Zimbabwe Journal of Science & Technology pp 8 – 14 Vol.12 [2017]
e-ISSN 2409-0360
Zimbabwej.sci.technol

Some Robust Liu Estimators

[1]Adewale F. Lukman, [1]Kayode Ayinde, [2]Ajiboye S. Adegoke and [1]Daramola Tosin

[1]Department of Statistics, Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria.

[2]Department of Statistics, Federal University of Technology, Akure

Abstract

In a classical linear regression model, Liu and Robust Estimators were developed to deal with the problem of multicollinearity and outliers respectively. This paper proposes some robust Liu estimators (RLEs) to jointly address the problem of multicollinearity and outliers and illustrates the proposed estimators with real life data sets. Based on the performances of these estimators using the Mean Square Error criterion, results show that the Robust Liu Estimators perform better than the ordinary least square (OLS), Liu estimator and Robust estimators when data sets suffer both problems. Furthermore, it is observed that the M Robust Liu Estimator (MRLE) is most efficient when outliers are in the y-direction; and when outliers are in the x or both y and x direction, the LTS Robust Liu Estimator (LTSRLE) is most efficient.

Key words: Ordinary Least Square Estimator, Liu Estimator, Robust Estimators, Robust Liu Estimators.

Received:01.11.2016. Accepted: 10.04.2017.

1. INTRODUCTION

Consider the standard linear regression model in matrix form:

$$Y = X\beta + U \quad (1)$$

where X is an $n\times p$ matrix of n observations of p explanatory variables with full rank, Y is a $n\times 1$ vector of dependent variable, β is a $p\times 1$ vector of unknown parameters, and U is $n\times 1$ vector of error term such that $E(U) = 0$ and $E(UU') = \sigma^2 I_n$.

The Ordinary Least Squares (OLS) estimator is the most popularly used estimator to estimate the parameters of the linear regression model and it is Best Linear Unbiased Estimator (BLUE) when all the assumptions of classical linear regression model are satisfied (Aitken, 1935). The estimator is defined as:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \quad (2)$$

The performance of this estimator depends on the validity of some assumptions, one of which is on the state of the $X'X$ matrix. If the matrix is ill-conditioned due to linear relationship among explanatory variables, it results into multicollinearity problem. The OLS estimator, even though unbiased, has large variances and covariances which in turn make precise estimation difficult (Gujarati, 2003). Consequently, regression coefficients may exhibit wrong sign, may be statistically insignificant and confidence intervals tend to be wider leading to wrong conclusion. Several estimators are available in literature to circumvent this problem. This includes the Ordinary Ridge Regression (ORR) estimator proposed by Hoerl and Kennard (1970),

$$\hat{\beta}_{ORR} = (X'X + KI)^{-1}X'X \quad (3)$$

where K is the ridge parameter such that 0≤K≤1.

Stein (1960) defined another estimator as a linear function of the Ordinary Least Square (OLS) estimator to still handle multicollinearity. This is given as:

$$\hat{\beta}_s = K\hat{\beta}_{OLS} \quad (4)$$

where 0<K<1.

Liu (1993) combined the stein estimator with ORR estimator to combat multicollinearity. Liu estimator (LE) is defined for the biased parameter $d \in (-\infty, \infty)$ as follows:

$$\hat{\beta}_{LE} = (X'X + I)^{-1}(X'X + dI)\hat{\beta}_{OLS} \quad (5)$$

Another problem that affects the popular OLS estimator is the presence of outliers or leverage points. Outliers can be in the y or x direction or both. Robust regression estimators have been developed as an alternative to OLS to dampen the influence of outliers. Lists of these estimators are the M, MM, Least Trimmed Square (LTS), Least Absolute Deviation (LAD), Least Median Square (LMS) and S estimator.

