Continental J. Applied Sciences
Usman *et al*. (2020) 15 (1): 14 - 23 DOI: 10.5281/zenodo.3764305

Research Article

Assessing the Performance of Ordinary Least Square and Kernel Regression

U. Usman[1], N. Garba[2], A. B. Zoramawa[1] and H. Usman[3]
[1]Statistics Unit Usmanu Danfodiyo University, Sokoto Nigeria
[2]Academic Planning Unit Sokoto State University, Sokoto Nigeria
[3]Mathematics Unit Usmanu Danfodiyo University, Sokoto Nigeria
Corresponding author: nasirug9@gmail.com

Abstract

The assessment of Ordinary Least Squares (OLS) and kernel regression on their predictive performance was studied. We used simulated data to assess the performance of estimators using small and large sample. However, the mean square error (MSE) and root mean square error (RMSE) was used to find out the most efficient among the estimated models. The results show that, when $n = 100$ the ordinary least square is more efficient than the kernel regression due to having the least MSE and RMSE in both distributions. Whereas for $n = 500$ the ordinary least square and the kernel regression have the same performance for normal distributed data while for lognormal, the result also shows that the kernel regression perform better than the ordinary least square. Finally, when, $n = 1000$ the kernel regression is more efficient than the ordinary least square for having the least MSE and RMSE in both distributions. The overall results show that the kernel regression estimate is more efficient than the ordinary least square (OLS) estimate.

Keywords: ordinary least square (OLS), Kernel regression, Root mean square error (RMSE), Mean square error (MSE)

Received: 28 Jan 2020 Accepted: 12 March 2020

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

Introduction

In general, there are visible differences between a parametric and nonparametric regression estimate. It is therefore quite natural to compare the two regression models Hardle and Mammen, (1993). Akpos and Jude (2018), Compared parametric and nonparametric regression using normal and non-normal data set with difference sample size in the paper, comparisons of Theil's and simple regression on normal and non-normal data set with different sample sizes. Shah *et al.* (2016) Compared parametric and nonparametric regression using normal distribution data, in their paper: comparative study of ordinary least squares regression and Theil-sen regression using outliers data, Jude *et al.* (2016) compared parametric and nonparametric regressions with a normal distribution, in the paper: nonparametric over its parametric counterparts with large sample size. Parametric regression analysis is a statistical process for estimating the relationship among variables, included many techniques for modeling and analyzing several variables, when the interest is to study the relationship between a dependent variable and one or more independent variables. More specifically regression analysis help one understand how the typical value of the dependent variable changes when any one of the independent variables is varied and other independent variable are held fixed Erilli and Alakus, (2014). The important special case of the general model is a nonparametric simple regression, where there is only one predictor:

$$y = m(x) + \varepsilon \qquad (1)$$

Non parametric regression is a category of regression analysis in which the predictor does not take free determine form, but is constructed according to the information drive from data. A nonparametric simple regression is often called "scatter plot smoothing" because an important application is to trace a smooth curve through a scatter plot of y against x. We frequently use John and Weisberg, (2010).

Assessment of parametric and nonparametric regressions is presented in terms of practical problems and measurement on theoretical considerations (Anderson, 1961). Al-Noor and Muhammad (2013), used Model of Robust Regression with parametric and nonparametric method. They evaluated the performance of the classical parametric estimation method ordinary least squares with classical nonparametric estimation Theil's regression method, some robust estimation methods and two suggested methods for conditions in which varying degrees and direction of outliers are presented

