METHODS OF ESTIMATING LINEAR REGRESSION MODEL PARAMETERS IN THE PRESENCE OF OUTLIERS

(i) Olufemi-ojo,Olukemi B.PhD and (ii) Okechukwu Ijeoma E.PhD.

(i) Statistics Department, Federal Polytechnic, Oko. olufemiolukemi@gmail.com 07065118003

(ii)Statistics Department, Federal Polytechnic, Oko. okechukwuijeoma41@gmail.com 08060256823

ABSTRACT

In Linear regression model, Ordinary Least Square estimate is considered the best method to estimate the parameter if all the assumptions are met. The violation of some of these assumptions may give misleading result due to the presence of outliers, hence, the need for the use of robust regression methods. The work investigated seven robust regression methods (Least Trimmed Squares estimates, Tukey Bisquare Estimator, Yohai MM Estimates, S- estimator, Least Absolute Value, Robust Weighted Least Squares Estimator and Least Winsorized Square for estimating regression parameters in the presence of outliers. Cases with two and five variables were compared. Simulation which covered data sets with 2%, and 10% outlying Contamination Rate and 20, 50, 100, 200, and 500 as sample sizes were performed using the R –Package. The Root Mean Square was used as the performance measure of accuracy for the estimators in predicting each of the parameters. The results obtained showed that the number of variables does not affect the performances of the robust methods considered. Robust Weighted Least Square and Least Winsorized Estimator performed best both in simulation and the real life data application. Hence, they are perfect substitute for Ordinary Least Square estimate when data are contaminated with outliers.

Keywords: Robust, Outliers, Estimating, Parameters, Regression model.

Introduction

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon that the heights of descendants of tall ancestors tend to regress down towards a normal average (Mogull, 2004).Regression analysis is a statistical process for estimating the relationships among variables. It helps one to understand how the typical value of the dependent variable changes when any of the independent variables is varied, while the other independent variables are held constant. Regression analysis estimates the average value of the dependent variable when the independent variables are fixed. It is very useful for prediction and forecasting. It is also used to explore the relationship between the independent and the dependent variables. Scott (2012) wrote that regression analysis can be used to infer causal relationships between the independent and dependent variables in restricted circumstances. This can lead to illusions or false relationships, hence, caution should be taken.In practice, the performance of regression analysis methods depends on the form of data generating process. Most often the true form of the data-generating process is generally unknown and such regression analysis often depends on making assumptions about this process to some extent. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are slightly violated, although, they may not perform optimally. (Freedman, 2005).

Recently, new methods have been developed for robust regression. These methods include regression methods which involve various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations and causal inference with regression.In classical statistical theory, most of the statistical estimation methods are based on the model with certain assumptions, such as variables follow normal distribution, have constant variance and are mutually independent. Though in practice, these are not often met. Statistical models are just approximations to the actual events at a certain degree and when the actual data do not fulfill the model assumptions, the estimators may no longer be the best. The optimal estimators could be through these methods only if all the assumptions are satisfied.

A robust estimation procedure dampens the effect of observations that would be highly influential if least square are used. It should produce essentially the same results as least squares when the underlying distribution is normal and there are no outliers. It is a form of weighted and reweighted least squares regression (Holland & Welsch 1977). Robust methods seek to provide ways with optimal performance when the basic assumptions for data sets are not fully fulfilled or violated.

When data are contaminated with outliers or influential observations, the alternative to least squares regression is robust regression. Robust regression is a good substitute in any situation in which Least Squares Regression is used. When fitting a Least Squares Regression, one might find some outliers or high leverage data points. These data points are neither necessarily data entry errors, nor from a different population than our data. Hence, there is no reason excluding them from the analysis. Based on this, Robust regression becomes the remedy, since it is a compromise between excluding these points entirely from the analysis and including all the data points and treating them all equally in Ordinary Least Squares regression.

Statement of the Problem

In linear regression model, Ordinary Least Squares (OLS) method is considered the best method to estimate the regression parameter if the assumptions are met. However, when the data does not satisfy the assumptions, the results will be misleading. The violation of some of these assumptions is caused by the presence of outliers in the data thereby necessitating the use of robust regression methods. In this regards, It is necessary to review and compare the performance of these robust estimators under different sample sizes and percentages of contamination by outliers to determine which method is best under what conditions.

Purpose of the Study

The aim of this study is to determine the best method of estimating parameters of a linear regression model in the presence of outlying values.

