FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 6 No. 1, March, 2022, pp 368 - 376
DOI: https://doi.org/10.33003/fjs-2022-0601-908

COMPARATIVE STUDY OF SOME ESTIMATORS OF LINEAR REGRESSION MODELS IN THE PRESENCE OF OUTLIERS

***[1]Ibrahim, A., [2]Dike, I. J., [1]Badawaire, A. B.**
[1]Department of Mathematics and Statistics Federal University, Wukari
[2]Department of Statistics and Operations Research Modibbo Adama University, Yola

*Corresponding authors' email: **abdulmudallib@fuwukari.edu.ng** Phone: (+234) 07062418865

ABSTRACT
The paper examined the performance of five estimation methods using six different outlier percentages (0%, 5%, 10%, 20%, 30% and 40%) and five different sample sizes (20, 40, 60, 100 and 200) were used to investigate effect of sample size on the performance of each of the estimation methods. The study adopted absolute bias, variances, relative efficiency and root mean square errors as comparison criteria through Monte-Carlo experiment and real life data was used to validate the simulation results. The study found that, under 5%, 10%, 20% and 30% outlying condition Robust-MM is the most preferred estimator across all criteria and sample size except using relative efficiency criterion and when the sample size is 40, 200 and 200 under 5%, 20% and 30% outlying condition and using absolute bias criterion respectively while Robust-LTS is the least preferred estimator except when the sample size is 40, 20 ; 40, 20 ; 20, 200 under 5%, 20% and 30% outliers and using absolute bias, variance and root mean square error respectively. Under 40% outlying condition Robust-MM is the most preferred estimator across all criteria and sample size except using relative efficiency and when the sample size is 20. Furthermore, Robust-MM is the most consistent estimator across the comparison criteria except when using relative efficiency and sample size has little or no effect on the performance of the estimators across all the different outlier levels. R Statistical package was used for the data analysis. This study therefore recommends the used of Robust-MM estimator.

Keywords: Estimation, Estimator, Outliers, Performance, Regression, Robust

INTRODUCTION
Stephen and Senthamarai (2017) defined Regression analysis as a statistical technique for analyzing and modeling the relationship between dependent variable and one or more independent variables. This technique uses the mathematical equation to establish the relationship between variables. It is a predictive modeling technique used for forecasting and to find causal effect relationship between the variables. Outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error, the latter are sometimes excluded from the dataset. An outlier can cause serious problem in statistical analysis (Zimek and Filzsomer 2018). Outlier detection has many applications, such as data cleaning, fraud detection and network intrusion. The existence of outliers can indicate individuals or groups that have behavior very different from most of the dataset. Frequently, outliers are removed to improve accuracy of the estimators. But sometimes the presence of an outlier has a certain meaning, which explanation can be lost if the outlier is deleted (Hawkins 1980).

The Classical Linear Regression Model
The classical linear regression model is a statistical model that describes a data generation process.
The classical linear regression model of the form

$$Y = X\beta + \varepsilon \qquad (1)$$

Where Y is an $n \times 1$ vector of observed response values, X is the $n \times p$ matrix of the predictor variables, β is the $p \times 1$ vector which contains the unknown parameters and needs to be estimated, and ε is the $n \times 1$ vector of random error terms.

Assumptions of classical linear regression model
(1) The dependent variable is linearly related to the coefficients of the model and the model is correctly specified. i.e. $\mathbf{Y = X\beta + \varepsilon}$
(2) The independent variable(s) is/are uncorrelated with the equation error term. i.e. Cov($\mathbf{X,\varepsilon}$) =0
(3) The mean of the error term is zero. i.e. E($\boldsymbol{\varepsilon}$) = 0
(4) The error term has a constant variance (Homoscedastic error). i.e. var ($\boldsymbol{\varepsilon}$) = $\sigma^2\mathbf{I}$
(5) The error terms are uncorrelated with each other. No autocorrelation or serial correlation. i.e. cov($\varepsilon_i,\varepsilon_j$) = 0 for all $i \neq j$
(6) There is no perfect linear relationship between the independent variables.
(7) The error term is normally distributed.
(8) There is absence of outlier in the dataset

Applying Ordinary Least Squares Estimators (OLSE) in simple or multiple linear regressions always calls for some assumptions: normality of the error terms; equal variance of the error terms, and absence of outliers, leverage points and Multicollinearity. According to Hampel (2001) and Huber (1972), normality of the error distributions finds its basis from the central limit theorem; which is a limit theorem based on approximations. Additionally, outliers in the dependent variable, lead to large residual values which further results in the failure of the normality assumption of the error terms. Therefore, in regression analysis, the ordinary Least Squares estimation is the best method if the assumptions are met. However, if these assumptions are not satisfied, the results can easily be affected (Alma, 2011). One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis (Williams*et.al* 2002; Liu *et.al* 2004).

