International Journal of Innovation in Science and Mathematics
Volume 4, Issue 5, ISSN (Online): 2347–9051

Parametric Versus Non-Parametric Simple Linear Regression on Data with and Without Outliers

Okenwe Idochi
Department of Statistics,
School of Applied Sciences, Rivers State Polytechnic
PMB 20, Bori, Rivers State Nigeria
Email ID: nwonda@yahoo.com

Opara Jude
Department of Statistics,
Imo State University
PMB 2000, Owerri Nigeria
E-mail ID: judend88@yahoo.com

Ononogbu Amarachi C.
Department of Statistics, Michael Okpara
University of Agriculture, Umudike
P.M.B. 7267, Umuahia Abia State Nigeria
Email ID: micable2016@gmail.com

Bassey Uwabunkonye
Department of Mathematics and Statistics,
Akanu Ibiam Federal Polytechnic, Uwana
P.M.B 1007, Uwana, Afikpo, Ebonyi State, Nigeria
Email ID: okoriewo@gmail.com

Abstract **– This study is on Parametric versus Non-Parametric Simple Linear Regression on Data With and without Outliers. Data used for this study were collected from the department of Mass Communication, Imo State University Owerri Imo State Nigeria. Twenty five (25) students were selected at random to determine the Cumulative Grade Point Average (CGPA) at the end of 2014/2015 Academic session (Y) and their respective Joint Admission Matriculation Board (JAMB) score (X). The use of a programming language software known as "R Development" was used in this study. The set of data was subjected to normality test, and it was concluded that all residuals in the y-direction are not normally distributed via the Anderson-Darling technique. The procedures for the parametric Theil's and that of its non-parametric OLS regression were highlighted. The data were analyzed for both parametric and non-parametric techniques; thereafter outliers were detected and expunged from the data. The data after removing outliers were re-analyzed. From the analysis, the result revealed that there is a significant relationship between students CGPA and their JAMB scores for both the parametric OLS regression and non-parametric Theil's regression with and without outliers.**

It was concluded that the parametric OLS is better than its non-parametric Theil's regression for both data with and without outliers since their standard error, AIC and BIC are lower than that of Theil's regression. It was also concluded that the standard error for the parametric regression with outliers which is 0.3405 reduced to 0.1962 for the parametric regression without outliers. On the other hand, the standard error for the non-parametric regression with outliers which is 0.3609 reduced to 0.2087 for the non-parametric regression without outliers. This implies that the model for the data without outliers is more efficient than the model for the data with outliers for both the parametric and non-parametric regression. Therefore the researchers recommend that future researchers should look into a similar work with large sample size to examine the differences between the parametric and nonparametric Regression.

Keywords **– Non-Parametric Theil's Regression, Parametric OLS, Akaike Information Criterion, Bayesian Information Criterion, Residual Standard Error, Outliers.**

I. INTRODUCTION

The simple linear regression model is the ordinary or traditional equation representing the relationship between two variables; the response and the explanatory variables. Sometimes the residuals in a regression analysis may deviate far from the others. In this case, an outlier occurs. It is obvious that no observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. Therefore, the probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions. Observations that in the opinion of the investigator stand apart from the bulk of the data have been called "outliers", "extreme observations" "discordant observations", "rouge values", "contaminants", "surprising values", "mavericks" or "dirty data" by Ranjit [11]. An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs. An outlier is a data point that is located far from the rest of the data.

Again, the presence of outliers may contribute to non-normal distribution. Consider a situation where the distribution of the errors is not normal. If the errors are coming from a population that has a mean of zero, then the OLS estimates may not be optimal, but they at least have the property of being unbiased. If we further assume that the variance of the error population is finite, then the OLS estimates have the property of being consistent and asymptotically normal. However, under these conditions, the OLS estimates and tests may lose much of their efficiency and they can result in poor performance by Mutan [8]. To deal with these situations, two approaches can be applied. One is to try to correct non-normality, if non-normality is determined and the other is to use alternative regression methods, which do not depend on the assumption of the normality according to Birkes and Dodge [2].

