The Performance of Robust Latent Root Regression Based on MM and modified GM estimators

HABSHAH MIDI
Universiti Putra Malaysia
Faculty of Science and,
Institute For Mathematical Research
43400 UPM, Serdang, Selangor
MALAYSIA
habshahmidi@gmail.com

MOHAMMED A. MOHAMMED
Universiti Putra Malaysia
Institute For Mathematical Research and,
Al-Dewanyia Technical Institute, Iraq
43400 UPM, Serdang, Selangor
MALAYSIA
mohammedam23@yahoo.com

Abstract: It is now evident that the Ordinary Least Squares (OLS) estimator suffers a huge set back in the presence of multicollinearity. As an alternative, the Latent Root Regression (LRR) is put forward to remedy this problem. Nevertheless, it is now evident that the LRR performs poorly when outliers exist in a data. In this paper, we propose an improved version of the LRR to rectify the problem of multicollinearity which comes together with the existence of outliers. The proposed method is formulated by incorporating robust MM - estimator and modified generalized M- estimator (MGM) in the LRR algorithm. We call these methods the Latent Root MM-based (LRMMB) and the Latent Root MGM-based (LRMGMB) methods. The performance of our developed methods are compared with some existing methods such as the OLS, LRR, and the Latent Root M-based (LRMB). The numerical results indicate that the LRR performs very well in the presence of multicollinearity, but performs poorly in the presence of outliers. The proposed methods (LRMMB and LRMGMB) are more efficient than the OLS and the LRR estimators for data having both problems of multicollinearity and outliers.

Key–Words: Multicollinearity, OLS , Latent Root Regression , MM-estimator, GM-estimator

1 Introduction

Consider a multiple linear regression model:

$$Y = X\beta + \varepsilon, \quad (1)$$

where Y is an $n \times 1$ vector of observation of dependant variables, X is an $n \times p$ matrix of independent variables, β is $p \times 1$ vector of unknown regression parameters, ε is an $n \times 1$ vector of random errors which follow the classical assumptions, namely, $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon^T) = \sigma^2 I$, and p is the number of independent variables. For model fitting, the Y and $X's$ are in standardized form that make the design matrix $X^T X$ equivalent to the correlation matrix of independent variables. The Ordinary Least Squares (OLS) method is widely used to estimate the parameters of model in Eqn.7. The OLS estimates is written as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad (2)$$

Under the Gaussian Markov assumptions, this ordinary least squares estimator has minimum variance and it is the best linear unbiased estimator (BLUE) if all underlying assumptions of this model are met. In practice, multicollinearity is a common problem in regression model. This problem occurs when two or more independent variables are highly correlated. It may be due to the data collection employed, constraints on the models, model specification and overdetermined model. This problem may produce inflated standard errors for the coefficients that will lead to misleading parameter inferences. To rectify this problem, Webster et al. [19] proposed a new biased procedure which is called Latent Root Regression (LRR) to recover accuracy of regression estimates. However, this method is inefficient if the errors are not normally distributed which is often due to outliers. In regression, outliers can occurs in Y, X and in both Y and X directions. Habshah and Lau [3] suggested the Latent Root M-based Regression (LRMB) to overcome the multicollinearity problem in the presence of outliers in Y direction. Unfortunately, this method is inefficient when the outliers are located in the X- direction. Bagheri [1] modified the Generalized M-estimator (GM-estimator) to overcome the problem of high leverage points (outliers in the X direction) in multiple linear regression model. They proposed two modified GM procedures, namely MGM2 and MGM3. In this study a Robust Latent Root Regression (RLRR) method is formulated by incorporating the high efficiency and high breakdown point MM-estimator [17] and MGM2 procedure in the latent root regression.

2 Latent Root Regression (LRR)

Multicollinearity occurs when there is an almost exact linear dependency among the explanatory variables, where the coefficient of determination (R^2) will be very close to one. This type of ill-conditioning among the explanatory variables is referred to as near singularity and the OLS estimators can be very poor in this situation. Webster et al. [19], suggested a bias alternative method, namely the Latent Roots Regression (LRR). The LRR can identify near singularities and determine whether or not these near singularities have predictive value. Those latent roots and latent vectors that are non-predictive near singularities be removed. Subsequently, a stepwise backward elimination of variables is performed. For the linear regression model in Eqn.7, let $\Omega_{n\times(p+1)} = (\,Y^*_{n\times 1} : X^*_{n\times p})$ be an augment matrix of the standardized response variable, and standardized regressors variables X

$$\Omega = \begin{bmatrix} y_1^* & x_{11}^* & x_{12}^* & \cdots & x_{1p}^* \\ y_2^* & x_{21}^* & x_{22}^* & \cdots & x_{2p}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n^* & x_{n1}^* & x_{n2}^* & \cdots & x_{np}^* \end{bmatrix}, \tag{3}$$

where

$$y_i^* = (y_i - \bar{y})/\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \tag{4}$$

$$x_i^* = (x_i - \bar{x})/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{5}$$

The $(\Omega^T\Omega)$ is the correlation matrix of the response and regressors variables [18], and it has latent roots (eigenvalues) and latent vectors (eigenvecors) defined by: $|\Omega^T\Omega - \lambda_j I| = 0$ and $(\Omega^T\Omega - \lambda_j I)\gamma_j = 0$ for $j = 0, 1, \ldots, p$.

