RIDGE REGRESSION FOR HANDLING DIFFERENT LEVELS OF MULTICOLLINEARITY

By Netti Herawati

RIDGE REGRESSION FOR HANDLING DIFFERENT LEVELS OF MULTICOLLINEARITY

Herawati, N.[1], Nisa, K.[1], Azis, D.[1], and Nabila, S.U.[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Indonesia

e-mail : netti.herawati@fmipa.unila.ac.id

ABSTRACT: *Ridge regression (RR) is a method that can solves multicollinearity by adding a bias constant to a diagonal* $X'X$ *matrix. The purpose of this study is to evaluate the effectiveness of RR for handling different levels of multicollinearity compare to ordinary least square (OLS). Using simulation data with p=8; n = 25, 50, 75, 100, 200;* $\beta_0 = 0$ *and* $\beta_1 = \beta_2 = \ldots = \beta_8 = 1$ *repeated 100 times, full and partial multicollinearity among independent variables are designed. The existence of multicollinearity evaluated using VIF values. The results show that RR can solve multicollinearity at different levels and provides better estimator compare to OLS based on the value of Average Mean Square Error (AMSE).*

Keywords: Ridge regression, multicollinearity, AMSE

1. INTRODUCTION

One of basic assumptions of multiple regression model is the assumption of nonmulticollinearity among the independent variables in the model. This assumption requires that none of the independent variables in the model be correlated with any other independent variables nor with any linear combination of those independent variables. There are two types of multicollinearity. They are full/perfect/exact multicollinearity and partially/less than perfect multicollinearity. The presence of full/perfect/exact multicollinearity is when independent variables overlap completely. This conditoin can mean that no unique least squares solution to a multiple regression analysis can be computed [1]. Partially multicollinearity exist when two or more independent variables correlated with each other but still contain independent variation. Partial multicollinearity can lead to unstable estimates of the coefficients for individual independent variables. The standar error and confidence intervals for the coefficients estimates will be inflated [2]. This can affect the accuracy of model predictions and lead to errors in decision making. The Variance Inflation Factor (VIF) is one popular measure of multicollinearity, although several other diagnostics are available [3,4].

Ridge regression is one of the alternative methods to overcome the problem of multicollinearity. This method was first introduced by [5] and developed by [6]. Though this technique is based on the addition of the bias constant k to the diagonal of the $\boldsymbol{X^TX}$ matrix, it obtains more accurate regression coefficients estimation than the least squares estimator [7]. In this research, ridge regression (RR) will be applied in different levels of multicollinearity using simulation data and compare its estimates with ordinary least square (OLS) based on average mean square error values. Generalized cross validation criteria will be used to seek the magnitude of the bias constant k [8].

2. ESTIMATION METHODS IN MULTIPLE REGRESSION

Multiple linear regression analysis is an extension of simple linear regression analysis used to assess the association between two or more independent variables and a single continuous dependent variable. A population model for a multiple linear regression model that relates a y-variable to p-1 x-variables is written as

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, i=1,2,...,n; $k = 1, , \ldots, p$ where $\varepsilon_i \sim iidN(0, \sigma^2)$. $\beta's$ are parameter or regression coefficients to be estimated [9].

To estimate the regression coefficients, the classic method is OLS. This methods requires the assumptions of independency among all independent variables in the model. If the independent variables have multicollinearity, the estimates of coefficient regression may be imprecise. This methods minimizing the sum of the error squares [10].

