EFFECT OF CLASSICAL AND ROBUST REGRESSION ESTIMATORS IN THE CONTEXT OF HIGH-DIMENSIONAL DATA WITH MULTICOLLINEARITY AND OUTLIERS

Muthukrishnan R
•
Department of Statistics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India
muthukrishnan1970@gmail.com

Karthika Ramakrishnan
•
Department of Statistics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India
karthikaramakrishnan45@gmail.com

Abstract

Regression methods are used for the estimation and prediction in various fields of statistical study. It is a statistical method commonly used for determining the degree of relationship between a response and a number of explanatory variables. These explanatory variables may correlate each other and lead to multicollinearity. More than two predictor variables with high correlation show the existence of multicollinearity which results in the estimator having a high variance. Ordinary Least Square estimation fails to give a better regression estimator, when the model's presumptions are not met. This paper explores the various methods which can tolerate the problems of multicollinearity and outliers. This study compares different types of regression estimators such as Ordinary Least Square, Robust, Ridge, and Liu by computing various error values such as Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error and R2 under real environment that has both multicollinearity and outliers. To compare the fit of the aforementioned regression models, the Akaike Information Criterion was also calculated. According to the error measures and AIC this study concludes that the Liu regression estimator performs well when compared with the other estimation methods.

Keywords: Regression, Multicollinearity, Outliers, Ridge, Liu

I. Introduction

OLS estimator is the commonly used method to predict the parameters of a regression model when all the assumptions of the model are satisfied. The problems that would be affected the results of this method are multicollinearity and outliers. Multicollinearity is the situation where the explanatory variables have highly interdependent. It will increase the error values and thus the estimator may unreliable. Hoerl and Kennard [1] develop a regression procedure to control multicollinearity.

An outlier is a data observation that is unusual. It results the estimator to be not efficient

and changes the sign of the regression coefficients. Mendenhall and Sincich [7] give the definition of outlier as value with absolute standardized error greater than 3. Robust regression methods are usually used to obtain a better result when there are outliers. The main purpose of this paper is to compare different regression methods and identify the good one with better estimator in the presence of both multicollinearity and outlier.

The rest of the paper is structured as follows. In section 2, various regression estimators like OLS, Robust regression, Ridge regression and Liu regression are explained briefly. The performance of these regression procedures is studied under real environments and the results are summarized in section 3 and conclusion of the study is presented in the last section.

II. **Regression Methods**

Regression analysis is used to draw inferences from data when there is a connection between the response and the predictor variables, according to Draper and Smith [9]. These approaches in machine learning come in a variety of forms, and their use depends on the type of data being used. It is the primary method to solve the problems in machine learning using data modeling. This paper includes the methods OLS, Robust, Ridge and Liu with the comparison of error measures under different real datasets having the presence of both outliers and multicollinearity. Outliers are identified by the Cook's distance procedure and the analysis has been carried out using R software.

Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is a technique used to predict the dependent variable (y) with the help of a number of predictor variables (X). It is the popularly used and Best Linear Unbiased Estimator (BLUE) when all the suppositions of the classical regression model are satisfied [Aitken [3]]. The general model of an OLS method with k independent variables is given by

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

where $\boldsymbol{y}$ is the $(m \times 1)$ vector of response variable, $\boldsymbol{X}$ is a $(m \times k)$ matrix, $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of an unknown regression parameters and $\boldsymbol{\varepsilon}$ is a $(m \times 1)$ vector of residual term that is considered to be independently and identically distributed as normal with mean zero and fixed variance σ^2. The OLS estimator for the unknown parameter is

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

The performance of $\widehat{\beta_{OLS}}$ will be statistically insignificant when multicollinearity exists between the explanatory variables.

Robust Regression

Robust regression is an alternative approach to the classical regression model, when the nature of the data deviates from the key assumptions. The goal of robust regression is to get beyond some of the drawbacks of conventional regression analysis. Under normal distribution with no outliers, this robust method should produce approximately the similar results as OLS. In this section robust regression method like Least Trimmed Square (LTS), Least Median Square (LMS) and M were described.

Least Trimmed Square (LTS)

Least Trimmed Square (LTS) is a robust regression method developed by Rousseeuw [11].

This method has an objective function of the lowest trimmed of squared residuals as

$$\sum_{i=1}^{h} r_i^2(\beta) \tag{3}$$

where $r_1^{\,2}(\beta) \leq \cdots \leq r_n^{\,2}(\beta)$ are the ordered residuals for $i = 1,2,\ldots,n$. The number of observations that are not trimmed from the dataset is denoted by h. LTS minimizes the trimmed sum of these squared residuals. Its estimator is equal to the OLS estimator only when $h = n$. The estimator $\widehat{\beta_{LTS}}$ is estimated by minimizing the sum of residuals over β is given by

$$\widehat{\beta_{LTS}} = arg\min_{\beta} \sum_{i=1}^{h} r_i^{\,2}(\beta) \tag{4}$$

LTS can be calculated for α, the trimming proportion tending to 50%. It attains the maximum possible breakdown point at $h = \left({}^{n}/_{2}\right) + [(p+1)/2]$. The computation of LTS estimator uses an algorithm called FAST-LTS of Rousseeuw and Van Driessen [10].