Let $\gamma_j^T = (\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{pj})$ be the elements of jth latent vector and $\gamma_j^{0T} = (\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{pj})$. Assume the $\lambda_0 \le \lambda_1 \le \ldots \le \lambda_p$ be the ordered latent roots and corresponding latent vectors. The augment matrix for latent roots and latent vectors for dependent and independent variables is:

$$M(\lambda,\gamma) = \begin{bmatrix} \lambda_0 & \gamma_{00} & \gamma_{01} & \cdots & \gamma_{0p} \\ \lambda_1 & \gamma_{10} & \gamma_{11} & \cdots & \gamma_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_p & \gamma_{p0} & \gamma_{p1} & \cdots & \gamma_{pp} \end{bmatrix}, \tag{6}$$

Webster et al.[19], Gunst et al.[18] and Lawrence and Arthur [15] pointed out that small latent roots correspond to non-predictive near singularities and they suggested a cut-off value in which $\lambda_j \le 0.3$ and $|\gamma_{0j} \le 0.1|$. Subsequently, they found out that a stricter cut-off value of $\lambda_j \le 0.2$ and $|\gamma_{0j} \le 0.1|$ could reinforce the analysis. The least squares estimator of coefficients can be written as a form of latent roots and latent vectors $M(\lambda,\gamma)$ as follows:

$$\hat{\beta}_{OLS} = -\delta \sum_{j=0}^{p} \alpha_j \gamma_j^0, \tag{7}$$

where

$$\alpha_j = \gamma_{0j}\lambda_j^{-1}(\sum_{t=0}^{p} \gamma_{0t}^2/\lambda_t)^{-1}, \tag{8}$$

$$\delta^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2, \tag{9}$$

with sum squares of residuals

$$SSE_{ols} = \delta^2(\sum_{j=0}^{p}\gamma_{0j}^2/\lambda_j)^{-1}, \tag{10}$$

In the presence of multicollinearity in the data set, some values of α become sufficiently large relative to the other values of α_j, that leads to distortion of some of the coefficients due to large term $\alpha_0\gamma_0^0$ in (7). Let $\gamma_0, \gamma_1, \ldots, \gamma_{r-1}$ related with non-predictive near singularities, then the non-predictive values are eliminated and only the predictive values are retained. Webster et al. [19] proposed modified OLS estimator by setting $\alpha_0 = \alpha_1 = \ldots, \alpha_{r-1} = 0$, so the adjusted least squares estimator, namely Latent Root Regression (LRR) is defined as:

$$\hat{\beta}_{LRR} = -\delta \sum_{j=r}^{p} \alpha_j \gamma_j^0, \tag{11}$$

where

$$\alpha_j = \gamma_{0j}\lambda_j^{-1}(\sum_{t=r}^{p} \gamma_{0t}^2/\lambda_t)^{-1}; \quad j = r, r+1, \ldots, p, \tag{12}$$

The residuals sum of squares for latent root is:

$$SSE_{LRR} = \delta^2(\sum_{j=r}^{p}\gamma_{0j}^2/\lambda_j)^{-1}. \tag{13}$$

If all near singularities have predictive value, none of the $\alpha_j s$ equal to zero and then the least squares coefficients and the latent root coefficient will be equivalent.

The mean square errors for $\hat{\beta}_{LRR}$ is not known exactly but it can be approximately computed by similar way as the mean square error of the Principal Components Regression (PCR) on the latent vectors $(X^T X)$ of (see [14], [19]), then:

$$MSE(\hat{\beta}_{LRR}) \approx \sigma^2 \sum l_i^{-1} + (\alpha_i^T \beta)^2 \quad (14)$$

where $l_1 \leq l_2 \leq \ldots \leq l_p$ are the latent roots of the design matrix $(X^T X)$.

3 Robust Methods

Huber [12] established the classes of bias robust M-estimator which is nearly as efficient as the ordinary least squares and it is resistant to outliers in the Y- direction. The OLS method aim to minimize the sum of squares residuals $\sum_{i=1}^{n}(\hat{\beta})^2$ which is inefficient in the presence of outliers (see [5],[7],[9]). The M-estimator tries to reduce the impact of unusual data by changing the sum of squares error by another function of error:

$$\min \sum_{i=1}^{n} \rho(\frac{r_i(\hat{\beta})}{\hat{\sigma}}) = \min \sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}}), \quad (15)$$

where ρ is a symmetric function and positive definite with a unique minimum at zero and it represents the contribution of each error to the minimize function, and $\hat{\sigma}$ is an estimate of scale parameter. Huber [12] showed that the M-estimator is sensitive to high leverage points in X- direction. In other words, it does not have bounded influence. Schweppe [8] suggested a new robust method called bounded influence Generalized M-estimator (GM-estimator) to overcome this drawback of M-estimator. GM-estimators try to downweigh the points that have high residuals in the X- direction which is called high leverage points. The coefficients for GM-estimators $(\tilde{\beta})$ in convergence can be written as:

$$\tilde{\beta} = (X^T W X)^{-1} X W y, \quad (16)$$

where the diagonal elements of W are the weights w_i, defined as:

$$w_i = \frac{\psi[(y_i - x_i^T \tilde{\beta})/\pi_i s]}{(y_i - x_i^T \tilde{\beta})/\pi_i s)}, \quad (17)$$

where s is the estimated scale and ψ is an influence function (see [8]). Different π-weight functions exist in the literature of GM-estimator, such as Krasker and Welsch's weight function (see [8]), [13] which depend on the hat matrix (h) and written as:

$$\pi_i = [(1 - h_{ii})/h_{ii}]^{1/2}, \quad (18)$$

$$h_{ii} = x_i^T (x^T x)^{-1} x_i \,; i = 1, 2, \ldots, n, \quad (19)$$

MM-estimator is a robust method introduced by Yohai in 1987 (see [13]). It combines an elevated breakdown point 50% and supreme efficiency (95% of OLS efficiency under normal assumptions). The MM-estimates include three-stage procedures (see [2]):

- Computes S-estimate as an initial consistent estimate with high breakdown point and bisquare influence function as:

$$\rho(x) = \begin{cases} 3(\frac{x}{c})^2 - (\frac{x}{c})^4 + (\frac{x}{c})^6 & \text{if } |x| < c, \\ 1 & \text{otherwise.} \end{cases}$$

  where c is the tuning constant selected as 1.548.

