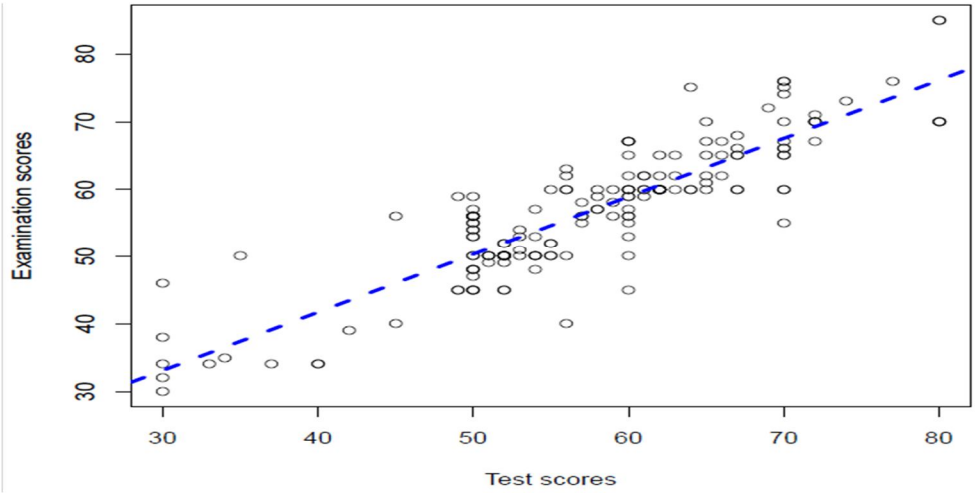
We examine each of the influential observations to be able to reason out why the observation turned out influential. It is likely that one of variables included in the model had extreme values. Observation 2, 26, 71 and 94 have very high examination scores. Observation 6, 8, 94 and 137 have very high test scores. Observation 6, 8, 137 and 165 have very low examination scores. Observation 2, 17, 30, 40 and 165 have very low test scores. It is very obvious that outliers get the extreme observation from the mean. Once outliers are identified, we have to decide what to do to these outliers.

The observations listed above are different from the usual observations that are observed in the data set. Outliers in the data set can be regarded as suspicious cases of examination and test malpractices. For instance, observation 2 (with examination score of 73 and test score of 30) might be associated with a case of examination malpractices with reasons that the student performed very poorly in the test and better in the examination. It might be assumed that the student must have copied from another student (giraffing) or inscription. This student should be called on to defend him or herself and probably be reassessed to check for accuracy. On the other hand, such student can also argue that since he knew that he had performed poorly in the test, he decided to study harder in as to make up with his examination. This same student can also defend himself or herself by saying that he did not understand the course as at the time of the test but studied to perform better in the examination. In both cases, the students should be reassessed. The 26^{th} and 94^{th} observations can also be treated in a similar way as the 2^{nd} observation.

In the case of the 8^{th} observation, the student had an examination score of 30 and test score of 77. This case should also be investigated and the student should be reassessed. This might be a case where the student sat with friends in the test and performed well due to giraffing but he was not able to copy from friends due to the security during the examination. The 6^{th} and 137^{th} observations can be treated in a similar way as the 8^{th} observation. The 17^{th} observation (with examination score of 53 and test score of 20) and 165^{th} observation (with examination score of 45 and test score of 20) are fair cases. This might not be real cases of outliers. These are due to the lowest extreme values of with test score of 20. In such case, the student must have struggled to improve a bit more in the examination. The student should not be reassessed. The 94^{th} observation (with examination score of 87 and test score of 70) might also not be a case of real outlier. It is an outlier due to the highest extreme value of examination score of 87. In this case, the student should not be reassessed because he or she has performed very well in both examination and test.