In a simple linear model, Theil [15] proposed the median of pairwise slopes as an estimator of the slope parameter. Sen [13] extended this estimator to handle ties. The Theil-Sen Estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses a high asymptotic efficiency. Thus it is very competitive to other slope estimators (e.g., the least squares estimators), see Sen [13], Dietz [3] and Wilcox [16].

The proposed estimators contain an integer variable which controls the amount of robustness and efficiency.

The maximal possible robustness (in terms of break-down point) is attained when the integer variable is chosen to be the number of the parameters to be estimated; while the maximal efficiency is achieved when the variable assumes the sample size; any value of the variable taking in between results in an estimator which gives a compromise between robustness and efficiency.

In straight-line regression, the least squares estimator of the slope is sensitive to outliers and the associated confidence interval is affected by non-normality of the dependent variable. A simple and robust alternative to least squares regression is Theil regression, first proposed by Theil (1950). Theil's method actually yields an estimate of the slope of the regression line. Several approaches exist for obtaining a nonparametric estimate of the intercept. In this paper, we shall use the R for estimating the parameters. This paper shall be of paramount significant to future researchers who may wish to carry out a similar research, knowing when and how to use the parametric and non-parametric methods.

II. Related Literature Review

There is need to review works done by past researchers in order to have a proper guide. Here are some recent works done by past researchers.

Opara [10] conducted a research on the comparison of parametric and non-parametric linear regression. First, the set of data was subjected to normality test, and it was concluded that all errors in the y-direction are normally distributed (i.e. they follow a Gaussian distribution) for the commonly used least squares regression method for fitting an equation into a set of (x,y)-data points using the Anderson-Darling technique. The algorithms for Theil's were stated in their work as well as its non-parametric counterpart. Data used for the study were collected from a trader in Dauglas Owerri Market in Imo State Nigeria who sales pears. The numbers of rotten pears (y) in 20 randomly selected boxes from a large consignment were counted after they have kept in storage for a studied number of days (x). The use of a programming language software known as "R Development" and Minitab were used in the study. From their analysis, the result revealed that there exists a significant relationship between the numbers of rotten pears and the number of days for both the ordinary least squares and the Theil's regression. It was concluded that the parametric OLS is better than its non-parametric Theil's regression since their AIC and BIC are both lower than that of Theil's regression. It was recommended that future researchers should embark on a similar research study using large sample size, and using non-normal data to examine the differences between the OLS and Theil's Regression.

Ohlson and Kim [9] conducted a work on Linear Valuation without OLS: The Theil-Sen Estimation Approach. According to them, OLS confronts two well-known problems in many archival accounting research settings. First, the presence of outliers tends to influence estimates excessively. Second, in the cross-sections, models often build in heteroscedasticity which suggests the need for scaling of all variables. Their study compared the relative efficacy of Theil [15] and Sen [13] (TS) estimation approach vs. OLS estimation in cross-sectional valuation settings. Next-year earnings or, alternatively, current market value determines the dependent variable. To assess the two methods' estimation performance the analysis relied on two criteria. The first focused on the inter-temporal stability of coefficient estimates. The second focused on the methods' goodness-of-fit, that is, the extent to which a particular model's projected values come close to actual values. On both criteria, results showed that TS performed much better than OLS. The dominance was most apparent when OLS estimates have the "wrong" sign. TS estimations, by contrast, never lead to such outcomes. Conclusions remained intact even when variables have been scaled for size.

Erilli and Alakus [5] conducted a study on non-parametric regression estimation for data with equal values. The study proposed a new method for the estimation of nonparametric regression parameters with sample data. The method proposed and other nonparametric methods such as Theil, Mood-Brown, Hodges-Lehmann methods and OLS method were compared with the sample data. In the data set which the independent variable had outliers, the OLS estimators gave incorrect values as expected. The proposed method produced more successful results like other nonparametric regression methods. In addition, the proposed methods' results were close to OLS results in the data set which were close to normal distribution and in the data set which the dependent variable had outliers. It showed that the proposed method can be among the alternative nonparametric regression family. They researchers concluded that since the analysis were made without searching if the data had the linear regression assumptions for the OLS method or not, the analysis results were in favor of OLS.

