Academic Journal of Nawroz University (AJNU), Vol.10, No.1, 2021

https://doi.org/10.25007/ajnu.v10n1a578

Using Robust Ridge Regression Diagnostic Method to Handle Multicollinearity Caused High Leverage Points

Kafi Dano Pati

Department of Computer Science, Duhok University, Dohuk, Kurdistan-region, Iraq

ABSTRACT

Statistics practitioners depends on the Ordinary Least Squares (OLS) method to estimate the parameters because of its optimal properties and simplicity to Computation. However, the OLS estimator can be strongly affected by the existence of multicollinearity. Even though in the presence of multicollinearity the OLS estimate still remained unbiased. But the standard errors of the estimated parameter become inaccurate. In this paper, we proposed some alternative methods to estimate the parameters in presence of multiple high leverage points which cause the multicollinearity problem. The procedure used the ordinary least squares method to estimate the parameters as the initial followed by ridge regression estimator. We incorporated the robust Least Trimmed Squares (LTS) estimator to down weight the effects of multiple high leverage points which leads to the reduction of the effects of multicollinearity. The result seemed that the suggested method Ridge Least Trimmed Square (RLTS) gives a substantial improvement over the Ridge Regression.

Keywords: Multicollinearity, Outlier, Ridge Regression, Robust Regression.

1. Introduction

The ordinary least squares method (OLS) is one of the oldest statistical method dating back to the age of slide rules. Today computer is abundant, high-quality statistical software are free, and statisticians have developed several estimation techniques making it easier to understand., The linear regression is still popular (Toutenburg 2008). The estimate β is determined by minimizing the function $\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 = \sum_{i=1}^{n}\left(r_i\right)^2$ and the estimator of the β is given by,

$$\hat{\beta}_{LS} = \left(x^T x\right)^{-1} x^T y$$

OLS estimate is unbiased and minimum variance among all unbiased linear estimators provided that the errors are independent and identical. The presences of multicollinearity in the data set produce poor estimates in the regression coefficient although OLS estimate is unbiased in the presence of multicollinearity.

Multicollinearity is a problem where two or more explanatory variables are correlated with each other or they are highly linearly related and give an inaccurate prediction about the dependent variable. The collinearity may involve more than two variables and are more difficult to detect and the consequences may be harder to explain (Hocking 2003). We may modify ridge regression estimator which may improve the reliability of the regression coefficients.

The ridge regression (RIDGE) method was first proposed by Hoerl (1962) and was improved by Hoerl and Kennard (1970). Whilst method of ridge regression is not very strong about deviation from the outliers. Outliers are a value in a data set that is far from the other values and far from the line implied by the rest of the data. Outlying data points may have inappropriate effect on the OLS and the ridge estimates. There are different types of outliers.

However, outliers can be categorized into two types; the observations with large residuals are also known as vertical outliers. The second type is high leverage points which are horizontal outliers. (Molina et al., 2009).

The problem has become more complicated when both multicollinearity and outliers are present in the data set. In the recent years many efforts have been made to obtain dependable estimates especially in the presence of heavy– tailed error distribution and multicollinearity.

Robust regression is an alternative procedure to ordinary least squares that dampen the influence of outlying cases, it can be properly used when there is evidence that the distribution of the error term is abnormal, and/ or there are outliers that affect the model. It is also resistant which is less affected by the outliers. Ridge regression method improved the estimates when there is multicollinearity problem.

(Midi and Zahari 2007) did a simulation study to test the robustness of six estimators on a multiple linear regression model with joint problems of multicollinearity and non-normal errors. The achievement of the six estimators, that is the Ordinary Least Squares (OLS), Ridge Regression (RIDGE), Ridge Least Absolute Value (RLAV), Weighted Ridge (WRID), MM and a robust ridge regression estimator based on MM estimators (RMM) are compared.

The RMM is a modification of the Ridge Regression (RIDGE) by combining robust MM estimator. The experimental evidence shows that RMM is the best between the six estimators for many combinations of disturbance distribution and degree of multicollinearity. They guess that the modified method would be less sensitive to the existence of outliers and has a high breakdown point. They have deleted the influence of outliers by the highly robust and efficient MM estimator and also deleted the problem of multicollinearity by ridge regression.

