Effects of Outliers and Multicollinearity on Some Estimators of Linear Regression Model

S. A. Ibrahim[1]; W. B. Yahya[2]

[1]Department of Physical Sciences,
Al-Hikmah University,
Ilorin, Nigeria.
e-mail: adeshinas2010@alhikmah.edu.ng[1]

[2]Department of Statistics,
University of Ilorin,
Ilorin, Nigeria.
e-mail: dr.yah2009@gmail.com[2]

***Abstract*— In this study, the effects of twin factors of outliers and multicollinearity on the efficiency of classical multiple linear regression model were investigated. In a Monte-Carlo study, data were simulated under different forms of outliers in the response variable and collinear structures among the independent variables at varying sample sizes. Four regression estimators were employed for models' estimations. In a Monte-Carlo study, datasets plagued with the twin problems of outliers and multicollinearity as earlier stated were simulated to examine the performances of the four chosen estimators over 1000 replications. The results showed that Humber-M (HM) estimator was quite efficient than the least squares (LS) estimator in the present of outliers only, while the Ridge regression (RR) is a better choice than LS in the presence of multicollinearity only. However, the Robust Ridge regression (RRR) and HM were both efficient when the data are plagued with both problems of outliers and multicollinearity simultaneously. In all cases, the mean square error was used as models' assessment criterion.**

Keywords-- *Least squares, Ridge regression, Robust ridge regression, Humber-M estimator, Outliers, Multicollinearity.*

I. Introduction

In classical (multiple) linear regression analysis, the well-known least squares (LS) method is used to estimate the parameters and it is only optimal when the error term in a regression model satisfies Gauss-Markov properties [1-6] among others. However, in practice, these assumptions may not hold due to outliers and multicollinearity in the observational or experimental data, which might render the LS to be less efficient [4,6].

Outlier is known as extreme observation that appears inconsistent with the rest of the data [1]. LS will be biased and/ or not efficient, when normality assumption on the model's error terms fails to hold due to outliers. Robust regression methods like Least absolute value, m-estimators, least median square etc. have been advocated in the literature as alternative to LS in order to model data with outliers[1-4].

Multicollinearity, as described in the study [7, 8] as a situation in which two or more predictors in a multiple regression model are highly correlated. Also, if some of the predictor variables in multiple linear regression models are correlated, least square estimator becomes less efficient and unstable, thereby rendering the resulting regression model unsuitable for meaningful inference [8]. Some bias estimators have been provided in literature to model data with multicollinearity. Bias estimators include the ridge regression, principal component regression [4, 8, 9, 10] among others.

Outliers and multicollinearity problems could occur simultaneously in a data set almost as often as each problem could occur separately. Robust-ridge regression has been suggested in literature to model data with both outliers and multicollinearity simultaneously. Therefore this study is aimed to investigate the efficiency of some estimators of multiple linear regression models in the presence of outliers and multicollinearity against least square estimator, using mean square error (MSE) as models' assessment criterion.

II. Materials and Methods

Brief descriptions of the mathematical development of methods used are provided in what follow.
Consider a multiple linear regression model of the form:

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

where Y is an $n \times 1$ response vector, X is an $n \times m$ matrix of data (i.e. n observation on m variables) with rank m, β is a $m \times 1$ vector of model's parameters, and ε is $n \times 1$ random vector with independent, identically and normally distributed elements, i.e. $\varepsilon_i \sim N(0, \sigma^2)$. We provide brief theoretical backgrounds of some of the estimators of model (1) in the presence or absence of problems of outliers and multicollinearity.

1. *Least Square (LS) Method*

One of the main goals in regression analysis is to find the best estimates of unknown parameters in the model. The traditional method commonly used is the LS and it has the best performance if the error term ε in (1) has a normal probability distribution. Hence, the LS estimator is obtained by minimizing the sum of squares errors.

$$min \sum \varepsilon^2 = min \sum (Y - X\beta)^2 \qquad (2)$$

which resulted to the LS estimator given by:

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

(3)

The equation (2) is the LS estimator of model (1). However, LS method is sensitive to the existence of outliers, because each data point is equally weighted in the process of determining the parameters of the model. Therefore, robust estimation techniques which weight each data point based on statistical criteria are desired for modeling data set that are plague with outliers.