The specific objectives are to;

1. compare the performance of seven robust methods using simulated and real life data.
2. ascertain from the comparison if the Contamination Rate of outlying values affect the performance of the methods.
3. determine if number of variables affect the performance of the methods.
4. ascertain if sample size affects the performance of the methods.

Materials and Methods

This study considered seven robust methods of estimating the parameters of regression model namely ; Least Trimmed Squares (LTS)estimates, Tukey Bisquare Estimator, Yohai MM Estimates, S- estimator, Least Absolute Value (LAV) , Robust Weighted Least Squares Estimator (RWLS) and Least Winsorized Square. Cases with two and five independent variables were considered. The simulation experiments covered data sets with 2%, and 10% contamination rate. The sample sizes that were considered are 20, 50 ,100,200 and 500.

Methods Considered for the Analysis\

Least Trimmed Squares (LTS) Estimate

Least Trimmed Squares (LTS) method was developed by Rousseeuw (1984).It is given from minimizing

$$\hat{\beta}_{LTS} = arg\ Min \qquad 1$$

Where $h = \left[\frac{n}{2}\right] + \left[\frac{(p+1)}{2}\right]$ with n and p being the given sample size and number of parameters involved in the model respectively.

The Least Winsorized Square (LWS) estimator

LWS was developed by Yale and Forsythe in 1976. The estimator is given as:

$$\min \sum_{i=1}^{n} e_i + (n-h)(e^2)\, where\, h = \left[\frac{n}{2}\right] + \left[\frac{(p+1)}{2}\right] \qquad 2$$

Tukey’s Bisquare Estimator (TB)

Tukey Bisquare Estimator was introduced by Tukey in 1977. Rather than minimize the sum of squared errors as the objective, the Tukey Bisquare estimate minimizes a function ρ of the errors. The objective function is,

$$\min \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i\beta}{s}\right),\ for\ all\ i = 1,2,\ldots,n \qquad 3$$

where s is an estimate of scale formed from the linear combination of the residuals. The ρ function is defined as

$$\rho(e_i)=\begin{cases}\frac{c^2}{6}\left[1-\left(1-\left[\frac{e_i}{c}\right]^2\right)^3\right] & for\,|e_i|\le c\\ \frac{c^2}{6} & for\,|e_i|>c\end{cases}$$ 4

(Arya et al., 2007)

Yohai MM Estimator

MM estimation is a special type of M-estimation developed by Yohai (1987). It is a combination of high breakdown value estimation and efficient estimation. M M estimates is the solution to the equation;

$$\frac{1}{n-\rho}\sum_{i=1}^{n}\rho\left[\frac{y_i-x_i\hat{\beta}_{YM}}{s}\right]=0.5,\quad for\;all\;i=1,2,...,n$$ 5

S-estimator

S estimation is a high breakdown value method introduced by Rousseeuw and Yohai (1984). The objective function is given by the solution to 6.

$$\frac{1}{n-\rho}\sum_{i=1}^{n}\rho\left[\frac{Y_i-\hat{Y}_i}{s}\right]=k,\quad for\;all\;i=1,...,n$$ 6

where K is a constant usually 0.1995

Least Absolute Value (LAV)

This estimator minimizes the sum of absolute values of errors instead of minimizing the sum of squares of error as in Ordinary Least Squares. The estimates is given by;

$$\hat{\beta}_{LAD}=\min\sum_{i=1}^{n}|e_i|,\;for\;all\;i=1,2,...,n.$$ (Chen et al, 2008). 7

ROBUST WEIGHTED LEAST SQUARES ESTIMATOR (RWLSE)

RWLSE was introduced by Yohai and Gervini (2002). It is estimate using

Method of Comparison

In this work, the Root Mean Square Error (RMSE) was used for the comparison of the results and it is given as

$$RMSE=\sqrt{\sum_{i=1}^{n}\frac{\left(\hat{B}_i-B_i\right)^2}{n}}$$

Where the lower values of RMSE indicate a better fit.

Also, Coefficient of Determination (R^2) was used to determine if the variability in B that is explained by the regression model.