Ekezie and Opara [4] researched on Estimation of Bivariate Regression Data via Theil's algorithm. The method was adopted since all errors in the y-direction are not normally distributed (i.e. the do not follow a Gaussian distribution) for the commonly used least squares regression method for fitting an equation into a set of (x,y)-data points using the Kolmogorov Smirnov test. The algorithms for Theils were stated in the study. The data used for their research were collected from selected primary schools in Owerri Municipal, Imo State Nigeria. The data were on weights and shoulder heights of 100 randomly selected pupils in primary four, five and six. The use of a programming language software known as "R Development" was used to write an appropriate expression in the study. From the analysis, the result revealed that there exist a significant relationship between weights and shoulder heights of primary school pupils, and the estimated fitted Theil's is $\hat{y}_i = 42.5833 + 0.1177\ x_i$ and it was observed that both the intercept and slope were significant.

In a research study carried out by Fernandes and Leblanc [6] on Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for

predicting biophysical parameters in the presence of measurement errors, Parametric (Modified Least Squares) and non-parametric (Theil-Sen) consistent predictors were given for linear regression in the presence of measurement errors together with analytical approximations of their prediction confidence intervals. Three case studies involving estimation of leaf area index from nadir reflectance estimates were used to compare these unbiased estimators with OLS linear regression. A comparison to Geometric Mean regression, a standardized version of Reduced Major Axis regression, was also performed. The Theil–Sen approach was suggested as a potential replacement of OLS for linear regression in remote sensing applications. It offered simplicity in computation, analytical estimates of confidence intervals, robustness to outliers, testable assumptions regarding residuals and requires limited a priori information regarding measurement errors.

Having reviewed some of these past researches, we shall embark on Parametric Versus Non-Parametric Simple Linear Regression on Data With and without Outliers using real life data of CGPA and JAMB scores.

Regression analysis is a statistical technique that express mathematically the relationship between two or more quantitative variables such that one variable (the dependent variable) can be predicted from the other or others (independent variables). Regression analysis is very useful in predicting or forecasting by Inyama and Iheagwam [7]. It can also be used to examine the effects that some variables exert on others. However, regression analysis may be simple linear, multiple linear or non linear. In this study, simple linear regression is applicable.

III. Simple Linear Regression

This is a regression line that involves only two variables as it is applicable in this research study. A widely used procedure for obtaining the regression line of y on x is the Least Squares Method.

The linear regression line or y on x is

$$y = \alpha + \beta x + e \quad \ldots \quad (1)$$

where y is the response or dependent variable, x is the predictor or independent variable. α is the intercept, β is the slope, while e is the error term.

Using the least squares method, the parameters are estimated as shown in equations (2) and (3);

$$\hat{\beta} = \frac{n\Sigma x_i y_i - \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad \ldots \quad (2)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \ldots \quad (3)$$

The calculation is usually set out in Analysis of Variance (ANOVA) table as shown in Table 1

Table 1: Regression Table

Variance	Degree of freedom	Sum of square	Mean square
Regression	1	RSS = $\beta\Sigma xy$	RMS = $\frac{RSS}{1}$
Error	n – 2	ESS = TSS – RSS	EMS = $\frac{ESS}{n-2}$
Total	n – 1	TSS = Σy^2	

The test statistic is given by

$$F_{cal} = \frac{RMS}{EMS} \quad \ldots \quad (4)$$

The F_{cal} is now compared with the F-value obtained from the F-table or F-tabulated with 1 and (n – 2) degree of freedom.

IV. Theil's Regression Method

Theil's regression is a nonparametric method which is used as an alternative to robust methods for data sets with outliers. Although the nonparametric procedures perform reasonably well for almost any possible distribution of errors and they lead to robust regression lines, they require a lot of computation. This method is suggested by Theil [15], and it is proved to be useful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

Sprent [14] states that for a simple linear regression model to obtain the slope of a line that fits the data points, the set of all slopes of lines joining pairs of data points (x_i,y_i) and (x_j, y_j), $x_j \neq x_i$, for $1 \leq i < j \leq n$ should be calculated by;

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \quad \ldots (5)$$

Thus b* is the median of all Equation (5)

Hence, in this study, for n observations, we have $\frac{n(n-1)}{2}$ algebraic distinct $b_{ij} = b_{ji}$

But a^* is the median of all $a_i = y_i - b^* x_i$

The mean square error is given in equation (6)