  By using the results from the first stage, computes the robust Mestimate of the error standard deviation:

$$\min \sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \hat{\beta}_M}{\hat{\sigma_0}}),$$

  where $\rho(x)$ is the influence function with value of tuning constant 4.687 and $\hat{\sigma}_0$ is the estimate of standard deviation of the residuals.

- The final stage is calculating the MM-estimate as the solution to:

$$\frac{1}{n-p} \sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \hat{\beta}_M}{s}) = 0.5.$$

4 Robust latent root regression (RLRR)

The LRR performs well in the presence of multicollinearity. Habshah and Lau [3] pointed out when both multicolinearity and outliers are present in a data set, the LRR is inefficient and they proposed a new method, namely Robust Latent Root M-Based Regression (RLMB). This new suggested method does well with outliers in Y-directions and it is not robust to high leverage points. In this case, we suggest a robust method that can remedy the problems of multicollinearity in the presence of high leverage points. Since the MM-estimator and GM-estimator are robust in both Y and X-direction, we will incorporate both estimators in the establishment of the Robust Latent Root Regression (RLRR). The RLRR starts with imposing weights to the modified correlation matrix between the dependent and independent variables (Augmented matrix). The modified pair-wise correlation coefficient (r_w) can be written as:

$$r_w = \frac{\sum_{i=1}^{n} w_i(Y_i - \bar{Y}_w)(X_i - \bar{X}_w)}{\sqrt{\Big[\sum_{i=1}^{n} w_i(Y_i - \bar{Y}_w)^2\Big]\Big[\sum_{i=1}^{n} w_i X_i - \bar{X}_w)^2\Big]}}, \tag{20}$$

where

$$\bar{Y}_w = \Big(\frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}\Big),\ \bar{X}_w = \Big(\frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}\Big).$$

The weights, w_i in Eqn.20 can be chosen from the final stage of any robust method. Habshah and Lau [3] suggested using the Tukey's biweight function to compute the robust correlation coefficients for the augmented matrix (Ω). Subsequently, the latent roots and latent vectors are calculated. Following the same procedure, we suggested using the final weights of the MM-estimator and modified GM-estimator in the development of the RLRR and called them LRMMB and LRMGMB, respectively. The algorithm of modified GM-estimator (MGM) is summarized as followes (see [1]).

Step 1 Calculate the residuals $(e_i,\ i = 1, 2, \dots, n)$ of S-estimator and scale of residuals $(\hat{\tau})$ by applying:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}, \tag{21}$$

$$\hat{\tau} = 1.4826(1 + 5/(n - p))\text{Median}|e_i|. \tag{22}$$

Step 2 : Compute weight function w_i as:

$$w_i = \min\Big[1, \Big\{\frac{\chi^2_{0.95,k}}{RMD_i^2}\Big\}\Big],\ i = 1, 2, \dots, n \tag{23}$$

where the degree of freedom (k) is the number of independent variables including the constant terms. RMD_i^2 is the squared of Robust Mahalanobis Distance based on Minimum Volume Ellipsoid (MVE) which can be formulated as follows: (see [16])

$$RMD_i^2 = \sqrt{(x_i - M(X))S(X)^{-1}(x_i - M(X))^T}$$

where $M(X)$ and $S(X)$ are the robust locations and shape estimates of the MVE, respectively.

Step 3 Compute $Q = diag\Big[\psi'(\frac{e_i}{\hat{\tau}\times w_i})\Big]$, where ψ' is a derivative of Huber's function (ψ).

Step 4 Letting β_0 be the S-estimator, the MGM-estimator can be derived from One-Step Newton Raphson as:

$$\hat{\beta}_{MGM} = \hat{\beta}_0 + (X^T QX)^{-1} XW\psi(\frac{e_i}{w_i\hat{\tau}})\hat{\tau} \tag{24}$$

where W is an $n \times n$ diagonal matrix with w_i , $i = 1, 2, \dots, n$, obtained from Step 2

5 Examples

Two sets of real data are used to compare the efficiency of the suggested methods LRMMB and LRMGMB with the other existing methods such as OLS, LRR and LRMB. The first data set is the Consumption Income Expenditure (CIE) which is taken from Gujarati (see [11]). The Consumption Expenditure (CE) is the response variable and the Income and Wealth are the explanatory variables. The data set is modified to have one high leverage point by replacing the first observation of each variable by multiplying the observation by 10 . The Diagnostic Robust Generalized Potential (DRGP) and Variance inflation factor (VIF) were used to identify the high leverage point (hlp) and multicollinearity, respectively (see [6], [1]). It can be observed from Table 1 that the original data has no hlp but it has problem of multicollinearity, while the modified data has combined problem of hlp and multicollinearity. However, the degree of multicollinearity has increased by adding one high leverage point to the original data as shown in Table 1 and 2

Table 1: Hlp and Multicollinearity Diagnostics for Gujarati Data

Diagnostic	Data	hlp Detected	x_1	x_2
DRGP	Original	-	-	-
	Modified	Case 1	-	-
VIF >10	Original	-	482.12	482.12
	Modified	-	5448.12	5448.12

The second empirical example is the Body Fat data set. Kutner et al.[10] introduced this data set with 20 observations and three independent variables which are, the Triceps Skinold Thickness (TST), Thigh Circumference (TC), and Midarm Circumference (MC). To study the multicollinearity problem in the presence of outliers, we modified the data set by replacing the first observation for each variable by 100 to get the combined problem of multicollinearity and outliers. Table 2 shows that the original data has no hlp but it has problem of multicollinearity, while the modified data has combined problem of hlp and multicollinearity.

