PAPER • **OPEN ACCESS**

Robust Ridge regression to solve a multicollinearity and outlier

Robust Ridge regression to solve a multicollinearity and outlier

N E Jeremia, S Nurrohmah and I Fithriani

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

Corresponding author's e-mail: snurrohmah@sci.ui.ac.id

Abstract. Regression analysis is one of many methods used for analysing data. Method that used for estimating parameter in linear regression model is ordinary least square (OLS). OLS will give best estimator when all the assumptions are met. But in reality, sometimes not all the assumptions are met. Assumptions that usually violated are multicollinearity and outlier. Ridge regression is a regression method that give constrain on the parameters that used to deal with multicollinearity, meanwhile Robust regression is used to overcome the presence of outlier. Robust regression is a regression method that has robust property that achieved by using S-estimation is used. Ridge regression and Robust regression combined into Robust Ridge regression to overcome multicollinearity and outlier simultaneously.

Keywords: multicollinearity, outlier, Ridge, Robust, regression.

1. Introduction

There are some assumptions that should be satisfied in regression analysis using ordinary least square (OLS) method so that the estimator results from the process become best linear unbiased. The estimator from OLS method is considered as bad estimator when the assumptions are not fulfilled. The violation of assumptions in this paper is the presence of multicollinearity and outlier.

Multicollinearity is a condition where there are two or more independent variables have medium to high correlation. Multicollinearity will make the estimator to have large variance. Large variance will increase the mean square error, thus making the estimator bad.

In 1970, Hoerl and Kennard develop Ridge regression to overcome the presence of multicollinearity [1]. Ridge regression will give a bias estimator. The bias will depend on the ridge constant k so that it is required to choose the optimal Ridge constant k to minimize the bias. For choosing the optimal Ridge constant k, Hoerl *et al.* [2] develop the formula in 1975.

Outlier is an unusual observation from the data. Outlier often tends to pull the regression line to itself, resulting bad estimator and change sign of the coefficient parameter. Mendenhall and Sincich define outlier as an observation that has absolute standardized residual more than three [3].

Robust regression is used to overcome the presence outlier. Robust regression is a regression method that using robust estimation to achieve robust property that is resistance to outlier. S-estimation is one of well-known robust estimation and will be used in Robust regression.

Ridge regression and Robust regression are combined resulting Robust Ridge regression to address the problem where multicollinearity and outlier exists. Robust coefficient will be computed using S-estimation, and then applying Ridge method to construct parameter estimation of Robust Ridge regression.

This paper is organized as follows: section 2 present brief explanation about OLS method. Section 3 gives explanation about Ridge regression and generalized Ridge regression. Section 4 gives the explanation about Robust regression and procedure of S-estimation. Section 5 explaining Robust Ridge regression. Section 6 gives its application on data. Finally, Section 7 will give conclusion.

2. Linear regression model

Linear regression is a method to identify relationship between a response variable (y) and one or more independent variables (**X**). The multiple linear regression model with p independent variables is:

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

where $\boldsymbol{y}$ is a $(n\times 1)$ vector of response variable, $\boldsymbol{X}$ is a $(n\times p)$ matrix, $\boldsymbol{\beta}$ is a $(p\times 1)$ vector of an unknown regression parameters and $\boldsymbol{\varepsilon}$ is a $(n\times 1)$ vector of error term that is assumed to be independent and identically normally distributed with mean 0 and constant variance σ^2. The estimator for unknown parameters using OLS is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = (\boldsymbol{X^tX})^{-1}\boldsymbol{X^ty} \tag{2}$$

will give BLUE properties under the assumptions in Gaussian-Markov Theorem. The canonical form of regression found by using Spectral Decomposition is:

$$\boldsymbol{y} = \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{Z} = \boldsymbol{X\gamma}$ and $\boldsymbol{\alpha} = \boldsymbol{\gamma^t\beta}$, $\boldsymbol{\gamma}$ is a matrix of eigenvectors whose columns are normalized and corresponding with $\boldsymbol{\Lambda}$, a diagonal matrix with entry eigenvalues of $\boldsymbol{X^tX}$. By Spectral decomposition, $\boldsymbol{\Lambda} = \boldsymbol{Z^tZ} = \boldsymbol{\gamma^tX^tX\gamma}$. The OLS estimator for canonical form of regression is:

