A NEW ROBUST LIU REGRESSION ESTIMATOR FOR HIGH-DIMENSIONAL DATA

Muthukrishnan. R, Karthika Ramakrishnan

•

Department of Statistics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India
muthukrishnan70@buc.edu.in, karthikaramakrishnan45@gmail.com

Abstract

Aim: *To provide a new Liu regression procedure for predictive modeling in cases of multicollinearity and with/without outliers.* ***Methods:*** *Regression analysis is employed in many statistical research domains for both estimation and prediction. Liu and Robust Estimators were developed in a classical linear regression model to address the issues of multicollinearity and outliers, respectively. In order to jointly handle the issues of multicollinearity and outliers, this research paper explores a new Robust Liu regression estimator based on the MM estimator, which is then demonstrated using real and simulated data sets. The performances of various regression estimators such as Least Square, Ridge, Liu and the Robust Liu are compared based on the Mean Square Error criterion.* ***Findings:*** *According to the computed error measure, the study concludes that the Robust Liu regression estimator provides more reliable results than the other mentioned regression procedures in situations where datasets have both multicollinearity and outliers.*

Keywords: Regression, Multicollinearity, Outliers, Liu, Robust Liu

I. Introduction

The Least Squares estimator is frequently utilized to predict the parameters of a regression model, provided that all the assumptions of the model are fully satisfied. Multicollinearity and outliers are the issues that could skew the outcomes of this approach. When there is a significant correlation between the independent variables, the situation is called multicollinearity, defined by Farrar and Glauber [7]. It will increase the error values and thus making the estimator not good. An outlier is a unique observation in the data. It causes the estimator to become inefficient and modifies the regression coefficients' sign. The existence of outliers, according to Chatterjee and Hadi [4], may leads to influence the parameter estimation and inaccurate predictions for traditional approaches. Ridge and Liu regression procedures were developed to overcome multicollinearity. When the data deviates from key assumptions, robust regression offers an alternative to the classical regression model. This research described a regression technique with a better estimate when multicollinearity and outliers are present in the dataset.

The rest of the paper is organized as follows. Various regression estimators like Least Squares, Ridge, Liu and Robust Liu are explained briefly in section 2. A numerical study is carried out based on real and simulated datasets to compare the Mean Square Error of different regression estimators in section 3 and section 4 will give the conclusion.

II. Regression Procedures

Regression analysis serves as a means to glean insights from data by identifying relationships between the response and predictor variables, as outlined by Draper and Smith [6]. These methodologies within machine learning manifest in various forms, selected based on the nature of the dataset. It stands as the primary approach for addressing machine learning challenges through data modeling. This study encompasses Least Squares, Ridge, Liu and Robust Liu methods, comparing mean square error measures of different real and simulated datasets having the presence of both outliers and multicollinearity. Outliers in the actual data are detected and removed using Cook's distance, with the analysis conducted utilizing R software.

Least Squares Estimator

The Least Squares (LS) is a standard approach in regression analysis for estimating the parameters of a linear model. This method is employed to forecast the dependent variable (y) using several predictor variables (X). It stands as the widely utilized and optimal linear unbiased estimator, when all the suppositions of the classical regression model are satisfied. The standard model of LS with k independent variables is represented as follows.

$$y = X\beta + \varepsilon \tag{1}$$

here, y is an $(m\times1)$ vector of response variables, X is an $(m\times k)$ matrix of predictors, β is a $(k\times1)$ vector of unknown regression parameters, and ϵ is an $(m\times1)$ vector of residuals assumed to be independently and identically distributed as normal with a mean of zero and a fixed variance σ^2. The LS estimator for the unknown parameter is given by

$$\widehat{\beta_{LS}} = (X'X)^{-1}(X'y) \tag{2}$$

The performance of the LS estimator $\widehat{\beta_{LS}}$ becomes statistically insignificant when multicollinearity exists among the explanatory variables.

Ridge Estimator

Ridge regression estimator was provided by Hoerl and Kennard [9] to deal with the problem of multicollinearity. It gives a biased estimator and will depend on the ridge constant k which is used for minimizing the bias. The complexity parameter k needs to be selected appropriately in order to optimize the prediction accuracy. Hoerl et al. [10] find out a formula for the calculation of an optimal ridge constant k such that

$$k = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \widehat{\beta_\iota}^2} \tag{3}$$

where p is number of independent variables, $\hat{\sigma}^2$ is the estimated variance and $\widehat{\beta_\iota}$ is an LS regression parameter of canonical form. Ridge regression is depend on this constant k and will give a biased estimator as given below.

