International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

COMPARISON OF LEAST MEDIAN SQUARE AND ORDINARY LEAST SQUARE METHODS IN THE PRESENCE OF OUTLIERS

Awariefe, Christopher
Department of Statistics, Delta State Polytechnic, Ozoro
awariefec@gmail.com

Ekerikevwe, Kennedy
Department of Statistics, Delta State Polytechnic, Otefe-Oghara

Abstract
The general assumption concerned with linear regression model is that under ideal conditions the ordinary least square performs better than other regression methods. However under some non-ideal conditions, that is, when the general assumption of normality is violated the ordinary least square breaks down. Thus, this study aimed to compare the efficiency of ordinary least squares (OLS) and least median squares (LMS) estimators by subjecting both estimators to dataset with and without the presence of outliers. We made use of real and simulated data. The simulated data were obtained from R program. The data was analyzed with multiple linear regression methods (Ordinary Least Square and Least Median Squares). Also, the residual standard error of both models and the standard error of the coefficients (intercept and slope) were used to assess and compare their performances. The result of the regression analysis shows that the OLS perform better when normality of the data is not violated; however, the OLS perform poorly compared to LMS when the normality is violated due to the presence of outliers as revealed by its higher residual standard errors and parameters standard errors.

Keywords: Outliers, Multiple regressions, breakdown point, Ordinary Least Square and Least Median Squares

1. Introduction

Regression analysis is a statistical tool for investigating the relationship between two or more variables, such that one variable can be predicted from the other(s). We use regression to establish causal (cause and effect) relationship through a mathematical model which shows how one variable depends on the other. In modeling a regression, researchers usually used ordinary least squares (OLS) method, due to the simplicity of the idea of minimizing the sum of squared residuals and the interpretability of the final model parameter estimates. Cankaya and Abaci (2015), in their study they compared some estimation methods, such as; Least square, Least trimmed square, M-Estimation, MM-Estimation and S-Estimation for estimating the parameters of simple linear regression model in the presence of outlier and different sample sizes. Mean square error (MSE) and coefficient of determination (R^2) values were used as criteria to evaluate the estimator performance. They found that LTS estimator is the best models with minimum MSE and maximum (R^2) values for different size of sample in the presence of outliers. The research conducted by Çankaya et al., (2011) identifies that under ideal conditions OLS method achieves optimum results when the underlying error distribution is normal, it brings some shortcomings. One of these is that, OLS method is sensitive to outliers which can disturb the assumption of normality, one of the most important components of statistical analysis. This condition reduces the predictive power of the method. Ordinary Least squares perform poorly in terms of robustness because a single, aberrant data point, or outlier, can throw the fitted line way off. The smallest percentage of bad data that can cause the fitted line to explode is defined as the breakdown point. Since a single bad data point can destroy the least squares line, OLS is said to have a zero breakdown point. Montgomery (2012) opined that the application of regression is well appreciated when real life problems and results that typically arise when the method is adopted for both theoretical and real life data. Best linear unbiased estimates (BLUE) produced by the ordinary least squares (OLS) regression technique under the normal error distribution. Birkes and Dodge. (1993). However, many researchers have noted that the optimal condition is rarely met in real data analyses. In classical multiple regressions, the ordinary least square estimation is the best method if assumptions are met to obtain regression weights when analyzing data. Though, if some of these assumptions are not satisfied, then sample estimates and results from the data can be misleading. Especially, outliers violate the assumption of normally distributed residuals in the least squares regression. The problem of outliers, in both directions of the dependent and explanatory variables and to the least squares regression is that they can have a strong adverse impact on the estimate and they may remain unnoticed. The Least Median Squares estimator is simple to describe and is very robust against outliers. Robust regression such as Least Median of Squares is an important method for analyzing data that are contaminated with outliers and can be used to detect outlying observations and provide resistant results in the presence of dataset with outliers.

Other scholars have done similar works related to the impact of outliers in regression, Jianju (1999) illustrated several advantages of the least median squares (LMS) regression by adopting a user-friendly software program, "Program for RObust reGRESSion" (PROGRESS),in his research for the comparison between least square and least median square he observed that the LMS technique results in a smaller average error of prediction and the LMS also perform better for real data without outliers than the simple LS fit. Other similar works are by Zhu and Zhilin (2000); Birkes and Dodge (1993); Rousseeuw (1984); Weisberg (1985).