$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}} = (\boldsymbol{Z^tZ})^{-1}\boldsymbol{Z^ty} = \boldsymbol{\Lambda}^{-1}\boldsymbol{Z^ty} \tag{4}$$

and:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = \boldsymbol{\gamma}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}} \tag{5}$$

The MSE for OLS estimator is:

$$\boldsymbol{MSE}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\sigma}^2(\boldsymbol{X^tX})^{-1} \tag{6}$$

3. Ridge regression and generalized Ridge regression

3.1. Ridge regression

Ridge Regression is a method to overcome the presence multicollinearity. The principal of Ridge regression is to bind the value of estimated parameters by some constant P, where P is a finite positive constant. By using this principal, it is same as adding a constant *k* at the diagonal of $\boldsymbol{X^tX}$ matrix. This constant *k* is known as Ridge constant. Ridge regression will give biased estimator depend on the constant *k*. The estimator for Ridge regression is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X^tX} + \boldsymbol{kI})^{-1}\boldsymbol{X^ty} \tag{7}$$

and for the canonical form of regression, the estimator for Ridge regression is:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\boldsymbol{R}} &= (\boldsymbol{\Lambda} + \boldsymbol{kI})^{-1}\boldsymbol{Z^ty}\\ &= \boldsymbol{B}^{-1}\boldsymbol{Z^ty}\\ &= \boldsymbol{B}^{-1}\boldsymbol{\Lambda}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}}\\ &= \boldsymbol{B}^{-1}\boldsymbol{\Lambda}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}}\end{aligned} \tag{8}$$

with $\boldsymbol{B} = \boldsymbol{\Lambda} + k\boldsymbol{I}$.

One method for choosing the optimal Ridge constant *k* is proposed by Hoerl *et al.* [2] that is:

$$\boldsymbol{k} = \frac{\boldsymbol{p}\,\widehat{\boldsymbol{\sigma}}^2}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}}\widehat{\boldsymbol{\alpha}}_{\boldsymbol{i}}^2} \tag{9}$$

where p is number of independent variable, $\hat{\sigma}^2$ is estimate variance and $\hat{\alpha}_i$ is a parameter of canonical form of regression using OLS.

Journal of Physics: Conference Series **1442** (2020) 012030 doi:10.1088/1742-6596/1442/1/012030

The bias, variance and mean square error (MSE) of Ridge estimator are:

$$\boldsymbol{Bias}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = (\boldsymbol{R}_{\boldsymbol{k}} - \boldsymbol{I})\boldsymbol{\beta} \tag{10}$$

$$\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = \boldsymbol{\sigma}^{2}\boldsymbol{R}_{\boldsymbol{k}}(\boldsymbol{X}^{t}\boldsymbol{X})^{-1}\boldsymbol{R}_{\boldsymbol{k}}^{t} \tag{11}$$

$$\boldsymbol{MSE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = \boldsymbol{\sigma}^{2}\boldsymbol{R}_{\boldsymbol{k}}(\boldsymbol{X}^{t}\boldsymbol{X})^{-1}\boldsymbol{R}_{\boldsymbol{k}}^{t} + (\boldsymbol{R}_{\boldsymbol{k}} - \boldsymbol{I})\boldsymbol{\beta}\boldsymbol{\beta}^{t}(\boldsymbol{R}_{\boldsymbol{k}} - \boldsymbol{I})^{t} \tag{12}$$

with $\boldsymbol{R}_{\boldsymbol{k}} = (\boldsymbol{X}^{t}\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^{t}\boldsymbol{X}$.