2. *Robust M-Estimator*

The most common general method of robust regression is the M-estimation, proposed by Huber (1964) which was claimed to be nearly as efficient as LS under the violation of model's Gaussian error. Instead of minimizing the sum of squared errors as the objective function, the M-estimate of the objective function is given as:

$$min \sum \rho \left(\frac{r_i}{s}\right) = min \sum \rho \left(\frac{Y_i - X_i'\beta}{s}\right) \qquad (4)$$

where $r_i = (Y_i - X_i'\beta)$ and S is a robust estimate of scale parameter σ and can be estimated by using the formula

$$S = \frac{median|\, r_i - median(r_i)|}{h} \qquad (5)$$

In (5), h is suggested to be 0.6745 which make S an approximately unbiased estimator of σ if n is large and error distribution is normal (Draper and Smith, 1998). A reasonable ρ should satisfy the following properties:

$\rho(r_i) \geq 0, \rho(0) = 0, \rho(r_i) = \rho(-r_i), \rho(r_i) \geq \rho(r_i')$ for $|r_i| \geq |r_i'|$ (6)

To minimize equation (4), equate the first partial derivatives of ρ with respect to $\beta_j\ (j = 0,1)$ to zero, yielding a necessary condition for a minimum. This gives the system of $p = k + 1$ equation.

$$\sum X_{ij} \psi \left(\frac{Y_i - X_i'\beta}{s}\right) = 0, \quad j = 0,1 \qquad (7)$$

where ψ is the $\rho'$and X_{ij} is the i^{th} observation on the j^{th} predictor. The equation (7) does not have an explicit solution. In general, an iterative methods or nonlinear optimization techniques will be used to solve them. In fact, iterative reweighted least squares (IRLS) might be used. To use (IRLS), first the weight has to be defined normal (Draper and Smith, 1998) as

$$w_{iB} = \begin{cases} \dfrac{\psi\left(\frac{Y_i - X_i'\beta}{S}\right)}{\left(\frac{Y_i - X_i'\beta}{S}\right)} & ,Y \neq X\beta \\ 1 & Y = X\beta \end{cases} \qquad (8)$$

Then equation (8) becomes

$$\sum X_{ij}\, w_{iB}(Y_i - X_i'\beta) = 0 \qquad (9)$$

The equation (9) can be written in matrix notation as:

$$X'\, w_{iB} X\beta = X'\, w_{iB} \mathrm{Y} \qquad (10)$$

where w_{iB} is an $n \times n$ diagonal matrix of weights with diagonal elements: $w_{iB}, w_{2B}, \ldots, w_{nB}$ is given by equation (8). The equation (9) and (10) can be considered as the usual least square normal equation. Meanwhile, Several choices of ρ have been proposed in the literature as shown in Table 1. One of these was used in this study, which is Huber m-estimation.

Table 1: Table of Huber objective, influential and weight functions.

Objective Function	Influential Function	Weight Function
$\rho(r_i) = \begin{cases} \frac{r_i^2}{2}, & for\ \lvert r_i \rvert \le k \\ k\lvert r_i \rvert - \frac{k^2}{2}, & for\ \lvert r_i \rvert > k \end{cases}$	$\psi(r_i) = \begin{cases} k & for\ \ r_i > \ k \\ r_i & for\ \ \lvert r_i \rvert \le k \\ -k & for\ \ r_i < -k \end{cases}$	$w(r_i) = \begin{cases} 1, & for\ \lvert r_i \rvert \le k \\ k/\lvert r_i \rvert, & for\ \lvert r_i \rvert > k \end{cases}$

The value of k = 1.345 and is refer to as tuning constant, chosen to achieve desired efficiency.

3. *Ridge Regression*

The ridge regression (RR) was originally proposed by [9,10]. It is an advance tool to model data with multicollinearity. The objective of ridge regression is to reduce the size and variance of the least squares estimates by introducing a slight amount of bias. The ridge regression estimator is given as:

$$\hat{\beta}_R = (X'X + kI)^{-1}\, X'Y \qquad (11)$$

where I is known as $q \times q$ identity matrix and k is called biasing parameter. Meanwhile, various methods for determine the value of k have been proposed in the literature. Thus, this study employed a method described by Hoerl, Kennard and Baldwin (1975) to obtained the value of k where $k = \hat{k}_{HKB}$ given by