$$\widehat{\beta_{Ridge}} = (X'X + kI)^{-1} (X'y) \tag{4}$$

Liu Estimator

Liu Estimator is a class of biased estimators used to deal with datasets having multicollinearity. It was introduced by Liu [12]. These estimators are depending upon a biasing parameter d called the Liu parameter which lies between 0 and 1. The estimator of Liu regression is given by

$$\widehat{\beta_{Liu}} = (X'X \ + I_p\ )^{-1}(X'y + d\ \widehat{\beta_{LS}}) \tag{5}$$

where $0 \leq d \leq 1$, I_p is the identity matrix of order $p \times p$ and $\widehat{\beta_{LS}}$ is the LS estimator. The biasing parameter d of Liu is computed by the formula,

$$\hat{d} = 1 - \hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i\,(\lambda_i+1)}}{\sum_{i=1}^{p} \frac{\widehat{\beta_i}^2}{(\lambda_i+1)^2}} \right] \tag{6}$$

where, $\hat{\sigma}^2$ and $\widehat{\beta_i}^2$are the mean square error and the regression estimates computed via LS respectively. λ_1, λ_2, …, λ_p are the eigen values of the matrix $X'X$. $\widehat{\beta_{Liu}}$ is named as the Liu estimator by Akdeniz and Kaciranlar [2]. The d value with minimum mean square error gives an efficient estimator as compared to other values . The R package liureg was developed by Muhammad Imdadullah et al. [15] provides the tools for the computation of the Liu estimator and the biasing parameter.

MM Estimator

The MM estimator is a robust regression technique introduced by Yohai [23], used to estimate parameters in the presence of outliers. It is a modification of the M-estimator, designed to provide robustness and high efficiency. The construction of this estimator starts with an initial robust estimator as S estimator obtained using a robust method such as minimizing the scale of residuals. The MM estimator is defined as follows.

$$\widehat{\beta_{MM}} = \ arg\ \min_{\beta} \sum_{i=1}^{n} \rho(\frac{r_i(\beta)}{s}) \tag{7}$$

where $r_i(\beta)$ are the residuals, s is a scale estimate based on the initial robust estimator, and ϱ is a loss function like Tukey's biweight.

Robust Liu Estimator

The objective of robust regression is to overcome some of the limitations of traditional regression analysis. The estimation and reference methods in robust regression should be straight forward to implement. Under a normal distribution without outliers, this robust method should yield results similar to those of LS. In this section, a new Robust Liu (RLiu) regression estimator was described to deal with the datasets having both multicollinearity and outliers by incorporating the properties of both Liu and MM regression procedures. The estimator of RLiu regression is given by

$$\widehat{\beta_{RLiu}} = \left(X'X \ + I_p\right)^{-1}(X'y + d_{MM}\widehat{\beta_{MM}}) \tag{8}$$

where $0 \leq d_{MM} \leq 1$, I_p is the identity matrix of order $p \times p$, $\widehat{\beta_{MM}}$ is the MM estimator. The biasing parameter $d_{MM}$of RLiu is computed by the formula,

$$\widehat{d_{MM}} = 1 - \widehat{\sigma_{MM}}^2 \left[\frac{\sum_{i=1}^{p}\frac{1}{\lambda_i\,(\lambda_i+1)}}{\sum_{i=1}^{p}\frac{\widehat{\beta_i}^2}{(\lambda_i+1)^2}}\right] \tag{9}$$

where, $\widehat{\sigma_{MM}}^2$ and $\widehat{\beta_i}^2$ are the mean square error and the regression estimates computed via MM respectively. $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigen values of the matrix $X'X$.

III. Experimental Results

Table 1: *Computed MSE under various regression methods (Real Data)*

	Methods			
Datasets	LS	Ridge	Liu	RLiu
Case 1:Prostate Cancer	0.46	0.46	0.16	0.14
	(1.5)	(0.33)	(0.13)	0.12
Case 2:Hald	3.68	3.32	1.19	0.14
	(2.57)	(2.46)	(1.02)	(0.89)

(.)Without outlier

Table 2: *Computed MSE under various regression methods (Simulation Data)*

		Methods			
n	Contamination	LS	Ridge	Liu	RLiu
50	0%	11.15	10.59	9.24	8.44
	5%	15.20	15.08	10.46	4.88
	10%	17.40	17.17	14.31	13.42
	15%	19.17	19.07	12.19	11.99
100	0%	12.31	12.13	4.53	4.41
	5%	12.78	12.68	5.96	5.42
	10%	12.62	12.44	4.99	4.82
	15%	18.88	18.42	7.06	5.73
200	0%	15.43	15.41	3.17	3.04
	5%	17.01	16.67	3.14	2.99
	10%	14.76	14.29	3.17	2.87
	15%	15.68	15.40	3.02	2.98

This section presents numerical analyses conducted on both real and simulated datasets. The first real dataset explained in case 1 exhibits moderate multicollinearity with outliers. The second dataset presented in case 2 displays high multicollinearity along with the presence of outliers. Outliers in the actual datasets were detected and eliminated using Cook's distance method introduced by Cook [5], and the analyses were performed using R software. A statistical technique called the Variance Inflation Factor (VIF) by Frisch [8] can detect and measure the amount of multicollinearity in a multiple regression model. The VIF assesses how much the regressors collectively impact the variance of each term within the model. The computed MSE measures of different regression estimators are summarized and presented in tables.