Least Median Square (LMS)

The Least Median Square (LMS) estimator was suggested by Rousseeuw [15]. Being a robust regression technique, the least median of squares method is not sensitive to outliers or other breaches of the normal model's assumption. In this method the sum is replaced by median in the method of least squares. Here the parameters are estimated by reducing the median of the squared residuals. The least median square estimator can be given by

$$\widehat{\beta_{LMS}} = arg\ \min_{\beta}\ r_h^{\,2}(\beta) \tag{5}$$

where $r_h^{\,2}(\beta)$ is a median. LMS is robust due to its breakdown value of 50%.

M Estimator (M)

M-estimators and their asymptotic properties were introduced by Huber [5]. Here the M stands for *"maximum likelihood type"*. This method laid the foundation for the growth of the other robust methods in the context of regression estimators. M-estimation attempts to reduce the squared residuals r_h^2 in OLS by another function of these r_h^2

$$\min_i \sum_{h=1}^{n} \rho(r_h) \tag{6}$$

$\rho(r_h)$ is introduced for reducing the effect of outliers, where ρ is a definite positive, symmetric function with zero as its unique minimum. An algorithm was developed by Susanti et al. [16] for computing the M estimator.

Ridge Regression (RR)

Ridge Regression (RR) developed by Hoerl and Kennard [1] to give a reliable regression estimates even in the presence of multicollinearity. It produces an estimator that is biased and will be associated with the constant k that is used to reduce the bias. Hoerl et al. [2] find out a formula for the calculation of an optimal ridge constant k such that

$$k = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \widehat{\alpha_i}^{\,2}} \tag{7}$$

where p denotes the number of predictor variables, $\hat{\sigma}^2$ is the estimated variance and $\widehat{\alpha_i}$ is a conventional OLS regression parameter. Ridge regression depends on this constant k and will give a biased estimator as given below.

$$\widehat{\beta_{RR}} = (X'X + kI)^{-1}(X'y) \tag{8}$$

Liu Regression (LR)

Liu Regression (LR) is used to deal with datasets having multicollinearity. It was proposed by Liu [6]. It forms a new class of biased estimators called Liu estimators. 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

$$\widehat{\beta_{LR}} = (X'X + I_p)^{-1} (X'y + d\,\widehat{\beta_{OLS}}) \quad (9)$$

where $0 \leq d \leq 1$, I_p is the identity matrix of order $p \times p$ and $\widehat{\beta_{OLS}}$ is the OLS estimator. $\widehat{\beta_{LR}}$ is the Liu estimator named by Akdeniz and Kaciranlar [4]. The d value with the minimum Mean Square Error (MSE) gives an efficient estimator. The R package *liureg* developed by Muhammad Imdadullah et al. [8] provides the tools for the computation of the estimator and the biasing parameter.

III. Experimental Results

Table 1: *Computed error measures and AIC under various regression methods (Acetylene Data)*

	Regression Methods					
Errors	OLS	LTS	LMS	M	RR	LR
MAPE	0.284	0.195	0.194	0.262	0.210	0.195
MAE	0.008	0.009	0.009	0.007	0.034	0.007
RMSE	0.009	0.016	0.016	0.008	0.044	0.009
R^2	0.900	0.711	0.728	0.918	0.990	0.992
AIC	93.24	383.09	331.99	183.86	42.10	42.09

Table 2: *Computed error measures and AIC under various regression methods (Prostate Cancer Data)*

	Regression Methods					
Errors	OLS	LTS	LMS	M	RR	LR
MAPE	2.33 (2.17)	2.51 (2.27)	2.59 (2.62)	2.34 (2.32)	2.23 (2.11)	2.21 (2.10)
MAE	0.56 (0.48)	0.64 (0.97)	0.69 (0.93)	0.56 (0.90)	0.56 (0.47)	0.55 (0.47)
RMSE	0.68 (0.58)	0.86 (1.32)	0.97 (1.30)	0.68 (1.25)	0.68 (0.58)	0.67 (0.57)
R^2	0.66 (0.73)	0.46 (0.70)	0.31 (0.67)	0.66 (0.70)	0.67 (0.73)	0.67 (0.74)
AIC	219.54	91.35	117.08	62.51	62.99	60.29

(.)Without outlier

Table 3: *Computed error measures and AIC under various regression methods (Hald Data)*