Among methods used in detecting the presence of outlier are graphical methods and scatter plot. In the situation where the assumptions of the linear regression are not met, robust regression estimator is an important estimation technique for analyzing data that are contaminated with outliers or data with non normal error term.

Many estimation methods such as Least Trimmed Squares Estimator (LTSE), M-Estimator (ME), S-Estimator (SE), Modified Maximum Likelihood Estimator (MMLE) have been proposed which are more efficient than the Ordinary Least Squares (OLS) when there is outlier.

In regression analysis, the application of ordinary least squares method works well if the assumptions of the regression model, variables and the error terms are met. However, the presence of outliers or failure of the assumptions renders the ordinary least squares method of estimation unreliable. This is because bad leverage points, vertical outliers and good leverage points can influence the coefficients in the model, the residuals, as well as the standard errors of the model and the coefficients (David, 2014). Several estimation procedures have been proposed in literature to handle the problem of outliers during parameter estimation. Therefore, the paper examined the performances of five robust estimators using different percentages of outlier conditions with varying sample sizes say; 20, 40, 60, 100 and 200 and identified the most preferred estimator based on each of the selected comparison criteria.

Least Absolute Deviation (LAD)

This estimator obtains a higher efficiency, instead of minimizing the sum of squared errors; it minimizes the sum of absolute values of errors. The LAD method is not sensitive to outliers and produces robust estimates, (DasGupta and Mishra, 2004).

$$\min\sum_{i=1}^{n}|e_i| \quad (2)$$

$$\min\sum_{i=1}^{n}|y_i - x_i b| \quad (3)$$

M-Estimator (ME)

One of the robust regression estimation methods is the M estimation. M estimation is an estimation of the maximum likelihood type. M-estimation is an extension of the maximum likelihood estimate method and a robust estimation.

$$\hat{\beta}_1 = (X^T W X)^{-1} X^T W y \quad (4)$$

Least Trimmed Squares Estimator (LTSE)

Rousseeuw (1984) developed the least trimmed squares estimator (LTSE) given by,

$$\hat{\beta} = \min\sum_{i=1}^{h}(e_i^2) \quad (5)$$

Where $q = [n\ (1 - \alpha) + 1]$ is the number of observations included in the calculation of the estimator, and $\boldsymbol{\alpha}$ is the proportion of trimming that is performed. Using $q = \left(\frac{n}{2}\right) + 2$ ensures that the estimator has a breakdown point of 50%.

S – Estimation (SE)

The S-estimation is a high breakdown method introduced by Rousseeuw and Yohai (1984) that minimizes the dispersion of the residuals. The S-estimator was introduced to take care of the low breakdown point of the M-estimators.

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right) = k \quad (6)$$

MM- Estimation (MME)

MM estimation is a special type of M-estimation developed by Yohai (1987). MM-estimation is a combination of high breakdown value estimation and efficient estimation. MM estimator was the first estimate with a high breakdown point and high efficiency under normal error.

$$\frac{1}{n-1}\sum_{i=1}^{n}\rho\left(\frac{y_i - X_i\beta}{s}\right) = 0.5 \quad (7)$$

MATERIALS AND METHOD

Data Generation and Model Formulation Procedure

The dataset used for this study was simulated using Monte-Carlo in the environment of R statistical package (www.cran.org).

Mechanism for Generating the Independent Variables

In this study, three regressors were simulated, where two of the regressors were simulated to be normally distributed with mean zero and variance 1 and the other one was simulated with different percentages of outlier. The procedure is:

$$X_{ti} \sim N(0,1)$$

Where t = 1,2, 3,...,*n*;*i*= 1,2.

$$X_{t3} \sim (1 - n1\%)N(0,1) + n1\%N(0,500)$$

Where $n1\% \in \{0\%, 5\%, 10\%, 20\%, 30\%\ ,40\%\}$ is the percentage of outliers to be injected in the third predictor variable.