Huber (1973) proposed the M estimator and is used extensively in analyzing data when there is outlier in the y-direction but it is not robust with respect to leverage points. The M-estimate objective function is

$$\min\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right)=\min\sum_{i=1}^{n}\rho\left(\frac{y_i\text{-}X'\hat{\beta}_i}{s}\right) \quad (6)$$

Where *s* is an estimate of scale often formed from linear combination of the residuals. The function ρ gives the contribution of each residual to the objective function.

Dielman (1984) introduced the LAD estimator which minimizes the sum of the absolute values of the residuals with respect to the coefficient vector β:

$$\min\sum_{i=1}^{n}|y_i - x_i\hat{\beta}|. \quad (7)$$

LAD is robust to an outlier in the y-direction. However, LAD estimator does not protect against outlying x (leverages).

S estimation is a high breakdown value method introduced by Rousseeuw and Yohai (1984). It minimizes the dispersion of the residuals. The dispersion $e_1(\theta), \ldots, e_n(\hat{\theta})$ is defined as the solution of:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right) = k \quad (8)$$

k is a constant and $\rho\left(\frac{e_i}{s}\right)$ is the residual function.

Yohai (1987) proposed the MM estimator by combining the high breakdown value estimation and M estimation. This estimator estimate the regression parameter using S estimation which minimize the scale of the residual from M estimation and then proceed with M estimation.

Rousseeuw (1998) introduced LTS estimator which is a high breakdown value method. LTS regression minimizes the sum of trimmed squared residuals. This estimator is given as:

$$\hat{\beta}_{LTS} = \operatorname{argmin} Q_{LTS}(\beta) \quad (9)$$

where $Q_{LTS}(\beta) = \sum_{i=1}^{h} e_i^2$ such that $e_{(1)}^2 \leq e_{(2)}^2 \leq e_{(3)}^2 \leq \cdots \leq e_{(n)}^2$ are the ordered squares residuals and h is defined in the range $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, with n and p being sample size and number of parameters respectively. The largest squared residuals are excluded from the summation in this method, which allows those outlier data points to be excluded completely.

In most econometric works, both problems jointly exist especially with time series data. Holland (1973) proposed robust M-estimator for ridge regression to handle the problem of multicollinearity and outliers. Samkar and Alpu (2010) proposed robust ridge regression methods based on M, S, MM and GM estimators. Lukman et al (2014) combined the ridge regression with some robust estimators such as M, MM, LTS, S, LAD and LMS estimator to handle these problems jointly. Ozlem and Hattice (2009) combined the Liu estimator with the M estimator to handle both problems.

In this study, to circumvent both problems jointly, some robust Liu estimators are proposed.

2. MATERIALS AND METHODS

2.1 Liu Estimator

The regression model in equation (1) can be written in the canonical form as:

$$Y=Z\alpha+\varepsilon \quad (10)$$

where Z=XP, $\alpha=P'\beta$. $X'X$ is symmetric matrix such that there exists a p×p orthogonal matrix P where $P'X'XP = \Lambda$, $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_p)$ such that $\lambda_1>\lambda_2>\cdots>\lambda_p$ (the eigenvalues). The OLS and LIU estimators for equation (10) in canonical form can be written respectively as:

$$\hat{\alpha}_{OLS} = \Lambda^{-1}Z'Y \quad (11)$$

and

$$\hat{\alpha}_{LE} = (\Lambda+I)^{-1}(\Lambda+dI)\,\hat{\alpha}_{OLS} \quad (12)$$

where d is the biasing parameter. Liu (1993) obtained this parameter by minimizing the mean square error of Liu estimator. This is defined as:

$$\hat{d}=1-\hat{\sigma}^2\left[\frac{\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}}{\sum_{i=1}^{p}\frac{\hat{\alpha}_i^2}{(\lambda_i+1)^2}}\right] \quad (13)$$

Where $\hat{\sigma}^2$ and $\hat{\alpha}$ are the mean square error and the regression estimates compute via OLS respectively.