3.2. Generalized Ridge regression

Generalized Ridge regression is a method found by generalizing Ridge regression. In Ridge regression, the Ridge constant k is add to diagonal of matrix $\boldsymbol{X}^{t}\boldsymbol{X}$, while at generalized Ridge regression, the constant can be different for every diagonal entry of matrix $\boldsymbol{X}^{t}\boldsymbol{X}$ so that there will be $k_1, k_2, \dots, k_p$. It can be said that we add matrix $\boldsymbol{K}$ (diagonal matrix with entry $k_1, k_2, \dots, k_p$) instead of matrix $k\boldsymbol{I}$. The estimator for generalized Ridge regression in canonical form can be written as:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\boldsymbol{GR}} &= (\boldsymbol{Z}^{t}\boldsymbol{Z} + \boldsymbol{K})^{-1}\boldsymbol{Z}^{t}\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}\boldsymbol{Z}^{t}\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}\boldsymbol{\Lambda}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}} \\ &= (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1})\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{LS}}\end{aligned} \tag{13}$$

where $\boldsymbol{A} = \boldsymbol{Z}^{t}\boldsymbol{Z} + \boldsymbol{K}$ and the estimator for generalized Ridge regression is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GR}} = \boldsymbol{\gamma}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{GR}} \tag{14}$$

One method for choosing the optimal Ridge constant $k_1, k_2, \dots, k_p$ is proposed by Hoerl and Kennard [1] that is:

$$\boldsymbol{k}_{\boldsymbol{i}} = \frac{\widehat{\boldsymbol{\sigma}}^{2}}{\widehat{\boldsymbol{\alpha}}_{\boldsymbol{i}}^{2}} \tag{15}$$

where $\hat{\sigma}^2$ is estimate variance and $\hat{\alpha}_i$ is a parameter of canonical form of regression using OLS.

The bias, variance and mean square error (MSE) of generalized Ridge estimator are:

$$\boldsymbol{Bias}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{GR}}) = (\boldsymbol{G}_{\boldsymbol{k}} - \boldsymbol{I})\boldsymbol{\alpha} \tag{16}$$

$$\boldsymbol{Var}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{GR}}) = \boldsymbol{\sigma}^{2}(\boldsymbol{A}^{-1}\boldsymbol{Z}^{t})(\boldsymbol{A}^{-1}\boldsymbol{Z}^{t})^{t} \tag{17}$$

$$\boldsymbol{MSE}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{GR}}) = \boldsymbol{\sigma}^{2}(\boldsymbol{A}^{-1}\boldsymbol{Z}^{t})(\boldsymbol{A}^{-1}\boldsymbol{Z}^{t})^{t} + (\boldsymbol{G}_{\boldsymbol{k}} - \boldsymbol{I})\boldsymbol{\alpha}((\boldsymbol{G}_{\boldsymbol{k}} - \boldsymbol{I})\boldsymbol{\alpha})^{t} \tag{18}$$

with $\boldsymbol{G}_{\boldsymbol{k}} = (\boldsymbol{Z}^{t}\boldsymbol{Z} + K)^{-1}\boldsymbol{Z}^{t}\boldsymbol{Z}$ and $\boldsymbol{A} = \boldsymbol{Z}^{t}\boldsymbol{Z} + \boldsymbol{K}$.

4. Robust regression

Robust regression is one method to overcome the presence of outlier. Differ from OLS which is giving same weight to all observation, Robust regression will give different weight for all observation. The observation that has large standardized residual will be given small weight so that it gives small influence.

S-estimation is one of method to achieve robust property on regression. S-estimation is proposed by Rousseeuw and Leroy [4] in 1987 as alternative method of M-estimation. While in M-estimation only using median as estimator of σ, S-estimation define the estimator of σ as dispersion of residual. The dispersion of residual s define by Rousseeuw and Leroy as solution of:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right) = K \tag{19}$$

where K is a constant and $\rho(u)$ is a objective function.

S-estimation using s define at equation 3 to minimize:

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) \tag{20}$$

The $\widehat{\boldsymbol{\beta}}$ that satisfied (4) will be S-estimator and has robust property.