Table 2: Hlp and Multicollinearity Diagnostics for Body Fat Data

Diagnostic	Data	hlp Detected	x_1	x_2	x_3
DRGP	Original	-	-	-	-
	Modified	Case 1	-	-	-
VIF >10	Original	-	708.84	564.34	104.60
	Modified	-	3437.3	1157.5	689.34

6 Simulation Study

A simulation study was conducted to compare the efficiency of the preceding methods. Following Lawrence and Arthur [15], two independent variables are generated as follows.

$$y_i = \beta_0 X_{i1} + \beta_1 X_{i2} + e_i,\ i = 1, 2, \ldots, n \quad (25)$$

where the coefficients are fixed and equal to one. The predictor variables were generated as below:

$$x_{ij} = (1 - \rho^2) z_{ij} + \rho z_{ij},\ j = 1, 2 \quad (26)$$

where z_{ij} are independent standard normal random numbers. The correlation coefficient between the explanatory variables were chosen as 0.0, 0.5, and 0.99 with different samples of size 20, 40,100 and 200. Two different distributions for the error terms were considered:

- Standard normal distribution.
- Cauchy distribution with mean zero and scale parameter one.

The Cauchy distribution is a heavy tail distribution and it has symmetrical bell shape which tend to produce a significant amount of outliers. The standard error (SE), bias, and root mean square errors (RMSE) over 1000 runs were used to compare the performance of the estimation methods. Another measure for comparison is the efficiency of the regression estimators by comparing the MSE ratio of two estimators. The ratio less than one indicates that the first estimator is more efficient than the second, while the ratio which is more than one indicates that the second estimator is more efficient and the ratio equal to one shows that both estimators are of the same efficiency.

7 Discussion

Let as first focus to the first set of real data, i.e. consumption income expenditure data. Table 3 exhibited the parameter estimates and the standard errors of the OLS, LRR, LRMB, LRMMB and LRMGMB for the original data. It can be observed that when only multicollinearity is present in the data, all methods are approximately equally good, except the OLS because it failed to remedy the problem of multicollinearity. It is interesting to see the situation when both hlp and multicollinearity are present in the data. In the presence of outliers in the Y-direction shows that the OLS and the LRR are heavily affected by outliers, not only the standard errors of their estimates become large, but the sign of $\hat{\beta}_2$ for LRR had changed. The LRMGMB has the least standard errors followed by both LRMB and LRMMB. We can see all our robust methods give good results since they down weight outliers that were created in the Y-direction (see Table 5).

Table 3: The Estimate Values and Standard Errors (in parenthesis) for Original Gujarati Data

	Original Data		Modified Data (outlier in Y direction		Modified Data (outlier in Y and X direction	
	b_1	b_2	b_1	b_2	b_1	b_2
OLS	1.814	-0.834	3.491	-3.886	1.999	-1.006
	(1.585)	(1.585)	(7.493)	(7.493)	(3.178)	(3.178)
LRR	0.437	0.542	-0.202	-0.192	0.481	0.519
	(0.160)	(0.198)	(0.986)	(0.984)	(1.180)	(1.191)
LRMB	0.448	0.583	0.558	0.637	0.491	0.519
	(0.158)	(0.205)	(0.714)	(0.881)	(0.190)	(0.201)
LRMMB	0.448	0.583	0.523	0.652	1.751	2.162
	(0.158)	(0.205)	(0.714)	(0.881)	(0.090)	(0.111)
LRMGMB	0.446	0.573	0.347	0.357	1.756	2.065
	(0.159)	(0.203)	(0.209)	(0.211)	(0.079)	(0.093)

The results of Table 3 also show that the OLS and LRR were more affected by outliers and have large values for SE when multicollinearity comes together with the presence of outliers in both Y and X- directions. The standard error of LRMB is also high since it is well known that the M-estimator is robust only in the Y-direction. The LRMGMB is slightly better than LRMMB. The confidence interval lengths in Table 4 support the earlier findings that the LRMGMB and LRMMB provide the best results, with their confidence intervals lengths being the shortest compared to other estimators.

Now let us focus to the body fat data set. As

Table 4: The Estimators (Est),Standard Error (SE) and Confidenct Intervals (C.I) for Modified Gujarati Data (outliers in both Y and X- directions)

		b_1	b_2
OLS	Est	1.999	-1.006
	S.E	3.178	3.178
	C.I	(-5.51,9.51)[15.02]	(-8.52,6.50)[15.02]
LRR	Est	0.481	0.512
	S.E	0.180	0.191
	C.I	(0.05,0.90)[0.85]	(0.05,0.96)[0.91]
LRMB	Est	0.491	0.519
	S.E	0.190	0.201
	C.E	(0.04,0.94)[0.90]	(0.04,0.99)[0.95]
LRMMB	Est	1.751	2.162
	S.E	0.090	0.111
	C.I	(1.56,1.94)[0.38]	(1.94,2.38)[0.44]
LRMGMB	Est	1.756	2.065
	S.E	0.079	0.093
	C.I	(1.56,1.94)[0.38]	(1.84,2.28)[0.44]

Table 5: Robust Bisquare Weights for Robust Methods (M, MM and MGM) for Gujarati data

