Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
 ISSN 1118 - 1931

APPLICATION OF RANGE-TEST IN MULTIPLE LINEAR REGRESSION ANALYSIS IN PRESENCE OF OUTLIERS

[1] E. O. Biu, [2]U. P. Ogoke and [3] S. I. Iwueze

[1,2]Department of Mathematics /Statistics,
University of Port Harcourt, Rivers State, Nigeria.

[3]Department of Statistics,
Federal University of Technology, Owerri, Imo State, Nigeria
Email: [1]emmaunelbiu@yahoo.com,[2]uchedubem@yahoo.comand [3]isiwueze@yahoo.com

Received: 27-11-13
Accepted: 28-04-14

ABSTRACT

Application of range-test in multiple linear regression analysis in the presence of outliers is studied in this paper. First, the plot of the explanatory variables (i.e. Administration, Social/Commercial, Economic services and Transfer) on the dependent variable (i.e. GDP) was done to identify the statistical trend over the years. The identified trend is linear and positive upward. Secondly, a multiple linear regression model is constructed to describe the relationship between the dependent variable and independent variables. Thirdly, it is shown how outliers could be handled using the Range Test, because outliers can have deleterious effects on statistical analyses. When researchers ignore such abnormal observations, especially with respect to dependent variables, the empirical results can be misleading. From our findings, we conclude that treating outliers from regression models give better fit of the model in terms of R-square.

INTRODUCTION

Data collected by research workers commonly contain outliers, and it is important that these outliers be identified in the course of a thorough and correct statistical analysis. Outliers are observations that appear to be inconsistent with the remainder of the collected data set (Iglewiez and Hoaglin, 1993). An outlier is generally considered to be a data point that is far outside the data range for a variable or population Jarrell (1994). Hadi and Simonoff (1993) provided results for testing multiple outliers in regression. Beckman and Cook (1983) encountered a serious problem of "Masking". If there are several outliers, the least squares estimation of the parameters of the model may leads to small residuals for the outlying observations.

A number of procedures have been proposed in recent years for detecting outliers in linear regression, yet their detection still may be difficult, especially when there are multiple outliers in the data.

A comprehensive text on the study of outliers is that of Barnett (1993). A recent review article by Chatterjee and Hadi (1986) describes many of the well-known outlier-detection procedures and model diagnostics and their interrelationships in the context of linear regression.

In the early years, many statisticians and practitioners viewed outlier-identification methodologies largely as ways to legitimize deleting observations which, though not necessarily erroneous, fell outside the pattern seen in the bulk of the data and were perhaps troublesome in the analysis.

Nowadays, outlier identification is viewed more broadly. It is widely recognized that, in some applications, outliers are of interest in their own right and may be the most important observations in the data set; identifying them may help chart future research.

And the literature on influential observations has expanded our understanding of the need to identify certain points as candidates for special treatment (perhaps downweighting, or deletion), lest they warp our impression of relationships in the body of the data.

The focus has moved away from viewing these procedures as providing support for automatic deletion of points, and toward seeing them as aids in identifying points for more careful scrutiny.

The latter perspective suggests interest in identifying moderate as well as extreme outliers. Data analysts routinely encounter data sets which potentially contain one or more outliers. When, as is usually the case, there is no a priori reason to suspect that particular observations are the outliers, an outlier test based on the sequential (perhaps better called "repeated") application of a single-outlier test statistic is commonly used.

The linear regression model can be expressed in terms of matrices as $y = X\beta + \varepsilon$, where y is $n \times 1$ vector of observed response values, X is $n \times p$ design matrix of p repressors, β is $p \times 1$ vector of errors terms. The most widely used technique in finding the best estimates of β is the method of ordinary least squares (OLS) which minimizes the sum of squared distances for all points from the actual observations to the regression surface. When the errors are not normally and independently distributed (NID), distortion of the fit of the regression model can occur and consequently the parameter estimates and inferences can be flawed. The presence of one or more outliers is one of the common causes of non-normal errors terms. As defined by Barnett and Lewis (1994), outliers are observations that appear inconsistent with the rest of the data set. Such outliers can have a profound destructive influence on the statistical analysis. It is not unusual to find an average of 10% outlying observations in data set of some processes [Hampel et al (1986)]

Informally, an outlier is any data value that seems to be out of place with respect to the rest of the data. Consider the single attribute of height of a boy who is considerably taller than everyone else in his class is said to be outlier. However, outlier need not be extreme values only. For example, if a woman of average height were to walk into a room filled with only basketball players and jockeys, then she would be an outlier because her height would differ noticeably from the other individuals.