The weight for every observation is define as:

$$\boldsymbol{w_i} = \frac{\boldsymbol{\rho(u_i)}}{\boldsymbol{u_i^2}} \tag{21}$$

Susanti *et al.* [5] in 2013 propose the algorithm to solve equation 4, that is:

1. Use the OLS method to compute estimator $\boldsymbol{\beta_s^0}$ that will be used as initialization.
2. Compute the residual e_i^0 using $\boldsymbol{\beta_s^0}$.
3. Compute the $\hat{\sigma}_0 = \frac{median|e_i - median(e_i)|}{0.6745}$
4. Compute $u_i = \frac{e_i^0}{\hat{\sigma}_0}$.
5. Compute weight value w_i^0 define at (5) and construct diagonal matrix $\boldsymbol{W}$ where the diagonal is the weight value.
6. Compute $\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}^{\boldsymbol{1}} = \left(\boldsymbol{X^t W^0 X}\right)^{\boldsymbol{-1}} \boldsymbol{X^t W^0 y}$.
7. Repeat 2-6 until obtain convergent value of $\widehat{\boldsymbol{\beta}}$ but at step 3 use $\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i e_i^2}$.

The resulting $\widehat{\boldsymbol{\beta}}$ at step 7 is the Robust estimated parameter.

5. Robust Ridge regression

Assume that Robust estimated parameter obtained using S-estimation is $\widetilde{\boldsymbol{\beta}}$ and by using $\boldsymbol{\alpha} = \boldsymbol{\gamma^t \beta}$ obtain the Robust estimated parameter of canonical form $\widetilde{\boldsymbol{\alpha}}$. By using (1), it can be said that when the Ridge method is applied, it is same as multiplied it by $\left(\boldsymbol{I} - k\boldsymbol{B^{-1}}\right)$. The Robust Ridge estimator is obtained by applying Ridge method to estimated parameter obtained using S-estimation. The estimated parameter of Robust Ridge regression for canonical form is:

$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}} = \left(\boldsymbol{I} - \boldsymbol{kB^{-1}}\right) \widetilde{\boldsymbol{\alpha}} \tag{22}$$

and the estimated parameter of Robust Ridge regression is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{RR}} = \boldsymbol{\gamma}\, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}} \tag{23}$$

The bias, variance and MSE of the Robust Ridge estimator are:

$$\boldsymbol{Bias}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = -\boldsymbol{kB^{-1}\alpha} \tag{24}$$

$$\boldsymbol{Var}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = \left(\boldsymbol{I} - \boldsymbol{kB^{-1}}\right)\boldsymbol{\Omega}\left(\boldsymbol{I} - \boldsymbol{kB^{-1}}\right)^{\boldsymbol{t}} \tag{25}$$

$$\boldsymbol{MSE}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = \left(\boldsymbol{I} - \boldsymbol{kB^{-1}}\right)\boldsymbol{\Omega}\left(\boldsymbol{I} - \boldsymbol{kB^{-1}}\right)^{\boldsymbol{t}} + \boldsymbol{k^2 B^{-1} \alpha \alpha^t B^{-1}} \tag{26}$$

where Ω is a covariance matrix obtained by using S-estimation [1].
Choosing k for Robust Ridge regression is same as when choosing k for Ridge regression. Using method that Hoerl and Kennard [1] proposed that is:

$$\boldsymbol{k} = \frac{\boldsymbol{p\,\hat{\sigma}^2}}{\sum_{i=1}^{p} \boldsymbol{\hat{\alpha}_i^2}} \tag{27}$$

Generalized Ridge regression can also be used to replace Ridge regression. By using equation (2), it can be said that when the generalized Ridge regression is used, it is same as multiplied it by $\left(\boldsymbol{I} - \boldsymbol{KA^{-1}}\right)$. The Robust Ridge estimator is obtained by using generalized Ridge regression instead of Ridge regression to estimate parameter obtained using S-estimation. The estimated parameter of Robust Ridge regression for canonical form is:

