ENHANCING PREDICTIVE ACCURACY WITH ROBUST LIU AND RANDOM FOREST REGRESSION

Muthukrishnan. R and Karthika Ramakrishnan
•
Department of Statistics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India
muthukrishnan70@buc.edu.in, karthikaramakrishnan45@gmail.com

Abstract

Robust and Random Forest regression procedures are powerful machine learning algorithms that enhance predictive accuracy and handle complex datasets with diverse characteristics. The Robust Liu regression algorithm is designed to address the limitations of conventional regression methods. The robust Liu regression method is considered more reliable in the presence of outliers and multicollinearity. On the other hand, Random Forest is an ensemble learning method that constructs multiple decision trees to improve classification and regression tasks by aggregating predictions from individual trees, thereby mitigating overfitting and increasing robustness to noise. Its ability to capture nonlinear relationships and handle high-dimensional data makes it suitable for many real-world applications. This study compares the performance of Least Squares, Liu, Robust Liu, and Random Forest regression methods using prediction error measures in real and simulated environments. This synergy offers new possibilities for researchers to improve prediction accuracy in the presence of heterogeneous non-normal, multicollinear data and outliers.

Keywords: Liu, Multicollinearity, Outliers, Random Forest, Robust

I. Introduction

The Least Squares regression method is commonly employed to estimate the parameters of a regression model, assuming all model assumptions are met. However, issues such as multicollinearity and outliers can distort the results of this method. Multicollinearity, as described by Farrar and Glauber [8], occurs when there is a high correlation among the independent variables. This increases error values and reduces the effectiveness of the estimator. On the other hand, an outlier is an unusual data point that can reduce the estimator's efficiency and alter the direction of regression coefficients. Chatterjee and Hadi [5] noted that outliers could impact parameter estimation and lead to inaccurate predictions in traditional methods. To address multicollinearity, the Liu regression procedure was introduced. For data that deviates from fundamental assumptions, robust regression provides an alternative to conventional models. Additionally, Random Forest, a supervised learning algorithm, can be applied to both classification and regression problems. This study compares different regression methods in the presence of multicollinearity and outliers in datasets.

The rest of the paper is structured as follows. Various regression procedures like Least Squares, Liu, Random Forest and Robust Liu are explained briefly in section 2. An experimental study is conducted under real and simulated datasets to compare different error measures such as Mean Square Error, Mean Absolute Error, Median Absolute Error and Mean Absolute Percentage Error for various regression methods in section 3 and section 4 will give the conclusion.

II. Regression Methods

Regression analysis, as stated by Draper and Smith [7], is a method used to derive insights from data by examining the relationship between response and predictor variables. Various forms of regression techniques exist in machine learning, and their application depends on the nature of the data involved. It serves as a key approach for addressing problems in machine learning through data modelling. This study focuses on methods such as Least Squares, Liu, Random Forest, Ridge, and Liu regression, comparing their error measures across different real-world datasets that include both outliers and multicollinearity. Outliers are identified using Cook's distance procedure, and the analysis is conducted with the help of R software.

Least Squares Regression

The Least Squares (LS) is a standard procedure in regression modelling for estimating the parameters of a linear model. This method is used to estimate the dependent variable (y) using a number of predictor variables (X). It is a widely utilized and optimal linear unbiased estimator when all the assumptions of the classical regression model are satisfied. The standard model of LS with r independent variables is represented as follows.

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

where y is a $(qx1)$ vector of response variables, X is an (qxr) matrix of predictors, β is a $(rx1)$ vector of unknown regression parameters, and ϵ is an $(rx1)$ 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.