In brief, the aim of this research work was to apply the range-test in multiple linear regression analysis to handling outliers. From this handling of outliers, we can gain an insight if treat of outliers in a regression model can sometimes give completely different results. The aim of this research was achieved by the following objectives:

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

- To identify the statistical trend and examine the condition of the function
- By building a suitable multiple regression models relationship between the dependent variable and independent variables.
- Then, the range test was applied to handle the suspected outliers on the explanatory variables.
- Check the effect of the outliers to the co-efficient of ordinary least square variables and *R-square.*

MATERIALS AND METHOD

This study employs annual data from 1961 to 2010, obtained from two sources: Central Bank of Nigeria (CBN) Statistical Bulletin and National Bureau of Statistics (2010). The study employs indicators such as the Annual Federal Government Budget Estimates on Administration, Social/Commercial services, Economic services and Transfer as the independent (explanatory) variables, while the real Nigeria Gross Domestic product (GDP) as the dependent (explained) variableuse in measuring economic growth. This research work was analyzed with the help some statistical packages particularly designed for analysis. The statistical softwares are Micro-Excel and Minitab 16.

In order to test the effectiveness of the applied method, a multiple linear regression model is constructed to describe the relationship between the dependent variable (GDP) and independent variables (Administration, Social/Commercial services, Economic services and Transfer). Then, the range test method was applied to confirm that a suspected outlier is as extreme as it appears.

Model Specification

The functional relationship between the dependent and the independent variables in our study are established as follows:

GDP = f(GBA, GBSC, GBES, GBT) (1.0)

where,

GDP = Gross Domestic Product

GBA = Federal Government Budget Estimates on Administration

GBSC = Government Budget on Social/Commercial services

GBES = Government Budget on Economic Services

GBT = Government Budget on Transfer

Thus we have:

$$GDP = \beta_0 + \beta_1 GBA + \beta_2 GBSC + \beta_3 GBES + \beta_4 GBT + \varepsilon \quad (2.0)$$

where,

β_1 , β_2 , β_3 , and β_4 = Parameters of the model and ε = Stochastic Disturbance term

Coefficients of Multiple Determination (Correlation) $R^2_{Y, X_1\ X_2\ \ldots, X_K}$.

R^2 shows the percentage of the total variations of y explained (accounted or) by the regression (i.e. by changes in X_1, X_2, X_3 and X_4). This can be defined by

$$R^2 = \frac{SS_r}{SS_{yy}},\ 0 < R^2 \leq 1 \quad (3.0)$$

The coefficient of determination or multiple determinations (in multiple linear regression). $\mathbf{R^2}$ is the percentage of total variation in the response that is explained by predictors or factors in the model. In general, the higher the R^2 , the better the model fits your data. $\mathbf{R^2}$is always between 0 and 1.The closer R^2 is to zero the worse the fit (i.e. the model is inadequate).

Methods of Handling Outliers

Some of the methods of handling outliers are Data Trimming (or screening or cleaning), Winsorization and Range-Test, but this research work will focus on the Range-Test.

The Range Test

A statistical test should be applied to confirm that a suspected outlier is as extreme as it appears. A number of these tests have been devised. One such test is the range test which is the quickest to calculate, and is given by

$$R_T = \frac{Extreme\ Value - Overall\ Mean}{Overall\ Standard\ Deviation},$$

[Nduka, 1999] (4.0)

That is, x (Extreme value) is an outlier if

$$\frac{abs(x - mean)}{Std\ dev} > 3 \qquad (5.0)$$

Then, we call anything that falls more than three standard deviations away from the mean an outlier. We reject the outlier if its ratio exceeds 3 (three), otherwise accept the outlier. When some value is discovered as an outlier being different from other observations, it may safety be replaced by the next lowest extremes values.