Mechanism for simulating Model with Outlier(s) in the Error Term.

The error term ε_t was simulated to be distributed according to a Gaussian mixture, i.e $\varepsilon_t \sim (1 - n1\%)N(0,1) + n1\%N(0,500)$.

Where $n1\% \in \{0\%, 5\%, 10\%, 20\%, 30\%, 40\%\}$ is the percentage of outliers injected in the error term.

Mechanism for Generating the Dependent Variable

The dependent variable Y_t was simulated to be distributed according to a Gaussian mixture, i.e $Y_t \sim (1 - n1\%)N(0,1) + n1\%N(0,500)$

The response variable is obtained from the relation given by:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_p X_{tp} + \varepsilon_t \quad (8)$$

$t = 1, \ldots, n\ and\ i = 1,2,3$

Data Simulation

A Monte-Carlo experiment of 1000 trials was carried out for five sample sizes (20, 40, 60, 100 and 200) each with different percentage of outliers (0%, 5%, 10%, 20%, 30%, 40%). Five Robust estimators were used to estimate parameters that were fitted to this simulated data. Real life data was used to validate our findings from the simulation study.

Criteria for Evaluating the Estimators

The assessments of the estimators considered in this work was based on the following criteria.

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). The formula is given by:

$$\text{RMSE}=\sqrt{E(\hat{\beta}-\beta)^2}=\sqrt{MSE}\ (9) \quad (9)$$

Bias

The bias is measured by

Bias= $\hat{\beta}-\beta$

The smaller the bias of an estimator the better.

Variance

The variance of an estimator $\hat{\beta}$ for β is defined as

$$Var\left(\hat{\beta}\right)=E\left[\hat{\beta}-E(\hat{\beta})\right]^2 \quad (10)$$

Relative efficiency of an estimator

The efficiency of an estimator is its minimum possible variance to its actual variance, and this estimator is efficient when the ratio gives one 1.

$$Efficiency(T_1 T_2)=\frac{E(T_1-\beta)^2}{E(T_2-\beta)^2} \quad (11)$$

RESULTS AND DISCUSSION

Table 1: RMSE of Coefficient estimates from the various estimators under different outlying condition

Outlier percentage	Sample Size	Estimator Robust -M	Robust MM	Robust-S	Robust-LTS	Robust-LAD
0	20	0.07004922	0.07270832	0.20603092	0.28988819	0.09925639
	40	0.03058571	0.03111030	0.09679149	0.14748303	0.04427831
	60	0.02084533	0.02109373	0.07648414	0.11972958	0.03112109
	100	0.01157941	0.01166778	0.04406762	0.07257024	0.01709462
	200	0.005610022	0.005630804	0.026097072	0.045671852	0.008255684
5	20	0.26966038	0.06385605	0.15927430	0.25197156	0.31758160
	40	0.07241582	0.02659153	0.08147891	0.13592261	0.07735246
	60	0.02268040	0.01901485	0.06164533	0.10200748	0.02958008
	100	1.166247e-02	9.726048e-03	3.802088e-02	6.413856e-02	1.537842e-02
	200	0.004944823	0.004451795	0.022944731	0.035267729	0.006745539
10	20	1.464509e-01	6.517039e-02	1.417662e-01	2.225894e-01	1.625383e-01
	40	0.04416410	0.02641450	0.07199517	0.11976373	0.05211716
	60	0.02448715	0.01746641	0.05508193	0.09259277	0.02934745
	100	0.01436604	0.01038389	0.03555881	0.05877127	0.01761216
	200	0.006166376	0.004519597	0.022055986	0.031988469	0.007735793
20	20	5.294326e+01	7.083667e-02	1.191625e-01	1.917030e-01	1.062558e+00
	40	8.425618e-02	2.870143e-02	6.404264e-02	1.060874e-01	5.918696e-02
	60	0.04045459	0.01928232	0.04777722	0.07835684	0.03753909
	100	0.02227780	0.01067995	0.03406066	0.05376324	0.02160283
	200	1.017204e-02	5.174712e-03	2.136545e-02	2.937123e-02	1.006707e-02
30	20	4.214707e+02	8.618141e-02	1.097780e-01	1.686087e-01	3.582085e+01
	40	1.357361e+01	3.357638e-02	6.392138e-02	8.931977e-02	1.286937e+00
	60	4.152435e-01	4.152435e-01	4.673393e-02	7.085362e-02	5.534586e-02
	100	0.05794300	0.01278831	0.03481236	0.04555237	0.02959935
	200	1.995946e-02	5.918878e-03	2.412938e-02	2.568796e-02	1.337968e-02
40	20	1920.8596992	444.4912523	19.3950962	0.1101777	438.8606465
	40	2.458574e+02	4.861703e-02	7.229440e-02	6.901251e-02	2.994321e+00
	60	5.435109e+01	3.171891e-02	5.615444e-02	5.299854e-02	1.019711e-01
	100	6.27156950	0.01584252	0.04143840	0.03632329	0.04352504
	200	6.032946e-02	7.280338e-03	2.931028e-02	2.199554e-02	1.850706e-02