If data consists of n observations $\{y_i, x_i\}_{i=1}^n$ and each observation i includes a scalar response y_i and a vector of p predictors (regressors) x_{ij} for $j=1,\ldots,p$, a multiple linear regression model can be written as n the matrix form the model as $\boldsymbol{Y = X\beta + \varepsilon}$ where $\boldsymbol{Y_{nx1}}$ is the dependent variable, $\boldsymbol{X_{nx2}}$ represents the independent variables, $\boldsymbol{\beta_{2x1}}$ is the regression coefficients to be estimated, $\boldsymbol{\varepsilon_{nx1}}$ represents the errors or residuals. $\boldsymbol{\hat{\beta}^{LS}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ is estimated regression coefficients using OLS by minimizing the squared distances between the observed and the predicted dependent variable [11]. To have unbiased OLS estimation of the model, some assumptions should be satisfied. Those assumptions are that the errors have an expected value of zero, that the independent variables are non-random, that the independent variables are linearly independent (nonmulticollinearity), that the disturbance are homoscedastic and not autocorrelated. If the independent variables have multicollinearity the estimates of coefficient regression may be imprecise.

Ridge Regression (RR)

Ridge regression which introduced by [5] is one methods to handle multicollinearity. Difference from OLS, ridge regression provides a biased regression coefficient estimate by modifying the least squares method to obtain variance reduction by adding a k constant in stabilizing coefficients [12].The ridge regression coefficients estimator is

$\boldsymbol{\hat{\beta}_R} = (\boldsymbol{X^TX} + k\boldsymbol{I})^{-1}\boldsymbol{X^Ty}$ where **Y**=dependent variable (n x 1), **X**= independent variable (n x p), $\boldsymbol{\hat{\beta}_R}$ = ridge coefficients (k+1) x 1, **I**= identity matrix (n x n), k= scalar. It shows that ridge regression is based on the addition of the bias constant k to the diagonal of the $X^T X$ matrix, so that the ridge coefficient estimation is influenced by the magnitude of the bias constant k, where k values are between 0 and 1[7]. To choose an appropriate value of k, a graphical method called ridge trace is suggested by [6]. The plot graph is based on the individual component value of β (k) with the sequence of k ($0 < k < 1$). The k reflects the amount of bias in the $\hat{\boldsymbol{\beta}}_R$ estimator when $k = 0$ then the $\hat{\boldsymbol{\beta}}_R$ estimator will be equal to $\hat{\boldsymbol{\beta}}_{OLS}$. If $k > 0$ the ridge estimator will be biased against the $\hat{\boldsymbol{\beta}}_{OLS}$ but tends to be more accurate than the least squares estimator.

As suggested by [8]and [9] the value of k can be obtained by using the generalized cross validation criteria method. The simplest benefit of this procedure is to select the best model and more stable estimation coefficients by minimizing visible GCV through a simple plots between generalized cross-validation and k.The GCV formula is defined as

$$GCV = \frac{SSE\ k}{\{n - [1 - trace\ \boldsymbol{H}_k]\}^2} \text{ with}$$

$$\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\,\boldsymbol{X}' \equiv \boldsymbol{H}_k \text{ and}$$

$$H_k = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + k}$$

where SSE_k =Sum Square Error of ridge regression, λ_j = eigen value j- th, k = constanta between 0 and 1, n= sample sizes.

Average Mean Square Error (AMSE)

The efficiency of the method for handling multicollinearity will be evaluated with the average of Mean Square Error (MSE) from the estimated β parameters, defined as

$$AMSE(\hat{\boldsymbol{\beta}}) = \frac{1}{m}\sum_{j=1}^{m} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}\|^2$$

where $\hat{\boldsymbol{\beta}}_j$ denotes the estimated parameter in the jth simulation. The AMSE indicates to what extent the slope and intercept are correctly estimated, therefore the aim is to obtain an AMSE value close to zero.

3. METHODS

The data used in this research is simulation data with p=8, n= 25, 50, 75, 100 and $\beta_0 = 0$; $\beta_1 = \beta_2 = \ldots = \beta_8 = 1$ with the true model $Y = X\beta + \varepsilon$. Following [13], to obtain the multicollinearity in each data set, X_p is generated using Monte Carlo's simulation using formula $X_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho z_{i(p+1)}, i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p$ where $z_{i1}, z_{i2}, \ldots, z_{i(p+1)}$ is generated normally distributed (0, 1) and $\rho = 0.99$ repeated 100 times. The multicollinearity simulation is done partially and fully in the independent variables and evaluate using VIF. Dependent variable (Y) for each p independent variable is obtained based on the model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with ε generated based on the normal distribution N (0, 1) so that Y is a linear combination of the independent variable p plus the error. Generalized cross validation criteria method is used to select the best value of k. The performance is identified by AMSE of the $\hat{\boldsymbol{\beta}}$ for RR and OLS.