The *hypothesis* of interest would be

H_0: There is no outlier

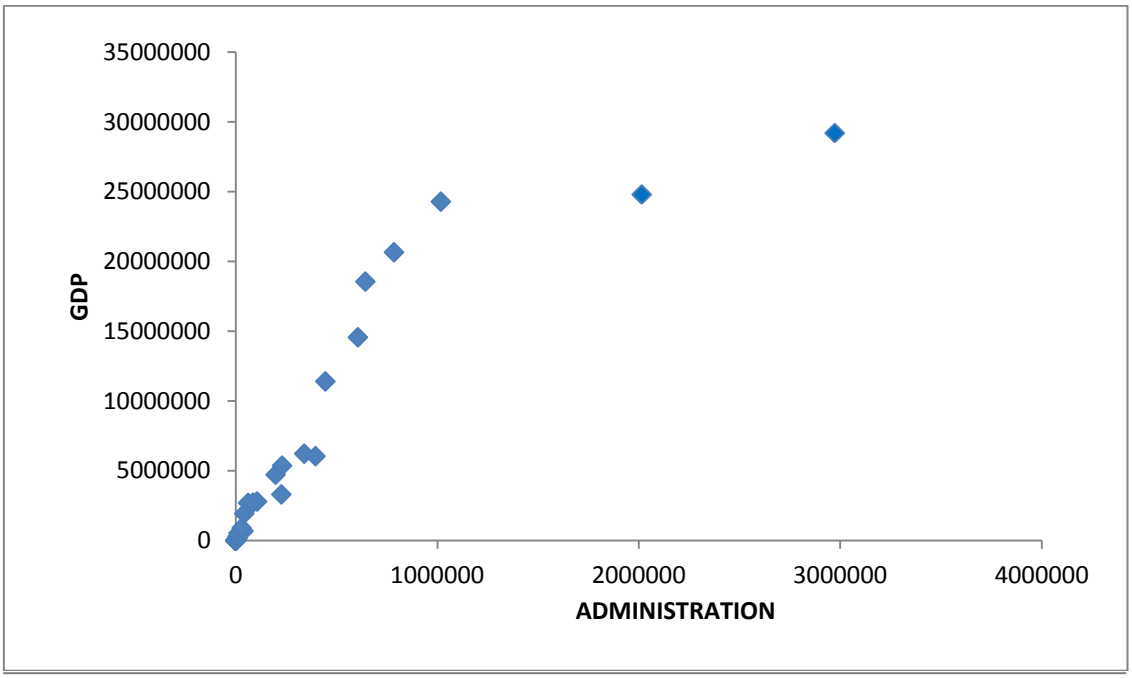
H_1: There is the presence of outlier

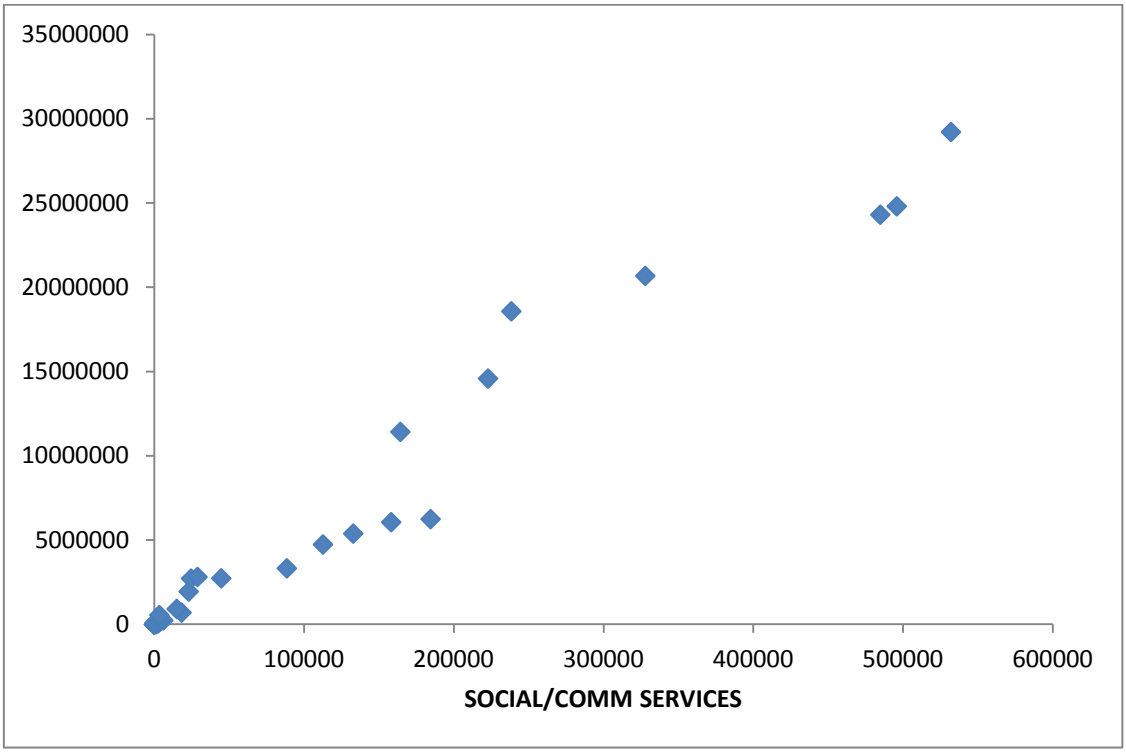