Table 2: Bias of Coefficient estimates from the various estimators under different outlying condition

Outlier percentage	Sample Size	Estimator Robust -M	Robust MM	Robust-S	Robust-LTS	Robust-LAD
0	20	0.2093655	0.2126916	0.3539176	0.4124872	0.2499738
	40	0.1395238	0.1403076	0.2455689	0.3006825	0.1674017
	60	0.1160409	0.1167313	0.2193111	0.2757462	0.1413239
	100	0.08580933	0.08605832	0.16643276	0.21315400	0.10401642
	200	0.05969014	0.05977360	0.12845390	0.17035070	0.07242573
	20	0.2092223	0.1808144	0.2851223	0.3687487	0.2375808
	40	0.1351381	0.1154182	0.2021793	0.2687385	0.1519090

5	60	0.10205656	0.09747658	0.17450052	0.23092976	0.11648561
	100	0.07356755	0.06856203	0.13554200	0.18086474	0.08546150
	200	0.04856831	0.04603198	0.10627630	0.13217748	0.05652200
10	20	0.2337176	0.1769880	0.2606601	0.3342456	0.2522402
	40	0.1385725	0.1122347	0.1854661	0.2429825	0.1524827
	60	0.10750955	0.09079325	0.16198870	0.21243171	0.21083653
	100	0.08296108	0.07093100	0.13078178	0.16680495	0.09198731
	200	0.05485178	0.04676106	0.10385016	0.12453398	0.06115979
20	20	1.2997506	0.1828762	0.2376238	0.3048431	0.3645741
	40	0.1859330	0.1179816	0.1746408	0.2234484	0.1671327
	60	0.13742655	0.09609987	0.15031883	0.19183251	0.13330738
	100	0.10258125	0.07051206	0.12733506	0.15971273	0.10067426
	200	0.06963586	0.04967092	0.10172848	0.11863076	0.06953510
30	20	7.5544496	0.2001701	0.2243135	0.2747414	1.0190594
	40	0.7290462	0.1259629	0.1740215	0.2057689	0.2341343
	60	0.2558604	0.1037535	0.1510383	0.1827968	0.1626338
	100	0.15344382	0.07810156	0.12822900	0.14684268	0.11888091
	200	0.09701464	0.05338986	0.10800531	0.11096562	0.07965919
40	20	23.1747599	9.0524219	0.7865547	0.2272805	5.6218269
	40	6.2673208	0.1491436	0.1854998	0.1797782	0.3621040
	60	2.1566678	0.1220350	0.1623518	0.1573657	0.2019440
	100	0.53797981	0.08636101	0.14056027	0.13142098	0.14224803
	200	0.16250669	0.05885506	0.11940789	0.10224310	0.09379611

Table 3: Variance of Coefficient estimates from the various estimators under different outlying condition

