International Journal of Contemporary Mathematical Sciences
Vol. 10, 2015, no. 3, 119 - 130
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ijcms.2015.5210

Alternative Ridge Robust Regression Estimator for Dealing with Collinear Influential Data Points

Moawad El-Fallah Abd El-Salam

Department of Statistics & Mathematics and Insurance
Faculty of Commerce, Zagazig University, Egypt

Abstract

The multicollinearity in multiple linear regression models and the existence of influential data points are common problems. These problems exert undesirable effects on the least squares estimators. So, it is very important to introduce some alternative biased estimators of the robust ridge regression to overcome the influence of these problems simultaneously. In this paper, alternative biased robust regression estimator is defined by mixing the ridge estimation technique into the robust least median squares estimation to obtain the Ridge Least Median Squares (RLMS). The efficiency of the combined estimator (RLMS) is compared with some existing regression estimators, which namely, the Ordinary Least Squares (LS); Ridge Regression (RR) and Ridge Least Absolute Deviation (RLAD). The numerical results of this study show that, the RLMS regression estimator is more efficient than other estimators, based on, Bias and mean squared error criteria for many combinations of influential data points and degree of multicollinearity.

Keywords: Influential Data Points; Multicollinearity; Ridge regression; Ridge Least Absolute Deviation; Ridge Least Median Squares estimation; Bias and Mean Squared Error criteria

1. Introduction

Two important problems are considered in regression analysis; multicollinearity and the existence of influential data points. The ordinary least squares estimators (LS) of coefficients are known to possess certain optimal proper-

ties when explanatory variables are not correlated among themselves, and the disturbances of the regression equation are independent, identically distributed normal random variables. The presence of correlation among the explanatory variables may result in imprecise information being available about the regression coefficients. In addition, the least squares estimator may produce extremely poor estimates in the presence of leverage or influential data points. Thus, various remedial techniques have been suggested for these problems separately. One such remedial technique is ridge regression to deal with multicollinearity, and the robust estimation techniques are not as strongly affected by the presence of influential data points. However, although, we usually think of these two problems separately, but in practical situations, these problems occur simultaneously. To remedy these two problems simultaneously, several robust ridge regression estimators have been put forward that are much less influenced by the influential data points and multicollinearity. Askin and Montgomery (1980), suggested combining the ridge and the least absolute deviation (LAD) robust regression techniques. Montgomery and peck (1982), have suggested that either robust or ridge estimation methods alone may be sufficient for dealing with the combined problem. In this paper, we take the initiative to develop a more robust technique to remedy these two problems. We proposed combining the ridge regression with the highly efficient and high breakdown point estimator, namely the Ridge Least Median Squares (RLMS) estimator. We call this modified method, the robust ridge regression based on Least Median Squares estimation (RLMS). We expect that, the modified method would be less sensitive to the presence of influential points and multicollinearity. So, the aim of this paper is devoted to examine some estimators which are resistant to the combined problems of multicollinearity and influential points. Exactly, can the ridge estimators and some robust estimation techniques be combined to produce a robust ridge regression estimator?. The remainder of the paper is organized as follows. In section (2), the ridge regression estimator will be reviewed. The robust regression estimation will be discussed in section (3). In section (4), we discuss the augmented ridge robust estimators as a way of combining biased and robust regression techniques, while, Section (5) introduces the proposed combined ridge robust estimator (RLMS). Section (6) presents the results of a Monte Carlo simulation study to investigate how such estimators perform well, and some concluding remarks are presented in section (7).