		1	2	3	4	5	6	7	8	9	10
	M	0.97	0.68	0.92	0.99	0.97	1.00	0.91	0.99	0.86	0.94
Original Data	MM	0.97	0.75	0.93	0.99	0.97	1.00	0.93	0.99	0.89	0.95
	MGM	1.00	0.68	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Data with	M	0.00	0.91	0.94	1.00	0.97	1.00	0.96	0.99	0.94	0.95
outlier in	MM	0.00	0.93	0.95	1.00	0.98	1.00	0.96	0.99	0.95	0.96
Y-direction	MGM	0.02	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Data with	M	0.97	0.94	0.86	0.97	0.98	1.00	0.96	0.95	0.95	0.79
outlier in Y	MM	0.00	0.93	0.95	1.00	0.98	1.00	0.97	0.99	0.95	0.96
and X direction	MGM	0.02	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

already mentioned, the modified data has multicollinearity and outliers. It can be clearly seen from Table 6 that the OLS has poor estimates resulting in inflated standard error and wider confidence intervals, while the LRR is less affected by this problem. The procedure of combining the classical latent root regression with robust methods such as LRMB,LRMMB and LRMGMB able to remedy the multicollinearity problem in the presence of outliers. The results of Table 6 show that the standard error for the LRMGMB is slightly smaller than the LRMMB. A reasonable explanation up to this point is that the LRMGMB is the best estimation method among the existing methods for data that has multicollinearity and outliers.

Further, the results of the simulation study are discussed here. As previously mentioned, we had generated the error term with two different distributions: standard normal and Cauchy. Table 7 shows the results for regression estimators when the error term is distributed with standard normal without outliers. When there is no and low degree of multicollinearity problem ($\rho = 0.0$ and 0.5) the performances of all five methods are approximately the same but OLS and LRR are slightly better than the robust methods. Also we can see the OLS and LRR performances are equivalently good since all near singularities have predictive values and none of these are equal to zero. With high multicollinearity ($\rho = 0.99$), the OLS is poorly estimated and it has inflated standard error while the LRR, LRMB, LRMMB and LRGMB have good results. This results are supported by the MSE ratio in Table 8 which shows the efficiency for OLS and LRR is equally good where the values of MSE ratio are greater than one . In the case of high collinearity among regressor variables ($\rho = 0.99$), the LRR is more efficient than other estimation methods. It can be observed from Table 8 that increasing the sample size, improves the performance for all methods. Let us now focus to Table 9 and 10 when the error terms are distributed as Cauchy distribution with (0,1) which prone to create some outliers in the data. When there is no or low multicollinearity ($\rho = 0.0$ and 0.5), the OLS and LRR have poor estimates due to outliers created in the data set. On the other hand, all the robust methods (LRMB, LRMMB and LRMGMB) are much better and more efficient than the OLS and the LRR.

Table 6: The Estimators, Standard error and Length of Coefficients Interval (L.C.I) for modified Body Fat Data (outliers in both Y and X directions)

		b_1	b_2	b_3
OLS	Est.	31.62	-16.82	-15.77
	S.E.	6.31	3.66	2.82
	L.C.I	[26.77]	[15.53]	[11.99]
LRR	Est.	0.397	0.520	-0.082
	S.E	0.104	0.137	0.104
	L.C.I	[0.444]	[0.585]	[0.443]
LRMB	Est.	0.396	0.529	-0.089
	S.E	0.11	0.147	0.108
	L.C.E	[0.470]	[0.626]	[0.461]
LRMMB	Est.	-0.074	-0.218	0.137
	S.E	0.038	0.043	0.037
	L.C.I	[0.160]	[0.185]	[0.160]
LRMGMB	Est.	-0.073	-0.219	0.138
	S.E	0.036	0.040	0.033
	L.C.I	[0.161]	[0.184]	[0.161]

With high multicollinearity ($\rho = 0.99$), the performance of OLS and LRR become very bad compared to LRMB, LRMMB and LRGMMB. In this situation, the RMSE of the OLS becomes very large. It is interesting to see that the LRMGMB is slightly better than the LRMMB having the smallest bias and RMSE, followed by the LRMB estimator.

8 Conclusion

The main focus of this article was to develop a reliable alternative methods for rectifying the problem of multicollinearity and outliers. In the presence of multicollinearity, the OLS performs very poorly. On the other hand, the performance of LRR, LRMB, LRMMB and LRMGMB are equally good. The LRR is slightly better than the other three methods when only multicollinearity problem occurs in the data set. However, the performance of LRR deteriorates very badly in the presence of both multicollinearity and outliers. With the combined problem of multicollinearity and outliers in the Y- direction, the three robust methods are reasonably close to each other. However, the LRMB is much affected by multicollinearity and outliers in both Y and X- directions. In this situation, the LRMB is not efficient. The results of the study show that the LRMGMB and LRMMB estimates are more efficient and more reliable because they are not much affected by the presence of outliers in Y as well as outliers in both Y and X- directions. Hence, we can consider the LRMGMB and LRMMB as a better estimation methods for handling the problems of outliers and multicollinearity.

References:

[1] A. Bagheri, *Robust Estimation Methods and Robust Multicollinearity Diagnostics For Multiple Regression Model in The Presince of High Leverage Collinearity -Influential Observation* Thesis submitted to the School of Graduate Studies– UPM 2011.

[2] G. A. Ozlem , Comparison of robust regression methods, *Int. J. Contemp. Math. Sciences*, Vol. 6, No. 9, 2011, pp 409-421.

[3] M. Habshah and Lau Ung Hua, The Performance of Latent Root-M based Regression, *Journal of Mathematics and Statistics*, Vol. 5 No.1, 2009, pp 1-9.