4. RESULTS AND DISCUSSION.

The initial VIF values of simulated data is designed to have a high correlation ($\rho = 0.99$) between 2, 4, 6, and 8 independent variables. As a result, VIF of the corresponding variables is greater than 10 indicates the presence of multicollinearity in the variables. After applying ridge regression for partial or full multicollinearity of independent variables,the VIF drop drastically to be less than 10. It indicates that multicollinearity has been very well resolved by ridge regression. On the other hand, the OLS methods still have partial or full multicollinearity between the corresponding variables designed.

To compare the performance of OLS and RR, AMSE of both methods are calculated. The results of AMSE values for OLS and RR can be seen in Table 1 and Figure 1-4.

Table 1. AMSE of OLS and RR for n=25, 50, 75, 100, 200

Number of Multicollinearity	Method	AMSE				
		n				
		25	50	75	100	200
2 independent variables (x_1, x_2)	OLS	1.5	0.6	0.5	0.3	0.1
	RR	0.6	0.3	0.2	0.2	0.1
4 independent variables (x_1, x_2, x_3, x_4)	OLS	4.2	1.8	1.1	0.7	0.3
	RR	1.2	0.5	0.3	0.3	0.1
6 independent variables (x_1, x_2, x_3, x_4, x_5, x_6)	OLS	7.8	2.6	1.7	0.9	0.5
	RR	1.4	0.6	0.4	0.2	0.2
8 independent variables (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)	OLS	9.2	3.7	2.2	1.3	0.7
	RR	1.3	0.7	0.4	0.2	0.2