DATA ANALYSIS

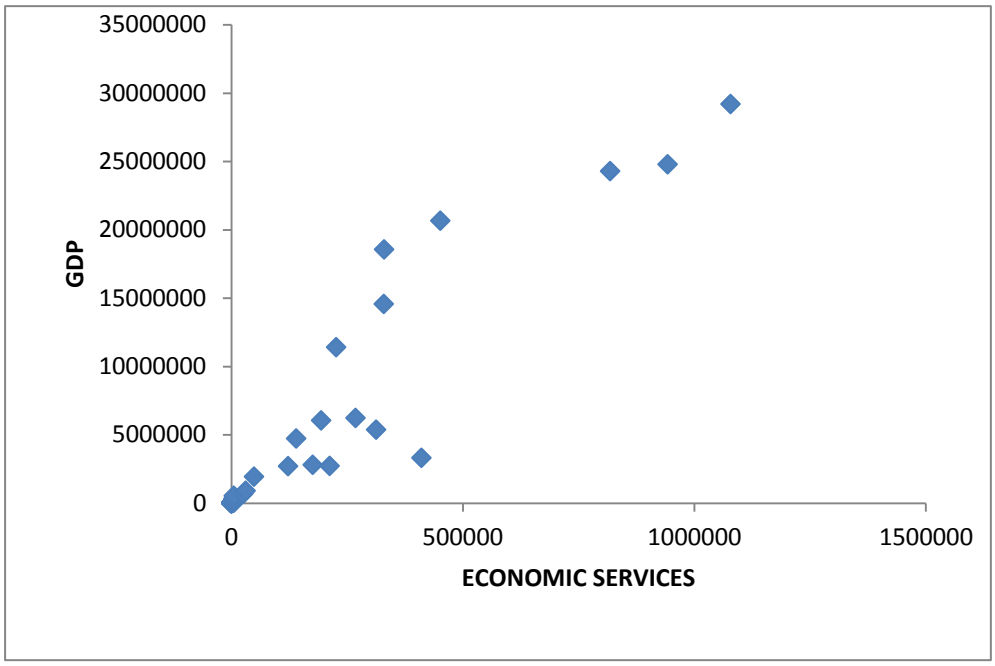
Identification of Trend

The scatter plots of the explanatory variables on GDP are showed in Figure 1.0 and Figure 2.0