This paper reports yet another contribution to the nature of research efforts illustrated above; that is, research efforts directed towards the study of the performance of least square and least median square by subjecting both methods to real, simulated contaminated and skew data using R statistical software.

2. Review of Regression and ordinary least square.

The usual regression model, in matrix notation is

$$Y=X\beta + \varepsilon \qquad 2.01$$

The regression coefficients can be estimated by solving:

$$\beta=(X'X)^{-1}X'Y. \qquad 2.02$$

A regression with n cases and K-1 with X variables can be stated in the following matrices form:
Y (n x 1), a column vector of Y observed values:

$$Y=\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

$\hat{Y}$ (n x 1), a column vector of Y predicted values:

$$\hat{Y}=\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \hat{Y}_3 \\ . \\ . \\ . \\ \hat{Y}_n \end{bmatrix}$$

(k x 1) matrix column vector of the estimated regression coefficients:

$\beta = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ . \\ b_{k-1} \end{bmatrix}$ (n x 1) column vector of ε sample residuals for the respective n cases is given below

:

$\varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix}$

The general matrix format for a regression model with any number of variables is given below

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13}\ldots\ldots & X_{1K-1} \\ 1 & X_{21} & X_{22} & X_{23}\ldots\ldots & X_{2K-1} \\ 1 & X_{31} & X_{32} & X_{33}\ldots\ldots & X_{3K-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & X_{n3}\ldots\ldots & X_{nK-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ . \\ b_{k-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix} \qquad 2.03$$

(n x 1) (n x k) (k x 1) (n x 1)

The elements of ε are assumed to be independent and identically distributed with V (ε) = $\sigma^2 I_n$ where I_n is an n x n *identity* matrix and σ^2 = (>0) is a constant. The ordinary least squares (OLS) give better and reliable estimate of β when the required data is well behaved, that is, when it is free of outliers.

2.1 Estimation of the Regression Coefficients

The ordinary least square can be used to solve for the coefficients

The sum of squared residual is given by

$$\sum_{i=1}^{n} e_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

Expanding the RHS of the equation we get:

= Y'Y -2$\beta' X' Y$ + $\beta' X' X \beta$

Applying partial differentiation with respect to β and equating to zero we get

$\frac{\delta \sum_i^n e_i^2}{\delta \beta}$= -2X'Y + 2X'X$\hat{\beta}$=0

$2X'X\hat{\beta} = 2X'Y$

By simplification we get: $\hat{\beta} = (X'X)^{-1}X'Y$

By solving the above equation, gives the estimates of the regression coefficients assuming X is full rank. We now obtain the expectation and the variance of $\hat{\beta}$.

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

The matrix model, $Y=X\beta + \varepsilon$ is substituted in $\hat{\beta}$

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

$= \frac{1}{X'X}X'X\beta + (X'X)^{-1}X'\varepsilon = \beta + (X'X)^{-1}X'\varepsilon$ 2.11

The expectation becomes,

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

From equation 4.21

$\hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon$ 2.12

The variance of $\hat{\beta}$ becomes

$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1}X'\varepsilon\,(X'X)^{-1}X\varepsilon']$

$= (X'X)^{-1}X'(X'X)^{-1}X\,E(\varepsilon\varepsilon')$

$= (X'X)^{-1}X'\sigma^2 I_n(X'X)^{-1}X = \sigma^2 (X'X)^{-1}$ 2.13

$= \sigma^2 (X'X)^{-1}$ which is the variance-covariance matrix of $\hat{\beta}$

3. Review of the Least Median Square

Least Median Square is one of method of creating resistance line, procedure that is resistant to some percentage of arbitrarily large outliers is a resistance statistics, and robustness means the procedure is not greatly affected by slight deviations in the assumptions. There are various ways to create a resistant regression line; the mean and standard deviation of the regression line is very sensitive to outliers. The median which is less sensitive to outliers replaces the mean in the estimation of the parameters.

3.1 Least median square Estimation

The Least Median of Squares fit is determined by solving the following optimization problem:

$$\min_{b_0,b_1} \min_{i} SR_i = \text{Median}\,\{Y_1 - (b_0 + b_1X_1)^2, (Y_2 - (b_0 + b_1X_2)^2, \ldots . (Y_n - (b_0 + b_1X_n)^2\}$$

Since Least Square is based on minimizing the sample mean and means are sensitive to extreme values, it makes sense that least median square (LMS), replaces the mean by the median which is less sensitive to outliers and will generate a more robust estimator.