$$\hat{k}_{HKB} = \frac{PS_{LS}^2}{\sum \beta_{LS\,i}^2} \quad , i = 1,2, \dots, p \qquad (12)$$

and $$S_{LS}^2 = \frac{(Y - X\beta_{LS})'(Y - X\beta_{LS})}{n-p} \qquad (13)$$

4. *Robust–Ridge Regression Huber m-estimator*

The Robust–Ridge Regression Huber m-estimator (RR-Hub) is given as follow:

$$\hat{\beta}_{RR-HUB} = (X'X + k_{HUB}\, I)^{-1}\, X'Y \qquad (14)$$

where k_{HUB} is called the robust ridge bias parameter and it is obtained from the Huber m-estimator, instead of using $\hat{k}_{HKB}$ obtained from LS. The k_{HUB} is determined from data using

$$\hat{k}_{HUB} = \frac{PS_{HUB}^2}{\sum \beta_{HUB\,i}^2} \quad , i = 1,2, \dots .\, p \qquad (15)$$

The robust scale for RR Hub-M-estimator is defined as:

$$S_{HUB}^2 = \frac{(Y - X\beta_{HUB})'(Y - X\beta_{HUB})}{n-p} \qquad (16)$$

The $\hat{k}_{HUB}$ is aimed to reduce the effect of outlier on the value chosen for the biasing parameter [11].

III. Simulation Study

The simulation scheme adopted here was adapted from those employed by [3, 4, 11, 12] to evaluate the performance of the four estimators considered here: LS, Hub, RR, and RR-Hub.

The simulation was designed to allow both the outliers and multicollinearity problems to be present in data simultaneously. The following linear regression model is used for modelling.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2\, X_{i2} + \beta_3 X_{i3} \ + \varepsilon_i \quad (17)$$

$i = 1,2,3, \dots , n$

The values of the model's parameters were set as: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$. The predictor variables were generated [3, 4] from

$$X_{ij} = (1-\rho)z_{ij} + \rho z_{ij}, \quad j = 1,2,3 \qquad (18)$$

where z_{ij} is independent standard normal random variable and was generated from the standard normal distribution. ρ represents correlation between two explanatory variables and its values were chosen as: 0.0, 0.95, and 0.99. Two sample sizes n = 20 and n = 100 were used for the study.

The error term of model (17) were simulated from the following three distributions:

$Case\ I$: $\varepsilon \sim N(0,1)$ (19)

$Case\ II$: $\varepsilon \sim N(0,1)$ with identical outliers in Y direction (where we let the first two value of $Y's$ equal to 20) (20)

$Case\ III$: $\varepsilon \sim 0.9N(0,1) + 0.1N(0,100)$ (21)

which is mixture of two normal populations.

For comparison purpose, the Mean Square Error (MSE) was used as models' assessment criterion. R statistical packages was employed for all analysis.

IV. Analysis And Results

Each of the estimator was used to fit the regression model (17) at 1,000 replications over the two sample sizes n = 20,100. Thus, the estimates of the parameters provided by each estimator were the average values obtained over 1000 iterations.

The performances of all the four estimators were assessed using the MSE at the two chosen sample sizes at varying levels of collinearity and specified error term distributions. These results are presented in Tables 2 through Table 10.

Table 2: Table of MSEs of the estimated parameters of multiple linear regression models with different distributions of error term and no multicollinearity among the predictor variables at sample sizes $n = 20, 100$.The true parameter values are in parentheses.

