International Journal of Statistics and Applied Mathematics 2021; 6(4): 131-136

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2021; 6(4): 131-136

www.mathsjournal.com
Received: 03-05-2021
Accepted: 10-06-2021

Opara Jude
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B. 5047, Port Harcourt, Rivers, Nigeria

George Isobeye
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B 5047, Port Harcourt, Rivers, Nigeria

Corresponding Author:
Opara Jude
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B. 5047, Port Harcourt, Rivers, Nigeria

Effect of non-normal error distribution on simple Linear/non-parametric regression models

Opara Jude and George Isobeye

Abstract
This study is on the effect of non-normal error distribution on simple linear regression versus its nonparametric equivalent. The error term for normality proved that it is not from a normal population using Ryan-Joiner, which violates the major assumption of simple linear regression. Hence, estimating its slope becomes immaterial and any inference drawn from the OLS won't be reliable. Since, there is no need of employing the technique, due to its poor performance in the presence of error non-normality, then a feasible alternative technique which performs consistently and robust to non-normality residual is required. The simulation study conducted in this study suggested that the nonparametric Theil's simple linear regression is an alternative to OLS when there is existence of non-normal error in a data set. The study recommended among others that further studies on simple linear regression should ensure that the underlying assumptions of OLS are fulfilled before estimation; otherwise its non-parametric equivalent should be employed, but if the researcher must continue with OLS after failure of assumption, then outliers should be checked and if detected, should be removed and re-examine the underlying assumptions.

Keywords: Non-normality, error distribution, simple linear regression, non-parametric simple regression, bias

Introduction
Ordinary Least Squares (OLS) is one of the most known techniques of measuring relationship between variables, due to its simplicity in application. However, the implementation of OLS estimators requires some conditions to be met, one of which is that the residual is assumed to be independently, identically distributed random variables with zero mean and a fixed variance (σ^2).

The traditional model such as a simple linear regression equation represents an association between the dependent and predictor variables. The primary interest lies on the parameter estimation of the model, in which the OLS technique is being employed. However, when the error term fails the normality assumption to the data set, then obtaining a valid estimate from this parametric technique won't be feasible (Ekezie & Opara, 2014). A very effective method being a nonparametric method becomes an alternative. Those statistical techniques that do not make assumptions about the population distribution are known as Non-parametric or distribution-free.

If the distribution of errors isn't normal and probably from a population whose mean is zero, then the estimates from OLS will be far from being optimal, but at least possesses the unbiased property. Furthermore, if the variance of the residual population is assumed to be finite, then the property of OLS estimates have consistent and asymptotically normal distribution. These conditions however, may affect the efficiency and performance of estimates and test of the OLS technique (Mutan, 2004) [4]. To remedy these situations, two possible solutions can be employed. According to Birkes & Dodge (1993) [1], the first is to ensure that the non-normal error term becomes normal using any possible means, while the second is to employ its non-parametric equivalent, which probably do recognize the normality assumption.

For bivariate linear model, the median of pair-wise slopes was proposed by Theil (1950) [10] as an estimator of slope parameter, even though it was extended by Sen (1968) [8] to tackle ties.

In a simple linear regression, the slope of the OLS estimators is sensitive to outliers and respective confidence interval is thereby disturbed by the response variable being non-normal. This study is aimed at investigating the non-parametric Theil's regression technique for error non-normal condition.

Statement of Problem

Once there is an established relationship between two variables, what comes into so many researchers' mind is the simple linear regression. However, there is nothing wrong in using a simple linear regression to relate variables say, a dependent variable and an independent variable, but it becomes a serious error when the underlying assumptions are not thoroughly examined before employing it, or perhaps violating the assumptions after testing. Because it is a parametric test and any attempt to employ the simple linear regression model, after assumption failures, will adversely affect the interpretation of the slope. It is as result of these researchers who are not too strong statistically that led to this present study. The study proved the inconsistency of the OLS estimator when the error term is not normally distributed and also displayed an appropriate statistical tool adequate to tackle the problem.