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

in the observed data. They concluded that nonparametric introduce better performance in the presence of the outliers x-direction and xy-direction compared to Ordinary least square (OLS), LAD and M-estimators. However the OLS of parametric was considered with data that contain outliers which has a strong influence on the method of OLS. Shah *et al.* (2016) Compared ordinary least squares regression and Theil-Sen Regression through simulation in the presence of outliers. In practice the experimental data obtained from multi-location trials contains outliers. In the conclusion the OLS estimates are highly affected by the presence of outlier and become less efficient in the presence of outlier and as a result lead to wrong conclusions. Theil-sen simple regression is a nonparametric estimation method which is robust to outliers present in the data. Theil-sen regression not only shows consistent performance in the presence of outliers but also a competitor of OLS, considering ordinary least square (OLS) from parametric with data that contain outliers which have a strong influence on (OLS) method. Jude *et al.,* (2016) compared ordinary least squares regression and Theil-Sen Regression with the presence of outliers data, where they used Ordinary Least Square (OLS) for parametric estimation method and TLS for nonparametric regression. First the set of data was subjected to normality test, in the conclusion the OLS estimates of parametric are better than its non-parametric Theil's regression in both data with and without outliers since their AIC and BIC are both lower than that of Theil's regression. They considered AIC and BIC, the two information criterion favor only the OLS since the nonparametric has no likelihood function. Erilli and Alakus, (2014), Compared Parametric regression analysis depends on some assumptions. One of the most important of assumption is that the type of relationship between dependent and independent variable or variables is known. Under such circumstances, in order to make better assumptions, regression methods which enable flexibility in the linearity assumption of the parametric regression are needed. These methods are nonparametric methods known as semi parametric regression methods, estimation of parameters in a parametric regression which has independent variables of different values has been studied extensively in literature Sometimes, one or more observation series of independent variable values can be equal while dependent variable values are different.

This study offers a new method for the estimation of regression parameters under such data. Proposed method and other nonparametric methods such as Theil's Mood-Brown, Hodges-Lehmann methods and OLS method were compared with the sample data and the results is the method which gives more successful results in simple regression and more effective in cases when there are outliers. Conclude that the parametric method produces better result. Considering- mood-browon, OLS, THEIL's regressions; Jude *et al.* (2016) Compared parametric (OLS) and Nonparametric (THEIL's) regression,

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

Concluded that the parametric (OLS) is better than Nonparametric (Theil's) regression. Considering the (OLS) estimation procedure base on some assumption, they avoided the data contains outliers and considered the Akaike Information criterion (AIC) and Bayesian information criterion (BIC) by using Gaussian distribution. The AIC and BIC the two information criterion favor only the OLS since the nonparametric has no likelihood. Several researchers have proposed the comparative study of parametric and nonparametric regression Such as: ((Al -Noor and Mohammad, 2013); (Shah *et al.*, 2016)), assessed the performance of classical parametric estimation OLS with the classical Nonparametric Theil's regression. They considered (OLS) method with outliers data, which has strong influence on the (OLS) method, also the MSE of the (OLS) become very sensitive to these outliers (Gad and Qura, 2016), and (Jude *et al.*, 2016; Akpos and Jude 2018), studied parametric (OLS) and nonparametric Theil's regression using AIC and BIC for measuring the predictive performance, these two traditional model selection criteria favor only the (OLS) since the maximum likelihood cannot be applied to complete nonparametric estimation (Geman and Hwang, 1982) and (Martin and Nils, 2017)

In this research Ordinary Least Square (OLS) and Kernel Regression (KE) assessed by applying MSE and RMSE to measure their efficiency where the data follow normal and Log-normal distributions. The main objectives were to simulate data that follow normal and log-normal distributions, apply MSE and RMSE for measuring the predictive performance of OLS and KE and assess the predictive performance of the (OLS and KE) with the view to finding the best of efficiency.

