

Comparison of Parametric (OLS) and Non-Parametric (THEIL'S) Linear Regression

Part -3 (Business, Economics & Statistics)

Chapter-I

June/Vol.2.0/Issue-I

Abstract: This paper is 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 this paper as well as its non-parametric counterpart. Data used for this 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 this paper. From the 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 is 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 is 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.

keywords: : Theil's Regression, OLS, Pairwise slope, Akaike Information Criterion, Bayesian Information Criterion.

ISSN CODE : 2456-1045 (Online)

(ICV-BES/Impact Value): 2.18

(GIF) Impact Factor : 1.121

Copyright@IJF 2016

Journal Code : ARJMD/BES/V-2.0/I-1/June-2016

website : www.journalresearchijf.com

Received : 26.06.2016

Accepted : 28.06.2016

Date of Publication: 30.06.2016

Page: 24-29



Name of the Authors:

OPARA, JUDE

Department of Statistics, Imo State University
PMB 2000, Owerri Nigeria

IHEAGWARA ANDREW I.

Procurement Officer/Director Planning, Research & Statistics, Nigeria Erosion & Watershed Management Project (World Bank-Assisted), Ministry of Petroleum & Environment, Plot 36, chief Executive Quarters, Area "B", New Owerri, Imo State Nigeria

OKENWE IDOCHI

Department of Statistics, School of Applied Sciences, Rivers State Polytechnic
PMB 20, Bori, Rivers State Nigeria

INTRODUCTION

The simple linear regression model is the ordinary or traditional equation representing the relationship between two variables; the response and the explanatory variables. The question becomes, what is of interest in the estimation of the parameters of the model? The most popular method is the Least Square Method (LSM). It is obvious that when the data deviate to satisfy the assumptions such as the normality, then this parametric method of estimation of the parameters fails to give a valid estimate. In the alternative, a non-parametric equivalent comes in as being effective. **Non-parametric** (or **distribution-free**) statistical methods are those, which make no assumptions about the population distribution from which the data are taken.

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 (Mutan; 2004). 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 (Birkes and Dodge; 1993).

In a simple linear model, Theil (1950) proposed the median of pairwise slopes as an estimator of the slope parameter. Sen (1968) 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 (1968), Dietz (1989)) and Wilcox (1998).

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.

Review of related Literatures.

It is of paramount importance to review some past works done by researchers which in one way or the other relates to this study, as it will add taste to it. Here are some recent works done by past researchers.

Ohlson and Kim (2014) carried a research 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 (1950) and Sen (1968) (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 (2014) worked 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 favour of OLS.

Ekezie and Opara (2014) worked 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. they 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.

Fernandes, and Leblanc (2005) worked on Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. Linear regression, where there are no nonlinear combinations of regressors, is a common approach to this prediction problem in the remote sensing community. 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 the comparison of parametric (OLS) and non-parametric (Theil's) linear regression.

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 (Inyama and Iheagwam, 2006). 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.

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 of y on x is

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

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);

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

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

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 = $\beta \sum xy$	$RMS = \frac{RSS}{1}$
Error	n - 2	ESS = TSS - RSS	$EMS = \frac{ESS}{n - 2}$
Total	n - 1	TSS = $\sum y^2$	

The test statistic is given by

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

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.

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 (1950), and it is proved to be useful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

Sprent (1993) 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;

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

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)

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

Akaike Information Criterion (AIC)

The Akaike’s information criterion AIC (Akaike, 1974) 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;

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

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;

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

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

Bayesian Information Criterion (BIC)

Bayesian Information Criterion BIC (Schwarz, 1978) 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

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

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

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

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

Data Analysis

Data used for this study were collected from a trader in Dauglas Owerri Market in Imo State Nigeria who sales pear. 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).

x	2	4	6	7	9	10	13	15	17	18	20	21	23	24	25	26	27	29	30	31
y	3	5	8	12	15	17	27	24	38	36	42	49	67	63	73	74	95	98	101	107

The R Software package was used to run the data for this study, and the output is shown below.

