Volume 4, Issue 3, 2020 ISSN: 2617-4537
DOI: https://doi.org/10.31058/j.data.2020.43002

Estimation of Parameters of Parametric/Non-Parametric Simple Linear Regression Via Simulated Data

Esemokumo Perewarebo Akpos[1*], Bekesuoyeibo Rebecca[1], Opara Jude[2]

[1] Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe, Yenagoa, Bayelsa State, Nigeria
[2] Department of Mathematics and Statistics, School of Science, Sure Foundation Polytechnic Ikot Akai, Ukanafun, Akwa Ibom State, Nigeria

Email Address

contactperes4good@gmail.com (Esemokumo Perewarebo Akpos), fzimughan@gmail.com (Bekesuoyeibo Rebecca), judend88@yahoo.com (Opara Jude)
*Correspondence: judend88@yahoo.com

Received: 10 January 2020**; Accepted:** 22 March 2020**; Published:** 9 May 2020

Abstract:

This study is on the estimation of parameters of Parametric/non-parametric simple linear regression via simulated data. Data of different sample sizes for both cases of residuals be normally distributed and non-normally distributed were simulated via"R Development" was used in this study. The different data sets were tested for normality using Anderson-Darling technique. The algorithms for the parametric Theil's and that of its non-parametric OLS regression were stated. It was concluded that the parametric OLS regression was better than its non-parametric Theil's regression for both data whose residuals are normal and non-normal since their AIC and BIC are lower than that of Theil's regression. Therefore the researchers recommended that future researchers should look at a similar work by comparing with other non parametric regression models, to know if the parametric OLS regression will still outperforms its non-parametric equivalents.

Keywords:

Non-Parametric Theil's Regression, Parametric Regression, Akaike Information Criterion, Bayesian Information Criterion, Simulation

1. Introduction

The simple linear regression model is the ordinary or traditional equation representing the relationship between two variables; the response and the explanatory variables [1]. 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. 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 [2]. 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 [3].

In a simple linear model, [4] proposed the median of pairwise slopes as an estimator of the slope parameter. [5] 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 ([5,6,7].

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 [4]. 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.

2. Review of Related Literature

[8] 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.

[9] carried out a work on Non-parametric regression with a circular response variable and a unidimensional linear regressor, which was a study examined in the literature. In the work, they extended the results to the case of multivariate linear explanatory variables. Nonparametric procedures to estimate the circular regression function were formulated. A simulation study was carried out to study the sample performance of the proposed estimators justifying the correct performance of the proposed estimators.

[10] 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 [4] and [5] (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.

[1] carried out a research on parametric versus non-parametric simple linear regression on data with and without outliers. Data used for the 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 the 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. The researchers recommended that future researchers should look into a similar work with large sample size to examine the differences between the parametric and nonparametric Regression.

[11] researched on empirical performance of nonparametric regression over LRM and IGRM addressing influential observations. In their study, they described nonparametric regression as commonly used for summarizing the relationship between variables without requiring the assumptions of model. Generalized linear model and linear regression model are usually used to examine the relationship of variables, but both are badly affected by influential observations. Due to this, detection and removal of outliers attain a lot of attention of researchers to obtain reliable estimates. They focused on such robust technique whose performance was acceptable in the presence of outliers. The study empirically compared the performance of linear regression model and generalized linear model with multivariate nonparametric kernel regression. Multivariate nonparametric kernel regression was used with Gaussian kernel and six different bandwidths on Aerial biomass data. The performance of nonparametric regression with Bayesian bandwidth was found to be better as compared with other methods.

The study shall examine the estimation of parametric/non-parametric simple linear regression via simulated data, having reviewed past works.

3. Methodology

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 [12]. 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.

3.1. Parametric 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

$$y = \alpha + \beta x + e \tag{1}$$

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 technique, the parameters of the linear regression model in equation (1) are estimated as shown in equations (2) and (3);

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{3}$$

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

Table 1. *ANOVA Table for Regression.*

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

The test statistic is given by

$$F_{cal} = \frac{RMS}{EMS} \tag{4}$$

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.

3.2. Non-parametric 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 [4] and it is proved to be useful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

[13] 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;

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \tag{5}$$

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)

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

3.4. Akaike Information Criterion (AIC)

The Akaike's information criterion AIC [14] 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;

$$AIC = e^{\frac{2k}{n}} \frac{\sum \hat{u}_i^2}{n} = e^{\frac{2k}{n}} \frac{RSS}{n} \tag{7}$$

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;

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

where ln (AIC) = natural log of AIC and $\frac{2k}{n}$ = penalty factor.