2. Ridge Regression Estimators

Consider the following linear regression model:

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

where : y is an $(n \times 1)$ vector of observations on the dependent variable , X is an $(n \times p)$ matrix of observations on the explanatory variables, β is a $(p \times 1)$ vector of regression coefficients to be estimated , and ε is an $(n \times 1)$ vector of disturbances. The least squares estimator of β can be written as:

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

This method gives unbiased and minimum variance among all unbiased linear estimators provided that the errors are independent and identically, normally distributed. However, in the presence of multicollinearity, the singularities present in $(X'X)$ matrix and this ill-conditioned X matrix can result in very poor estimates. The degree of multicollinearity is often indicated by conditioned number (CN) of the matrix X $(or\ X'X)$. *CN* is defined as the ratio of the largest singular values of X to the smallest,

$$CN\ (X) = \frac{\lambda_{max}}{\lambda_{min}} \geq 1\ , \tag{3}$$

where: λ are the eigenvalues of the matrix $(X'X)$.

Belsley et al. (1980) have empirically shown that weak dependencies are linked to CN around 5 to 10, whereas moderate to strong relations are linked to CN of 30 to 100. Hoerl and Kennard (1976) pointed out that adding a small constant to the diagonal of a matrix, will improve the conditioning of a matrix as this would dramatically reduced its CN. The ridge regression is defined as follows:

$$\hat{\beta}_{RR} = (X'X + KI)^{-1} X'Y\ , \tag{4}$$

where**:** I is the $(p \times p)$ identity matrix and K is the biasing constant. Various methods for determining K value been introduced in the literature such as described by Hoerl and Kennard (1976) and Gibbons (1981) as:

$$\hat{K}_H = \frac{PS^2_{LS}}{\beta'_{LS}\hat{\beta}_{LS}}\ , \tag{5}$$

where,

$$S^2_{LS} = \frac{(Y - X\hat{\beta}_{LS})'(Y - X\hat{\beta}_{LS})}{n - p} \tag{6}$$

when $k = 0$, $\hat{\beta}_{RR} = \hat{\beta}_{LS}$, when $K > 0$, $\hat{\beta}_{RR}$ is biased but more stable and precise than LS estimator and when $K \to \infty$, $\hat{\beta}_{RR} \to 0$. Hoerl and Kennard (1976) have shown that, there always exist a value $K > 0$ such that $MSE(\hat{\beta}_{RR})$ is Less than $MSE(\hat{\beta}_{LS})$.

3. Robust Regression Estimators

Robust regression estimators have been proven to be more reliable and efficient than least squares estimator especially when the data are contaminated with influential observations. Since the influential data points greatly influence the estimated coefficients, standard errors and test statistics, the usual statistical procedure may be most inefficient as the precision of the estimator has been affected. Several different robust regression estimators exist. Two of the most commonly considered are: LAD-estimators and LMS-estimators.

3.1 The Least Absolute Deviation Estimator (LAD):

The LAD estimator, $\hat{\beta}_{LAD}$, can be defined as the solution to the following minimization problem :

$$\min \sum_{i=1}^{n} \left| Y_1 - X'_i \beta_{LAD} \right| \tag{7}$$

Rather than minimizing the sum of squared residuals as in least squares estimation, the sum of the absolute values of the residuals is minimized. Thus, the effect of influential data points on the LAD estimates will be less than that on LS estimates.

3.2 The Least Median Squares Estimator (LMS):

The Least Median Squares (LMS) estimator was proposed by Rousseeuw and Leroy (1987) and has the advantage of minimizing the influence of influential data points. According to (Venables and Ripley, 1999), this estimation minimizes the median of ordered squares of residuals in order to get the estimated of the regression coefficients $\hat{\beta}_{LMS}$, which can be defined by:

$$\min \text{median} \quad \left| Y_1 - X'_i \beta_{LMS} \right|^2 \tag{8}$$

4. Robust Ridge Regression Estimators

There are many studies that have been related using the robust ridge regression estimators in literature such as: Pariente and Welsch (1977); Askin and Montgomery (1980); Pfaffenberger and Dieman (1984); Moawad El-Fallah (2013) and Mal and Dul (2014). In this section, we present some combinations of ridge and robust regression estimation discussed in sections (2) and (3) respectively. In this respect, the RLAD estimator, which is based on the LAD and

