EVALUATING THE PERFORMANCE OF ORDINARY LEAST SQUARE AND POLYNOMIAL REGRESSION WITH RESPECT TO SAMPLE SIZE

N. Garba[1], N. S. Danchadi [2] and M. K. Abdulmumin[2]

[1]Academic Planning Unit Sokoto State University, Sokoto Nigeria

[2]Department of Mathematics and Statistics Umaru Ali Shinkafi Polytechnic, Sokoto Nigeria

Corresponding Author's Email & Phone No.: kabiru97.mk@gmail.com, +234 8036592276

ABSTRACT

The evaluation of Ordinary Least Squares (OLS) and polynomial regression (PR) on their predictive performance was studied. We used simulated data to evaluate the performance of estimators using small and large sample. However, the mean square error (MSE ($\hat{\beta}_0$); MSE$(\hat{\beta}_1)$ and MSE$(\hat{y})$) ware used to find out the most efficient among the estimated models. The results show that, for $n = 1000$ the OLS is efficient than the PR due to having the least MSE ($\hat{\beta}_0$); MSE$(\hat{\beta}_1)$ and MSE$(\hat{y})$ on both normal and log-normal distributions. Whereas for $n = 3000$ the values of MSE ($\hat{\beta}_0$); MSE$(\hat{\beta}_1)$ and MSE$(\hat{y})$ of PR are little bit lower than that of OLS which indicates the efficiency of PR over OLS on both distributions. Finally, For, $n = 5000$ the values of MSE ($\hat{\beta}_0$); MSE$(\hat{\beta}_1)$ and MSE$(\hat{y})$ of PR are much more lower than that of OLS which shows that PR is efficient than OLS on both distributions. Overall, the results suggest that OLS is more efficient than PR for smaller sample sizes and PR is more efficient for bigger sample sizes.

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

1.0 INTRODUCTION

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables Erilli and Alakus (2014). Regression analysis depends on some assumptions. The most important of these assumptions is that the type of relationship between dependent and independent variables is known. Estimates which are made when there are no assumptions cannot be good estimates. Under such circumstances, in order to make better assumptions, regression methods which enable flexibility in the linearity assumption of the parametric regression are needed. Therefore, these methods are regression models known as nonparametric and semi-parametric regression methods Erilli and Alakus (2014).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 & Mammen, (1993), Assessment of parametric and nonparametric regressions is presented in terms of practical problems and measurement on theoretical considerations Anderson, (1961), U. Usman, *et.al* (2020) assessed the predictive performance of ordinary least square (OLS) and kernel regression with the increase of the sample size which shows that as the sample size gets higher kernel regression become more efficient than ordinary least square, Akpos & 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. Al-Noor & 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 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 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 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 is 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 & 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 result is the method which give more successful results in simple regression and more effective in cases when there are outliers. Conclude that the parametric method produce better result. Considering- mood-browon, OLS, THEIL's regressions; Jude *et al.,* (2016) Compared parametric (OLS) and Nonparametric (THEIL's) regression, 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 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 researches 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 criterions 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 polynomial regression (PR) and ordinary least square regression (OLS) were assessed by applying MSE and RMSE to measure their efficiency where the data follows normal and Log-normal distributions. The main objectives was to simulate data that follows normal and log-normal distributions, apply MSE and RMSE for measuring the predictive performance of OL and KE and assess the predictive performance of the (PR and OLS) with the view to finding the best of efficiency.

2.0 MATERIALS AND METHOD

In this research work, Monte Carlo *simulations with R was used to* evaluate the predictive ability of the parametric and nonparametric regression we considered random simulation, the Normal N($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\sigma} = \mathbf{1}$) and log-normal($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\sigma} = \mathbf{1}$) distributions for the dependent and independent variables was also used.

2.1 POLYNOMIAL REGRESSION

The polynomial regression of y is defined as

$$yi = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_m x_i^m + \in_i \ (i = 1, 2, \ldots . n) \ (1)$$

It can be expressed in matrix form in terms of a design matrix X, a response vector $\vec{y}$, a parameter vector $\vec{\beta}$, and a vector $\vec{\in}$, of random errors. The *i*-th row of X and $\vec{y}$ will contain the *x* and *y* value for the *i*-th data sample.