Outlier percentage	Sample Size	Estimator				
		Robust -M	**Robust MM**	**Robust-S**	**Robust-LTS**	**Robust-LAD**
0	20	0.02578261	0.02704752	0.07948785	0.11809840	0.03613602
	40	0.01106449	0.01136920	0.03637458	0.05682909	0.01619366
	60	0.007346661	0.007432610	0.028297225	0.043510126	0.011103152
	100	0.004214050	0.004259672	0.016327368	0.027113597	0.006268764
	200	0.002043017	0.002053880	0.009593546	0.016641168	0.003007219
5	20	0.22092921	0.02658536	0.06723951	0.10814025	0.25342874
	40	0.05255633	0.01095720	0.03374572	0.05720576	0.05165415
	60	0.009607724	0.007524900	0.024471860	0.043088983	0.012414408
	100	0.004690252	0.003698088	0.014454423	0.026015922	0.005908462
	200	0.001819106	0.001684846	0.008252968	0.013908275	0.002509478
10	20	0.08416814	0.02690145	0.05757122	0.09385618	0.08944693
	40	0.02086267	0.01043085	0.02780874	0.05017417	0.02369002
	60	9.324749e-03	6.709055e-03	2.066250e-02	3.615056e-02	1.103618e-02
	100	0.005246410	0.003694931	0.012827771	0.022814382	0.006401544
	200	0.002170412	0.001616775	0.007762443	0.011431618	0.002766718
20	20	5.071262e+01	2.716386e-02	4.484080e-02	7.316037e-02	9.310093e-01
	40	0.03918089	0.01014939	0.02340542	0.04092913	0.02287583
	60	1.536502e-02	6.948244e-03	1.760075e-02	2.922820e-02	1.391265e-02
	100	8.269337e-03	4.058920e-03	1.245086e-02	1.976530e-02	8.120719e-03
	200	0.003718949	0.001890064	0.007604865	0.010627911	0.003633944
30	20	3.439819e+02	3.238792e-02	4.235182e-02	6.808967e-02	3.443174e+01
	40	12.85508379	0.01244392	0.02352072	0.03278415	1.21380707
	60	3.275425e-01	8.176777e-03	1.633352e-02	2.618610e-02	2.014467e-02
	100	2.658090e-02	4.633651e-03	1.284705e-02	1.672554e-02	1.072177e-02
	200	7.411217e-03	2.122629e-03	8.599894e-03	9.203125e-03	4.917510e-03
40	20	1.199181e+03	3.346673e+02	1.853301e+01	4.098227e-02	3.961937e+02
	40	193.17325883	0.01890083	0.02642862	0.02562976	2.81868165
	60	48.08441510	0.01179076	0.02092649	0.01969546	0.04738662
	100	5.886092e+00	5.893713e-03	1.506998e-02	1.303252e-02	1.656233e-02
	200	2.509650e-02	2.662694e-03	1.033502e-02	7.878468e-03	6.773354e-03

Table 4: Relative efficiency of Coefficient estimates from the various estimators under different outlying condition

Outlier percentage	Sample Size	Robust -M	Robust MM	Robust-S	Robust-LTS	Robust-LAD
	20	0.9502803	0.9142981	0.3234201	0.2313309	0.6695599
	40	0.9499412	0.9340235	0.3005521	0.1976194	0.6566619
0	60	0.9310824	0.9202790	0.2543356	0.1629528	0.6236101
	100	0.9468383	0.9396493	0.2495488	0.1512752	0.6428818
	200	0.9467297	0.9432245	0.2035174	0.1162836	0.6426031
	20	4446.582	8857.413	3584.929	2427.893	3512.680
	40	7359.158	9558.256	3106.604	1943.479	5867.569
5	60	7048.575	8156.033	2470.017	1602.889	5261.541
	100	7025.968	8146.707	2071.548	1315.383	5182.173
	200	8910.908	8543.236	1652.990	1133.492	6819.636
	20	8532.333	14553.032	6591.848	4437.340	7353.182
	40	11051.584	16458.735	5926.395	3781.082	9119.084
10	60	11941.264	16964.892	5337.219	3265.661	9986.442
	100	10638.954	19357.338	5502.181	2629.700	8696.458
	200	14610.677	20372.211	3300.578	2275.875	11798.374
	20	38.57275	27782.48307	16464.34242	10312.08388	4430.82078
	40	9585.038	32851.025	14841.514	7579.035	13760.967
20	60	14288.304	38160.312	14981.560	7449.403	15389.818
	100	17693.580	38436.654	12256.200	7586.027	18759.457
	200	19233.635	38689.305	9109.595	6981.017	19763.103
	20	7.020023	31729.801559	24866.910084	16497.875374	79.019580
	40	109.8216	45961.3023	23741.3530	16671.9663	1919.3973
30	60	2483.55	48242.75	22932.21	15045.23	15311.28
	100	10886.24	50759.73	18258.09	13970.58	21124.41
	200	14628.87	49873.48	12132.63	11681.28	21876.89
	20	2.446423	9.903955	244.248127	34767.274756	9.580380
	40	8.20698	43732.11948	29897.32603	30891.11193	758.49617
40	60	24.21141	46000.41793	25561.36393	26549.30855	11046.33007
	100	148.2724	53936.2892	21078.4905	23644.7305	18537.1715
	200	6119.044	53661.753	13215.588	17823.239	20404.410