ridge estimators denoted by , $\hat{\beta}_{RLAD}$, can be computed using the following:

$$\hat{\beta}_{RLAD} = (X'X + K^{*}I)^{-1}X'Y \quad , \tag{9}$$

where the value of K^{*} is determined from data using:

$$K^{*} = \frac{PS^{2}_{LAD}}{\hat{\beta}'_{LAD}\hat{\beta}_{LAD}} \tag{10}$$

and,

$$S^{2} = \frac{(Y - X\hat{\beta}_{LAD})'(Y - X\hat{\beta}_{LAD})}{n-p} \quad , \tag{11}$$

where , $\hat{\beta}_{LAD}$ is the LAD estimator defined as the solution to equation (7). It be noted that the value of K^{*} is the estimator of K presented by equation (5) with two changes. First, the LAD estimator of β is used rather than LS estimator. Second, the estimator of σ^{2} used in equation (11) is modified by the LAD coefficient estimates rather than the least squares estimates. These changes are aimed to reduce the effect of extreme points on the value chosen for the biasing parameter.

5. The Proposed Augmented Estimator

Instead of $\hat{\beta}_{RLAD}$ estimator which was aimed to reduce the effect of influential data points on the value chosen for the biasing parameter k. Another alternative augmented estimator between robust and ridge regression estimation is $\hat{\beta}_{RLMS}$. In this respect, it is hoped that, the problems of multicollinearity and influential data points can be solved simultaneously. The $\hat{\beta}_{RLMS}$ estimator can be calculated by the following.

$$\hat{\beta}_{RLMS} = (X'X + K^{*}_{LMS}\, I)^{-1}X'Y \quad , \tag{12}$$

where the value of K^{*} is given by:

$$K^{*}\text{LMS} = \frac{PS^{2}_{LMS}}{\hat{\beta}'_{LMS}\hat{\beta}_{LMS}} \tag{13}$$

and

$$S^2_{\text{LMS}} = \frac{(Y - X\hat{\beta}_{LMS})'(Y - X\hat{\beta}_{LMS})}{n-p} \quad , \qquad (14)$$

where, p is the number of independent variables, n is the number of observations in the data and $\hat{\beta}_{LMS}$ is the estimator of β using the least median of squares.

6. Simulation Study

(6.1) Design of the Experiment:

We carry out a Monte Carlo simulation study to compare the performance of the different estimators $\hat{\beta}_{LS}$, $\hat{\beta}_{RR}$ and $\hat{\beta}_{RLAD}$ with the proposed estimator $\hat{\beta}_{RLMS}$. The simulation is designed to allow both multicollinearity and influential data points to exist simultaneously.

Suppose, we have the following linear regression model:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad \text{where } i = 1,2,..,n \qquad (15)$$

Siti et. al (2012), pointed out that, the parameter values of β_o , β_1 and β_2 are set equal to one . The explanatory variables x_{i1} and x_{i2} are generated as:

$$x_{ij} = (1-\rho^2)\, z_{ij} + \rho\, z_{ij} \quad , \; i = 1,2,\ldots,n \; , \; j = 1,2 \; , \qquad (16)$$

where, z_{ij} are independent standard normal random numbers generated by the normal distribution. The value of ρ representing the correlation between the two explanatory variables and its values were chosen as : 0.0 and 0.99. Once, for a given sample size n, the explanatory variables values were generated. The sample sizes which will be examined in this study are: 20 and 60 and the percentage of influential points present in this data is 30 %. The number of replications used is 1000 .The statistics computed are the bias (B) , mean squared error(MSE), standard error (SE), and 6 pairwise MSE ratios of the estimators. The bias and MSE are given as:

Bias= ($\bar{\beta}_j - \beta_j$) ,

$$\bar{\beta}_j = \frac{\sum_{l=1}^{1000} (\hat{\beta}_{jl})}{1000} \quad , \qquad \text{j= 1,2 , l=1,2,…,1000,}$$

and

$$MSE_j = \frac{1}{1000} \sum_{l=1}^{1000} (\hat{\beta}_j - \beta_j)^2 \qquad (17)$$

(6.2) The Results of comparisons:

Table (1) presents the CN results for the simulated data, in that case when ρ = 0.99, which representing high correlation between the two explanatory variables.