Case1. Prostate Cancer Dataset: The data come from a study that looked at how males undergoing radial prostatectomy correlated their level of prostate-specific antigen with several clinical measures. This data set has 97 observations. There are seven independent variables namely lweight (log of prostate weight), age, lbph (log of benign prostatic hyperplasia amount), svi (seminal vesicle invasion), lcp (log of capsular penetration), gleason (Gleason score), lpsa (log of prostate specific antigen) and one dependent variable lcavol (log of cancer volume). Seven outliers are found in this dataset. Since the VIFs of the independent variables are in between 1 and 5, there is an indication of moderate multicollinearity.

Case 2. Hald Data: Woods et al [22] was introduced the Hald or Portland Cement Data. This data frame contains 13 observations with four independent variables. They are tricalcium aluminate (X1), tricalcium silicate (X2), tetracalcium aluminoferrite (X3) and β-dicalcium silicate (X4). The response variable Y is the evolved heat after 180 days in a cement mix. Since the VIFs of this Hald data set was greater than 10, the explanatory variables are highly correlated. As a result, the dataset has high multicollinearity. Also this data set has one outlier. The computed Mean Square Error (MSE) based on with and without outliers of different estimators in Cases 1 and 2 is given in Table 1.

Simulation studies were carried out to examine the efficiency of different regression estimators. In the study, the data was generated from a multivariate normal distribution with mean $\mu = [0]_{px1}$and the variance $\sum = [\sigma_{ij}]$ for the level of correlation, $\rho = 0.90$ and number of variables $p = 5$. Different levels of contamination (0%, 5%, 10%, and 15%) were studied for sample size $n = 50, 100,200$. The performance of various regression procedures were compared using the MSE criterion and the results obtained for different number of observations with various levels of contamination are shown in Table 2.

The results obtained from Table 1 and Table 2 show that the error values for different estimators are slightly different from each other. Also the RLiu estimator has the smallest MSE of all others. Hence Robust Liu (RLiu) estimator is more efficient than the other estimators in the case of datasets having indication of multicollinearity and has outliers.

IV. Conclusion

Statistical learning techniques are crucial in various research fields, with regression analysis being a prominent method. Traditional linear regression often falls short when data deviates from its assumptions, necessitating alternative approaches. This paper explores several regression methods, including Least Squares, Ridge, Liu, and Robust Liu, and assesses their performance across different real and simulated datasets. The study addresses issues of multicollinearity and outliers by calculating Mean Square Error (MSE). The findings indicate that the Robust Liu regression method provides better estimates for datasets having both multicollinearity and/or outliers. This approach can be particularly advantageous for researchers employing machine learning techniques that need to account for these factors.

References

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

[2]Akdeniz, F. and Kacıranlar, S. (1995). on the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE. *Communications in Statistics, Theory and Methods*, 24: 1789–1797.

[3]Arslan, O. and Billor, N. (2000). Robust Liu estimator for regression based on an M-estimator. *Journal of applied statistics*, 27: 39-47.

[4] Chatterjee, S. and Hadi, A. S. Sensitivity Analysis in Linear regression. John Wiley & Sons, New York, 2009.

[5]Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics*, 42: 65-68.

[6] Draper, N. R. and Smith, H. Applied Regression Analysis. John Wiley & Sons, New York, 1998.

[7]Farrar, D. E. and Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, 49: 92-107.

[8]Frisch, R. Statistical confluence analysis by means of complete regression systems. University of Oslo, 1934.

[9]Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12: 55–67.

[10]Hoerl, A. E., Kennard, R. W. and Baldwin K. F. (1975). Ridge regression: Some simulation. *Communications in Statistics*, 4.

[11]Huber, P. H. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35: 7-101.

[12]Liu, K. J. (1993). A new class of biased estimate in linear regression. *Communications in Statistics*, 22: 393-402.

[13]Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, 53: 44–53.

[14]Mendenhall, W. and Sincich, A. (2014). Second Course in Statistics: Regression Analysis, 7th ed (Harlow: Pearson), 105–123.

[15]Muhammad Imdadullah, Muhammad Aslam and Saima Altaf. (2017). liureg: A Comprehensive R Package for the Liu Estimation of Linear Regression Model with Collinear Regressors. *The R Journal 9*.

[16]Muthukrishnan, R. and Kalaivani, S. (2022). Data depth approach in fitting linear regression models, *Materials Today: Proceedings*, 57: 2212-2215.

[17]Muthukrishnan, R. and Karthika Ramakrishnan (2024). Effect of Classical and Robust Regression Estimators in the context of High dimensional data with Multicollinearity and Outliers. *Reliability: Theory & Applications*, 19: 335-341.

[18]Muthukrishnan, R. & Maryam Jamila, S. (2020). Predictive Modeling Using Support Vector Regression, *International Journal of Scientific and Technology Research*, 9: 4863-4865.

[19]Rousseeuw, P. J.(1983). Multivariate estimation with high breakdown point, The Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdorf, Austria.

[20]Rousseeuw, P. J and Leroy, A. M. Robust Regression and Outlier Detection , John Wiley & Sons, 1987.

[21]Susanti, Y., Pratiwi, H., Sri Sulistijowati, H. and Liana, T. (2014). M estimation, S estimation and MM estimation in Robust Regression. *International Journal of Pure and Applied Mathematics*, 91: 349-360.

[22]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.

[23]Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, 642-656.