Assessing the Performances of the Estimators Using Various Criteria

Table 5: Rank of Performances of the estimators using RMSE criterion

Sample size	Estimator	outlier 0%	5%	10%	20%	30%	40%
	Robust-M	1	4	3	5	5	5
	Robust-MM	2	1	1	1	1	4
n=20	Robust-S	4	2	2	2	2	2
	Robust-LTS	5	3	5	3	3	1
	Robust-LAD	3	5	4	4	4	3
	Robust-M	1	2	2	4	5	5
	Robust-MM	2	1	1	1	1	1
n=40	Robust-S	4	4	4	3	2	3
	Robust-LTS	5	5	5	5	3	2
	Robust-LAD	3	3	3	2	4	4
	Robust-M	1	2	2	3	5	5
	Robust-MM	2	1	1	1	1	1
n=60	Robust-S	4	4	4	4	2	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	3	4
	Robust-M	1	2	2	3	5	5

	Robust-MM	2	1	1	1	1	1
n=100	Robust-S	4	4	4	4	3	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	2	4
	Robust-M	1	2	2	3	3	5
n=200	Robust-MM	2	1	1	1	1	1
	Robust-S	4	4	4	4	4	4
	Robust-LTS	5	5	5	5	5	3
	Robust-LAD	3	3	3	2	2	2

Table 6: Rank of performances of the estimators using absolute bias criterion

Sample size	Estimators	Outlier					
		0%	5%	10%	20%	30%	40%
	Robust-M	1	2	2	5	5	5
	Robust-MM	2	1	1	1	1	4
n=20	Robust-S	4	4	4	2	2	2
	Robust-LTS	5	5	5	3	3	1
	Robust-LAD	3	3	3	4	4	3
	Robust-M	1	1	2	4	5	5
	Robust-MM	2	4	1	1	1	1
n=40	Robust-S	4	5	4	3	2	3
	Robust-LTS	5	2	5	5	3	2
	Robust-LAD	3	3	3	2	4	4
n=60	Robust-M	1	2	2	3	5	5
	Robust-MM	2	1	1	1	1	1
	Robust-S	4	4	4	4	2	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	3	4
	Robust-M	1	2	2	3	5	5
	Robust-MM	2	1	1	1	1	1
n=100	Robust-S	4	4	4	4	3	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	2	4
	Robust-M	1	2	2	1	3	5
n=200	Robust-MM	2	1	1	4	1	1
	Robust-S	4	4	4	3	4	4
	Robust-LTS	5	5	5	5	5	3
	Robust-LAD	3	3	3	2	2	2

Table 7: Rank of performance of the estimators using Variance criterion

Sample size	Estimator	outlier					
		0%	5%	10%	20%	30%	40%
	Robust-M	1	4	3	5	5	5
	Robust-MM	2	1	1	1	1	3
	Robust-S	4	2	2	2	2	2
n=20	Robust-LTS	5	3	5	3	3	1
	Robust-LAD	3	5	4	4	4	4
	Robust-M	1	4	2	4	5	5
n=40	Robust-MM	2	1	1	1	1	1
	Robust-S	4	2	4	3	2	3
	Robust-LTS	5	5	5	5	3	2
	Robust-LAD	3	3	3	2	4	4
	Robust-M	1	2	2	3	5	5
n=60	Robust-MM	2	1	1	1	1	1

	Robust-S	4	4	4	4	2	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	3	4
n=100	Robust-M	1	2	2	3	5	5
	Robust-MM	2	1	1	1	1	1
	Robust-S	4	4	4	4	3	3
	Robust-LTS	5	5	5	5	4	2
	Robust-LAD	3	3	3	2	2	4
n=200	Robust-M	1	2	2	3	3	5
	Robust-MM	2	1	1	1	1	1
	Robust-S	4	4	4	4	4	4
	Robust-LTS	5	5	5	5	5	3
	Robust-LAD	3	3	3	2	2	2

Table 8: Rank of performances of the estimators using Relative Efficiency criterion