	Regression Methods					
Errors	OLS	LTS	LMS	M	RR	LR
MAPE	0.04 (0.01)	0.04 (0.01)	0.12 (0.02)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
MAE	0.78 (0.96)	0.91 (0.89)	3.89 (2.53)	0.75 (0.85)	0.81 (0.95)	0.75 (0.90)
RMSE	0.96 (1.29)	1.51 (1.64)	6.43 (3.70)	0.97 (1.28)	0.99 (1.32)	0.95 (1.25)
R^2	0.996 (0.99)	0.991 (0.99)	0.841 (0.93)	0.996 (0.99)	0.996 (0.99)	0.997 (0.99)
AIC	47.67	39.69	46.95	22.19	10.13	6.72

(.)Without outlier

The experimental studies were carried out under real environments to study and compare the performance of the various regression procedures and thus obtained results were discussed in this section. The numerical studies have been conducted by considering three different case studies under real datasets in which the first data has the presence of multicollinearity and no outliers. The second one has a presence of moderate multicollinearity and outliers. And the third dataset has the presence of high multicollinearity and outliers. The outliers in the real data sets were detected and removed by using cook's distance, Cook [12] and the analysis has been carried out using R software. The presence and absence of multicollinearity has been identified by computing Variance Inflation Factors (VIF). The overall impact of the regressors' dependencies on each term's variance is measured by the VIF for each term in the model. VIF is 1 indicates that there are no correlation between the variables. Moderate correlation is indicated by a VIF between 1 and 5. VIF more than 5 is an indication of high multicollinearity between the variables. The error measures under OLS, LTS, LMS, M, RR and LR estimators were calculated by considering with and without outliers and are summarized in tables.

The Acetylene data set contains the percentage of n-heptane that is converted to acetylene, together with three independent variables. These are typical data from a chemical process, and a full quadratic response surface in each of the three regressors is sometimes regarded as a suitable preliminary model. It has 16 observations and 4 variables in which conversion of n-Heptane to Acetylene (y) is considered as the dependent variable and Reactor Temperature (X1), Ratio of H2 to n- Heptane (X2), Contact Time (X3) are taken as the independent variables. Cook's distance is used to check the presence of outliers in the dataset and there are none to be found. The VIF measures are higher than 10 and hence the dataset has high multicollinearity. The computed error measures and AIC value under various estimators of the dataset are given in Table 1.

The second data come from a study that looked at how males undergoing radial prostatectomy correlated their level of prostate-specific antigen with several clinical measures. The data is available in the R-package "lasso2". It has 97 observations, and 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 and eliminated using Cook's distance. Since the VIFs of the independent variables are in between 1 and 5, there is an indication of moderate multicollinearity. The error measures and AIC calculated under different estimators are shown in Table 2.

Woods et al [17] was introduced the Hald or Portland Cement Data. It has been extensively analysed by Hald (1952), Hamaker (1962) and Kaciranlar et al (1999). 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 Variance Inflation Factors (VIF) 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 three outliers, which are found and eliminated by using Cook's distance. The computed error measures and AIC under various estimators are given in Table 3.

The results from Table 1, Table 2 and Table 3 demonstrate that the error levels for various estimators differ from one another, with LR having the lowest of all these. Also the AIC value of LR is minimum compared to the other estimators. Thus, for a dataset having high multicollinearity and outliers, the Liu (LR) regression estimator is more effective than the other estimators.

IV. Conclusion

Statistical learning techniques play a vital role in almost all the field of research study. Regression analysis is one of the statistical learning techniques. In general, the commonly used linear regression procedure will not be sufficient to build a regression model when data deviates from the modelling assumptions. Hence, there is a need of alternatives to build a good model for the given dataset. This paper explores various regression procedures such as OLS, Robust, Ridge and Liu. Further, evaluates their performance on different real datasets by considering the problems of multicollinearity and outliers by computing various error measures along with AIC value. On the basis of error and AIC values, the study concluded that the Liu regression procedure gives a better estimator for modelling the data when the dataset having multicollinearity and/or outliers. Further, this regression procedure can be beneficial to researchers, who work on machine learning techniques by considering the factors such as multicollinearity, outliers and high dimensionality.

References

[1] Aitken, A.C. (1936). On least squares and linear combination 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] Cook, R.D. (2000). Detection of influential observation in linear regression, *Technometrics*. 42: 65-68.

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

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

[6] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: Some simulations, *Communications in Statistics- Theory and Methods*. 4:105-123.

[7] Huber, P.J. (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics*. 35:492-518.

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

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

[10] Mendenhall, W. and Sincich, A. Second Course in Statistics: Regression Analysis, 2014.

[11]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*.9:232.

[12]Rousseeuw, P. J. (1984). Least Median of Squares Regression, *Journal of the American Statistical Association*.79:871–880.

[13]Rousseeuw, P.J. (1985), Multivariate estimation with high breakdown point, *Mathematical statistics and applications*. 37:283-297.

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

[15] Rousseeuw, P.J. and Van Driessen, K. (2006). Computing LTS regression for large data sets, *Data mining and knowledge discovery*. 12:29-45.

[16] 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.

[17] 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.