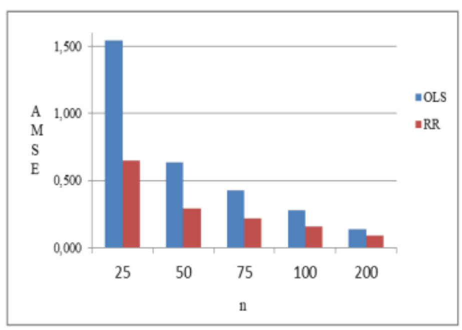

Fig.1. AMSE of OLS and RR contain multicollinearity in 2 independent variables

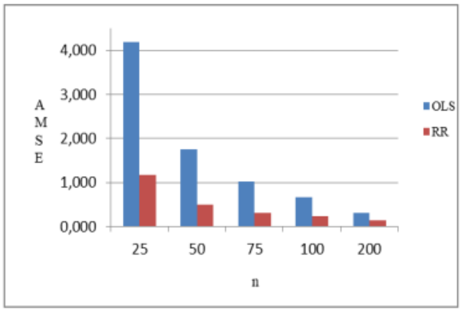

Fig.2.AMSE of OLS and RR contain multicollinearity in 4 independent variables

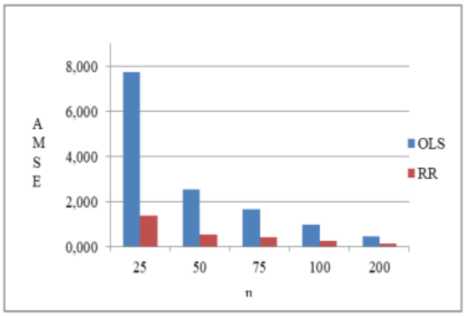

Fig.3.AMSE of OLS and RR contain multicollinearity in 6 independent variables

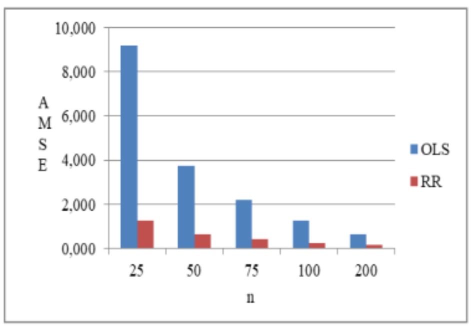

Fig.4. AMSE of OLS and RR contain multicollinearity in 8 independent variables

As seen in Table 1 and Figure 1-4, the AMSE values of the RR are smaller than OLS for n = 25, 50, 75, 100, 200. Ridge regression clearly gives better coefficients regression estimate than OLS. Moreover, the sample sizes seem to affect the value of AMSE as well. The AMSE decreases as the number of data increases. Ridge regression performs superior to OLS in higher sample sizes even when multicollinearity present. This shows that RR exceeds OLS in dealing with either partial or full multicollinearity in the multiple regression models being studied.

These results are consistent with previous studies such as studies by [14] who compare the performance of OLS, LASSO, RR and PCR, [15] who applied RR in different sample sizes and studies by [16] which showed that RR estimation method performed better than OLS in handling multicollinearity. Also studies by [17] who applied the ridge regression method to the unemployment rate in Iraq. the researchers recommended the ridge regression method rather than OLS because it provides a better estimate than OLS when independent variables are related without omitting any of the independent variables. In addition, RR was found to be a better method when the number of observations and the number of multicollinearity was considerable [18].

5. CONCLUSION

This study shows that ridge regression is a reliable method in dealing with partial or full multicollinearity between independent variables in multiple regression models. The method exceeds OLS in all cases studied. Ridge regression provides a better estimate of regression coefficients particularly in large sample size.

REFERENCES

[1] Slinker, B.K. and Glantz, S.A.,"Multiple regression for physiological data analysis: the problem of multicollinearity," *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, **249**(1):R1–R12,(1985).

[2] Belsley, D.A.,Kuh, E. and Welsch,R.E., Regression diagnostics: Identifying influential data and sources of collinearity. New York, John Wiley & Sons, (1980).

[3] Cohen, J., Cohen,P., West,S.G. and Aiken,L.S., Applied multiple regression/ correlation analysis for the behavioral sciences, 3rd ed., Mahwah, NJ, Lawrence Erlbaum Associates, (2003).

[4] Dormann, et al., C.F., "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, **36**: 27–46, (2013).

[5] Hoerl, A.E., "Application of ridge analysis to regression problems," *Chem. Eng. Prog*., **58**: 54-59, (1962).

[6] Hoerl, A.E. and Kennard, R.W.,Ridge "Regression:Biased Estimator to Nonorthogonal Problems,"*Technometrics,* **12**(1): 68-82, (1970).

[7] Dereny, M.El. and Rashwan, N.I.,"Solving Multicollinearity Problem Using Ridge Regression Models," *Int. J. Contemp. Math. Sciences*, **6**(12): 585-600, (2011).

[8] Myers,R.H., Classical and Modern Regression With Application, PWSKENT publishing Company, Boston, (1990).

[9] Montgomery, D.C. and Peck, E.A. and Vining, G.G., Introduction to Linear Regression Analysis, Wiley and Sons, Inc., New York, (2006).

[10] Hastie, T.,Tibshirani,R. and Wainwright,M., Statistical learning with Sparsity The LASSO and Generalization, Chapman and Hall/CRC Press, USA, (2015).

[11] Draper, N.R. and Smith, H., Applied Regression Analysis, 3rd edition, New York : Wiley, (1998).

[12] Mardikyan, S. and Cetin, E.,"Efficient Choice of Biasing Constant for Ridge Regression," *Int. J. Contemp. Math. Sciences*, **3**(11):527 – 536, (2008).