2.2. Robust Liu Estimators

This method combines the liu and robust estimators such as M, MM, LTS, LAD to handle the problem of multicollinearity and outliers/leverage point simultaneously. This is supposed to dampen the effects of both problems in a classical linear regression model. The Robust Liu Estimators based on M, MM, LTS and LAD are defined respectively as:

$$\hat{\alpha}_{MRLE} = (\Lambda+I)^{-1}(\Lambda+d_M I)\,\hat{\alpha}_M \quad (14)$$

$$\hat{\alpha}_{MMRLE} = (\Lambda+I)^{-1}(\Lambda+d_{MM} I)\,\hat{\alpha}_{MM} \quad (15)$$

$$\hat{\alpha}_{LTSRLE} = (\Lambda+I)^{-1}(\Lambda+d_{LTS} I)\,\hat{\alpha}_{LTS} \quad (16)$$

$$\hat{\alpha}_{SRLE} = (\Lambda+I)^{-1}(\Lambda+d_S I)\,\hat{\alpha}_S \quad (17)$$

$$\hat{\alpha}_{LADRLE} = (\Lambda+I)^{-1}(\Lambda+d_{LAD} I)\,\hat{\alpha}_{LAD} \quad (18)$$

where each of the regression estimates and the biasing parameters are obtained using the robust estimates as alternative to OLS estimates. For instance, the robust biasing parameter for equation (14) is defined as:

$$\hat{d}_M=1-\hat{\sigma}_M^2\left[\frac{\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}}{\sum_{i=1}^{p}\frac{\hat{\alpha}_{Mi}^2}{(\lambda_i+1)^2}}\right] \quad (19)$$

3. CRITERION FOR COMPARISON

The mean square error is used to compare the estimators together to identify the most efficient of them.

$$\text{MSE}(\hat{\alpha}_{OLS}) = \hat{\sigma}^2\sum_{i=1}^{p}\frac{1}{\lambda_i} \quad (20)$$

$$\text{MSE}(\hat{\alpha}_{LE}) = \hat{\sigma}^2\sum_{i=1}^{P}\frac{(\lambda_i+d)^2}{(\lambda_i+1)^2} + (d-1)^2\sum_{i=1}^{P}\frac{\hat{\alpha}_i^2}{(\lambda_i+1)^2} \quad (21)$$

The mean square error of each of the robust estimators and robust Liu estimators are obtained by replacing $\hat{\sigma}^2$ and $\hat{\alpha}_i^2$ in equation (20) and (21) with their respective robust version.

4. NUMERICAL EXAMPLES

Example 1. Longley data

To evaluate the performance of these estimators, we consider the widely analyzed Longley dataset (Longley, 1967). It consists of six economic variables related to total derived employment from 1947 to 1962. The data has been used by some authors to explain the effect of multicollinearity on OLS estimator and also to check influential points. Aboobacker and Jianbao (2011) concluded that the data suffers multicollinearity since the condition number is 43275. Cook (1977) used the same dataset in detecting influential observation in a linear regression model using the method of Cook's *D* and found that cases 5, 16, 4, 10 and 15 (in this order) were the most influential observations. Ayinde et al (2015) carried out diagnostic checks on the presence of outlier and presented the summary as given in Table 1:

Table 1. Summary of outlier results in terms of standardized residual using Longley data

Estimators	Outliers
OLS	10
M	10, 14,15,16
MM	14,15,16
S	14,15,16
LTS	5,14,15,16

Source: Ayinde et al*,* 2015.

The result revealed that there are outliers in the y-direction but no leverage point. Hence, it is evident that the dataset suffers both the problem of multicollinearity and outlier simultaneously. The results of the OLS and robust estimators are given in Table 2 while that of the Liu and Robust Liu estimators are given in Table 3.