Distribution of Error Term & outlier defined on Y	Estimator	$\rho = 0.0$							
		n = 20				n=100			
		$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$	$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$
$\varepsilon \sim N(0,1)$	LS	**0.0566**	**0.0696**	**0.0681**	**0.064**	**0.0105**	**0.0102**	**0.0099**	**0.0102**
	HM	0.0624	0.0759	0.0704	0.0668	0.0113	0.0109	0.0104	0.011
	Ridge	0.1925	0.0813	0.0848	0.0703	0.0436	0.0151	0.0145	0.0153
	RR-Hub	0.1925	0.0814	0.085	0.0705	0.0436	0.0151	0.0145	0.0153
$\varepsilon \sim N(0,1)$ & 10% outliers in Y	LS	3.9114	2.317	2.5842	2.2579	3.6475	0.3635	0.3814	0.3682
	HM	**0.145**	**0.1787**	**0.1851**	**0.1401**	**0.0546**	**0.0178**	**0.0173**	**0.0168**
	Ridge	3.8227	1.4981	1.6143	1.4796	3.6705	0.309	0.3331	0.311
	RR-Hub	3.8227	0.7825	0.8204	0.7762	3.6705	0.2965	0.3172	0.2955
$\varepsilon \sim 0.9N(0,1)$ $+ 0.1N(0,100)$	LS	59.0539	66.1498	67.0868	70.4422	10.2378	11.8012	9.8567	10.4803
	HM	**0.1017**	**0.1475**	**0.1763**	**0.1311**	**0.0167**	**0.0171**	**0.0154**	**0.0163**
	Ridge	50.335	47.9418	48.7244	48.4467	9.8721	4.408	3.8876	4.164
	RR-Hub	50.335	0.9804	0.9795	0.984	9.8721	0.9277	0.9437	0.9501

Table 3: Table of MSEs of the estimated parameters of multiple linear regression models with different distributions of error term and 95% multicollinearity among the predictor variables at sample sizes $n = 20, 100$.The true parameter values are in parentheses.

Distribution of Error Term & outlier defined on Y	Estimator	$\rho = 0.95$							
		n = 20				n=100			
		$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$	$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$
$\varepsilon \sim N(0,1)$	LS	**0.0574**	4.6681	4.7363	4.2086	**0.0107**	0.6966	0.693	0.7454
	HM	0.0627	5.0486	4.9857	4.4617	0.0115	0.7602	0.7298	0.8001
	Ridge	0.4545	**0.5936**	**0.5867**	**0.5298**	0.0883	**0.1442**	**0.1448**	**0.1522**
	RR-Hub	0.4545	0.8439	0.8352	0.7516	0.0883	0.1513	0.1519	0.1598
$\varepsilon \sim N(0,1)$ & 10% outliers in Y	LS	4.0233	166.6803	181.8337	169.1736	3.672	25.2337	26.1652	25.4915
	HM	**0.1445**	11.4819	11.7982	9.483	**0.0552**	1.2495	1.1901	1.1979
	Ridge	4.1605	15.9247	17.2198	17.1586	3.7141	5.5452	5.7038	5.6498
	RR-Hub	4.1605	**0.302**	**0.3159**	**0.3217**	3.7141	**0.1804**	**0.1696**	**0.1755**
$\varepsilon \sim 0.9N(0,1)$ $+ 0.1N(0,100)$	LS	59.7033	4737.745	4778.024	4816.7566	10.363	792.5691	729.4988	752.0988
	HM	**0.1251**	10.8605	12.5775	14.5781	**0.017**	1.1559	1.1051	1.1965
	Ridge	50.6239	667.8211	628.8046	677.9194	9.9194	196.2786	182.5951	186.6809
	RR-Hub	50.6239	**0.8877**	**0.8864**	**0.8893**	9.9194	**0.7986**	**0.8006**	**0.8011**

Table 4: Table of MSEs of the estimated parameters of multiple linear regression models with different distributions of error term and 99% multicollinearity among the predictor variables at sample sizes $n = 20, 100$.The true parameter values are in parentheses.

Distribution of Error Term & outlier defined on Y	Estimator	$\rho = 0.99$							
		n = 20				n=100			
		$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$	$\hat{\beta}_0(1)$	$\hat{\beta}_1(1)$	$\hat{\beta}_2(1)$	$\hat{\beta}_3(1)$
$\varepsilon \sim N(0,1)$	LS	**0.0573**	111.5954	113.791	101.0158	**0.0107**	16.5687	16.6838	17.9134
	HM	0.0627	120.7104	119.6564	106.9792	0.0115	18.084	17.5628	19.2371
	Ridge	0.4856	**8.5304**	**8.3871**	**7.6185**	0.0951	**0.3086**	**0.3092**	**0.3263**
	RR-Hub	0.4856	14.9931	14.7933	13.3675	0.0951	0.5455	0.5494	0.5795
$\varepsilon \sim N(0,1)$ & 10% outliers in Y	LS	4.0278	4007.161	4373.86	4044.6441	3.6734	603.8317	629.2534	612.3157
	HM	**0.1444**	273.3966	284.7591	225.8291	**0.0552**	29.7516	28.6522	28.7816
	Ridge	4.1916	332.9719	358.4006	354.7005	3.7198	119.6328	125.2069	123.15
	RR-Hub	4.1916	**0.3896**	**0.4284**	**0.4459**	3.7198	**1.0157**	**1.0105**	**1.0311**
$\varepsilon \sim 0.9N(0,1) + 0.1N(0,100)$	LS	59.793	113566.3	115431.8	115092.	10.36	19014.58	17473.38	18026.14
	HM	**0.1213**	255.586	282.604	321.0227	**0.017**	27.508	26.5915	28.7669
	Ridge	50.662	16134.5	15358.6	16215.76	9.9277	4660.53	4329.08	4438.991
	RR-Hub	50.6623	**2.8343**	**2.8156**	**2.8561**	9.9277	**0.6037**	**0.6047**	**0.6048**