Table (1): The CN for the simulated data.

ρ	ρ = 0.99	
Var	x_1	x_2
CN (n=20)	62.101	67.192
CN (n=60)	76.118	82.532

The results of Table (1) show that, the maximum value of CN is 82.532 when the correlation between the two explanatory variables was very high with different sample size. So, it is clear that the multicollinearity problem exists.

On the other hand, Tables (2) and (3) show the summary statistics such as: bias (B), mean squared error (MSE) and standard error (SE) for all estimators of the normal distribution for different two sample sizes (20 and 60) with 0% and 30 % influential observations and different values of ρ.

In addition, for the purpose of comparison, Table (4) shows the relative efficiency of the proposed estimator (RLMS) to other estimators (OLS RR and RLAD) based on the MSE ratios, which are written as:

$$MSE_{\text{ratios}} = \frac{MSE\ of\ (RLMS)}{MSE\ of\ (others)} \qquad (18)$$

In this case, values of MSE_{ratios} less than 1, denotes that, the RLMS estimator is more efficient, however, values greater than 1 denotes that, other estimator is more efficient.

Table (2): Bias, MSE and SE of $\hat{\beta}_1$ and $\hat{\beta}_2$ using error normal distribution of (n=20 and 60)) with 0% influential observations and ρ = 0.0, 0.99.

ρ	N	Coef.	Statist.	0 % Influential Observations			
				OLS	Ridge	RLAD	RLMS
0.00	20	$\hat{\beta}_1$	Bias	-.002	-0.301	-0.441	-0.331
			MSE	0.169	0.251	0.254	0.243
			SE	0.169	0.233	0.229	0.221
		$\hat{\beta}_2$	Bias	-.004	-0.362	-0.446	-0.392
			MSE	0.173	0.255	0.259	0.248
			SE	0.165	0.246	0.247	0.244
	60	$\hat{\beta}_1$	Bias	0.006	-0.229	-0.439	-0.287
			MSE	0.158	0.243	0.251	0.240
			SE	0.156	0.041	0.237	0.231
		$\hat{\beta}_2$	Bias	0.013	-0.295	-0.442	-0.371
			MSE	0.156	0.249	0.250	0.241
			SE	0.152	0.244	0.241	0.232
0.99	20	$\hat{\beta}_1$	Bias	1.011	-0.721	-0.672	-0.671
			MSE	12.16	1.011	0.992	0.971
			SE	10.12	0.932	0.332	0.212
		$\hat{\beta}_2$	Bias	-.963	-0.852	-0.732	-0.681
			MSE	10.22	0.493	0.364	0.353
			SE	9.244	1.009	0.126	0.104
	60	$\hat{\beta}_1$	Bias	0.853	-0.328	-0.541	-0.478
			MSE	4.211	0.531	0.368	0.302
			SE	3.031	0.326	0.261	0.231
		$\hat{\beta}_2$	Bias	2.467	-0.931	-0.811	-0.731
			MSE	8.116	0.514	0.469	0.398
			SE	8.001	1.984	0.158	0.153

Table (3): Bias, MSE and SE of $\hat{\beta}_1$ and $\hat{\beta}_2$ using error normal distribution of (n=20 and 60) with 30% influential observations and ρ = 0.0, 0.99.