Table 2. Estimates of OLS and robust estimators of Longley data

Coefficient	OLS	M	MM	LTS	S	LAD
$\hat{\alpha}_1$	0.1548	0.1547	0.1547	0.1547	0.1549	0.1549
$\hat{\alpha}_2$	-0.5494	-0.5496	-0.5495	-0.5458	-0.5448	-0.5503
$\hat{\alpha}_3$	0.8455	0.8156	0.8351	0.7562	0.7092	0.8639
$\hat{\alpha}_4$	1.0138	0.9347	0.9784	0.9897	1.0413	0.9347
$\hat{\alpha}_5$	42.6115	37.3772	40.4433	19.1757	13.2611	33.8234
$\hat{\alpha}_6$	-57.7536	-25.0120	-42.7171	-81.9413	-72.3343	-74.8478
$\hat{\sigma}^2$	225783	97706.63	198916	99874.96	198737.6	249594.16
MSE($\hat{\alpha}$)	17095.12	7397.87	15060.94	7562.04	15047.43	18898.04

Table 3. Estimates of OLS, Liu and Robust Liu estimators of Longley data

Coefficient	Liu	MRLE	MMRLE	LTSRLE	SRLE	LADRLE
$\hat{\alpha}_1$	0.1548	0.1547	0.1547	0.1547	0.1547	0.15490
$\hat{\alpha}_2$	-0.5494	-0.5496	-0.5496	-0.5458	-0.5496	-0.55030
$\hat{\alpha}_3$	0.8455	0.8156	0.8225	0.7520	0.8157	0.86390
$\hat{\alpha}_4$	1.0138	0.9347	0.9481	0.9897	0.9348	0.93470
$\hat{\alpha}_5$	42.5870	37.3557	38.3467	19.1647	37.3698	33.80376
$\hat{\alpha}_6$	-53.7192	-23.2670	-28.1515	-76.2155	-23.3325	-69.61868
D	0.00045865	0.00017184	0.00058920	0.00016013	0.0017088	0.0002055
MSE($\hat{\alpha}$)	14818.97	6412.16	13054.71	6582.04	13058.91	16395.90

From Table 2 and 3, the regression estimates of OLS and Liu are not too different except in two of the regression estimates, $\hat{\alpha}_5$ and $\hat{\alpha}_6$. However, in terms of the mean square error (MSE), Liu estimator is preferred. The results of the joint estimators of robust Liu estimators are more efficient than either the OLS or the Liu estimator; they have smaller MSEs. Moreover, MRLE is most efficient.

Example 2. Portland cement data

Portland dataset was introduced by Woods et al (1932), and has been widely analysed by Hald (1952), Hamaker (1962) and Kaciranlar et al (1999). The dataset contains four explanatory varaiables which are tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium aluminoferrite (X_3) and β-dicalcium silicate (X_4). The heat evolved after 180 days of curing is the dependent variable (Y). The dataset suffers multicollinearity since variance inflation factors, VIF(X_1)=38.496, VIF(X_2)=254.423, VIF(X_3)=46.868 and VIF(X_4)=282.513), are greater than 10. Mahalanobis distances of observation 3 and 10 are 2.4495 and 2.7353 which revealed that observation 3 and 10 are leverage. Also, the robust MCD distances are 3.6810 and 4.8610. Here, there is outlier in the x-direction and no outlier in the y-direction. Consequently, multicollinearity and leverage point jointly exist in the dataset.
The results of the OLS and robust estimators are given in Table 4 while that of the Liu and Robust Liu estimators are given in Table 5.

Table 4. Estimates of OLS and Robust estimators of Portland cement data

Coefficient	OLS	M	MM	LTS	S	LAD
$\hat{\alpha}_1$	1.6373	1.6371	1.6371	1.6377	1.6388	1.6447
$\hat{\alpha}_2$	-0.2097	-0.2032	-0.2027	-0.1806	-0.1831	-0.2147
$\hat{\alpha}_3$	0.9160	0.8905	0.8889	0.8205	0.8255	0.8463
$\hat{\alpha}_4$	-1.8405	-1.8672	-1.8693	-1.9697	-1.9623	-1.9257
$\hat{\sigma}^2$	5.8454	3.2671	5.8342	1.5561	5.8057	6.6564
MSE($\hat{\alpha}$)	0.0638	0.0356	0.0637	0.0170	0.0633	0.0726