Literature Review

Okenwe *et al.* (2016) [6] in their study on parametric against its non-parametric equivalent sourced for data in the department of mass communication, Imo State University with 25 randomly selected students to ascertain if cumulative grade point at the end of a particular session has any relationship with their Joint Admission and Matriculation Board score. The normality assumption for the residual was not met in the study using Anderson-Darling technique. Brief algorithms for both the parametric and nonparametric regression were outlined. The study went further to detect outliers in the data and thereafter it was expunged from the data set and re-analyzed. The result of their study revealed there was a relationship between the two variables used for both the OLS and its non-parametric equivalent, with both for outliers and non-outliers. Their study went further to conclude that the parametric OLS outperforms its nonparametric equivalent for data with outliers and does without outlier since the three goodness of fit measures were lower that of its parametric equation. The study recommended that further research should be conducted on large sample size with a similar work to subsequent examine the differences.

In the work of Opara *et al.* (2016) [7] whose work was comparison of parametric and non-parametric linear equation, the data were subjected to normality test, and it was deduced that the error term is normality. The data set employed in the study was collected from traders in Douglas market Owerri who were selling pears. Brief algorithms for both the parametric and nonparametric regression were outlined. The result of their study revealed there was a relationship between the two variables used for both the OLS and its non-parametric equivalent. Their study went further to conclude that the parametric OLS outperforms its nonparametric equivalent for data, since their goodness of fit measures was lower that of its parametric equation. The study recommended that further research should be conducted on large sample size with a similar work to subsequent examine the differences.

Ekezie & Opara (2016) [7] in their study titled "estimation of bivariate regression data using Theil's algorithm" adopted the technique of Kolmogorov Smirnov test to examine the normality test and concluded the error term was not normally distributed. The steps for nonparametric regression were stated in the study. Data set was collected for the study. The use of R package was employed to write codes. The result of their study revealed that there was a significant association between shoulder heights and weights pupils in the primary school, and the estimated fitted Theil's was $\hat{y}_i = 42.5833 + 0.1177\ z_i$ and both the slope and intercept were significant.

Having reviewed these few works, it becomes necessary to embark on this study to examine the effect of non-normal residual on OLS and its parametric equivalent and to simulate data set of different sizes to probably know the behavior of the slope for both techniques.

Methodology

Simple Linear Regression

It is mathematically defined (Inyama & Iheagwam, 2006) as stated in (1)

$$y_i = \theta + \lambda z_i + e_i \qquad \ldots(1)$$

If there are m pairs of sample observations $(z_1, y_1), (z_2, y_2), \cdots, (z_w, y_w),$ then we get

$$y_i = \theta + \lambda z_i + e_i, i = 1,2,\cdots, w \qquad \ldots(2)$$

Then seeking for the estimators $\hat{\theta}$ and $\hat{\lambda}$ of θ and λ respectively in such a way that V is minimized.

Let $V = \sum_{i=1}^{w} e_i^2 = \sum_{i=1}^{w} (y_i - \theta - \lambda z_i)^2 \qquad \ldots(3)$

(3) is differentiated partially w.r.t θ & λ , we get Equations (4) and (5) respectively

$$\sum_{i=1}^{w} y_i - w\theta - \lambda \sum_{i=1}^{w} z_i = 0 \quad \ldots(4)$$

$$\sum_{i=1}^{w} z_i y_i - \theta \sum_{i=1}^{w} z_i - \lambda \sum_{i=1}^{w} z_i^2 = 0 \quad \ldots(5)$$