They concluded that the simulation studies obviously display that ridge MM-estimator RMM estimator shows the general practical option over other estimators when both multicollinearity and outliers exist. (Meriam et al., 2012) improved a robust ridge regression estimator by using Weighted Ridge MM-Estimator (WRMM) and that is possible to handle the multicollinearity problem.

Thus they suggested to compare many present estimators with this technique called ordinary least squares (OLS), robust regression based on MM estimator, ridge regression (RIDGE), weighted ridge (WRID) and ridge MM-estimator (RMM) using two norms to compare root mean square error (RMSE) and biasness. Generally, it has been found that the suggested estimator scores will be good in contrast the five present estimators when the error term is abnormal.

Finally, they conclude that the OLS estimator is better than other estimators when there is no multicollinearity in the data. On the other hand WRMM and WRID estimators prefer the other estimator when multicollinearity is moderate and high, thus WRMM is the most efficient to all existing estimators when multicollinearity is high.

2. Methodology

2.1 Ridge Regression Estimators

It is an analytic technique to be used when the predictor variables in a multiple linear regression are highly correlated a situation which may result in unstable regression coefficients and difficulties in interpretation.

Ridge regression is one of several methods that have been proposed to handle multicollinearity problems by modifying the method of least squares to allow biased estimators of the regression coefficients. When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be the preferred estimator since it will have a larger probability of being close to the true parameter value. (Nachtsheim, 2004). The ridge regression estimator is defined as follows.

$$\hat{\boldsymbol{\beta}}_{\mathbf{Ridge}} = \mathbf{(X'X + KI)^{-1} X'Y}$$

where K is the biasing constant and I is the pxp identity matrix.

In the exercise, the optimal value of K is unknown constant adding to the diagonal of the correlation matrix. Different methods in finding K have appeared in the literature like described by Hoerl and kenard (1970). The estimator of K by Hoerl et al. (1975) is given by

$$\mathbf{K} = \frac{\rho \mathbf{S}^2_{\mathbf{OLS}}}{\hat{\boldsymbol{\beta}}'_{\mathbf{OLS}}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}}$$

where ρ is the correlation coefficient?

$$\mathbf{S}^2_{\mathbf{OLS}} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}})}{n-p}$$

Where n is the sample size, p is the number of parameters and if K= 0, $\hat{\boldsymbol{\beta}}_{\mathbf{RID}} = \hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$, when K> 0 , $\hat{\beta}_{RID}$ is biased but more stable and precise than the OLS estimator and when Hoerl and Kennard (1970) have shown that there always exist a value K> 0such that MSE $\hat{\beta}_{RID}$ is less than MSE $\hat{\beta}_{OLS}$.

2.2 Robust Regression Estimators

Robust regression estimators have been confirmed to be more reliable and efficient than least squares estimator especially when disturbances are non-normal distribution. Non-normal disturbances are disturbance distributions that have heavy or fatter tails than the normal distribution and are tends to produce outliers. Habsha Midi et al. (2007), Kafi, (2020).

Since outliers greatly affect the estimated coefficients, standard errors and test statistics, Therefore, the usual statistical procedure is more effective in estimating parameters with the presence of outliers.

So the robust procedure it is better process to reduce the effect in presence of outliers. Robust procedure fit a regression by using estimators that dampen the impact of influential points. To detect outliers, we look for those points that lying far away from the pattern formed and have large residuals from the line regression. Several works on robust estimation have been proposed in the literature.

Among them Habshah Midi et al. (2007) who proposed the Least Absolute Values (LAV) estimator. But we can use least Trimmed squares (LTS) to replace LAV since it has a highest possible breakdown point, that is 50%. LTS is a robust estimator having the highest possible breakdown point, that is 50% .

2.3 Least Trimmed Squares (LTS)

The formula of LTS is the same as OLS and the only difference being that the largest squared residuals are not used in the summation, thus allowing the fit to stay away from the outliers. (Rousseeuw 2003), (Kafi et al 2014).