$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}} = \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1}\right)\widetilde{\boldsymbol{\alpha}} \tag{28}$$

and the estimated parameter of Robust Ridge regression is:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{RR}} = \boldsymbol{\gamma}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}} \tag{29}$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{RR}} = \boldsymbol{\gamma}\,\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}$$

The bias, variance and MSE of the Robust Ridge estimator are:

$$\boldsymbol{Bias}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = -\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{\alpha} \tag{30}$$

$$\boldsymbol{Var}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1}\right)\boldsymbol{\Omega}\left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1}\right)^{\boldsymbol{t}} \tag{31}$$

$$\boldsymbol{MSE}(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{RR}}) = \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1}\right)\boldsymbol{\Omega}\left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}^{-1}\right)^{\boldsymbol{t}} + \left(-\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{\alpha}\right)\left(-\boldsymbol{K}\boldsymbol{A}^{-1}\boldsymbol{\alpha}\right)^{\boldsymbol{t}} \tag{32}$$

where Ω is a covariance matrix obtained by using S-estimation.

6. Application on dataset

The dataset that will be used is Boston dataset. It is about median house price at housing in Boston. This dataset is already available on R software. This dataset was made by Harrison and Rubinfeld [6] in 1978. It has 506 observations and 14 variables, and its 14 variables are criminal rate (x_1), proportion of residential land zoned for lots over 25,000 square feet (x_2), proportion of non-retail business acres per town (x_3), Charles River dummy variable (1 if tract bounds river; 0 otherwise) (x_4), nitric oxides concentration (parts per 10 million) (x_5), average number of rooms per dwelling (x_6), proportion of owner-occupied units built prior to 1940 (x_7), weighted distances to five Boston employment centers (x_8), index of accessibility to radial highways(x_9), full-value property-tax rate per $10,000 ($x_{10}$), pupil-teacher ratio by town (x_{11}), $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town (x_{12}), % lower status of the population (x_{13}) and median value of owner-occupied homes in $1000's (y). This dataset has multicollinearity on x_9 and x_{10} with VIF 7.484496 and 9.008554 and has outlier on seven observations. Because there are multicollinearity and outlier, Robust Ridge regression can be used. Robust Ridge regression using Ridge method resulting MSE 105.265, while using generalized Ridge resulting MSE 13.40343. It means it is better to use generalized Ridge instead of Ridge method on Robust Ridge regression, even though more calculation is needed because of value of $k_1, k_2, \dots, k_{13}$. Robust Ridge regression using Ridge method resulting bias 22.88475, while using generalized Ridge resulting bias 2.20294.

7. Conclusions

When there are multicollinearity and outlier on the data, it will not be enough to use just Robust regression or Ridge regression. Robust Ridge regression is a combined method from Robust regression and Ridge regression that can overcome the presence of multicollinearity and outlier simultaneously. Robust Ridge regression can be made by combining Ridge regression with Robust regression or combining generalized Ridge regression with Robust regression. The results of applying Robust Ridge regression on the data indicates that combining with generalized Ridge regression gives lower MSE than Ridge regression. Lower MSE means better estimator, so it can be concluded that combining using generalized Ridge is better than using Ridge regression.

References

[1] Hoerl A E and Kennard R W 1970 *Technometrics* 12 55–67.
[2] Hoerl A E, Kennard R W and Baldwin K F 1975 *Commun. Stat.* 4 105–23.
[3] Mendenhall W and Sincich 2014 *A Second Course in Statistics: Regression Analysis* 7th ed (Harlow: Pearson)
[4] Rousseeuw P J and Leroy A M 1987 *Robust Regression and Outlier Detection* (Hoboken: John Wiley & Sons Inc)
[5] Susanti Y, Pratiwi H, Sulistijowati S and Liana T 2014 *International Journal of Pure and Applied Mathematics* (*IJPAM)* 91 349–60.
[6] Harrison D and Rubinfeld D L 1978 *J. Environ. Econ. Manag.* 5 81–102.