Table 5. Estimates of OLS, Liu and Robust Liu estimators of Portland cement data

Coefficient	Liu	MRLE	MMRLE	TSRLE	SRLE	LADRLE
$\hat{\alpha}_1$	1.63745	1.63677	1.6371	1.6377	1.6388	1.6447
$\hat{\alpha}_2$	-0.2099	-0.2032	-0.2027	-0.1806	-0.1831	-0.2147
$\hat{\alpha}_3$	0.9196	0.8897	0.8881	0.8198	0.8247	0.8456
$\hat{\alpha}_4$	-1.8952	-1.8548	-1.8569	-1.9561	-1.9488	-1.9126
D	4.1604	0.2934	0.2928	0.2639	0.2659	0.2760
MSE($\hat{\alpha}$)	0.0674	0.0354	0.0631	0.0170	0.0628	0.0719

From Table 4 and 5, it can be seen that the regression estimates are not too different from each other. However, in terms of MSE criterion of the estimators, the LTSRLE, MRLE, SRLE and MMRLE, in this order, are more efficient than the OLS. Thus, LTSRLE is most efficient.

Example 3. Hussein and Abdalla data

This dataset was used by Hussein and Abdalla (2012) and it covers the products in the manufacturing sector of Iraq in the period of 1960 to 1990. The variables used are the product value in the manufacturing sector**(**Y**),** value of imported intermediate (X_1), imported capital commodities (X_2) and value of imported raw materials (X_3). Hussein and Abdalla (2012) showed that the dataset suffers the problem of multicollinearity since VIF(Max)>10. Lukman et al (2014) identified case number: 12, 14, 15, 16, 17, 18, 19, 20 and 21 as outliers in the y-direction and also identified case number 12, 14 and 15 as leverages.

Therefore, outliers exist in the y and x direction.
The results of the OLS and robust estimators are given in Table 6 while that of the Liu and Robust Liu estimators are given in Table 7.

Table 6. Estimates of OLS and Robust estimator of Hussein and Abdalla data

Coefficient	OLS	M	MM	LTS	S	LAD
$\hat{\alpha}_1$	1.3143	1.3948	1.3821	1.3803	1.3807	1.3862
$\hat{\alpha}_2$	-1.5151	-1.8513	-4.9978	-5.7278	-5.8198	-2.5380
$\hat{\alpha}_3$	2.0164	1.7145	-3.6142	-4.9724	-5.2153	-0.2247
$\hat{\sigma}^2$	37736	7851.32	5316.35	4017.17	5297.70	54336.6
MSE$(\hat{\alpha})$	4.7230	0.9827	0.6654	0.5028	0.6631	6.8007

Table 7. Estimates of OLS, Liu and Robust Liu estimators of Hussein and Abdalla data

Coefficient	Liu	MRLE	MMRLE	LTSRLE	SRLE	LADRLE
$\hat{\alpha}_1$	1.3143	1.3948	1.3821	1.3803	1.3807	1.3862
$\hat{\alpha}_2$	-1.5151	-1.8513	-4.9977	-5.7277	-5.8197	-2.5382
$\hat{\alpha}_3$	2.0162	1.7144	-3.6138	-4.9719	-5.2147	-0.2250
D	0.4395	0.3396	0.0758	0.0403	0.0367	8.4843
MSE$(\hat{\alpha})$	4.7225	0.9825	0.6653	0.5027	0.6629	6.8113

From Table 7, it can be seen that LADRLE has bigger MSE than other RLEs when outliers are in both y and x-direction. Thus, the results reveal that Robust Liu estimates based on LTSRLE, SRLE and MMRLE, in this order, are more efficient and preferred. Thus, the LTSRLE is most efficient of them.