3.5. Bayesian Information Criterion (BIC)

Bayesian Information Criterion BIC [15] 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

$$BIC = n^{\frac{k}{n}} \frac{\sum \hat{u}_i^2}{n} = n^{\frac{k}{n}} \frac{RSS}{n} \tag{9}$$

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

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

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

3.6. Result

According to Table 2, it is shown that performance criteria values obtained by OLS regression are smaller than results of the Theil's regression. The residuals in the y-direction for all the different sample sizes are normally distributed via the Anderson-Darling technique. Hence; it can be said that OLS regression is better than Theil's regression for prediction of these models.

According to Table 3, it is shown that performance criteria values obtained by OLS regression are also smaller than results of the Theil's regression. The residuals in the y-direction for all the different sample sizes are not normally distributed via the Anderson-Darling technique. Hence; it can be said that OLS regression is better than Theil's regression for prediction of these models.

Table 2. *Performance Values of the Models for Simulated Data ($\varepsilon \sim N(0, 4)$, assume $x \sim N(0, 1)$, $a = 0.5$ and $\beta = 2$, set.seed (20) for R package).*

Techniques/Performance Criteria				
Sample Sizes	OLS Regression		Theil's Regression	
	AIC	BIC	AIC	BIC
20	92.70131	95.68851	93.20606	96.19326
30	127.3543	131.5579	132.7773	136.9809
50	205.7195	211.4556	206.6025	212.3386
200	854.3652	864.2602	854.652	864.547
500	2156.489	2169.132	2157.175	2169.819
700	2941.332	2954.985	2942.33	2955.983
1000	4171.902	4186.626	4172.359	4187.082
1500	6275.486	6291.426	6278.688	6294.627
3000	12632.5	12650.52	12632.73	12650.75

Table 3. *Performance Values of the Models for Simulated Data ($_{\varepsilon}$ ~ Bin (n, 1, 0.5), assume x ~N (0, 1), a = 0.5 and β = 2, set.seed (126) for R package).*

Techniques/Performance Criteria				
Sample Sizes	OLS Regression		Theil's Regression	
	AIC	BIC	AIC	BIC
20	34.13123	37.11842	44.43173	47.41892
30	47.05599	51.25958	61.03725	65.24084
50	75.21257	80.94864	102.0793	107.8154
200	295.5479	305.4428	422.5709	432.4658
500	728.3599	741.0037	1038.843	1051.487
700	1020.342	1033.995	1476.649	1490.302
1000	1457.185	1471.909	2146.722	2161.445
1500	2180.189	2196.129	3153.505	3169.445
3000	4360.267	4378.286	6401.947	6419.966

4. Conclusion

From the result of the analysis of this study, it can be concluded that the parametric OLS regression is better for estimating the parameters than its non-parametric Theil's regression equivalent for both data whose residuals are from normal distribution and non-normal distribution. Therefore the researchers recommend that future researchers should look at a similar work by comparing with other non parametric regression models, to know if the parametric OLS regression will still outperforms its non-parametric equivalents.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial; or not for public sectors.

References

[1] Okenwe, I.; Opara, J.; Ononogbu, A.C.; 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), ISSN (Online): 2347–9051.

[2] Mutan, O.M. 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.

[3] Birkes, D.; Dodge, Y. Alternative Methods of Regression. New York, NY: Wiley, 1993.

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

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

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

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

[8] Opara, J.; Iheagwara, A.I.; Okenwe, I. Comparison of parametric and non-parametric linear regression. *Advance Research Journal of Multi-Disciplinary Discoveries,* 2016, 2(1).

[9] Meilán-Vila, A.; Francisco-Fernández, M.; Crujeiras, R.M.; Panzera, A. Nonparametric Regression Estimation for Circular Data. *Proceedings,* 2019, 21, 27.

[10] Ohlson, J.A.; Kim, S. Linear valuation without OLS: The Theil-Sen Estimation Approach, 2014. Available online: http://ssrn.com/abstract=2276927 (accessed on 31 January 2020)

[11] Khan, J.A.; Akbar, A. Empirical performance of nonparametric regression over LRM and IGRM addressing influential observations. *Journal of chemometrics,* 2019, 33(7), DOI: https://doi.org/10.1002/cem.3143.

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

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

[14] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, 19(6), 716-723, DOI: 10.1109/TAC.1974.1100705.

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