Evaluating Equations (4) and (5) simultaneously, we get

$$\hat{\lambda} = \frac{w\Sigma z_i y_i - \Sigma z_i y_i}{w\Sigma z_i^2 - (\Sigma z_i)^2} \quad \ldots(6)$$

$$\hat{\theta} = \bar{y} - \hat{\theta}_1 \bar{z} \quad \ldots(7)$$

Alternatively, Equation (7) can be stated as shown in Equation (8)

$$\hat{\theta} = \frac{\Sigma z_i^2 \Sigma y_i - \Sigma z \Sigma zy}{w\Sigma z_i^2 - (\Sigma z)^2} \quad \ldots (8)$$

The fitted regression model is:

$$\hat{y}_i = \hat{\theta} + \hat{\lambda} z_i \quad \ldots (9)$$

Table 1: Regression ANOVA Table

Variance	Degree of freedom	Sum of square	Mean square
Regression	1	$RSS = \lambda \sum zy$	$\text{RMS} = \frac{\text{RSS}}{1}$
Error	w – 2	ESS = TSS – RSS	$EMS = \frac{ESS}{w-2}$
Total	w – 1	TSS = $\sum y^2$	

Theil's Regression Method

Non-parametric Theil's regression has proven to be efficient and consistent, especially when the residual is not from a normal distribution. However, most times, the presence of non-normal error is as a result of presence of influential observations in a data set (Theil, 1950) [10].

According to Sprent & Smeeton (2001) [9], a linear regression in simple state is to obtain the gradient of a line that adequately suits the points in the data, the set of all slopes of lines joining pairs of data points (z_i, y_i) and (z_j, y_j), $z_j \neq z_i$, for $1 \le i < j \le w$ is to be computed as;

$$\lambda_{ij} = \frac{y_j - y_i}{z_j - z_i} \quad \ldots (10)$$

Thus λ^* is the median of all Equation (10)

Hence, this study has w observations of $\frac{w(w-1)}{2}$ algebraic distinct $\lambda_{ij} = \lambda_{ji}$

But θ^* is the median of all $\theta_i = y_i - \lambda^* z_i$

The mean square error is given as

$$MSE = \frac{\sum_{i=1}^{w} (y_i - \hat{y})^2}{w - k} \quad \ldots (11)$$

Table 2: Data on Systolic Blood Pressure (y_i) and age(x_i) of 60 patients randomly selected Federal Medical Centre, Owerri Imo State Nigeria

i	z_i	y_i	i	z_i	y_i	i	z_i	y_i
1	52	157	21	46	131	41	47	129
2	62	143	22	43	136	42	44	134
3	28	129	23	15	115	43	16	113
4	24	124	24	18	117	44	19	115
5	68	174	25	17	125	45	18	123
6	37	145	26	38	143	46	35	135
7	45	221	27	46	219	47	49	141
8	43	139	28	44	137	48	38	119
9	45	146	29	46	144	49	20	119
10	63	163	30	64	161	50	43	159
11	44	143	31	45	141	51	44	167
12	65	171	32	66	169	52	39	123
13	40	125	33	41	123	53	23	121
14	65	159	34	66	157	54	41	129
15	54	155	35	55	153	55	46	142
16	62	163	36	63	161	56	67	170
17	54	151	37	55	149	57	42	124
18	57	141	38	58	139	58	67	158
19	32	111	39	33	109	59	56	154
20	40	129	40	41	127	60	59	140