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-k} \quad \ldots \quad (6)$$

V. Akaike Information Criterion (AIC)

The Akaike's information criterion AIC by Akaike [1] is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. Thus, the AIC is defined as;

$$AIC = e^{\frac{2k}{n}} \frac{\sum \hat{u}_i^2}{n} = e^{\frac{2k}{n}} \frac{RSS}{n} \quad \ldots (7)$$

where k is the number of regressors (including the intercept) and n is the number of observations. For mathematical convenience, Equation (7) is written as;

$$\ln(AIC) = \left(\frac{2k}{n}\right) + \ln\left(\frac{RSS}{n}\right) \quad \ldots (8)$$

where ln (AIC) = natural log of AIC and 2k/n = penalty factor.

VI. Bayesian Information Criterion (BIC)

Bayesian Information Criterion BIC by Schwarz [12] is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. It is defined as

$$BIC = n^{\frac{k}{n}} \frac{\sum \hat{u}_i^2}{n} = n^{\frac{k}{n}} \frac{RSS}{n} \quad \ldots (9)$$

Transforming Equation (3) in natural logarithm form, it becomes (See Equation (9));

$$\ln(BIC) = \frac{k}{n}\ln(n) + \ln\left(\frac{RSS}{n}\right) \quad \ldots (10)$$

where $\frac{k}{n}\ln(n)$ is the penalty factor. For model comparison, the model with the lowest AIC and BIC score is preferred.

VII. Data Analysis

Data used for this study were collected from the department of Mass Communication, Imo State University Owerri Imo State Nigeria. Twenty five (25) students were selected at random to determine the Cumulative Grade Point Average (CGPA) at the end of 2014/2015 Academic session (Y) and their respective Joint Admission Matriculation Board (JAMB) score (X). The data for the 25 selected students are shown in Table 2.

Table 2: CGPA (Y) and JAMB Score (X) of 25 Selected Students

i	Y	X	i	Y	X	i	Y	X
1	3.21	215	10	2.45	198	19	3.11	221
2	2.86	196	11	3.67	234	20	3.17	235
3	2.58	211	12	3.82	218	21	3.81	253
4	3.37	228	13	3.78	256	22	4.01	274
5	3.68	245	14	3.48	248	23	3.89	265
6	4.25	289	15	3.56	239	24	2.56	233
7	3.45	238	16	2.89	197	25	3.77	255
8	3.16	201	17	2.19	204			
9	2.85	241	18	3.28	219			

The data set was subjected to normality test using Anderson-Darling Technique via R Software package, and the output is shown below;

```
jude=lm(CGPA~JAMB)
>summary(jude)
>resid(jude)
>amara=resid(jude)
> ad.test(amara)
```

Anderson-Darling normality test

```
data:  amara
A = 1.0609, p-value = 0.007127
```

The result showed that the residuals are not from a normal distribution.

Having not rejected the null hypothesis which implies the absence of normality, we can say that there is presence of outliers in the data set. However, let us analyze the data with outliers for the parametric and non-parametric regression.

VIII. Output for Parametric Ordinary OLS

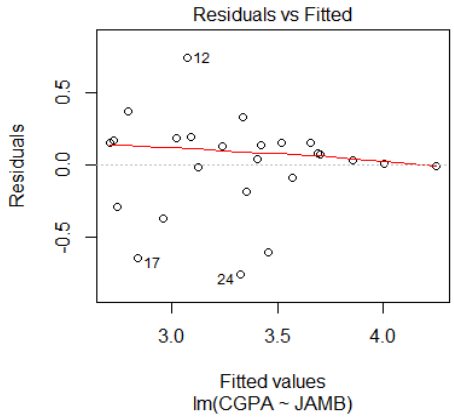
```
Residuals:
    Min      1Q  Median      3Q     Max
-0.76200 -0.09201  0.07465  0.15799  0.74801

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.56150 0.65220  -0.861  0.398
JAMB    0.01667   0.00279   5.975 4.31e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  Residual standard error: 0.3405 on 23 degrees of freedom
Multiple R-squared:  0.6082,    Adjusted R-squared:  0.5911
F-statistic:  35.7 on 1 and 23 DF, p-value: 4.309e-06
> AIC(jude)
[1] 20.99785
> BIC(jude)
[1] 24.65447
```