3.2 Breakdown Point

Let T be a functional defined on some subfamily P_T of the family P of all distributions on a sample space (X, B(X)) which takes its values in some metric space (θ, D) with

$Sup_{\theta_1,\theta_{2\in\theta}} D(\theta_1,\theta_2)=\infty$ 3.21

The finite sample breakdown point of T at a sample x_n = (x_1,,……, xn_N), $x_i \in$ X, i=1, ..., n, is defined as

fsbp (T, x_n , D) = $\frac{1}{n}$ min{$k \in(1,\dots,n)$: D$\left[T(P_n),T(Q_{n,k})\right]=\infty$} 3.22

Where P_n= $\sum_{i=1}^{n} \delta_{xi}$/n and $Q_{n,k}$ and is the empirical distribution of a replacement sample with at least n−k points from the original sample x_n

50% is the highest likely breakdown point of an estimator, which indicates that as many as half the observations could be discounted. A breakdown point higher than 0.5 is undesirable because it would mean that the estimate could be pertains to less than half of the data. Andersen (2012).

4. Application

4.1 Description of Data

The data in table 4.1 consists of 20 New York river basins which were originally collected by Haith (1976) to explore the relationship between nonpoint source water pollution (nitrogen concentration) and various types of land use (% land in agriculture,% land forest,% land urban). See (Hamilton, 1992).

Table 4.1

	Basin	% Land in Agriculture	% Land Forest	% Land Urban	Nitrogen (mg/1)
1	**Olean**	26	63	1.49	1.10
2	**Cassadaga**	29	57	0.79	1.01
3	**Oatka**	54	26	2.38	1.90
4	**Neversink**	2	84	3.88	1.00
5	**Hackensack**	3	27	32.51	1.99
6	**Wappinger**	19	61	3.96	1.42
7	**Fishkill**	16	60	6.71	2.04
8	**Honenye**	40	43	1.54	1.65
9	**Susquehanna**	28	62	1.25	1.01
10	**Chenango**	26	60	1.13	1.21
11	**Tioughnioga**	26	53	1.08	1.33
12	**West Canada**	15	75	0.86	0.75

13	**East Canada**	6	84	0.62	0.73
14	**Saranac**	3	81	1.15	0.80
15	**Ausable**	2	89	1.05	0.76
16	**Black**	6	82	0.65	0.87
17	**Schohaire**	22	70	1.12	0.80
18	**Raquette**	4	75	0.58	0.87
19	**Oswegatchie**	21	56	0.63	0.66
20	**Cohocton**	40	49	1.23	1.25

X_1 = % Land in Agriculture, X_2 = % Land Forest, X_3 =% Land Urban, Y= Nitrogen Concentration in river water (mg/1)

4.2 Regression Model for the Data

The following regression model is proposed for the data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots\ldots\ldots \beta_{K-1} X_{i,K-1} + \varepsilon_i \qquad 4.21$$

i=1, 2……, 20.

Where: X_{i1}= the ith value of the variable Y_i = Dependent variable

β_0 = The intercept coefficient $\beta_1, \beta_2 and \beta_3$= the slopes of the regression coefficients

ε Is random error in Y assumed to have zero mean with constant variance, that is, V (ε) = σ^2

For inferential purposes, it is required to assume that~ $N(0, \sigma^2)$

Table 4.2 Regression Estimates for the real data of table 4.1 and their Corresponding Standard Errors in Brackets.

		Estimates			
Methods of Estimation	**Residual Standard error**	b_0	b_1	b_2	b_3
Ordinary Least Square	0.2802	1.428598 (1.292693)	0.008503 (0.015816)	-0.008449 (0.014468)	0.029401 (0.027638)
Least Median Square	0.1692	2.2110 (0.9077)	-0.0005 (0.0111)	-0.0173 (0.0102)	0.0105 (0.0194)

Table 4.3 Regression Estimates for simulated normal distribution data and their Corresponding Standard Errors in Brackets. rnorm (35,170, 15)

		Estimates			
Methods of Estimation	**Residual Standard error**	b_0	b_1	b_2	b_3
Ordinary Least Square	17.21	70.5016 (59.7509)	0.1852 (0.2134)	0.2581 (0.2343)	0.1275 (0.1768)
Least Median Square	20.44	64.1108 (61.9322)	0.1784 (0.2212)	0.2930 (0.2429)	0.1358 (0.1832)