The LTS estimates are obtained by finding the regression model parameters to achieve min $\sum_{\mathbf{i=1}}^{\mathbf{h}}(\mathbf{y_i}-\boldsymbol{\beta}_0-\boldsymbol{\beta}_\mathbf{i}\mathbf{x_i})^2$ .In other words, minimizing $\sum_{\mathbf{i=1}}^{\mathbf{h}}\mathbf{e_i}^2$ *where* $\mathbf{e_1}^2 \le \mathbf{e_2}^2 \le .... \le \mathbf{e_n}^2$ are the ordered squared residuals and the value of *h* must be defined by $h = \left[\frac{n}{2}\right] + \left[\left(\frac{p+1}{2}\right)\right]$ to obtain the number of observations in each subsample, select a subsample of *h* observation of the original data as follows $\binom{n}{h}$ with replacement of the data set.

After the summation squared residuals for each subsample to obtain the single value, select the minimum residuals for each subsample and use the parameter for these residuals for the original observation to obtain the fitted regression to remove the high leverage point.

3. Results and Discussion

In this paper we proceeding least trimmed squares method to estimate parameter using suitable data; the multicollinearity and outliers data taken from the body fat example are to be used. This example contains twenty observation of (y) with corresponding explanatory variables of (x1, x2, x3) is used. The analysis begins by effects of multicollinearity and outliers using diagnostic plot, The OLS, Ridge

Regression, LTS, RLTS were used in the original and transformed data. The results are presented in the graphs and the tables below.

In the first plot, OLS residual plot of the original data against the regression fitted values which are presented in Figure (1). The situation for existence of multicollinearity and outliers and this can be identified when the residuals are not randomly distributed around the zero residual, with an indication of systematic trend on the plot. Based on this concept, the plots clearly indicate multicollinearity and outliers, the data for the purpose to remedy the short Coming of OLS problems of multicollinearity and outliers.

To use this technique for this data, we first need to plot the residual against the explanatory variables with a data that contain multicollinearity and outliers. The plot of Fig. 2 gives the diagnostic plot of the residual against the fitted values obtains the transformed data, while Fig 3 gives the linear regression models obtain from the Ridge regression techniques. From this plot we notice that there are some differences between the estimators. This is evidence that the performance of the methods was satisfactory. So that to test the consequence of multicollinearity in the existence of outliers.

Fig 4 gives the linear regression models obtain from the robust estimation techniques is used. The plot of linear regression models obtains from the robust estimation techniques using the transformed data give a clear idea for comparison of Figure 1 of OLS plot.The result from the plot indicate the performance of this transformation can assist in producing a outliers.

From Table 1 we observe estimate of parameter is performance of the proposed ridge parameter is better than other OLS for all combinations of correlation between predictors, and give the summary results of statistics. The result of Table 1 reveal the influence of outleirs on the regression model, when OLS is used to estimate the regression parameter comepared with the regression parameter obtain from the LTS estimate.

On the other hand, when considering the estimate of the result obtain from the body fat data that involve three explanatory variables. These tables provide the breakdown point result of the estimated parameter using OLS and Ridge Regression, LTS, Ridge LTS regression.

The criterion used to evaluate the best regression model is using the standard error and t-value estimated from the body fat data involving all the explanatory variables. Base on the results obtain in Table 1 and Table 2, The RLTS posses the least standard errors with largest t-values compared to the t-value obtain from LTS and OLS.

TABLE 1: SUMMARY OF STATISTICS FOR THE BODY FAT EXAMPLE

Method	β_0	β_1	β_2	β_3
OLS/original	11.709	0.433	-0.286	-0.219
OLS/Transformed	0.001	0.426	-0.293	-0.156
Ridge	0.005	0.046	0.044	-0.010
LTS	0.003	0.471	-0.332	-0.174
RLTS	0.001	0.176	-0.069	-0.060

TABLE 2: SUMMARY STATISTICS OF STANDARD ERROR AND T-VALUE FOR THE BODY FAT EXAMPLE

Method	S.E. β_0	S.E. β_1	S.E. β_2	S.E. β_3	t-value β_0	t-value β_1	t-value β_2	t-value β_3
OLS/ original	9.978	0.302	0.258	0.160	0.117	0.144	-0.111	-0.137
OLS/ Transformed	0.002	0.297	0.265	0.114	0.000	0.144	-0.111	-0.137
LTS	0.003	0.337	0.301	0.130	-0.004	0.140	-0.110	-0.134
RLTS	0.003	0.297	0.265	0.114	0.000	0.144	-0.111	-0.137