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{B}_i - \overline{B}\right)^2}{\sum_{i=1}^{n} \left(B_i - \overline{B}\right)^2}$$

Where $\hat{B}_{\text{i}}$ are the estimated parameters, B_i are the true values of the parameter and $\overline{B}$ is the mean of the true values of the parameter. The value of R^2 ranges between 0 and 1.That is, $0 \le R^2 \le 1$. To assess the effects of the Contamination Rate (CR), CR was varied as 2%, and 10%. Let α be the probability of a contaminated observation (τ_i) occurring, then the number of contaminated points (Contamination Rate (CR)) is $n \times \alpha$, where n is the sample size. This applied to the contaminants of the Y variable as well. Data were generated with some percentage of outliers presence. To assess the effect of sample size on the accuracy of the estimators, the sample sizes of generated samples were varied as 20, 50, 100, 200 and 500. The comparison was done using simulation to back up the result from the real life data set. A real life data application of the seven robust methods was done on data extracted from (Helmut, 1991), where there were one dependent and five independent variables. Mahalanobis distance was used to detect the outliers present in the data set. The Mahalanobis distance (MD) of an observation x from a given set of observations with mean μ and covariance matrix S is given by $MD = \sqrt{(X-\mu)^T S^{-1} (X-\mu)}$

From χ^2_p table at $\alpha = 0.975$ and p = 5 (since there are 5 independent variables) , χ^2_p value is 11.07. Mahalanobis distance greater than this were classified as outliers. All computations and simulations were done using the R statistical package. The R-codes were presented in the appendix.

Data Analysis

The results of the simulation and real life data application are presented for each data category Below are the tables presenting the summary some of the simulated experiments of two and five independent variables at 2%, 5%, and 10% contamination rate with varied sample sizes. The results are presented in tables containing the root mean square error (RMSE) and the coefficient of determination for each of the methods considered. All computations and simulated experiments were done using the R package.

Table 1: Results For Two Independent Variables At 2 % Contamination Rate at varying sample sizes

Sample size	n=20	n=20	n=50	n=50	n=100	n=100	n=200	n=200	n=500	n=500
METHOD	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
LWS	0.8236	6.4077	0.8583	5.5201	0.77327	4.12355	0.85919	3.1926	0.78883	2.31763
RWLSE	0.8583	8.2945	0.9858	9.6937	0.9067	4.31141	0.76081	3.2243	0.72131	2.23128
LAV	0.9858	8.3067	0.9471	10.45	0.97327	7.53275	0.85919	5.1981	0.87943	3.33641
S-	0.9744	15.0164	10.464	10.464	0.9849	7.54697	0.98503	5.2014	0.98558	3.33963
TB	0.9841	15.0394	0.9856	10.45	0.98577	7.5316	0.98554	5.1981	0.92132	3.33643
MM	0.7428	15.0394	0.9045	10.45	0.87343	7.53275	0.94022	5.1981	0.9859	3.33642
LTS	0.8236	15.0394	0.8871	10.438	0.9871	7.53009	0.98737	5.1989	0.98771	3.33676

From Table 1 for contamination rate (CR) 2%, when sample size was small, (LWS) method produced the least root mean square error (RMSE), closely followed by the re-weighted least square estimator (RWSLE) while the method that produced the highest RMSE was the least trimmed square estimator. When n = 50, the least Winsorized square method still showed least error followed by the RWSLE while the S-estimator produced the highest error. When n = 100, The Least Winsorized Square Estimator showed the least error followed by the Robust weighted least square estimator, while the S-estimator produced the highest error. For n = 200 and 500, the Robust weighted least square, showed the least error, followed by The Least Winsorized Square Estimator, while the RMSE of the other methods were very close with the S-estimator showing the highest error.

Table 2 : Results for two Independent Variables At 10% Contamination Rate and varying sample sizes

Sample size	n=20	n=20	n=50	n=50	n=100	n=100	n=200	n=200	n=500	n=500
METHOD	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
RWLSE	0.71058	**7.52775**	0.8246	**7.60618**	0.79779	**7.27198**	0.74319	**6.67856**	0.71975	**4.8603**
LWS	0.84857	7.547	0.7547	7.7697	0.8856	7.3243	0.81665	6.8821	0.60127	4.86763

LAV	0.8979	7.89901	0.9044	7.884	0.98411	7.46851	0.98678	10.9728	0.7692	5.57296
TB	0.93342	7.67638	0.9764	24.8844	0.89904	17.2682	0.84349	11.9751	0.98564	7.57296
MM	0.91007	7.64702	0.824	24.8843	0.89779	17.2682	0.87431	11.9752	0.71978	7.57296
LTS	0.95855	10.1088	0.9856	25.1588	0.98589	17.2728	0.98548	11.9751	0.98724	7.57454
S-	0.97123	8.87683	0.9271	25.1664	0.97388	17.301	0.98169	11.9784	0.98448	7.57338

From Table 3.2, for CR = 10%, the Robust weighted least square produced the least root mean square error (RMSE), closely followed by the least Winzorized square methods while the method that produced the highest RMSE were the least trimmed square and S- estimators across the sample sizes considered

.