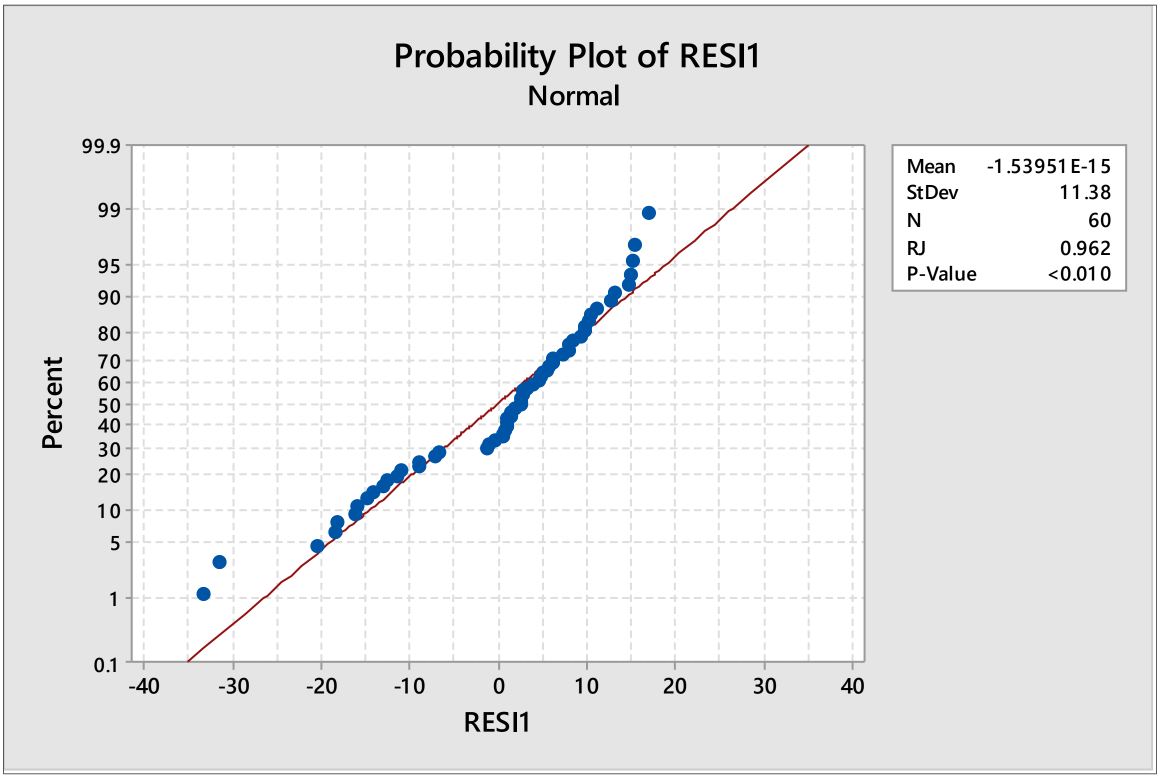

Fig 1: Test for Normality Assumption for the Residual

Using the MINITAB Software via Ryan-Joiner technique, the output displayed in Figure 1 shows that the p-value is less than 0.01, which implies that the data for normality assumption for the residual is not satisfied.

Table 3: R output for OLS Technique

Age<-c(Age)				
> SBP<-c(SPB)				
> jude<-lm(Age~SBP)				
> summary(jude)				
lm(formula = Age ~ SBP)				
Resids:				
Min.	1Q	Median	3Q	Max.
-33.372	-7.486	2.517	7.970	16.974
Coeffs:	Esti Std.	Error	t value	Pr.(>|t|)
(Inter)	-16.1087	9.6304	-1.673	0.0998
SPB	0.4275	0.0667	6.410	2.85e-08 ***

Output 3 is the parametric regression model from R-Studio software package. The systolic blood pressure is significant, but the slope is insignificant and the estimated regression equation is given as;

$$\hat{y}_i = -16.1087 + 0.4275\ z_i$$

Table 4: R Output for Non-Parametric Regression Technique

Library(Mblm)				
> Jude1<-Mblm(Age~SBP)				
> Summary(jude1)				
Mblm(formula = Age ~ SBP)				
Resids:				
Min.	1Q	Median	3Q	Max.
-53.534	-10.532	0.384	3.687	13.678
Coeffs:	Estimate	MAD V	value	Pr(>|V|)
(Int)	-43.7326	29.2270	95	1.61e-09 ***
SPB	0.6437	0.1740	1830	1.66e-11 ***