[4] A. Bagheri and M. Habshah, Robust estimations as a Remedy for Multicollinearity Caused by Multiple High Leverage Points, *Journal of Mathematics and Statistics*, Vol. 5, No.4, 2009, pp. 311-321.

[5] M. Habshah, R. Sohel and A. H. M. Imon, The Performance of Robust Weight Least Squares in the Presence of Outliers and Hetroscedastic Errors,*WSEAS Ttransactions on Mathematics*,Vol. 8, No. 7, 2009, pp 351-360.

[6] M Habshah, M. R. Norazan and A. H. M. Rahmatullah Imon, The performance of diagnostic robust generalized potentials for the identification of multiple high leverage points in linear regression, *Journal of Applied Statistics*, Vol. 36, No. 5, 2009, pp 507-520.

[7] I. Paolo, Application of multivariate robust statistical hypothesis testing to data from semiconductor environments, *WSEAS Ttransactions on Mathematics*,Vol. 6, No. 3, 2007, pp 506-512.

[8] R. A. Maronna, R. Douglas Martin and V. J. Yohai., *Robust Statistics: Theory and Methods*, 1st Edn, Wiley– New York– USA 2006.

[9] Z. Azami, M. Ibrahim, A Shahrum and Z.O. Mohd, Performances test statistical for single outliers detection in bilinear (1,1,1,1) models,*WSEAS Ttransactions on Mathematics*,Vol. 5, No. 12, 2006, pp 1359-1364.

[10] M. H. Kutner, C.J. Nachtsheim, J. Neter and W. Li, *Applied Linear Regression Models*, 5th Edn, MacGRAW Hill–New York –USA 2005.

[11] D. N. Gujarati, *Basic Econometrics*, McGraw Hill– 5th Edn.–New York– USA 2003.

[12] P. J. Huber, *Robust Statistics*, 1st Edn., Wiley–New York–USA 2003.

[13] P. J. Rousseeuw, P. J. and A.M. Leory, *Regression and Outlier Detection* 1st Edn., Wiley–New York–USA 2003.

[14] A. Bastlevsky, *Statistical Factor Analysis And Related Methods*, John Wilay and Sons–INC 1994.

[15] K. D. Lawrence and J. L. Arthur, *Robust Regression: Analysis and Applications*, 1st Edn., Marcel Dekker– New Yourk, 1989.

[16] P. J. Rousseeuw and Annick M. Leroy, *Robust Regression and Outlier Detection*, John Wilay and Sons–New York–USA 1987.

[17] V. Yohai, High Breakdown Point and High Efficiency Robust Estimates for Regression, *Annals of Statistics*, Vol. 15, No. 2, 1987, pp 642-656.

[18] R. F. Gunst, J. T. Webster and R. L. Mason, A comparison of least squares and latent root regression estimators, *Technometrics*, Vol .18, 1976,PP 75-83.

[19] J. T. Webster, R. F. Gunst and R. L. Mason, Latent Root Regression Analysis, *Technometrics*, Vol.16, No.4, 1974, pp 513-22.

Table 7: Bias, RMSE and SE for $\hat{\beta}_1$ and $\hat{\beta}_2$ with error term distributed normal (0,1)

			20			40			100			200	
Mehod	ρ	Bais	SE	RMSE	Bais	SE	RMSE	Bais	SE	RMSE	Bais	SE	RMSE
OLS	*	0.0041	0.2243	0.2242	-0.0040	0.1570	0.1570	-0.0065	0.1019	0.1017	0.0000	0.0719	0.0719
	**	0.0021	0.2099	0.2099	0.0020	0.1809	0.1808	-0.0067	0.1229	0.1228	-0.0004	0.0716	0.0716
LRR		0.0041	0.2243	0.2242	-0.0040	0.1570	0.1570	-0.0065	0.1019	0.1017	0.0000	0.0719	0.0719
		0.0021	0.2099	0.2099	0.0020	0.1809	0.1808	-0.0067	0.1229	0.1228	-0.0004	0.0716	0.0716
LRMB	0.00	0.0040	0.2351	0.2350	-0.0047	0.1639	0.1638	-0.0052	0.1047	0.1046	-0.0011	0.0735	0.0735
		0.0049	0.2226	0.2225	0.0010	0.1892	0.1892	-0.0070	0.1289	0.1287	-0.0010	0.0738	0.0738
LRMMB		0.0027	0.2338	0.2338	-0.0065	0.1623	0.1622	-0.0072	0.1047	0.1045	-0.0011	0.0735	0.0735
		0.0037	0.2206	0.2206	-0.0026	0.1877	0.1877	-0.0089	0.1285	0.1282	-0.0010	0.0738	0.0738
LRMGMB		0.0047	0.2340	0.2340	-0.0046	0.1625	0.1624	-0.0052	0.1048	0.1047	-0.0011	0.0736	0.0736
		0.0057	0.2208	0.2207	0.0019	0.1879	0.1879	-0.0069	0.1286	0.1284	-0.0010	0.0737	0.0737
OLS		-0.0001	0.3591	0.3591	-0.0075	0.2391	0.2390	-0.0027	0.1466	0.1465	0.0000	0.1129	0.1129
		-0.0050	0.4069	0.4068	0.0049	0.2407	0.2406	0.0019	0.1941	0.1941	-0.0015	0.1102	0.1102
LRR		-0.0001	0.3591	0.3591	-0.0075	0.2391	0.2390	-0.0027	0.1466	0.1465	0.0000	0.1129	0.1129
		-0.0050	0.4069	0.4068	0.0049	0.2407	0.2406	0.0019	0.1941	0.1941	-0.0015	0.1102	0.1102
LRMB	0.05	-0.0008	0.3751	0.3751	-0.0070	0.2466	0.2465	0.0000	0.1514	0.1514	-0.0009	0.1152	0.1152
		-0.0018	0.4332	0.4332	0.0042	0.2507	0.2507	0.0000	0.1969	0.1969	-0.0019	0.1133	0.1133
LRMMB		-0.0034	0.3747	0.3747	-0.0094	0.2455	0.2453	-0.0026	0.1512	0.1512	-0.0009	0.1152	0.1152
		-0.0014	0.4330	0.4330	0.0031	0.2489	0.2489	-0.0018	0.1965	0.1964	-0.0019	0.1133	0.1133
LRMGMB		-0.0054	0.3743	0.3743	-0.0075	0.2457	0.2456	-0.0006	0.1514	0.1514	-0.0009	0.1152	0.1152
		-0.0005	0.4325	0.4325	0.0051	0.2492	0.2491	0.0001	0.1966	0.1966	-0.0019	0.1133	0.1133
OLS		0.2183	17.506	17.504	-0.3204	10.946	10.941	-0.1756	7.6163	7.6145	0.0512	5.0674	5.0671
		-0.2286	17.570	17.569	0.3188	10.938	10.934	0.1781	7.6482	7.6464	-0.0523	5.0652	5.0650
LRR		-0.0065	0.1361	0.1360	-0.0002	0.0762	0.0762	-0.0004	0.0509	0.0509	-0.0006	0.0365	0.0365
		-0.0028	0.1363	0.1363	-0.0012	0.0761	0.0761	0.00241	0.0503	0.0502	-0.0004	0.0367	0.0367
LRMB	0.99	-0.0855	0.1948	0.1750	-0.0354	0.0948	0.0879	-0.0139	0.0568	0.0551	-0.0074	0.0390	0.0383
		-0.0818	0.1941	0.1760	-0.0364	0.0961	0.0889	-0.0109	0.0557	0.0547	-0.0073	0.0390	0.0383
LRMMB		-0.0859	0.1859	0.1649	-0.0341	0.0935	0.0870	-0.0137	0.0566	0.0549	-0.0009	0.0377	0.0377
		-0.0823	0.1849	0.1655	-0.0351	0.0949	0.0881	-0.0108	0.0556	0.0545	-0.0008	0.0378	0.0378
LRMGMB		0.0783	0.1737	0.1550	-0.0020	0.0825	0.0825	-0.0017	0.0556	0.0556	-0.0055	0.0384	0.0380
		-0.0774	0.1741	0.1559	0.0007	0.0827	0.0827	0.0019	0.0553	0.0553	-0.0057	0.0384	0.0379