IV. Discussion

In this study, the efficiency of four estimators (LS, RR, HM, RR-Hub) for multiple linear regression models were examined in the presence of the twin problems of outlier and multicollinearity in the data.

In dataset for which the above two problems are absent, the LS estimator with the least MSE of parameters estimates was quite efficient among the four estimators considered as shown by the results in Table 2. However, when outliers are present in the response variable, the HM estimator was relatively more efficient than the other three estimators (see results in Table 2).

When only the multicollinearity problem is present in the data, the RR with minimum MSE was the best estimators among the four considered (see results in Tables 3 and 4).

In situations where the data is suffering from the two problems of outliers and multicollinearity, the RR-Hub estimator was found to be relatively most efficient among the four estimators considered as can be observed from the results of MSEs in Tables 3 and 4.

V. Conclusion

In multiple linear regression analysis, using the LS estimator for modelling in the presence of outliers and multicollinearity can result to less efficient results with large standard errors of parameters estimates. This may consequently lead to poor predictive power of the model and statistical inferences from such model might not be reliable.

Whenever any of the problems outliers in the response variable and multicollinearity in the predictor variables are suspected, it might be desirable to employ appropriate estimators, other than the LS, to fit the regression model for better results.

References

[1] Alma, O. G. (2011). Comparison of Robust Regression Methods in Linear Regression. Int. J. Contemp. Math. Sciences, Vol. 6 No. 9, 409 – 421.

[2] Andrews, D.F. (1974). A robust method for multiple linear regression. Technometrics, 16(4), 523 – 531.

[3] Kafi, D.P., Robiah, A., & Bello, A.R. (2014). Ridge Least Trimmed Squares Estimators in Presence of Multicollinearity and Outliers. Nature and Science, 12(12).

[4] Kafi, D.P., Robiah, A., & Bello, A.R. (2015). Using Riedge Least Median Squares to Estimate the Parameter by Solving Multicollinearity and Outliers Problems. Modern Applied Science, Vol. 9, No. 2.

[5] Nadia, H., & Mohammad, A.A. (2013). Model of Robust Regression with Parametric and Nonparametric Methods.

[6] Yahya, W.B., Adebayo, S.B., Jolayemi, E.T., Oyejola, B.A., & Sanni, O.O.M.(2008). Effects of non-orthogonality on the efficiency of seemingly unrelated regression (SUR) models. *InterStat Journal*, 1-29.

[7] Madhulika, D., & Isha. (2014). A Review on the Biasing Parameters of Ridge Regression Estimator in LRM. Mathematical Journal of Interdisciplinary Sciences, Vol. 3 No. 1.

[8] Yahya, W.B. & Olaifa, J.B.(2014) A note on ridge regression modelling techniques. *Electronic Journal of Applied Statistical Analysis*. Vol. 07, Issue 02, 2014, 343-361.

.

[9] Hoerl, A.E., & Kennard, R.W. (1970a). Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), 69-82.

[10] Hoerl, A.E., & Kennard, R.W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[11] El-Salam, M. E. – F.A. (2013). The Efficiency of Some Robust Ridge Regression for Handling Multicollinearity and Non-Normals Errors Problems. Applied Mathematical Sciences, 7(77), 3831-3846.

[12] Zahari, S.M., Zainol, M. S., Al-Banna, M.I.(2012) Proceeding of Mathematical models and methods in modern science. 124-129

.

.