Table 4.4 Regression Estimates for simulated contaminated normal distribution data and their Corresponding Standard Errors in Brackets. rnorm(35, 0, (1+2*rpois(35, 3)))

		Estimates			
Methods of Estimation	**Residual Standard error**	b_0	b_1	b_2	b_3
Ordinary Least Square	6.88	-0.3195 (1.2020)	-0.1406 (0.1276)	-0.2344 (0.1684)	-0.3640 (0.1897)
Least Median Square	5.367	-0.4870 (1.2205)	-0.1300 (0.1296)	-0.1687 (0.1710)	-0.3728 (0.1927)

Table 4.5 Regression Estimates for simulated exponential distribution data and their Corresponding Standard Errors in Brackets. rexp(35,rate=8)

		Estimates			
Methods of Estimation	**Residual Standard error**	b_0	b_1	b_2	b_3
Ordinary Least Square	0.1116	0.14283 (0.03857)	-0.09891 (0.20549)	-0.03173 (0.24224)	-0.06130 (0.10292)
Least Median Square	0.07293	0.1088 (0.0271)	-0.0369 (0.1443)	0.0012 (0.1702)	-0.0458 (0.0723)

5. Discussion of Results

The regression estimates entries in the table 4.2 and 4.5 shows that estimates of regression from the robust regression fitting method (Least Median Square) have uniformly smaller residual and

standard errors than those from the regression estimates of the ordinary least squares. This is so because the real data comprises of outliers that makes the OLS to break down, the result is an indication that can serve as an instrument for boosting the efficiency of robust regression, which in essence is the main aim of this paper.

The relevant entries in table 4.3 shows that estimate of regression from the robust regression fitting method (Least Median Square) have uniformly higher residual and standard errors than those from the regression estimates of the ordinary least squares. This is an indication that ordinary least squares provides optimum estimates when the data set is free from outlier and is normally distributed, this is supported by Gauss who introduced the normal (or Gaussian) distribution as the error distribution for which ordinary least squares is optimal (see the citations in Huber 1972 and Le Cam 1986).

6. Conclusion

This study aimed to compare the performance of ordinary least squares (OLS) and least median squares (LMS) estimators by subjecting both estimators to dataset with and without the presence of outliers. The results show that the least median squares (LMS) estimator is robust method, that is, it is resistance to the presence of outliers compared to the ordinary least squares (OLS) estimator. The study shows that OLSE performed poorly against the robust regression method (least median squares).the OLS returns the poorest result in all the criteria, that is, residual standard error of the model and standard error of slope and intercept.

Therefore, the robust regression method (least median squares) will be more efficient in performing test of hypothesis and prediction than the ordinary least squares estimator when outliers are present in dataset.

REFERENCES

Andersen, R. (2012). Modern methods for robust regression. University of Toronto.

Birkes, D. & Dodge, Y. (1993). Alternative methods of regression. New York, NY: Wiley.

Çankaya S, Eker S, Tahtali Y, Ceyhan A. (2011). Comparison of some estimation methods for parameters of simple regression model in the presence of outliers. 7th National Animal Science Congress, 14-16 September, 2011, Adana, Turkey, p: 136-141

Cankaya,S.and Abaci,S.H.(2015).A Comparative Study of Some Estimation Methods in Simple Linear Regression Model for Different Sample Sizes in Presence of Outliers. Turkish Journal of Agriculture Food Science and Technology 3(6):380-386.

Çankaya S. (2009). A comparative study of some estimation methods for parameters and effects of outliers in simple regression model for research on small ruminants. Trop Anim Health Pro, 41 (1): 35-41.

Huber P.J. (1973). Robust regression: asymptotic, conjectures and Monte Carlo. Ann. Stat., 1 (5), 799-821.

Hamilton, L.C**.** (1992), regression with Graphics: A second Course in Applied Statistics, Duxbury Press. California.

Jianju, W (1999). An Illustration of the Least Median Squares (LMS) Regression Using PROGRESS. Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada).

Le Cam, L. (1986). The central limit theorem around .1935, *Stut.* Sci., 1, 78-96.

Montgomery DC, Peck EA, Vining GG. (2012). Introduction to linear regression analysis (Vol. 821). Wiley.

Weisberg, S. (1985). Applied linear regression. New York, NY: Wiley.

Zhu, X and Zhilin,L (2000) .Least median of squares matching for automated detection of surface deformations. International Archives of Photogrammetry and Remote Sensing. Vol. XXXIII, Part B3. Amsterdam.