* results for $\hat{\beta}_1$,
** results for $\hat{\beta}_2$.

Table 8: MSE ratio with error term distributed normal (0,1)

ρ			0.00				0.05				0.99		
n		20	40	100	200	20	40	100	200	20	40	100	200
LRMGMB	OLS	1.09	1.07	1.06	1.05	1.09	1.06	1.07	1.04	0.00	0.00	0.00	0.00
		1.11	1.08	1.09	1.06	1.13	1.07	1.03	1.06	0.00	0.00	0.00	0.00
	LRR	1.09	1.07	1.06	1.05	1.09	1.06	1.07	1.04	1.63	1.17	1.19	1.11
		1.11	1.08	1.09	1.06	1.13	1.07	1.03	1.06	1.63	1.18	1.21	1.09
	LRMB	0.99	0.98	1.00	1.00	1.00	0.99	1.00	1.00	0.00	0.00	0.00	0.00
		0.98	0.99	1.00	1.00	1.00	0.99	1.00	1.00	0.00	0.00	0.00	0.00
	LRMMB	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.78	0.96	1.04
		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.76	0.99	1.03
LRMMB	OLS	1.09	1.07	1.06	1.05	1.09	1.05	1.06	1.04	0.00	0.00	0.00	0.00
		1.10	1.08	1.09	1.06	1.13	1.07	1.02	1.06	0.00	0.00	0.00	0.00
	LRR	1.09	1.07	1.06	1.05	1.09	1.05	1.06	1.04	1.87	1.51	1.24	1.07
		1.10	1.08	1.09	1.06	1.13	1.07	1.02	1.06	1.84	1.56	1.22	1.06
	LRMB	0.99	0.98	1.00	1.00	1.00	0.99	1.00	1.00	0.00	0.00	0.00	0.00
		0.98	0.98	0.99	1.00	1.00	0.99	1.00	1.00	0.00	0.00	0.00	0.00
LRMB	OLS	1.10	1.09	1.06	1.05	1.09	1.06	1.07	1.04	0.00	0.00	0.00	0.00
		1.12	1.09	1.10	1.06	1.13	1.08	1.03	1.06	0.00	0.00	0.00	0.00
	LRR	1.10	1.09	1.06	1.05	1.09	1.06	1.07	1.04	2.04	1.54	1.24	1.14
		1.12	1.09	1.10	1.06	1.13	1.08	1.03	1.06	2.02	1.59	1.22	1.29
LRR	OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00

Table 9: Bias, RMSE and SE for $\hat{\beta}_1$ and $\hat{\beta}_2$ with error term distributed Cauchy (0,1)