[13] McDonald, G.C. and Galarneau, D.I.,"AMonte Carlo Evaluation of some Ridge-type Estimators," *J. Amer. Statist. Asoc*.,**70**(350):407 – 416, (1975).

[14] Herawati, N., Nisa, K., Setiawan, E., Nusyirwan and Tiryono,"Regularized Multiple Regression Methods to Deal with Severe Multicollinearity," *International Journal of Statistics and Applications*, **8**(4): 167-172, (2018).

[15] Alibuhtto,M.C.,"Relationship between ridge regression estimator and sample size when multicollinearity present among regressors,"*World Scientific News*,**59**:12-23, (2016).

[16] Fitrianto,A. and Yik,L.C.,"Performance of Ridge Regression Estimator Methods On Small Sample Size By Varying Correlation Coefficients: A Simulation Study," *Journal of Mathematics and Statistics*, **10**(1): 25-29, (2014).

[17] Bager, A., Roman, Algedih, M., and Mohammed, B.,"Addressing multicollinearity in regression models: a ridge regression application," MPRA Paper No. 81390, posted 16 September 2017 09:04 UTC, (2017).

[18] Toka, O.,"A Comparative Study on Regression Methods in the presence of Multicollinearity, "*Journal of Statisticians: Statistics and Actuarial Sciences***2**: 47-53, (2016).

RIDGE REGRESSION FOR HANDLING DIFFERENT LEVELS OF MULTICOLLINEARITY

ORIGINALITY REPORT

15%

SIMILARITY INDEX

PRIMARY SOURCES

1	en.wikipedia.org Internet	34 words — 2%
2	medium.com Internet	30 words — 2%
3	Obubu Maxwell, C. Nwokike Chukwudike, O. Virtus Chinedu, C. Okoye Valentine, Obite Chukwudi Paul. "Comparison of Different Parametric Methods in Handling Critical Multicollinearity: Monte Carlo Simulation Study", Asian Journal of Probability and Statistics, 2019 Crossref	27 words — 1%
4	online.stat.psu.edu Internet	22 words — 1%
5	education.ucdavis.edu Internet	20 words — 1%
6	www.iaeme.com Internet	17 words — 1%
7	www.tandfonline.com Internet	10 words — 1%
8	TOKA, Onur. "A Comparative Study on Regression Methods in the presence of Multicollinearity", Aktüerya Derneği, 2016. Publications	10 words — 1%
9	digilib.unila.ac.id	

Internet

10 words — 1 %

10 mpra.ub.uni-muenchen.de
Internet
9 words — < 1%

11 ideas.repec.org
Internet
9 words — < 1%

12 studfile.net
Internet
9 words — < 1%

13 Ali O. Alnahit, Ashok.K. Mishra, Abdul A. Khan. "Quantifying climate, streamflow, and watershed control on water quality across Southeastern US watersheds", Science of The Total Environment, 2020
Crossref
9 words — < 1%

14 silo.pub
Internet
9 words — < 1%

15 calhoun.nps.edu
Internet
9 words — < 1%

16 D Suhandy, M Yulia. "Discrimination of several Indonesian specialty coffees using Fluorescence Spectroscopy combined with SIMCA method", IOP Conference Series: Materials Science and Engineering, 2018
Crossref
9 words — < 1%

17 allecottarze.it
Internet
8 words — < 1%

18 S. H. Deng, J. Zhang, F. Shen, H. Guo, Y.-w. Li, H. Xiao. "The Relationship Between Industry Structure, Household-number and Energy Consumption in China", Energy Sources, Part B: Economics, Planning, and Policy, 2013
Crossref
8 words — < 1%

19 edoc.pub
Internet

8 words — < 1%

20 pubs.asahq.org
Internet
8 words — < 1%

21 Zhe Liu, Bernard J. Jansen. "Factors influencing the response rate in social question and answering behavior", Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13, 2013
Crossref
7 words — < 1%

22 www.readbag.com
Internet
4 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		