5. CONCLUSION.

Ordinary Least Square (OLS) estimator and Liu estimator (LE) could not perform well in the presence of multicollinearity and outliers based on MSE criterion. The performances of both estimators are not too different. The robust Liu estimators except LADRLE performed well than their individual counterparts (OLS and LIU) when both problems exist. Finally, it is observed that the MRLE estimator is most efficient when the outlier is the y-direction and the LTSRLE is also most efficient when the outlier is either in the x-direction (leverage) or in both y and x-direction. It is therefore important to note that the performance of these estimators depend on the nature or direction of the outliers.

REFERENCES.

Aboobacker, J. and Jianbao, C. (2011). Measuring Local Influential Observations in Modified Ridge Regression. Journal of Data Science 9(3), 359-372.

Aitken, A.C. (1935). On least Squares and linear combinations of observations. Proceedings of the Royal Statistical Society. Edinburgh 55: 42-48.

Alpu, O. and Samkar, H. (2009): Liu Estimator based on an M Estimator. Tukiye Klinikleri Journal of Biostatistics 2(2), 49-53.

Ayinde, K., Lukman, A. F. and Arowolo, O. (2015). Robust regression diagnostics of influential observations in linear regression model. Open Journal of Statistics 5, 273-283.

Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics 19, 15-18.

Dielman, T. E. (1984). Least absolute value estimation in regression models: An annotated bibliography. Communications in Statistics - Theory and Methods 4, 513-541.

Gujarati, N. D. (2003). Basic Econometrics (4th Ed.). New Delhi: TataMcGraw-Hill 748, 807.

Hald, A. (1952). Statistical Theory with Engineering Applications. Wiley, New York.
Hamaker, H. C. (1962). On multiple regression analysis. Statistica Neerlandica16, 31–56.

Samkar, H. and Alpu, O. (2010). Ridge regression based on some robust estimators. Journal of Modern Applied Statistical Methods: 9(2), 495-501.

Hoerl, A. E. and Kennard, R.W. (1970). Ridge regression: biased estimation for non-orthogonal. problems. Technometrics 12, 55-67.

Holland, P. W. (1973). Weighted ridge regression: Combining ridge and robust regression methods. NBER Working Paper Series 11, 1-19.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo, Ann. Stat 1,799–821.

Hussein,Y.A and Abdalla, A. A. (2012). Generalized Two stages Ridge Regression Estimator for Multicollinearity and Autocorrelated errors. Canadian Journal on Science and Engineering Mathematics 3(3), 79-85.

Kacıranlar, S., Sakallioglu, S., Akdeniz, F., Styan, G.P.H., and Werner, H. J. (1999). A new biased estimator in linear regression and detailed analysis of the widely analysed dataset on Portland cement, Sankhya, Series B 61, 443-459.

Liu, K. (1993). A new class of biased estimate in linear regression. Communications in Statistics 22(2), 393-402.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of American Statistical Association 62, 819-841.

Lukman, A.F, Arowolo, O. and Ayinde, K. (2014). Some robust ridge regression for handling multicollinearity and outlier. International Journal of Sciences: Basic and Applied Research 16(2), 192-202.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust Regression by Means of S Estimators in Robust and Nonlinear Time Series Analysis, Franke, J., Härdle, W. and Martin, R.D. Lecture Notes in Statistics, 26, New York: Springer-Verlag 256–274.

Rousseeuw, P. J. and Van Driessen, K (1998). Computing LTS Regression for Large Data Sets,Technical Report, University of Antwerp, submitted.

Stein, C. M. (1960). Multiple Regression, Contributions to Probability and Statistics, Stanford University Press 424-443.

Woods, H., Steinour, H. H. and Starke, H. R. (1932). Effect of composition of Portland cement on heat evolved during hardening. Industrial and Engineering Chemistry 24, 1207–1214.

Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression, Annals of Statistics 15, 642–656.