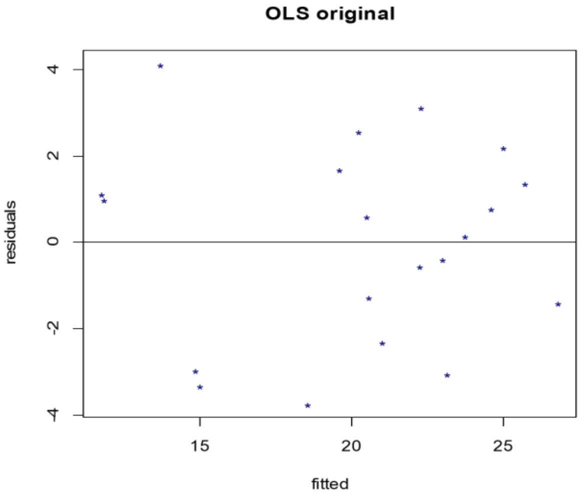

Fig. 1. Plot of OLS residual versus fitted values (original data)

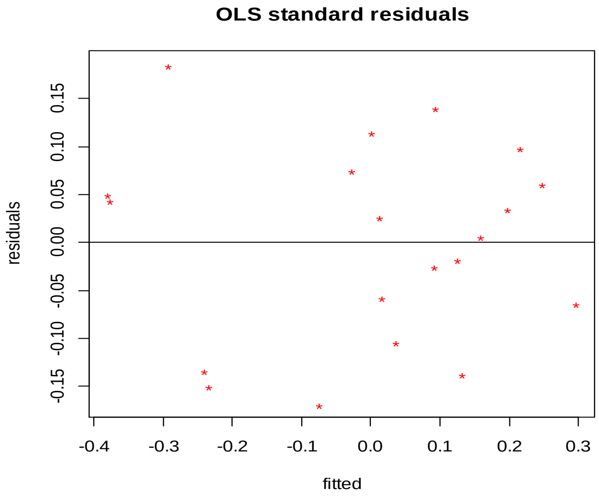

Fig. 2. Plot of OLS residual versus fitted values (Transformed data)

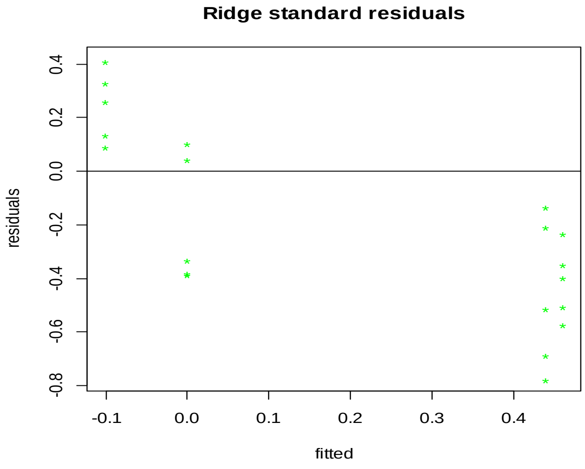

Fig. 3. Plot of Ridge Regression residual versus fitted values (Transformed data)

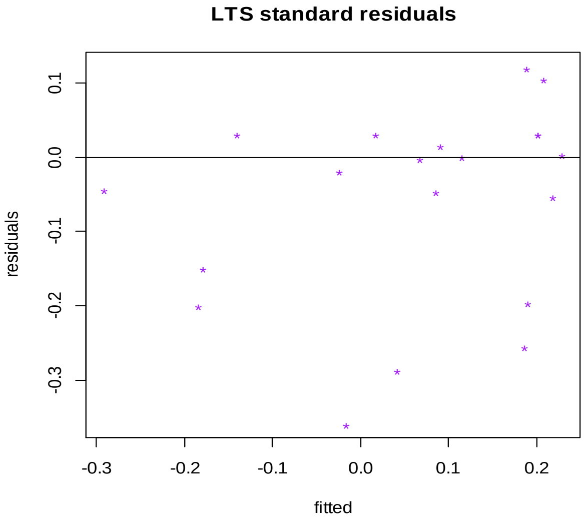

Fig. 4. Plot of LTS residual versus fitted values (Transformed data)