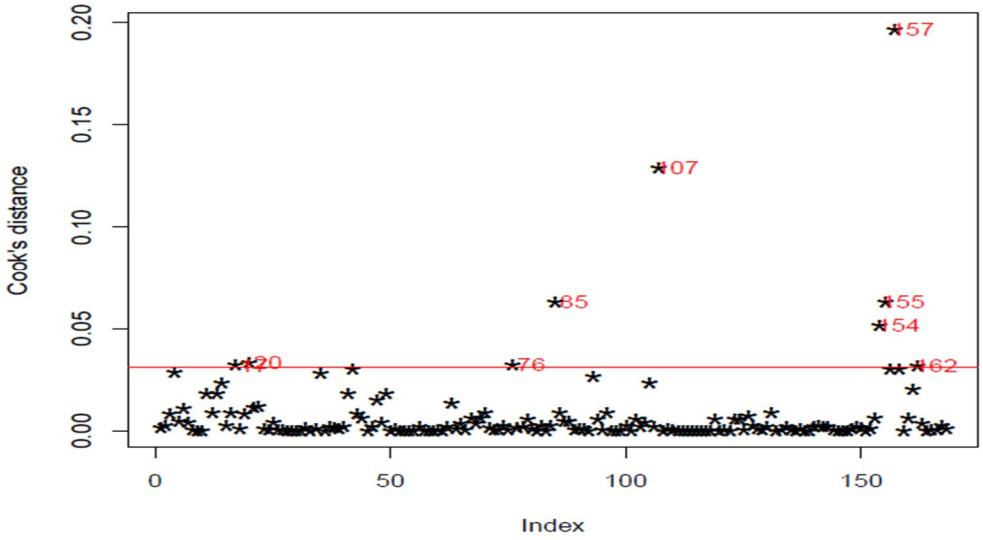
For us to understand the effects of outliers, we compare the fit of a simple linear regression model of the data set with and without the outliers. We fit a simple linear regression model using the data set without outliers. The regression parameters are given as $\hat{\beta}_0 = 7.33$ and $\hat{\beta}_1 = 0.86$. Hence, the regression line for the i^{th} observation is given as $\hat{y}_i = 7.33 + 0.86\, x_i$. The value of $R^2 = 0.77$. This implies that 77% of the total variation is accounted for by the regression model and the remaining 23% is due to error. We will note that the fit in this case is far better than the previous case including the outliers. Figure 5 gives a plot of the fit of the simple linear regression model when the outliers were excluded from the data set.

Figure 5: Plot of the fit of the simple linear regression model of examination scores against test scores without outliers.



We notice the change in the slope of the regression fit after removing the outliers. It is obvious that if we had used the outliers in the model, our predictions would have had high error. We would compute the Cook's distance with respect to the regression and check further for influential observations in the absence of the first detected outliers. Figure 6 shows a plot of influential observations by Cook's distance without including the first set of outliers.

Figure 6: A plot of influential observation by Cook's distance without including the first set of outliers.



It is obvious that we can still find some outliers in Figure 6. From Figure 6, the 21st, 24th, 82nd, 91st, 114th, 162nd, 163rd, 166th and 171st, 94th, 137th and 165th observations are now tagged as influential. The observations are as follows:

Observation number	y	x
21	34	40
24	56	45
82	34	40

91	85	80
114	50	35
162	55	70
163	85	80
166	46	30
171	40	56

Observation 91 and 163 are extreme observations with similar examination scores of 85 and test scores of 80 but the remaining observations are not necessarily extreme values with respect to either of the variables. Observation 91 and 163 should not be reassessed. Hadi (1992) defines masking effect as that which an outlier is undetected because of the presence of another adjacent outlying observations. Jung – Tsung (2007) defines the swamping effect as that which a good observation is incorrectly identified as an outlier because of the presence of another outlier. In our case, there is masking effect since observation 91 and 163 were undetected because of the presence of the 2^{nd}, 6^{th}, 8^{th}, 17^{th}, 26^{th}, 71^{st}, 94^{th}, 137^{th} and 165^{th} observations. Observation 91 and 163 can also be a case of swamping since these are good observations but incorrectly identified as an outlier because of the presence of other outliers. The 21st, 24^{th}, 82^{nd}, 114^{th}, 162^{nd}, 166^{th} and 171^{st}, observations are good observations which are incorrectly identified as an outlier because of the presence of the 91^{st} and 163^{rd} observations. For instance, the 82^{nd} observation (with examination score of 34 and test score of 40) could actually be the capability of the student. The student has shown likely the same scores in both examination and test. We will note that observation 94 (with examination score of 87 and test score of 70) was identified as an outlier in Figure 4 but observation 91 (with examination score of 85 and test score of 80) was not identified as an outlier in Figure 4 rather in Figure 6. This is a good case of masking effect since it appeared incorrectly as a good observation in Figure 4.