Liu Regression

Liu Estimator is a class of biased estimators used to deal with datasets having multicollinearity. It was described by Liu [11]. These estimators are depending upon a biasing parameter d called the Liu parameter which lies between 0 and 1. The regression estimator of this procedure is given by

$$\widehat{\beta_{Liu}} = (X'X + I_q)^{-1}(X'y + d\,\widehat{\beta_{LS}}) \tag{3}$$

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

$$\hat{d} = 1 - \hat{\sigma}^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)}}{\sum_{i=1}^{p} \frac{\widehat{\beta_i}^2}{(\lambda_i + 1)^2}} \right] \quad (4)$$

where, $\hat{\sigma}^2$ and $\widehat{\beta_i}^2$ are the mean square error and the regression estimates computed via LS respectively. $\lambda_1, \lambda_2, \ldots, \lambda_q$ are the eigenvalues of the matrix $X'X$. $\widehat{\beta_{Liu}}$ is called the Liu estimator by Akdeniz and Kaciranlar [2]. The estimator having a d value with minimum mean square error is considered an efficient estimator when compared with the other values.

Random Forest Regression

The Random Forest (RF) regression technique introduced by Breiman [4] is a versatile machine learning technique for predicting numerical values. It combines the predictions of multiple decision trees to reduce overfitting and improve accuracy. The development of Random Forests was a response to the limitations of traditional decision trees, which tend to overfit the data. It provides better predictive accuracy by averaging the results of multiple trees. The Random Forest model for regression is given in equation (5), where t is the total number of trees in the forest and $\hat{y}$ is the final predicted value.

$$\widehat{y_t} = f_t(x) \quad (5)$$

Robust Liu Regression

The robust regression methods were developed to overcome the limitations of traditional regression procedures. Under a normal distribution without outliers, this robust method should yield results similar to LS. A Robust Liu (RLiu) regression method described by Muthukrishnan and Karthika [14] was developed to deal with the datasets having both multicollinearity and outliers by incorporating the properties of both Liu and MM regression procedures. The MM estimator is a robust regression technique introduced by Yohai [19], 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 estimator of the RLiu regression is given by

$$\widehat{\beta_{RLiu}} = \left(X'X + I_q\right)^{-1} (X'y + d_{MM}\widehat{\beta_{MM}}) \quad (6)$$

where $0 \leq d_{MM} \leq 1$, I_p is the identity matrix of order $r \times r$, $\widehat{\beta_{MM}}$ is the MM estimator. The biasing parameter $d_{MM}$of RLiu is computed using the equation (4) by replacing $\hat{\sigma}^2$with $\widehat{\sigma_{MM}}^2$, the mean square error and $\widehat{\beta_i}^2$are the regression estimates computed via MM respectively.

III. Experimental Results

This section presents numerical analyses conducted on both real and simulated datasets. The first real dataset exhibits moderate multicollinearity with outliers. The second dataset displays high multicollinearity alongside outliers. Outliers in the real datasets were detected and eliminated using Cook's distance method introduced by Cook [6], and the analyses were performed using R software. A statistical technique called the Variance Inflation Factor (VIF) by Frisch [9] 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 error measures under real datasets, discussed in Cases 1 and 2, based on with and without outliers of various regression procedures are given in Table 1 and the corresponding results under simulation is presented in Table 2.

Table 1: *Computed error measures under various regression methods (Real Datasets)*

Methods	Prostate Cancer Dataset				Hald Dataset			
	MSE	MDAE	MAE	MAPE	MSE	MDAE	MAE	MAPE
LS	0.46 (1.5)	0.68 (0.58)	0.56 (0.48)	2.23 (2.17)	3.68 (2.57)	1.29 (2.31)	0.78 (1.23)	0.05 (0.26)
Liu	0.14 (0.13)	0.50 (0.41)	0.55 (0.47)	2.21 (2.10)	1.19 (1.02)	1.44 (2.23)	0.75 (0.96)	0.04 (0.25)
RF	0.15 (0.12)	0.52 (0.47)	0.33 (1.05)	0.87 (0.79)	0.16 (0.47)	1.31 (2.20)	0.33 (1.05)	0.12 (0.69)
RLiu	0.13 (0.12)	0.28 (0.26)	0.26 (0.90)	0.14 (0.44)	0.14 (0.44)	1.23 (0.88)	0.26 (0.90)	0.03 (0.07)

(.) Without Outliers

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 Dataset: Woods et al [18] introduced the Hald or Portland Cement Dataset. This data frame contains 13 observations with four independent variables. They are tricalcium aluminate (tca), tricalcium silicate (tcs), tetracalcium aluminoferrite (tcaf) and β-dicalcium silicate (bdcs). The response variable is the evolved heat after 180 days in a cement mix. Since the VIFs of this Hald data set were greater than 10, the explanatory variables are highly correlated. As a result, the dataset has high multicollinearity. Also, this data set has one outlier identified and removed using Cook's distance method.