Table 3: Results for five Independent Variables At 2% Contamination Rate and varying sample sizes

Sample size	n=20	n=20	n=50	n=50	n=100	n=100	n=200	n=200	n=500	n=500
METHOD	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
RWLSE	0.71231	7.40146	0.7051	6.30208	0.76741	5.63835	0.803	4.52476	0.89891	3.12379
LWS	0.73917	7.41438	0.77983	6.41334	0.8099	5.76959	0.7553	4.56407	0.88701	3.1475
LAV	0.84417	8.8059	0.98846	6.88158	0.9896	6.16819	0.9898	6.28498	0.87629	3.4753
TB	0.96985	18.8111	0.89838	13.0822	0.96353	9.16873	0.7296	6.28548	0.99009	3.9754
MM	0.91217	18.806	0.9051	13.0822	0.86741	9.16873	0.803	6.28548	0.89895	3.9753
LTS	0.98909	19.0679	0.9902	13.0995	0.99087	9.17508	0.991	6.29071	0.99135	3.9754
S-	0.93686	19.037	0.98301	13.1163	0.98788	9.1884	0.9891	6.30877	0.98987	3.981

From Table 3.3, at 2% contamination rate, when n= 20, RWLSE performed best with the lowest value of RMSE , it was closely followed by LWS. LTS performed least due to its high value of RMSE. The performance were the same when sample size was increased to 50, 100,200 and 500. RWLSE and LWS performed best with lowest RMSE values. The method that produced the highest RMSE here was S- estimator. Hence, the least performance.

Table 4 : Results For Five Independent Variables At 10% Contamination Rate and varying sample sizes

Sample size	n=20	n=20	n=50	n=50	n=100	n=100	n=200	n=200	n=500	n=500
METHOD	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RME**	**R^2**	**RMSE**
RWLSE	0.761	7.0272	0.4561	7.1472	0.5326	6.7488	0.67135	5.61	0.83437	4.10671
LWS	0.845	7.0303	0.70043	7.1537	0.5262	6.7995	0.6068	5.7607	0.7838	4.10881
LAV	0.834	7.5657	0.9899	7.1693	0.9788	7.2007	0.5622	9.182	0.76462	4.80268
TB	0.803	19.073	0.8266	19.069	0.9898	13.2016	0.9899	9.1831	0.99004	5.80283
MM	0.861	18.565	0.7561	19.0694	0.7326	13.2	0.67135	9.182	0.83437	5.80268
LTS	0.934	23.574	0.99112	19.075	0.9907	13.2025	0.9909	9.185	0.99125	5.80295
S-	0.796	22.036	0.9787	19.108	0.9865	13.204	0.98871	9.1833	0.9896	5.81044

From Table 4 at 10% contamination rate, RWLSE performed best with least RMSE. This performance was closely followed by LWS. LTS performed least in this category with highest RMSE. When the sample size increased to 50 and 100, RWLSE still had the best performance with lowest RMSE and still followed closely by LWS. S- estimator did not perform well in this category. When the sample sizes were 200 and 500, RWLSE and LWS performed best while LTS and S-estimator performed least .The value of Coefficient of Determination ($R^{2)}$ which is above 0.68 in all categories considered indicates that more than 68% of the variability in the dependent variable has been accounted for by the models.

Table 5 : Results from the real data application

METHODS	R-SQUARED	RMSE
Robust Weighted Least Squares Estimator	0.7734697	**7.74547**
Least Winsorized Square	0.7118966	7.74947
Least Trimmed Square	0.7728831	7.78537
Tukey's Bisquare	0.8854879	14.78539
MM Estimator	0.8734697	14.78539
Least Absolute Value	0.7775929	15.82905
S Estimator	0.706352	17.47265

From Table 3.5, Robust Weighted Least Squares performed best with lowest RMSE of 7.74547 and closely followed by Least Winsorized Square (LWS) with RMSE of 7.74947. S- estimator did not do well here with the highest RMSE of 17.47265. R^2 values of which none is less than 76% (that is, 0.7646) shows that at least 76% of total variation in the dependent variable which is death rate in this case can be explained by the independent variables which are the average annual precipitation, the average January temperature, the average July temperature, the size of the population older than 65 and the number of years of schooling for persons over 22 respectively.