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*







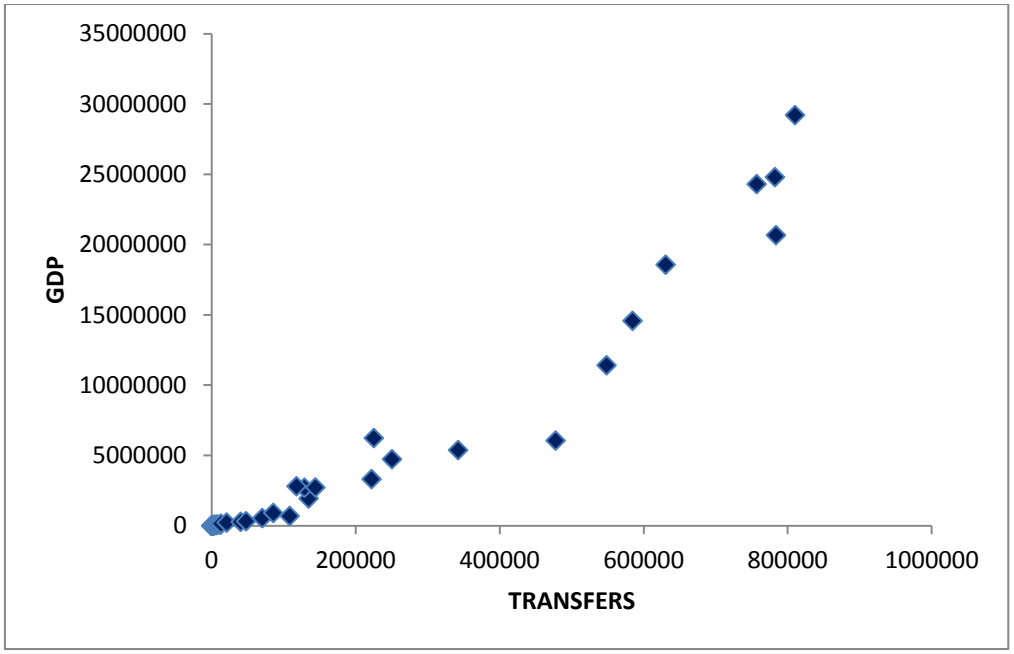


Figure 1.0: Each of explanatory variables on GDP

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

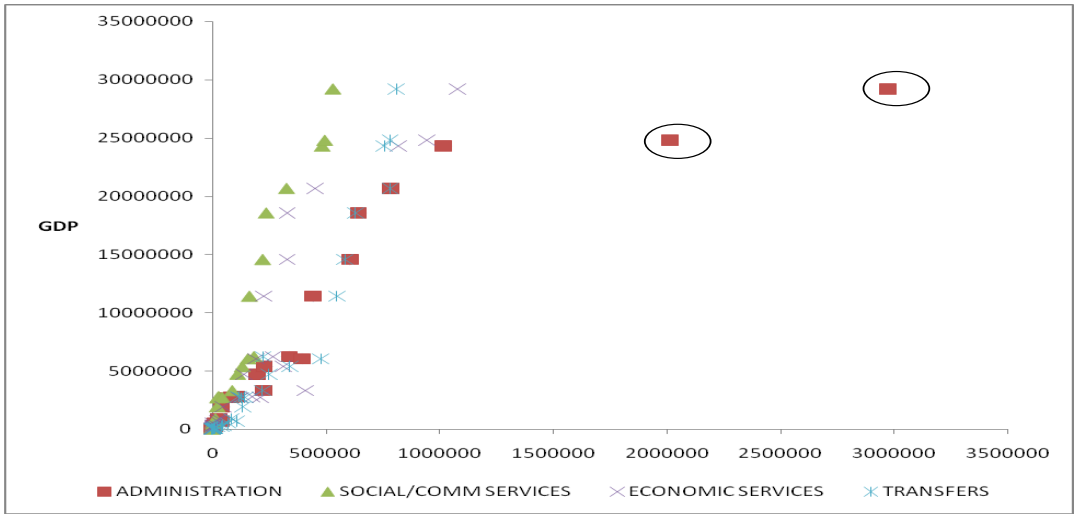


Figure 2.0: Explanatory variables on GDP

Each of explanatory variables on GDP is shown in Figure 1.0 while Figure 2.0 shows a combined plot of the explanatory variables on GDP. Examining Figure 1.0 and Figure 2.0, we notice that spending appreciate from 1961 to 2010 with some high and low extremes point (i.e. outliers), which seems to indicate a linear and positive upward growth of Nigeria economy. Also in Figure 2.0, two suspended outliers' data points are circled as extreme value.

According to our trend result, building suitable multiple regression model that will explain the mechanism that is involved in the growth of the Nigeria GDP is proper. These regression models were tested qualitatively and analytically in order to provide a better understanding of what and how each explanatory variables contributed to Nigeria GDP from 1961 to 2010.

Multiple Regression Models

Multiple Linear Regression Model before the Application of the Range Test

The results from the Minitab 16.0 statistical software output of the multiple linear regression model done are summary in terms of the P-value in the Analysis of Variance (ANOVA) table, R-square and Adjusted R-square are in Table 1.0 and they are discussed below. Also, discussed are the points of Large Standardized residuals, large influence and the estimated co-efficients of the regression model built (Table 2.0).

TABLE 1.0: Summary of the statistics Obtained for Multiple linear Regression Model [Before Apply the Range Test].

Statistics	P-value in the ANOVA	R^2 value	Adjusted R^2
Parameters	0.0001	98.00%	97.60%

From Table 1.0, the P-value in the ANOVA is (0.0001) for multiple linear regression models before the application of range test. This shows that the model estimated by the

regression procedure is significant at $\alpha - level$ of 0.05. This indicates that at least one variable contributed significantly to the regression model and some estimated parameter coefficients are significantly different from zero.

The R^2 value indicates that the predictors explain 98.0% of the variance in GDP. Additionally, the R^2 adjusted is 97.6%, which accounts for the number of predictors in the model. Since R^2 value is close to R^2 adjusted, it implied that the models do not appear to be overfit and has adequate predictive ability. It implies that the model fits the data well.