Output 4 is the nonparametric regression model from R-Studio software package. The systolic blood pressure is significant with the intercept as well, and the estimated regression equation is given as;

$$\hat{y}_i = -43.733 + 0.644\ z_i$$

Table 5: Bias for λ with Three Sample Sizes and Various Number of Simulations with Non-normal Residual and the Parametric Value of $\lambda = 0.428$ and Non-parametric Value of $\lambda = 0.644$

Simulation	w	Techniques	Bias
150	10	OLS	-0.0008
		Theil	0.0000
	20	OLS	-0.0040
		Theil	0.0000
	30	OLS	-0.0024
		Theil	0.0000
300	10	OLS	-0.0022
		Theil	0.0000
	20	OLS	-0.0052
		Theil	0.0000
	30	OLS	0.0002
		Theil	0.0000
500	10	OLS	0.0008
		Theil	0.0000
	20	OLS	-0.0011
		Theil	0.0000
	30	OLS	0.0011
		Theil	0.0000
1000	10	OLS	0.0006
		Theil	0.0000
	20	OLS	-0.0007
		Theil	0.0000
	30	OLS	-0.0020
		Theil	0.0000

From the results obtained in Table 5, it can be seen that the estimator (slope) for the non-parametric Theil-Sen regression is far more consistent in the presence of non-normal residual from the values of the bias. Hence, we can put it that the Theil's regression estimator is robust to non-normality residual in the data.

Conclusion

This study examined the effect of non-normal residual on simple linear regression versus its non-parametric equivalent. The error term for normality proved that it is not from a not population, which violates the major assumption of simple linear regression. Hence, estimating its slope becomes immaterial and any inference drawn from the OLS will be misleading. Since, there is no need of employing the technique, due to its poor performance in the presence of error non-normality, then a feasible alternative technique which performs consistently and robust to non-normality residual is required. The simulation study conducted in this study suggested that the nonparametric Theil's simple linear regression is an alternative to OLS when there is existence of non-normal error in a data set.

Having concluded the study, it is recommended that further studies on simple linear regression should ensure that the underlying assumptions of OLS are fulfilled before estimation; otherwise its non-parametric equivalent should be employed. However, if the researcher must continue with OLS after failure of assumption, then outliers should be checked and if detected, should be removed and re-examine the underlying assumptions. Again, further studies should look at a situation where the explanatory variable is more than one.

References

1. Birkes D, Dodge Y. Alternative Methods of Regression. New York, NY: Wiley 1993.
2. Ekezie DD, Opara J. Estimation of Bivariate Regression Data via Theil-Sen Algorithm. Journal of link Emerging Trends in Engineering and Applied Sciences (JETEAS) 2015;5(8):29-34.
3. Inyama SC, Iheagwam VA. Statistics and Probability: A Focus on Hypotheses Testing. Perfect Strokes Global Ventures, Imo State, Nigeria 2006.
4. Mutan OM. Comparison of Regression Techniques via Monte Carlo Simulation. A thesis submitted to the school of natural and applied sciences of middle-east technical University 2004.
5. Ohlson JA, Kim S. Linear valuation without OLS: The Theil-Sen Estimation Approach. Electronic copy 2014. available at: http://ssrn.com/abstract=2276927.
6. Okenwe I, Opara J, Ononogbu AC, Uwabunkonye B. Parametric Versus Non-Parametric Simple Linear Regression on Data with and Without Outliers. International Journal of Innovation in Science and Mathematics 2016;4(5):175-180.
7. Opara J, Iheagwara AI, Okenwe I. Comparison of parametric and non-parametric linear regression. Advance Research Journal of Multi-Disciplinary Discoveries 2016;2(1):24-29.
8. Sen PK. Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 1968;63(324):1379-1389.
9. Sprent P, Smeeton NC. Applied nonparametric statistical methods. Chapman & Hall/CRC, USA 2001.
10. Theil H. A rank-invariant method of linear and polynomial regression analysis. Nederlandse Akademie Wetenchappen Series 1950A;53:386-392.