Discussion of Results

The findings of the result revealed that for Contamination Rate (CR)of 2%,, with small sample size, the least Winsorized square (LWS) method produced the least root mean square error (RMSE), closely followed by the Re-Weighted Least Square Estimator (RWSLE) while the method that produced the highest RMSE was The Least Trimmed Square Estimator. The findings also revealed that at n = 50 and

100 the least Winsorized square method still showed least error followed by the RWSLE while the S-estimator produced the highest error. In the findings as the sample sizes increased to n = 200 and 500, the Robust weighted least square, showed the least error, followed by The Least Winsorized Square Estimator ,while the RMSE of the other methods were very close with the S-estimator showing the highest error.

The finding revealed that at contamination rate of 10%, the Robust weighted least square produced the least root mean square error (RMSE), closely followed by The Least Winzorized Square method, while the method that produced the highest RMSE was The Least Trimmed Square Estimator across the sample sizes considered. The Robust weighted least square and The Least Winsorized Square Estimator produced lesser R-squared than the rest of the methods across the sample sizes considered.

When the number of independent variables is 5, The Robust Weighted Least Square produced the least root mean square error (RMSE), closely followed by The Least Winzorized Square method, while the method that produced the highest RMSE was the least trimmed square estimator across the sample sizes considered. From the application to the real life data (n = 50, 5 independent variables), it can be seen that the results was backed up by the results from the simulation experiments. The Robust Weighted Least Square and The Least Winsorised Square method performed better than the rest of the methods, while the S-estimator did not perform well like the rest of the methods.

Conclusion

Based on the results from simulations and the real life data, it can be concluded that the best method for performing robust regression in presence of outliers, is The Robust Weighted Least Square method followed by The Least Winsorized Square method. Both methods produced minimal RMSE when compared to other methods considered for small and large samples and for 2 and 5 independent variables.

Recommendation

Based on the findings of the study, it was recommended

1. The Least Winsorized Square Estimator should be considered when the sample size is small with 2% contamination rate.

2. Robust weighted least square method should be preferred when considering robust regression in the presence of outliers when the sample size is from 200.

3. Also the Robust weighted least square should be considered when the variable is fro 5 upwards

REFERENCES

Arya,K.V., Gupta, P., Kalra P.K. & Mitra,P. (2007). Image registration using robust M-estimators. *Science direct pattern recognition letters*. 28, 1957-1968.

Bello A.R, Robiah Adnan Seyed E.S. & Kafi D.P. (2014). Robust Weighted Least of Regression Parameter in the presence of outliers and heteroscedastic errors. *Journal Teknologi* 7,(1), 11-20.

Chen D, Lu CT, Kou Y, & Chen F (2008). On detecting spatial outliers. *Geoinformatica* 12,(4), 24-33

Freedman, D. A. (2005). *Linear statistical models for causation:* A critical review. In B Everitt and D Howell, eds. Wiley Encyclopedia of Statistics in Behavioral Science,221-234

Gervini D. & Yohai V.J. (2002). A class of Robust and fully Efficient Regression Estimators. *The Annals of Statistics*.30, (2). 583-616.

Helmut, S. (1991). *Mathematical algorithm for Linear Regression*. Academic Press

Holland P.W & Welsch R. E. (1977).Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods* 6, (9).

Rousseeuw, P.J. & Yohai, V. (1984). *Robust Regression by Means of S estimators*, in Robustand Nonlinear Time Series Analysis, edited by J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics 26, Springer Verlag, New York,

Scott, A. J. (2012). Illusions on Regression Analysis . *International Journal of Forecasting(forthcoming).* 1, 001-009.

Tukey J.W., (1977). *Exploratory Data Analysis*. Addison – Wesley Publishers New York.

Yale C., & Forsythe, A.B. (1976). Winsorized Regression *Technometrics.* 2,(1), 291- 300.

Yohai, V.J. (1987), High Breakdown Point and High Efficiency Robust Estimates for Regression.

The *Annals of Statistics, 15,(2,)25-34*