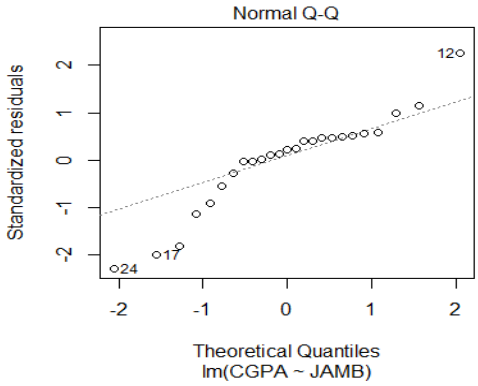
IX. Output for Non-Parametric Theil's Regression

```
> jude = mblm(CGPA~JAMB)
> summary(jude)

mblm(formula = CGPA ~ JAMB)

Residuals:
    Min      1Q  Median  3Q     Max
-0.86761 -0.17162 -0.01390  0.05119  0.61639

Coefficients:
   Estimate      MAD      V value  Pr(>|V|)
(Intercept) -0.051934  0.481272 126  0.339
JAMB       0.014934 0.002033  324   1.19e-07 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3609 on 23 degrees of freedom

> AIC(jude)
[1] 23.89809
> BIC(jude)
[1] 27.55472
```

Having carried out the analysis with outliers, we can conclude that from the result that student's CGPA can be predicted at the end of any academic session from the JAMB score. Thus, there is significance relationship between JAMB score and student's CGPA. Again, it can be concluded that the parametric OLS regression performs better than its

non-parametric Theil's regression since their residual standard error, AIC and BIC values are all smaller. Let us now detect outliers in the data set, and expunge them to enable us re-analyze the data. Using the R Software package, the output is shown below;





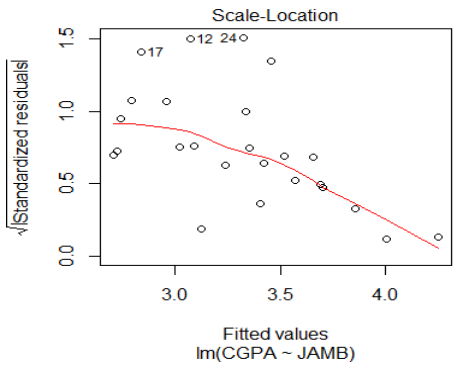
It can be observed that observations 6, 9, 12, 17, 22 and 24 are possibly problematic to our model. We shall now delete these observations and then re-analyze the data.

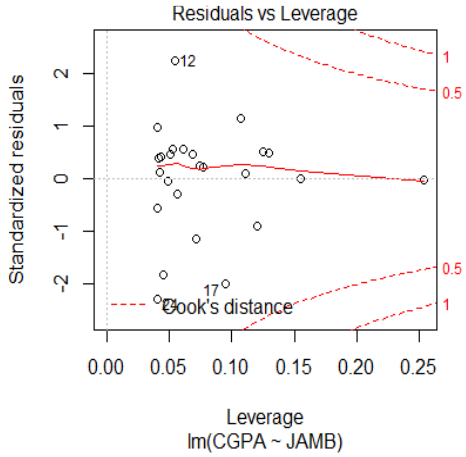
X. Anderson-Darling Normality Test

> amara = resid(jude)
> ad.test (amara)

data: amara
A = 0.62716, p-value = 0.08695

The result showed that the residuals are from a normal distribution.





XI. Output for Parametric Ordinary OLS

Residuals:
Min 1Q Median 3Q Max
-0.43741 -0.05902 0.06486 0.10300 0.31184

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.553806 0.484645 -1.143 0.269
JAMB 0.016925 0.002106 8.038 3.42e-07 ***

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1962 on 17 degrees of freedom
Multiple R-squared: 0.7917, Adjusted R-squared: 0.7794
F-statistic: 64.6 on 1 and 17 DF, p-value: 3.424e-07
> AIC(jude)
[1] -4.087108
> BIC(jude)
[1] -1.253791