However, some observations which are 39, 48 and 50 values (Hint: the Federal Government Budget Estimates for 1999, 2008 and 2010 respectively) on the model fitted for the data set are identified as unusual because the absolute values of the standardized residuals are greater than ± 2.00. This may indicate that they are outliers that need to be removedor handled.

This research work applies the range test to handle these outliers (See Table 3.0).

Also, the Minitab 16.0 software outputsheet shows that two observations "43" and "46" values (Hint: the Federal Government Budget Estimates for 2003 and 2006 respectively) indicating large influence points. Indicating that the federal government spending in these two years seems outrageous, because of the standard error values fitted in the regression model at this position are very large. These points were also handled by the range test (See Table 3.0).

Table 2.0: Estimated Parameters for Multiple linear Regression Model [Before Apply the Range Test].

Parameter Estimates	**(constant)**	**(Administration)**	**(Social/commercial Services)**	**(Economic Services)**	**(Transfers)**
Before Apply The Range Test	-187269 (0.747)	2.54(0.0005)	32.02(0.001)	-4.62(0.114)	11.86(0.001)

P-values for the above Table 2.0 are the parenthesis and the P-values for the estimated coefficient of administration, Social/commercial Services and Transfers is 0.000, indicating that they are significantly related to GDP at an $\alpha - level$ of 0.05, but the estimated coefficient of Economic services P-values is 0.114 for the model fitted, indicating that it has negative contributed to Nigeria GDP at an $\alpha - level$ of 0.05. This suggests that a model with administration, Social/commercial Services and Transfers may be more appropriate for data set in question. In addition, the estimated coefficientsindicate that three of the explanatory variables contributed positive to Nigeria GDP growth from 1961 to 2010.

Hence, the range test was used to handle the following extreme observations suspended by the Minitab 16.0 software as outliers. They are the explanatory variables in positions 39, 43, 46, 47, 48, 49 and 50.

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Table 3.0: Summary of theRange Test Statistic

Observations	Explanatory Variables	Range Test Value	Remark
39	ADMINISTRATION	0.04	Accept
	SOCIAL/COMM SERVICES	0.16	Accept
	ECONOMIC SERVICES	1.18	Accept
	TRANSFERS	0.30	Accept
43	ADMINISTRATION	-0.36	Accept
	SOCIAL/COMM SERVICES	-0.68	Accept
	ECONOMIC SERVICES	-0.29	Accept
	TRANSFERS	-1.33	Accept
46	ADMINISTRATION	0.82	Accept
	SOCIAL/COMM SERVICES	1.28	Accept
	ECONOMIC SERVICES	0.85	Accept
	TRANSFERS	1.95	Accept
47	ADMINISTRATION	1.09	Accept
	SOCIAL/COMM SERVICES	1.94	Accept
	ECONOMIC SERVICES	1.34	Accept
	TRANSFERS	2.57	Accept
48	ADMINISTRATION	1.53	Accept
	SOCIAL/COMM SERVICES	3.11	**Reject**
	ECONOMIC SERVICES	2.85	Accept
	TRANSFERS	2.46	Accept
49	ADMINISTRATION	3.40	**Reject**
	SOCIAL/COMM SERVICES	3.19	**Reject**
	ECONOMIC SERVICES	3.35	**Reject**
	TRANSFERS	2.56	Accept
50	ADMINISTRATION	5.20	**Reject**
	SOCIAL/COMM SERVICES	3.46	**Reject**
	ECONOMIC SERVICES	3.91	**Reject**
	TRANSFERS	2.67	Accept

Then, we call points that falls more than three standard deviations away from the mean outliers. We rejected the outliers that are different from other observations and replaced by the next lowest extremes values. In a similar manner, we built a suitable multiple linear regression model for the data set after the application of the range test.

Multiple Linear Regression Model after the Application of the Range Test

The results obtained from the Minitab 16.0 statistical software output sheet for the multiple linear regression model after the application of the range test was done is summary in terms of the P-value in the Analysis of Variance (ANOVA) table, R-square and Adjusted R-square are in Table 4.0 and its discussed below. Also, discussed are the points of Large Standardized residuals, large influence and the estimated co-efficients of the regression model built (Table 5.0).

TABLE 4.0: Summary of the Statistics Obtained for Multiple Regression Models [After Apply the Range Test].