Materials and Method

In this research work, Monte Carlo *simulations with R Programming* software using Normal N($\boldsymbol{\mu}, \boldsymbol{\sigma}$) and log-normal distributions for the dependent and independent variables *will be used to* evaluate the predictive ability of the parametric and nonparametric regression Nonparametric method (kernel regression) using normal and log-normal distribution

The Kernel estimate of f is defined as

$$f_h = \frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{x - x_j}{h}\right) \quad (4)$$

Or equivalently:

$$W_j = \frac{1}{h} k\left(\frac{x - x_j}{h}\right), \quad (5)$$

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

Where

$\hat{f}_h(x) = \frac{1}{n} \sum_{j=1}^{n} W_j \, y_j$, (6)

$K(.)$: is a Kernel function or shape function and it is non negative function.

h : is called bandwidth, window width, or smoothing parameter. In fact, if the space of X is very uneven, then the Kernel estimator will give poor. Result. This problem was solved by Nadaraya, (1964) and Watson, (1964), they proposed a kernel weight which is given by:

$$W_j = \frac{\sum_{i=1}^{n} k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j=1}^{n} k\left(\frac{x_i - x_i}{h}\right)} = \frac{k(u)}{\sum_{j=1}^{n} k(u)} \quad (7)$$

And the formula of a kernel estimator by using these weights is defined as; $\hat{f}_h$

$$(x_i) = \frac{\sum_{j=1}^{n} k\left(\frac{x_i - x_j}{h}\right) y_i}{\sum_{j=1}^{n} k\left(\frac{x_i - x_j}{h}\right)} \text{ , } I = 1, 2\ldots, n \quad (8)$$

The above estimator is called Nadaraya-Watson estimator, the shape of Kernel weights is determined by the Kernel function K, and the size of the weight is parameterized by the bandwidth h Hardle, (1994).

Ordinary Least Square Method

In Ordinary least square data a pair was observed and call them $\{(x_i, y_i), i = 1, \ldots, n\}$.We can describe the underlying relationship between y_i and x_i involving this error term ε_i Consider the model function

$$y = \beta_0 + \beta_1 x + \varepsilon_i \quad (9)$$

This relationship between the true (but unobserved) underlying parameters β_0 and $\beta_1$and the data points is called a standard linear regression model:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (10)$$

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

Or equivalently:

$$Y = X\beta + \varepsilon, \tag{11}$$

where Y is the vector of the n observed values of the response variable, X is a n x 2 matrix, β is a vector of the parameters, and ε is the vector of the errors. The estimates for β_0 and β_1 are $\hat{\beta}_0$ and $\hat{\beta}_1$ which can be find out by solving the two equations simultaneously

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \tag{12}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i = -\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0 \tag{13}$$

Finally result to

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{14}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15}$$

The design matrix is

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{16}$$

Mean Square Error

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations that is, the difference between the estimator and what is estimated. MSE is a risk function corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The MSE is a measure of the quality of an estimator, it is always non-negative, and values closer to zero are better.

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

if $\hat{Y}$is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

I.e., the MSE is the mean of the square of the errors. This is an easily computable quantity for a particular sample (and hence is sample-dependent).

Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modeled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

The RMSE of a model prediction with respect to the estimated $\hat{y}_i$ variable is defined as the square root of the mean squared error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

Where y_i is the observed value and $\hat{y}_i$ is the estimated value

Result and Discussions

Table 1: Normal & Log-normal when, n=100

Distributions		MSE	RMSE
N($\boldsymbol{\mu}, \boldsymbol{\sigma}$)	OLS	0.66	0.81
	Kernel	0.69	0.83
Lognormal($\boldsymbol{\mu}, \boldsymbol{\sigma}^2$)	OLS	3.36	1.83
	Kernel	4.06	2.02

Table 1 presents the MSE and RMSE when n = 100 it indicates that, the ordinary least square is more efficient than the kernel regression because it has the least MSE and RMSE in both distributions.

Table 2: Normal & Log-normal when, n=500

Distributions		MSE	RMSE
N($\boldsymbol{\mu}, \boldsymbol{\sigma}$)	OLS	0.96	0.98
	Kernel	0.96	0.98
Lognormal($\boldsymbol{\mu}, \boldsymbol{\sigma}^2$)	OLS	3.97	1.99
	Kernel	3.62	1.90

Table 2 presents the MSE and RMSE when n = 500 it indicate that, the ordinary least and the kernel regression have the same performance for normally distributed data while for lognormal the kernel regression perform better than ordinary least square.