Output for Ordinary OLS

Residuals:

	Min	1Q	Median	3Q	Max
	-13.5870	-6.7515	-0.5819	8.2626	13.7343

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.7807	4.5435	-3.913	0.00102 **
x	3.6684	0.2276	16.118	3.85e-12 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.099 on 18 degrees of freedom
 Multiple R-squared: 0.9352, Adjusted R-squared: 0.9316
 F-statistic: 259.8 on 1 and 18 DF, p-value: 3.847e-12

AIC(jude)
 [1] 148.9791
 > BIC(jude)
 [1] 151.9662

Output for Theil's Regression

Residuals:

	Min	1Q	Median	3Q	Max
	-8.00	-0.75	6.00	13.25	25.00

Coefficients:

	Estimate	MAD	V value	Pr(> V)
(Intercept)	-30.000	19.559	0	9.54e-05 ***
x	4.000	1.263	210	9.45e-05 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.69 on 18 degrees of freedom

> AIC(jude)
 [1] 159.0012
 > BIC(jude)
 [1] 161.9884

The normality test using the Minitab software is shown in Fig.1

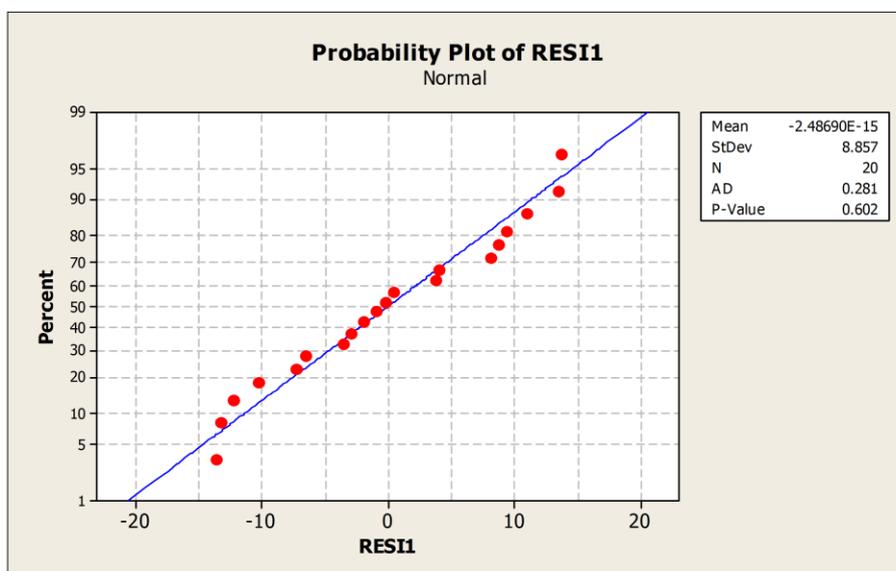


Fig. 1: Residual Normality test

The Anderson Darling test for normality shows that the residual is normally distributed.

Conclusion

From the 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 is 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. Therefore 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.

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 Focus 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) Schwarz, G. (1978). *Estimating the dimension of a model*. *The Annals of Statistics* 6, 461-464
- 11) Sen, P.K., (1968). *Estimates of the regression coefficient based on Kendall's tau*. *Journal of the American Statistical Association* 63 (324): 1379-1389.
- 12) Sprent, P. (1993). *Applied Nonparametric Statistical Methods*. London; New York: Chapman and Hall.
- 13) Theil, H., 1950. *A rank-invariant method of linear and polynomial regression analysis*. *Nederlandse Akademie Wetten chappen Series A* 53: 386-392.
- 14) Wilcox, R. (1998). *Simulations on the Theil-Sen regression estimator with right-censored data*. *Stat. & Prob. Letters* 39, 43-47.