			20			40			100			200	
Mehod	ρ	Bais	SE	RMSE	Bais	SE	RMSE	Bais	SE	RMSE	Bais	SE	RMSE
OLS		1.4948	34.042	34.009	4.1634	141.09	141.03	0.4153	50.326	50.324	-0.6411	40.130	40.125
		-2.659	68.322	68.271	5.1055	215.71	215.65	-1.6505	46.293	46.264	-0.9128	26.931	26.916
LRR		1.494	34.042	34.009	4.1634	141.09	141.03	0.4153	50.326	50.324	-0.6411	40.130	40.125
		-2.659	68.322	68.271	5.1055	215.71	215.65	-1.6505	46.293	46.264	-0.9128	26.931	26.916
LRMB	0.00	-0.0323	0.5968	0.5960	0.0123	0.3236	0.3234	-0.0006	0.1807	0.1807	0.0023	0.1300	0.1300
		0.0123	0.5292	0.5290	-0.0100	0.3140	0.3138	0.0094	0.1774	0.1772	0.0026	0.1214	0.1214
LRMMB		-0.0229	0.6039	0.6034	0.0128	0.3299	0.3296	-0.0005	0.1828	0.1828	0.0027	0.1322	0.1322
		0.0137	0.5258	0.5257	-0.0094	0.3167	0.3166	0.0099	0.1800	0.1797	0.0023	0.1227	0.1227
LRMGMB		-0.0299	0.5239	0.5231	0.0158	0.3052	0.3048	-0.0096	0.1811	0.1809	0.0029	0.1293	0.1294
		0.0064	0.5113	0.5113	-0.0114	0.2998	0.2996	0.0022	0.1706	0.1706	0.0023	0.1197	0.1127
OLS		3.2860	67.582	67.502	-1.8087	89.922	89.904	0.9286	74.635	74.629	-1.3063	46.259	46.241
		-2.8033	70.056	70.000	1.8074	135.43	135.42	-3.5184	100.40	100.34	-2.1963	42.880	42.824
LRR		1.4336	28.667	28.631	-1.8011	89.920	89.902	0.9286	74.635	74.629	-1.3063	46.259	46.241
		-0.706	28.026	28.017	1.8025	135.43	135.42	-3.5184	100.40	100.34	-2.1963	42.880	42.824
LRMB	0.05	-0.0566	0.9062	0.9045	0.0244	0.4845	0.4839	-0.0136	0.2767	0.2763	0.0011	0.1934	0.1934
		0.0373	0.8417	0.8409	-0.0276	0.4687	0.4679	0.0053	0.2823	0.2823	0.0011	0.1919	0.1919
LRMMB		-0.0556	0.8867	0.8849	0.0263	0.4912	0.4905	-0.0145	0.2806	0.2802	0.0020	0.1964	0.1964
		0.0298	0.8467	0.8462	-0.0260	0.4743	0.4736	0.0067	0.2872	0.2871	0.0003	0.1952	0.1952
LRMGMB		0.0023	0.8204	0.8204	-0.0097	0.4479	0.4478	-0.0012	0.2742	0.2743	0.0072	0.1850	0.1848
		-0.0218	0.8291	0.8289	0.0060	0.4405	0.4405	0.0011	0.2817	0.2814	0.0078	0.1878	0.1876
OLS		162.28	3459.9	3456.1	-94.909	5652.8	5652.0	110.30	3962.4	3960.9	18.165	1379.8	1379.7
		161.72	3441.0	3437.4	92.985	5615.7	5614.9	-112.46	4000.2	3998.6	-20.736	1386.0	1385.8
LRR		0.1039	10.995	10.995	-0.9965	23.768	23.747	-1.1316	30.931	30.910	-1.2876	25.519	25.487
		0.1333	11.206	11.205	-1.0047	24.417	24.396	-1.1065	30.284	30.264	-1.2838	25.484	25.451
LRMB	0.99	-0.0852	0.2830	0.2699	-0.0403	0.1605	0.1554	-0.0193	0.0926	0.0906	-0.0073	0.0612	0.0608
		-0.0855	0.2831	0.2699	-0.0395	0.1604	0.1554	-0.0195	0.0925	0.0905	-0.0073	0.0611	0.0609
LRMMB		-0.0064	0.2580	0.2579	-0.0005	0.1537	0.1537	-0.0041	0.0920	0.0919	0.0003	0.0612	0.0612
		-0.0068	0.2580	0.2579	0.0002	0.1544	0.1544	-0.0044	0.0919	0.0918	0.0003	0.0610	0.0610
LRMGMB		-0.0051	0.2508	0.2507	-0.0012	0.1506	0.1506	-0.0130	0.0904	0.0895	-0.0033	0.0607	0.0606
		-0.0051	0.2519	0.2518	-0.0010	0.1500	0.1500	-0.0136	0.0904	0.0894	-0.0031	0.0608	0.0607

Table 10: MSE ratio with error term distributed Cauchy (0,1)

ρ			0.00				0.05				0.99		
n		20	40	100	200	20	40	100	200	20	40	100	200
LRMGMB	OLS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LRR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LRMB	0.77	0.89	0.99	0.98	0.81	0.85	0.98	0.92	0.78	0.88	0.95	0.99
		0.93	0.91	0.92	0.97	0.97	0.89	0.99	0.96	0.79	0.87	0.95	0.99
	LRMMB	0.75	0.86	0.98	0.95	0.86	0.83	0.95	0.89	0.94	0.96	0.97	0.98
		0.94	0.90	0.89	0.95	0.96	0.86	0.96	0.93	0.95	0.94	0.97	0.99
LRMMB	OLS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LRR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LRMB	1.02	1.04	1.02	1.03	0.95	1.03	1.03	1.03	0.00	0.00	0.00	0.00
		0.99	1.02	1.03	1.02	1.01	1.03	1.03	1.03	0.00	0.00	0.00	0.00
LRMB	OLS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LRR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LRR	OLS	1.00	1.00	1.00	1.00	0.18	1.00	1.00	1.00	0.00	0.00	0.00	0.00
		1.00	1.00	1.00	1.00	0.16	1.00	1.00	1.00	0.00	0.00	0.00	0.00