Table 3: Normal and Log-normal when, n=1000

Distributions		MSE	RMSE
N($\boldsymbol{\mu}, \boldsymbol{\sigma}$)	OLS	0.99	0.98
	Kernel	0.98	0.96
Lognormal ($\boldsymbol{\mu}, \boldsymbol{\sigma}^2$)	OLS	5.50	2.35
	Kernel	0.52	0.35

Table 3 present the MSE and RMSE when n = 1000 it shows that, the kernel regression is more efficient than the kernel regression because it has the least MSE and RMSE in both distributions

Conclusions

In this study, from the results of the analysis when $n = 100$ the ordinary least square is efficient than the kernel regression because it has the least MSE and RMSE in both distributions while for $n = 500$ The results show that, the ordinary least square and the kernel regression have the same performance, for normal distributed data while for log-normal, the results also show kernel regression perform better than ordinary least square, when $n = 1000$, the kernel regression is efficient than the ordinary least square because it has the least MSE and RMSE in both distributions we concluded that the performances of the estimators improve with increased in the number of observations. However, with Increased in the number of observations, the kernel regression estimate is more efficient than the ordinary least square (OLS).

References

Akpos, *E. P.,* and Jude, O. *(2018).* Comparison of Theil's and simple regression on normal and non-normal data set with different sample size. *International Journal of Managment and Applied Science, 4(1), 70-74.*

Al-Noor. N., and Muhammad, A.(2013). Model of Robust Regression with parametric and Nonparametric Methods. *Mathematical Theory and Modelling, 3*(1), 27-39.

Anderson, H. N. (1961). Scale and statistics: Parametric and Nonparametric *Psychological Bullettin, 58*(4), 305-316.

Erilli, A. N., and Alakus, K. A.(2014). Nonparametric Regression Estimation for data with equal values *European Scientific Journal, 10*(4), 70-80.

Gad, A. M., and Qura, M. E. (2016). Regression estimation in the presence of outliers: A Comparative study, *International Journal of probability and Statistics, 9*(3), 56-72.

Hardle, W. (1994). Applied nonparametric method. *Economic Research Cambridge Univerty Press,* 426.

Hardle, W., and E, Mammen (1993). Comparing parametric versus nonparametric regression fits. *Annals of Statistics, 21*(4), 1926-1947.

John, F., and Weisberg, S (2010). Nonparametric Regression in R In V. knight (Ed.), *An R Companion To Applied Regression* (2nd ed, Vol. 2). SAGE Publications, Ltd 1 Oliver's yard 55 City Road London Oliver.

Jonathan, N., and Tam, H. P. (1998). Comparison of Rebust and Nonparametric Estimators Under Simple Linear Regression Model. *Multiple Linear Regression Viewpoint, 25*(1), 28-33.

Jude, O., Iheagwara, A. I., and Idoch, O. (2016). Comparison of parametric (OLS) and Nonparametric (THEIL'S) linear regression *Advance Research Journal of multi-Disciplinary Discoveries, 2*(1), 24-29.

Martin, J. and Nils L. H.(2017). Parametric or Nonparametric the FIC Approach *Annal. of Statistics 951-981.*

Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Applications, 9(1), 141-142.*

ISSN: 1597 – 9928. Science and Education Development Inst., Nigeria

Shah, S. H., Rashid, A. T., Kanim, J., and Shah, S. M. (2016). Comparative study of ordinary least square regression and Theil's regression in the presence of outliers. *Science and Technology, 5(11), 137-142.*

Watson, G.S., (1964). Smoothing Regression Analysis Sankhya *The Indian Journal of Statistics Series A., 26*(4), 359-372.