Then the model can be written as a system of linear equations:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \in_1 \\ \in_2 \\ \in_3 \\ \vdots \\ \in_n \end{bmatrix} \quad (2)$$

Which when using pure matrix notation is written as

$$\vec{Y} = X\vec{\beta} + \vec{\in} \quad (3)$$

The vector of estimated polynomial regression coefficients (using ordinary least squares estimation) is

$$\hat{\vec{\beta}} = (\mathrm{X}^{\mathrm{T}} \mathrm{X})^{-1} \mathrm{X}^{\mathrm{T}} \vec{y}, \quad (4)$$

Assuming $m < n$ which is required for the matrix to be invertible; then since X is a Vandermonde matrix, the invertibility condition is guaranteed to hold if all the x_i values are distinct. This is the unique least squares solution.

2.2 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

$$y = \beta_0 + \beta_1 x + \varepsilon_i \quad (5)$$

This relationship between the true (but unobserved) underlying parameters β_0 and β_1 and the data points is called standard linear regression model:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (6)$$

Or equivalently:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon, \quad (7)$$

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

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \quad (8)$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i = -\hat{\beta}_1 \sum_{i=1}^{n} x_i{}^2 = 0 \quad (9)$$

Finally result to

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i{}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \quad (10)$$

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

The design matrix is

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

2.3 SELECTION CARITERIA

2.3.1 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.

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 the

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (12)$$

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).

2.3.2 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:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (13)$$

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

3.0 RESULT AND DISCUSIONS

Table 1: The Result of MSE and RMSE of Point Estimates by Normal Distribution at n = 1000

Method	$\widehat{\boldsymbol{\beta}}_0$	MSE $(\widehat{\boldsymbol{\beta}}_0)$	RMSE$(\widehat{\boldsymbol{\beta}}_0)$	$\widehat{\boldsymbol{\beta}}_1$	MSE$(\widehat{\boldsymbol{\beta}}_1)$	RMSE$(\widehat{\boldsymbol{\beta}}_1)$	MSE$(\hat{y})$	RMSE$(\hat{y})$
OLS	0.00152	0.032590	0.180527	0.08076	0.03216	0.17933	0.6310	0.7943
PR	0.00927	0.03989	0.199730	0.80613	0.03217	0.17936	1.0110	1.0054

The result of the table (1) shows that the (OLS) method is efficient than the (PR), because it has the least estimate of MSE $(\hat{\beta}_0)$; MSE$(\hat{\beta}_1)$ and MSE$(\hat{y})$ with their root mean squares (RMSE) on normal distribution.

Table 2: The Result of MSE and RMSE of Point Estimates by Normal Distribution at n = 3000

Method	$\widehat{\boldsymbol{\beta}}_0$	MSE $(\widehat{\boldsymbol{\beta}}_0)$	RMSE$(\widehat{\boldsymbol{\beta}}_0)$	$\widehat{\boldsymbol{\beta}}_1$	MSE$(\widehat{\boldsymbol{\beta}}_1)$	RMSE$(\widehat{\boldsymbol{\beta}}_1)$	MSE $(\hat{y})$	RMSE$(\hat{y})$
OLS	0.018924	0.018264	0.13514	0.007965	0.01322	0.13535	1.2898	1.1357
PR	0.014668	0.01420	0.11916	0.007415	0.01049	0.10242	0.9930	0.9965

The result of the table (2) reveals that the (PR) is efficient than the (OLS) because it has the least estimates of MSE$(\hat{\beta}_0)$ and MSE$(\hat{\beta}_1)$ with their root mean squares errors (RMSE) on normal distribution.

Table 3: The Result of MSE and RMSE of Point Estimates by Normal Distribution at n = 5000

Method	$\widehat{\boldsymbol{\beta}}_0$	MSE $(\widehat{\boldsymbol{\beta}}_0)$	RMSE$(\widehat{\boldsymbol{\beta}}_0)$	$\widehat{\boldsymbol{\beta}}_1$	MSE $(\widehat{\boldsymbol{\beta}}_1)$	RMSE$(\widehat{\boldsymbol{\beta}}_1)$	MSE $(\hat{y})$	RMSE$(\hat{y})$
OLS	0.001704	0.015589	0.12485	0.004652	0.01526	0.12353	1.9857	1.4091
PR	0.001520	0.00145	0.03820	0.001781	0.00243	0.04929	0.9883	0.9941

The result of the table (3) indicates that the (PR) is efficient than the (OLS) because it has the least estimates of MSE$(\hat{\beta}_0)$ and MSE$(\hat{\beta}_1)$ with their root mean squares errors (RMSE) on normal distribution.