ρ	N	Coef.	Statist.	20 % Influential Observations			
				OLS	Ridge	RLAD	RLMS
0.0	20	$\hat{\beta}_1$	Bias	-1.342	-1.135	-0.826	-0.515
			MSE	11.341	0.720	0.665	0.573
			SE	11.232	0.652	0.657	0.492
		$\hat{\beta}_2$	Bias	-1.053	-1.212	-1.025	-1.013
			MSE	12.178	0.886	0.789	0.648
			SE	12.053	0.769	0.047	0.641
	60	$\hat{\beta}_1$	Bias	-1.981	-1.534	-1.223	-1.400
			MSE	8.618	0.979	0.950	0.783
			SE	8.304	0.876	0.845	0.679
		$\hat{\beta}_2$	Bias	-1.887	-1.980	-1.349	-1.201
			MSE	10.222	0.780	0.921	0.532
			SE	10.105	0.655	0.868	0.502
0.99	20	$\hat{\beta}_1$	Bias	-2.511	-1.134	-1.826	-1.515
			MSE	24.343	1.523	1.265	1.103
			SE	24.026	1.368	1.162	1.099
		$\hat{\beta}_2$	Bias	-2.004	-1.362	-1.216	-1.392
			MSE	26.175	0.758	0.659	0.548
			SE	26.123	0.649	0.597	0.538
	60	$\hat{\beta}_1$	Bias	-1.924	-1.273	-1.043	-1.279
			MSE	16.815	1.063	1.119	0.983
			SE	16.310	1.023	1.099	0.792
		$\hat{\beta}_2$	Bias	-1.772	-1.586	-1.314	-0.688
			MSE	12.186	0.910	0.762	0.652
			SE	12.074	0.879	0.747	0.597

Table (4): MSE ratios of 6 pairwise estimator of $\hat{\beta}_1$ only using normal distribution with 0 % and 30 % influential observations and ρ = 0.0, 0.99.

Estimator 1 vs Estimator 2 vs Estimator 3 vs Estimator 4	Values of ρ for 0 % influential observations.		Values of ρ for 30 % influential observations.	
	0.00	0.99	0.00	0.99

Table (4): (Continued): MSE ratios of 6 pairwise estimator of $\hat{\beta}_1$ only using normal distribution with 0 % and 30 % influential observations and ρ = 0.0, 0.99.

RLMS	OLS	1.012	0.671	0.791	0.595
	RR	1.023	1.924	0.611	0.843
	RLAD	1.002	0.547	0.923	0.648
RLAD	OLS	1.236	0.875	0.749	0.853
	RR	1.001	1.026	0.698	0.972
RR	OLS	1.153	0.732	1.002	0.820

The results of Table (2) show that, The MSE of the OLS estimators are relatively smaller than the other estimators when the errors are normally distributed without multicollinearity and no influential data points. However, for the case of high correlation between variables (ρ=0.99), the results of all other estimators are better than the OLS. While, the results of Table (3) show that, the RLMS estimator is more efficient than the others (OLS, RR and RLAD), especially in the presence of multicollinearity and influential data points. In addition, from the results of Table (4), (when ρ=0.00 and no influential points) it be noted that, the values of MSE ratios for the RLMS , RLAD and RR estimators to OLS are greater than 1.00, denoting that the OLS estimator is more efficient than other estimators when no multicollinearity and no influential data points. However, when multicollinearity and influential points are present in the data, the RLMS estimator is better than the OLS, RLAD and RR estimators. To conclude, the results from comparisons of RLMS estimator to OLS, RLAD and RR estimators denote that the RLMS estimator is more efficient than RR and RLAD, and much more efficient than OLS when multicollinearity and influential points are present in the data.

7. Concluding Remarks

The presence of influential data points and multicollinearity are considered two of the more frequent problems in regression analysis. Although, we usually think of these two problems separately, however, these problems occur simultaneously in applied situations. A Monte Carlo simulation was designed to compare the performance of some augmented ridge and robust regression estimators for dealing with these two problems. The results of comparisons indicated that, the ridge least median squares (RLMS) estimator is better than

other estimators (OLS, RR and RLAD) for many combinations of non-normal error distribution (which reflected the presence of influential data points) and when the degree of multicollinearity is high (Tables (2), (3) and (4)). Therefore, the RLMS estimator appears to be a suitable alternative to other estimators for the different combinations of multicollinearity and influential data points.