Statistics	P-value in the ANOVA	R^2 value	Adjusted R^2
Parameters	0.0001	98.60%	98.30%

From Table 4.0, the P-value in the ANOVA is (0.0001) for multiple linear regression models after the application of range test. This shows that the model estimated by the regression procedure is significant at $\alpha - level$ of 0.05, also like the model built before the application of the range test. However, the R^2 value indicates that the predictors explain 98.6% of the variance in GDP. Additionally, the R^2 adjusted is 98.3%, which accounts for the number of predictors in the model. Since, the R^2 value is closer to 1 than the model built before the application of the range test by 0.6%,it implies that this model does not appear to be overfit and it's has adequate predictive ability than the first model. We conclude that this model fits the data well.

Table 5.0: Estimated Parameters for Multiple Linear Regression Model [After Apply the Range Test].

Parameter Estimates	(constant)	(Administration)	(Social/commercial Services)	(Economic Services)	(Transfers)
Before Apply The Range Test	-32097 (0.857)	31.64(0.000)	-22.24(0.030)	-1.93(0.411)	3.43(0.298)

P-values for the above Table 5.0 are the parenthesis and the P-values for the estimated coefficients of administration and Social/commercial Services indicate that they are significantly related to GDP at an $\alpha - level$ of 0.05, but the estimated P-values coefficient for Economic services and Transfers are not significantly related to GDP at an $\alpha - level$ of 0.05. This suggests that a model with administration and Social/commercial Services may be more appropriate for data set in question.

Furthermore, the explanatory variables of Social/commercial Services and Economic services contributed negative to Nigeria GDP, while administration and Transfers contributed positive to Nigeria GDP growth from 1961 to 2010.

We concluded that an outlier will cause the coefficient of determination to be smaller than its actual value. Therefore, examining and possibly eliminating such highly suspected observations (or outliers) in a sample data will help the validity of the regression model.

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

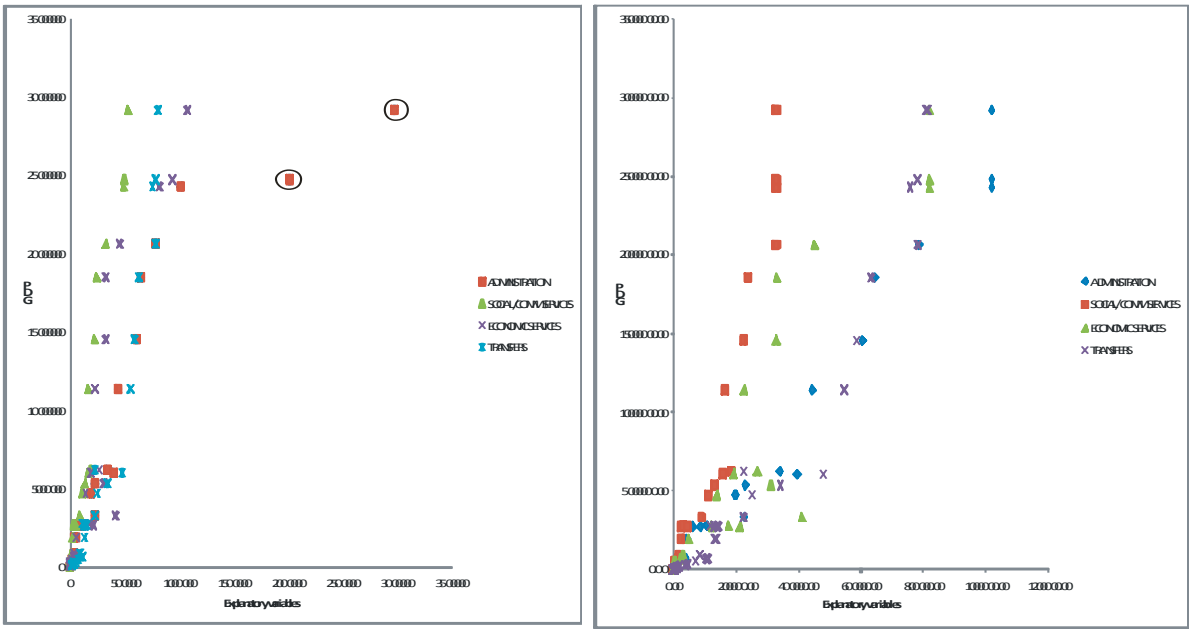


Figure 3.0a: With Outlier Figure 3.0b: With Outlier

Regression equation comparison:

$$y = -187269 + 2.54x_1 + 32.02x_2 - 4.62x_3 + 11.8x_4 :$$

$$y = -32097 + 31.6x_1 - 22.24x_2 + 1.93x_3 + 3.43x_4$$

Coefficient of determination: R^2 = 98.0% = 0.98Coefficient of determination: R^2 = 98.6% = 0.99

where,

y is the Gross Domestic Product (GDP), x_1 is the Administration expenditure, x_2 is the Social/Commercial Services expenditure, x_3 is the Economic Services expenditure, x_4 is the Transfers expenditure and e_i is the standard error of the model.