Sample size		Outlier					
	Estimator	0%	5%	10%	20%	30%	40%
n=20	Robust-M	5	4	4	1	1	1
	Robust-MM	4	5	5	5	5	5
	Robust-S	2	3	2	4	4	2
	Robust-LTS	1	1	1	3	3	4
	Robust-LAD	3	2	3	2	2	3
n=40	Robust-M	5	4	4	2	1	1
	Robust-MM	4	5	5	5	5	5
	Robust-S	2	2	2	4	4	3
	Robust-LTS	1	1	1	1	3	4
	Robust-LAD	3	3	3	3	2	2
n=60	Robust-M	5	4	4	2	1	1
	Robust-MM	4	5	5	5	5	5
	Robust-S	2	2	2	3	4	2
	Robust-LTS	1	1	1	1	2	4
	Robust-LAD	3	3	3	4	3	3
n=100	Robust-M	5	4	4	3	3	5
	Robust-MM	4	5	5	5	5	4
	Robust-S	2	2	2	2	2	2
	Robust-LTS	1	1	1	1	1	3
	Robust-LAD	3	3	3	4	4	1
n=200	Robust-M	5	5	4	3	3	1
	Robust-MM	4	4	5	5	5	5
	Robust-S	2	2	2	2	2	2
	Robust-LTS	1	1	1	1	1	3
	Robust-LAD	3	3	3	4	4	4

Table 9: Parameter Estimates of various Estimators under outlying condition using real life data

	Estimator				
	Robust -M	Robust –MM	Robust -S	Robust –LTS	Robust -LAD
β_0	173.3444913	133.7990838	63.3640509	78.3057784	141.1575281
β_1	0.9975838	1.3573958	1.7498547	1.5652388	1.2717276
β_2	1.1152763	0.9886164	0.8531579	0.8983310	1.0061706
β_3	-1.1158903	-1.0523911	-1.1621459	-0.9232399	-0.9359867

Table 10: Variance and Mean Square Error of Prediction (MSEP) of Coefficient estimates from the various estimators under outlying condition using real life data

	Estimator				
	Robust –M	Robust -MM	Robust –S	Robust -LTS	Robust –LAD
Variance	7484.3535	4447.8708	990.0747	987.0263	4950.8103
MSEP	21740.22	30862.24	35717.46	35415.3	29889.85

Assessing the Performance of Estimators under Outlying Condition Using Real Life Data

Table 11: Performance of estimators of real life data using variance and MSEP criterion

Criteria	Estimator	Rank
Variance	Robust-M	5
	Robust-MM	3
	Robust-S	2
	Robust-LTS	1
	Robust-LAD	4
MSEP	Robust-M	1
	Robust-MM	3
	Robust-S	5
	Robust-LTS	4
	Robust-LAD	2

Going by variance criterion, the estimator that best fit the dataset as evident from Table 11 is Robust-LTS estimator. Similarly, from Table 11 above it can be observed that Robust-M estimator is the best estimator in terms of predictions.

Based on the results about the estimators presented in Table 1 through Table 8, the paper found that under 0% outlying condition and using absolute bias, variance and root mean square error criterions, Robust-M is the most preferred estimator while Robust-LTS is the least Preferred at all sample sizes (20, 40, 60, 100, and 200). However, under 0% outlying condition and using Relative Efficiency criterion, Robust-LTS is the most preferred Estimator while Robust-M is the least preferred estimator at all sample sizes (20, 40, 60, 100, and 200).

In addition, under 5% outlying condition and using absolute bias, variance and root mean square error criterion, Robust –MM is the most preferred estimator except when the sample size is 40 using absolute bias criteria while Robust-LTS is the least preferred Estimator except when the sample size is 40 using absolute bias, 20 and 40 using variance and 20 using root mean square error criterion respectively.

However, under 5% outlying condition and using relative efficiency, Robust-LTS is the most preferred estimator across all sample size while Robust-MM is the least preferred estimator across all sample size except when the sample size is 200.

Under 10% outlying condition and using absolute bias, variance and root mean square error criteria, Robust-MM is the most preferred estimator while Robust-LTS is the least preferred Estimator across all sample size. However, under 10% outlying condition and using relative efficiency, Robust-LTS is the most preferred estimator while Robust-MM is the least preferred estimator across all sample size except when the sample size. However, under 20% outlying condition and using absolute bias, variance and root mean square error criteria, Robust –MM is the most preferred estimator across all sample size except when the sample size is 200 using absolute bias criterion while Robust-LTS is the least preferred Estimator across all sample size except when the sample size is 20.