There are limitations to the present study, however. First, since this is a simulation study, its limitations must be recognized. Data have been generated to try and allow generalization to practical situations, however. Second, other possible members of the robust regression approach may be used to construct the title of combined biased robust estimators.

References

[1] Askin, R. G., and D. C. Montgomery (1980) "Augmented robust estimators". *Technometrics,* 22, 333-341. http://dx.doi.org/10.2307/1268317

[2] Belsley, D., Kuh, E., and R.E. Welsh. (1980) "Regression diagnostics" *Wiley, New York.* http://dx.doi.org/10.1002/0471725153

[3] Chen, C. (2002) "Robust regression and outlier detection with the ROBUSTRER procedure" SUGI paper, *SAS Institute,* 265-277.

[4] Dempster, A.P., Schatzoff, M., and N. Wermuth (1977) "A simulation study of alternatives to Ordinary Least Squares" *J.S.A.*, 72, 77-91. http://dx.doi.org/10.1080/01621459.1977.10479910

[5] Gibbons, D. (1981) "A simulation study of some ridge estimators" *J.S.A.*, 76, 131-139. http://dx.doi.org/10.1080/01621459.1981.10477619

[6] Hoerl, A.E. and R. W.Kennard (1976) "Ridge Regression: Iterative Estimation of the Biasing parameter" *Communications in statistics: A Theory Methods*, 5, 77-88. http://dx.doi.org/10.1080/03610927608827333

[7] Koenker, R.W. (1982) "Robust methods in econometrics" *Econometric Rev.*, 1, 213-255. http://dx.doi.org/10.1080/07311768208800017

[8] Krasker, W.S. and R.E. Welsch (1982) "Efficient bounded influence regression estimation" *J.S.A.*, 77, 595-604. http://dx.doi.org/10.1080/01621459.1982.10477855

[9] Mal, C.Z., and Y.L. Dul (2014) "Generalized shrunken type-GM estimator and its application "International Conference on Applied Sciences (ICAS2013). http://dx.doi.org/10.1088/1757-899x/57/1/011001

[10] Martin, R.D. (2002) "Robust Statistics with the S-Plus "Robust Library and Financial Applications.

[11] Moawad, Al-Falah, A. (2013) "The Efficiency of Some Robust Ridge Regression for Handling Multicollinearity and Nonnormal errors problems" Applied Mathematical Science, 7, 77, pp. 3831-3846.
http://dx.doi.org/10.12988/ams.2013.36297

[12] Montgomery, D.C., and E.A. Peck (1982) "Introduction to Linear Regression Analysis "*Wiley, New York.*

[13] Pariente, S., and R.E. Welsch (1977) "Ridge and Robust Regression Using Parametric Linear Programming "Working Paper, MIT Alfred P Sloan School of Management.

[14] Pfaffenberger, R.C., and T.E. Dielman (1984) "A Modified Ridge Regression Estimator Using the Least Absolute Value Criterion in the Multiple Linear Regression Model" (pp.791-793) Proceedings of the American Institute for Decision Sciences, Toronto.

[15] Rousseeuw, P.J. and A.M. Leroy (1987) "Robust Regression and Outliers Detection "Wiley, New York. http://dx.doi.org/10.1002/0471725382.

[16] Siti, M. Z., Mohammad, S.Z., and Al. bin I. Muhammad (2012) "Weighted Ridge MM- Estimator in Robust Ridge Regression with Multicollinearity" Mathematical Models and Methods in Modern Science. Symp. Computational Statistics, 1, 471-476.

[17] Venables, W. N. and B.D. Ripley (1999) "Modern Applied Statistics with S-Plus" (4^{th} ed.) Springer: New York. 167-176.
http://dx.doi.org/10.1007/978-1-4757-3121-7

Received: February 17, 2015; Published: March 20, 2015