XII. Output for Non-Parametric Theil's Regression

Residuals:
Min 1Q Median 3Q Max
-0.52459 -0.10816 -0.00903 0.04844 0.21286

Coefficients:
Estimate MAD V value Pr(>|V|)
(Intercept) -0.129644 0.527037 66 0.258
JAMB 0.015328 0.002226 190 3.81e-06 ***

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2087 on 17 degrees of freedom

> AIC(jude)
[1] -1.728739
> BIC(jude)
[1] 1.104578

Having carried out the analysis without outliers, we can conclude that from the result that student's CGPA can be predicted at the end of any academic session from the JAMB score. Thus, there is significance relationship between JAMB score and student's CGPA. Again, it can be concluded that the parametric OLS regression performs better than its non-parametric Theil's regression since their residual standard error, AIC and BIC values are all smaller.

XIII. CONCLUSION

From the analysis, the result revealed that there is a significant relationship between students CGPA and their JAMB scores for both the parametric OLS regression and non-parametric Theil's regression with and without outliers. It was concluded that the parametric OLS is better than its non-parametric Theil's regression for both data with and without outliers since their standard error, AIC and BIC are both lower than that of Theil's regression. It was also concluded that the standard error for the parametric regression with outliers which is 0.3405 reduced to 0.1962 for the parametric regression without outliers. On the other hand, the standard error for the non-parametric regression with outliers which is 0.3609 reduced to 0.2087 for the non-parametric regression without outliers. This implies that the model for data without outliers is more efficient than the model for data with outliers for both the parametric and non-parametric regression. Therefore the researchers recommend that future researchers should look at a similar work with large sample size to examine the differences between the parametric and nonparametric Regression.

REFERENCES

[1] Akaike, H. *(1974),* "A new look at the statistical model identification" *(PDF), IEEE Transactions on Automatic Control 19 (6): 716–723,* doi:10.1109/TAC.1974.1100705, MR 042371

[2] Birkes, D., and Dodge, Y.(1993). Alternative Methods of Regression. New York, NY: Wiley.

[3] Dietz, E. J. (1989). Teaching Regression in a Nonparametric Statistics Course. The American Statistician. 43, 35-40.

[4] Ekezie, D. D., and Opara, J. (2015). Estimation of Bivariate Regression Data Via Theil's Algorithm. Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 5(8): 29-34© Scholarlink Research Institute Journals, 2014 (ISSN: 2141-7016).

[5] Erilli, N.A. and Alakus, K.A. (2014). Non-parametric regression estimation for data with equal values. European Scientific Journal February 2014 edition vol.10, No.4 ISSN: 1857–7881 (Print) e - ISSN 1857- 7431.

[6] Fernandes, R. and Leblanc, S.R. (2005). Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. Remote Sensing of Environment 95 (2005) 303–316

[7] Inyama, S.C. and Iheagwam, V.A. (2006): Statistics and Probability. A Focuse on Hypotheses Testing. Third edition. Strokes Global Ventures Owerri, Imo State, Nigeria.

[8] Mutan, O.M. (2004). Comparison of Regression techniques via monte carlo simulation. A thesis submitted to the school of natural and applied sciences of middle east technical University.

[9] Ohlson, J.A., and Kim, S. (2014). Linear valuation without OLS: The Theil-Sen Estimation Approach. Electronic copy available at: http://ssrn.com/abstract=2276927.

[10] Opara, J., Iheagwara, A.I., and Okenwe, I. (2016). Comparison of parametric and non-parametric linear regression. Advance Research Journal of Multi-Disciplinary Discoveries. Vol.2.0/Issue-I

[11] Ranjit, K.P. (2005). Some Methods of Detection of Outliers in Linear Regression Model. Ebook_2005_2006_MSc._trim1_4. Unpublished

[12] Schwarz, G. (1978). *Estimating the dimension of a model. The Annals of Statistics 6, 461-464*

[13] Sen, P.K., (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association 63 (324): 1379-1389.

[14] Sprent, P. (1993). Applied Nonparametric Statistical Methods. London; New York: Chapman and Hall.

[15] Theil, H., (1950). A rank-invariant method of linear and polynomial regression analysis. Nederlandse Akademie Weten chappen Series A 53: 386-392.

[16] Wilcox, R. (1998). Simulations on the Theil-Sen regression estimator with right-censored data. Stat. & Prob. Letters 39, 43-47.