Consequently, under 20% outlying condition and using relative efficiency, Robust-LTS is the most preferred estimator except when the sample size is 20 while Robust-MM is the least preferred estimator across all sample size. Using 30% outlying condition and using absolute bias, variance and root mean square error criteria, Robust –MM is the most preferred estimator across all sample size except when the sample size is 200 using absolute bias criterion while Robust-LTS is the least preferred Estimator across all sample size except when the sample size is 200 using root mean square error.

However, under 30% outlying condition and using relative efficiency, Robust-M and Robust-LTS are the most preferred estimators when the sample sizes are 20, 40 and 60 and 100 and 200 respectively while Robust-MM is the least preferred estimator across all sample size. Going 40% outlying condition and using absolute bias, variance and Root Mean Square Error criteria, Robust –MM is the most preferred estimator across all sample size except when the sample size is 20 using Root Mean Square Error criterion while Robust-M is the least preferred Estimator across all sample size.

Also, under 40% outlying condition and using relative efficiency, Robust-M is the most preferred estimators except when the sample size 100 while Robust-LTS and Robust-MM is the least preferred estimator at sample sizes 20 and 40, 60 and 200 respectively.

That sample size has little or no effect on the performance of the estimators across all the different outlier levels.

CONCLUSION

In conclusion, the study concludes that Robust-M is the most efficient estimator across all the comparison criteria in the absence of outlier except when using relative efficiency and that Robust-MM is the most consistent estimator across the comparison criteria except when using relative efficiency. Also, sample size has little or no effect on the performance of the estimators across all the different outlier levels.

REFERENCES

Alma, O. G. (2011). Comparison of robust regression methods in linear regression. *International Journal for contempMaths and Science*, 6: 409-421.

DasGupta, M. & Mishra, S.K. (2004). *Least absolute deviation estimation of linear econometric models: A literature review*. Available:http://mpra.ub.uni muenchen.de/7 81.

David D. (2014). *Comparison of Robust Regression Estimators*. M.phil thesis, Kwameh Nkurumah University of Science and technology, Ghana, Ghana.

Hampel, F. (2001). Robust statistics: A brief introduction and overview. Pages 1-5. In David D. (2014). Comparison of Robust Regression Estimators

Huber, P. J. (1972). The 1972 wald lecture: Robust statistics. The Annals of Mathematical Statistics, 43: 1041-1067. In David D. (2014). *Comparison of Robust Regression Estimator.*

Hawkins, D. (1980). Identification of Outliers. Chapman and Hall. London. In Edgar, A. & Caroline, R. (2004). On detection of outliers and their effect in supervised classification. *Conference paper.*

Liu H., Shah S. and Jiang W. (2004). On-line outlier detection and data cleaning. Computers and Chemical Engineering. In Maimon O. and Rockach L. (Eds.) (2005). Data Mining and Knowledge Discovery Handbook: *A Complete Guide for Practitioners and Researchers.*

Rousseeuw, P. J. (1984). *Robust Regression and Outlier Detection.*JOHN WILEY and SONS.

Rousseeuw, P. J. &Yohai, V. J. (1984). Robust Regression by Mean of S Estimators. *Robust and Nonlinear Time Series Analysis*, 256-274, doi: 10.1007/978-1-4615-7821-5-15.

Stephen, R. &Senthamarai, K. K. (2017). Detection of Outliers in Regression Model for Medical Data. *International Journal of Medical Research &Health Sciences*, **6**(7), 50-56.

Williams, G. J., Baxter, R. A., He H. X. & Hawkins S., Gu L. (2002). A Comparative Study of RNN for Outlier Detection in Data Mining. *IEEE International Conference on Data-mining.* In: Maimon O. andRockach L. (Eds.) (2005). Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.

Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, **15**(20), 642-656, doi:10.1214/aos/1176350366.

Zimek, A. &Filzsomer, P. (2018). "There and Back again: Outlier detection between statistical reasoning and data mining algorithm". *Wiley Interdisciplinary Reviews: Data mining and Knowledge Discovery*, **8**(6). Doi:10.1002/widm.1280