5.0 CONCLUSION

Examination malpractice has become so very rampant in Nigerian Universities. The University authorities should suggest proper measures for controlling examination malpractices of students during examination and test period. Students engage in examination malpractice because they want to pass. Examination malpractices lead to irreversible loss of credibility, dismissal, lack of self-confidence, embarrassment e.t.c. Guilty ones caught should be punished.

The Niger Delta University student's handbook has discussed procedures for handling suspected examination malpractice cases (Students handbook, (2015)). If during marking, moderation or collation of examination scripts or materials, an examiner or any member of staff suspects that malpractice had taken place, the examiner or member of the staff must report same to the Chief Examiner who is the Head of Department in writing within 24 hours. The affected student must be informed immediately of the allegation and made to submit a written statement. The examiner should give the student a new answer script and allow the student to write another examination. The student statement must be under confidential cover to the Chief Examiner to the Dean of the Faculty. The Dean should forward the report to the Senate Student Disciplinary Committee through the Vice chancellor. Any student suspected to have been engaged in examination malpractice will appear before the Senate Student Disciplinary Committee. The decision of the Senate Student Disciplinary Committee shall be conveyed to the student in writing. It is also advised that all cases of suspected examination malpractice must be disposed of within the shortest possible time. In cases where malpractice is proven, written materials or proofs are kept by the University until the punishment has been served.

Identification of outliers or influential points is an important part of regression modelling. It is also important that we improve the quality of our original data by reducing the impact of outliers and we know

observations that have great influence on the parameter estimate in the case of regression models. Although authors like Ka – Veng and He-Qing, suggested using robust regression since it provides parameter estimates that are insensitive to the presence of outliers. We suggest that outliers should be excluded from the sample or be corrected. We will note that a data set with multiple outliers might become complicated due to masking and swamping effect. The masking and swamping effect appear in our case where observations are incorrectly regarded as good observation and others observations are incorrectly regarded as outliers respectively.

In our application, we have associated examination and test malpractice when detected with outlier. In some suspicious cases, we suggested reassessment. If students are caught guilty of the offence, they should be punished accordingly.

REFERENCES

Adikaram K.K.L.B, Hussein M.A, Effenberger .M and Becker .T (2014). Outlier detection method in linear regression based on sum of Arithmetic progression, The Scientific World Journal, Vol. 2014, page 1-12.
Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data, 3rd ed. UK : Wiley, Chicester.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

Farid K. and Swallow .W. H. A comparison of some classical approaches to outlier detection in linear regression and an approach based on adaptively – ordered recursive residuals, Unpublished Dissertation, North Carolina State University.
Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples, Technometrics, Vol. 11, page 1-21.
Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society, Series B, **54**, 761-771.
Jung – Tsung Chiang (2007), The masking and swamping effects using the planted mean-shift outliers models, International Journal of Contemporary Mathematical Sciences, Vol. 2, No. 7, page 297 -307.
Ka – Veng Yuen and He-Qing Mu, Outlier detection using the outlier probability for robust linear regression
Marasinghe, M. G. (1985), A Multistage Procedure for Detecting Several Outliers in Linear Regression, *Tech*nometrics, 27, 395-399.

Moore, D.S and McCabe, G. P. (1999) Introduction to the practice of Statistics, 3rd edition, New York: W. H. Freeman.
Montgomery, D. C., Peck, E. A., and Vining, G. G. (2003). Introduction to Linear Regression Analysis, 3rd ed. New York: John Wiley & Sons.

Onyibe C.O, Uma U.U and Ibinaz .E (2015). Examination malpractice in Nigeria: Causes and Effects on National Development, Journal of Education and Practice, Vol. 6, No. 26, page 12- 18.
Oyeyemi .G.M, Bukoye .A and Akeyede .I (2015). Comparison of outlier detection procedures in multiple linear regressions, American Journal of Mathematics and Statistics, Vol. 5, No. 1, page 37 -41.
Pimpan A. and Prachoom S. (2009). A comparative study of outlier detection procedures in multiple linear regression, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol.1.

R Development Core Team (2008), R: A Language and Environment for Statistical Computing, Vienna, Austria

Stephen R.S and Senthamarai K.K (2017). Detection of outliers in Regression model for medical data, International Journal of Medical Research & Health Sciences, Vol. 6, No. 7, page 50-56.
Student Handbook (2015), Niger Delta University, Wilberforce Island, Bayelsa State.