The charts in Figure 3.0a and 3.0b compare regression statistics for sample data set with and without an outlier. Here the chart in Figure 3.0a has some outlier, located at the high end of the X axis. As a result of those outliers, the constant value of the regression line changes greatly, from -187,269 to -32,097 in the chart in Figure 3.0b; so the outliers would be considered an influential point.

Summary

Application of range-test in multiple linear regression analysis in presence of outliers was studied in this paper with the help of one of the method of handling outliers. Firstly, the plot of the explanatory variables (i.e. Administration, Social/Commercial Services, Economic Services and Transfer) was plotted against the dependent variable (i.e. GDP). This helped to show the statistical trend over the years. The identified trend was a linear and positive upward trend. Secondly, a suitable multiple linear regression was constructed to describe the relationship between the dependent variable and independent variables. Thirdly, this research shows how Outliers could be handled using the Range Test because outliers can have deleterious effects on statistical analyses. Finally, after handling the outlier on the explanatory variables by the range test method, then multiple linear regression model was built again for the data set. Examining and comparison was done to draw valid inference. From our result, we conclude that handling outliers from

regression models give better fit of the model in terms of R-square and the data sets in questions.

The presence of an outlier can affect the parameter estimates of regression models and even the direction of the coefficient signs (from positive to negative and vice versa). When researchers ignore such abnormal observations, especially with respect to dependent variables, the empirical results can be misleading.

Contribution

- Able to treat the high extreme points by the application of the range test method of handling outliers.
- Able to identify the effect of the outliers on the **e**stimated parameters coefficientsin Multiple linear Regression Model build (i.e. change coefficient signs from positive to negative).

Technical: This study could be extended to a non-linear case

Non-technical: Federal Government Budget estimates' spending on Social/commercial Services and Economic services seems to contributed negative to Nigeria GDP and does not affect the economy growth (i.e. GDP). Therefore, the budget expenditure on these explanatory variables should be supervised.

REFERENCES

Barnett, V. (1983); *Principles and methods for handling outliers in data sets*. Statistical Methods and the Improvement of Data Quality, pp. 131-166.

Barnett, V. and Lewis, T. (1994);*Outliers in Statistical Data*, 3nd ed. John Wiley: New York.

Beckman, R. and Cook R. (1983);*Outlier……s* (with discussion and response Technometrics, 25(2): 119-163.

Central Bank of Nigeria (CBN) Statistical Bulletin(2010).

Chatterjee, S., and Hadi, A. S. (1986);*"Influential Observations, High Leverage Points, and Outliers in Linear Regression,"* Statistical Science, 1, 379-416.

Hadi, A, and Simonoff, J (1993); Comment to Paul and Fung (1991). *Technometrics*, 34, pp. 373-374.

Hadi, A. (1992);*Identifying multiple outliers in multivariate data*. Journal of the Royal Statistical Society, B, 54, pp. 761-771.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986);*Robust Statistics*. *The Approach Based on Influence Functions*. J.Wiley, N.York. ISBN 0-471-82921-8.

Hawkins, D. (1980);*Identification of outliers*. Chapman and Hall, Londom.

Iglewicz B. and Hoaglin D.C. (1993);*How to Detect and Handle Outliers*. American Society for Quality Control M. Iwnkee, WI.

Iglewicz B. and Hoaglin D.C. (1993); *How to Detect and Handle Outliers.* Milwaukee, WI: ASQC Quality Press.

Scientia Africana, Vol. 13 (No.1), June 2014. Pp138-150
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Jarrell, M.G. (1994);*A Comparison of two procedures, the Mabalanobis Distiance and the Andrews – Pregibon Statistics, for identifying Multivariate Outliers.* Researches in the Schools, 1:49 – 58.

National Bureau of Statistics Journal (2010).

Nduka, E. C. (1999);*Principles of Applied statistics 1, Regression and Correlation Analysis*. Crystal Publishers, Okigwe, Imo State, 2nd edition. pp. 41-48.