Simulation studies were carried out to examine the efficiency of various regression procedures. 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, $\rho = 0.99$ and number of variables $p = 5$. Different levels of contamination, $\varepsilon = 0\%$, 5% and 10% were studied for sample size $n = 50, 100\ and\ 200$. The performance of various regression procedures were compared by computing different error measures and the results obtained for different number of observations with various levels of contamination are shown in Table 2.

The results obtained from Tables 1 and 2 demonstrates that the computed error measures under various regression procedures are slightly different from each other. Also the Robust Liu method has the smallest error measures compared with the others. Hence Robust Liu (RLiu) regression technique is more efficient than the others in the case of datasets having indication of multicollinearity and has outliers.

Table 2: *Computed error measures under various regression methods (Simulation Dataset)*

n	ε	Methods	MSE	MDAE	MAE	MAPE
50	0%	LS	6.20	3.11	4.03	0.58
		Liu	4.01	2.88	2.78	0.28
		RF	2.64	1.61	1.26	0.15
		RLiu	1.78	1.05	1.23	0.14
	5%	LS	7.05	2.88	4.13	0.48
		Liu	5.96	2.62	3.14	0.34
		RF	5.23	2.47	3.13	0.47
		RLiu	4.88	1.02	1.55	0.18
	10%	LS	10.29	4.83	3.74	0.70
		Liu	7.32	2.62	3.36	0.68
		RF	7.31	2.55	3.24	0.61
		RLiu	3.28	1.19	1.75	0.24
100	0%	LS	6.01	2.55	3.89	0.58
		Liu	4.73	2.38	3.02	0.52
		RF	2.41	2.25	2.99	0.49
		RLiu	2.36	1.78	1.07	0.15
	5%	LS	4.92	4.46	3.48	0.69
		Liu	3.91	2.31	2.97	0.62
		RF	2.87	2.25	2.87	0.51
		RLiu	2.42	1.84	1.22	0.36
	10%	LS	6.37	2.96	3.15	1.02
		Liu	4.99	2.35	2.76	0.85
		RF	4.82	2.21	2.73	0.72
		RLiu	2.78	1.81	1.15	0.63
200	0%	LS	7.07	3.14	3.63	1.92
		Liu	5.85	2.62	3.19	1.43
		RF	3.72	2.59	3.08	1.19
		RLiu	1.32	1.57	1.79	0.95
	5%	LS	9.33	4.67	3.88	0.78
		Liu	8.27	2.97	3.45	0.61
		RF	2.99	2.70	3.15	0.52
		RLiu	2.11	1.67	1.98	0.38
	10%	LS	9.69	5.24	4.01	0.92
		Liu	7.18	3.16	3.32	0.86
		RF	6.93	2.35	2.89	0.79
		RLiu	2.87	1.76	1.05	0.62

IV. Conclusion

Statistical learning techniques are important in various research fields, with regression analysis being a prominent method. The commonly used linear regression procedures 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 datasets. This paper explores several regression methods, including Least Squares, Liu, Random Forest and Robust Liu. Further, evaluates their performance on different real and simulated datasets by considering the problems of multicollinearity and outliers by computing various error measures. On the basis of the computed error measures, this study concludes 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] Breiman, L. (2001). Random forests. *Machine learning*, 45: 5-32.

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

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

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

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

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

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

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

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

[13]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.

[14] Muthukrishnan, R. and Karthika Ramakrishnan (2024). A New Robust Liu Regression Estimator for High-Dimensional Data. *Reliability: Theory & Applications*, 19:214-219.

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

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

[17] 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.

[18] 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.

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