Table 4: The Result of MSE and RMSE of Point Estimates by Log- Normal Distribution at n = 1000

Method	$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	RMSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	RMSE$(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	MSE $(\hat{\boldsymbol{y}})$	RMSE$(\hat{\boldsymbol{y}})$
OLS	1.65783	0.08565	0.29266	0.02879	0.00064	0.02529	2.1100	1.4525
PR	1.85403	0.09899	0.31463	0.13952	0.00519	0.22789	2.1950	1.4815

The (OLS) approach is more efficient than the (PR) method, as shown in table (4), since it has the lowest estimate of MSE$(\hat{\beta}_0)$; MSE $(\hat{\beta}_1)$and MSE$(\hat{y})$ and their RMSE on log-normal distribution.

.

Table 5: The Result of MSE and RMSE of Point Estimates by Log- Normal Distribution at n = 3000

Method	$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	RMSE$(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	RMSE$(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	MSE $(\hat{\boldsymbol{y}})$	RMSE$(\hat{\boldsymbol{y}})$
OLS	1.62749	0.04719	0.21724	0.004069	0.01736	0.13174	2.1050	1.4663
PR	1.62299	0.03595	0.18960	0.000341	0.00227	0.04762	2.0450	1.4300

The result on the table (5) indicates that the (PR) method is efficient than the (OLS) because it has the least estimate of MSE$\hat{\beta}_0$; MSE$(\hat{\beta}_1)$ and their RMSE on log-normal distribution.

Table 6: The Result of MSE and RMSE of Point Estimates by Log- Normal Distribution at n= 5000

Method	$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	RMSE$(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})$	$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	MSE $(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	RMSE$(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})$	MSE $(\hat{\boldsymbol{y}})$	RMSE$(\hat{\boldsymbol{y}})$
OLS	1.61091	0.03402	0.18444	0.01509	0.01191	0.10913	2.3291	1.5261
PR	1.60666	0.03143	0.17728	0.01479	0.00077	0.02778	1.9631	1.4011

The result on the table (6) shows that the (PR) method is efficient than the (OLS) and (TLS) because it has the least estimate of MSE$\hat{\beta}_0$; MSE $(\hat{\beta}_1)$ and their RMSE on log-normal distribution.

CONCLSIONS

According to the findings of this study, when n=1000, ordinary least square regression is more efficient than polynomial regression since it has the lowest MSE and RMSE in both distributions, however when n=3000, polynomial regression is more efficient. The results show that polynomial regression is more efficient than ordinary least square for both normal and log-normal distributions, and that when n=5000, polynomial regression is more efficient than ordinary least square because it has the lowest MSE and RMSE in both distributions. We concluded that the estimator's performance improves as the number of observations increases. However, as the number of observations grows, the polynomial regression estimate becomes more efficient than the ordinary least square.

.

REFERENCES

Akpos, E. P., & 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., & Muhammad, A.(2013). Model of Robust Regression with parametric and Nonparametric Methods. *Mathematical Theory and Modeling, 3*(1), 27-39.

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

Erilli, A. N., & Alakus, K. A.(2014). Nonparametric Regression Estimation

for data with equal values. *European Scientific Journal, 10*(4), 70-80.

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

Geman, S., & Hwang, c.-R.(1982). nonparametric Maximum Likelihood estimation by the method of sieves. *Anal. of Statistics, 10*(2), 401-414.

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

Jude, O., Iheagwara, A. I., & 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. & Nils L. H.(2017). Parametric or Nonparametric the FIC Approach *Anal. of Statistics 951-981.*

Shah, S. H., Rashid, A. T., Kanim, J., & 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.*

Usman U, Garba N, Zoramawa A.B, & Usman H. (2020). Assessing the Performance of Ordinary Least Square and Kernel Regression. *Continental J. Applied Sciences, 15(1), 14–23. https://doi.org/10.5281/zenodo.3764305*