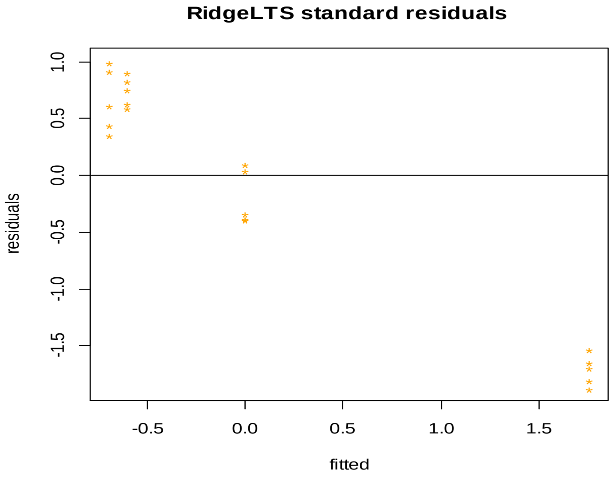

Fig. 5. Plot of RLTS residual versus fitted values (Transformed data)

4. Conclusion

First required in the data is to detection the multicollinearity. Ridge Regression (RR) is an alternative estimation method when has an extremely high of multicollinearity in the data therefore RR is a more advanced solution of multicollinearity but in general greatly reduces the standard error and giving more reliable estimates of β.

The robust ridge regression estimator allows to protect the data against outliers and shrink the regression coefficients. Replaced RR with the new proposed estimation method used ridge least trimmed squares-based number of the data points and strength of multicollinearity in it.

The performance of the proposed method is better than other estimators used. Then principal focus in this paper would be to produce reliable techniques for solve the problem of multicollinearity in the presence of outliers. Even though empirical study discloses the OLS estimation is often impacted by the outliers.

Conclude that OLS it isn't reliable in presence of multicollinearity. However, the RLTS arises to become basic all data efficiency and much reliable because it is less impacted by the outliers. Finally, the outcomes appeared the point out the LTS and RLTS methods provides a substantial improvement within the other existing technique for solve the problems of multicollinearity in the data.

5. References

1. Arthur E. H. and Robort W. K. (1970). "Ridge Regression: *Biased Estimation for Non-orthogonal Problems*". Technomertrics, Volume 12, Issue 1.
2. Rousseeuw. P; Daniels. B.; Leroy. A. (1984). "Applying robust regression to insurance". *Insurance: Mathematics and Economics* 3 -67-72.
3. Ronald R Hocking. (2003). "Methods and Application of Linear Models". Wiley Series in Probability and Statistics. DOI:10.1002/0471434159
4. Peter J. Rousseeuw, Annick M. Leroy. (2003). "Robust Regression and Outlier Detection". *(Second edition).*Wiley Book, 329 pages.
5. Kutner N. N. (2004). "Applied Linear Regression Models". *(fourth edition).*
6. Habshah M. and Marina Z. (2007). "A simulation study on ridge regression estimators in the presence of outliers and multicollinearity". *Jurnal Teknologi*, 47(C) Dis. 59-74.
7. Roa T..; Shalabh H. (2008). "Linear Models and Generalization". (Least Squares and Alternatives) (*third edition*).
8. Molina I., P., and Perez B. (2009). "Robust Estimation in Linear Regression Models with Fixed Effects". De Madrid Calle Madrid, 126 28903. Paper 09-88 (27)
9. Siti M. Z.; Mohammad S. Z. and Mohammad I. AL-Banna B. I. (2012). "Weighted Ridge MM-Estimator in Robust Ridge Regression with Multicollinearity". Pp. (124-129).
10. Kafi D. P., Robiah A., and Bello A.R. (2014). "Ridge Least Trimmed Squares Estimators in Presence of Multicollinearity and Outliers". Journal of Australian Journal of Basic and Applied Sciences, Volume 8(12), pp 429.
11. Kafi D. P. (2020). "Using Standard Error to Find the Best Robust Regression in Presence of Multicollinearity and Outliers", IEEE. **DOI:** 10.1109/CSASE48920.2020.9142066, Page(s): 266 - 271