HANDLING FULL MULTICOLLINEARITY AND VARIOUS NUMBERS OF OUTLIERS USING ROBUST RIDGE REGRESSION

[1]Setiawan, E.,[1]Herawati, N.,[1]Nisa, K.,[1]Nusyirwan,[1]Saidi, S.

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

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

ABSTRACT: *This study aims to examine ridge regression based on robust estimators S, M, MM when data contain full multicollinearity and various numbers of outliers. Simulation data with p= 10; n = 25, 50, 100;* $\beta_0 = 0$ *and* 1 *otherwise contain full multicollinearity (ρ=0.99) and various numbers of outliers (10%, 15%, 20%) was used and repeated 100 times. The existence of multicollinearity evaluated using VIF values. The empirical evidence shows that robust ridge regression based on MM-estimator (RMM) solves the problem of multicollinearity and various number of outliers very well compare to robust ridge regression based on M-estimator (RM) and S-estimator (RS). RMM provides the best estimator of regression coefficients and is more efficient because it has the smallest mean square error than RM and RS in any samples sizes and number of outliers.*

Keywords: Ridge Regression, M-Estimator, MM-Estimator, S-Estimator, MSE

1. INTRODUCTION

The existence of multicollinearity may have a negative impact, which will cause the estimated parameter variance to be greater than it should be, thus the precision of the estimate will decrease. Another consequence is the low ability to reject the null hypothesis (power of test). The cause of multicollinearity is data collection and small sample sizes or if the range of x is small [1,2]. Research on handling multicollinearity in multiple regression models has been done by many researchers using various methods. One of recommended method is ridge regression. This method has been proven outperformed to ordinary least square method in handling multicollinearity [3, 4, 5, 6, 7, 8].

In addition, the presence of outliers may also lead to other assumptions such as assumption of normality and uniformity of homogeneity. The presence of outliers may influence the estimation and results of inference tests in least square method. Least square is extremely sensitive to regression outliers, that is, observations that do not obey the linear pattern formed by the majority of the data [9]. Methods to overcome data containing outliers in a commonly used regression model are robust regression [10]. There are several types of robust regression methods. [11] introduced S-estimator that minimizes the dispersion of the residuals. [12] suggests. MM-estimation is a combination of high breakdown value estimation and efficient estimation and [13] introduced M-estimator that is nearly as efficient as OLS.

However, in the situation where multicollinearity and outliers are exist together in a data set, ridge regression or robust estimator cannot be used separately. The two methods has to be combined to handle the problems altogether. This combining methods is known as robust ridge regression estimator. Although several studies of handling multicollinearity and outliers has been done by some researchers [14, 15, 16, 17,18], the study of handling multicollinearity and various numbers of outliers which presence altogether in the multiple regression models has not been done thoroughly. Therefore, in this research will be discussed the performance of ridge regression method based on robust regression of S, M, and MM on data containing multicollinearity and various number of outliers.

2. ROBUST RIDGE REGRESSION

In the presence of multicollinearity**,** robust ridge regression methods provide an alternative to least squares regression by requiring less restrictive assumptions. This methods introduced by Hoerl (1962) attempt to dampen the influence of outlying cases in order to provide a better fit to the majority of the data. This method was introduced by [3] and developed by [4]. The estimator of the ridge regression coefficient is $\widehat{\beta}_{Ridge} = (X^TX + kI)^{-1}X^Ty$ with I = matrix identity $p\ x\ p$, k = bias constant $0 \leq k \leq 1$. **A**ugmented robust estimators as a way of combining biased and robust regression techniques is suggested by [16]. The combined procedure is based on the fact that robust estimates can be combined using a weighted least squares procedure. When, both outliers and multicollinearity occur in a data set, it seems preferable to combine methods to deal with these problems simultaneously. According to [14], robust r*idge-robust* regression is a combination of ridge regression and robust regression methods to overcome the problem of multicollinearity and outliers. The resulting ridge robust regression estimator will be stable and resistant to outliers. Parameter estimation formula ridge robust regression is $\widehat{\beta}_{RR} = (X^TX + C^*I)^{-1}X^TX\hat{\beta}_{Robust}$ with $C^* = \frac{p(\sigma^2_{Robust})}{\beta^T_{Robust}\ \beta_{Robust}}$ and p = the number of independent variables.

S-Estimator

Let ρ be a symmetric, continuously differentiable function such that ρ(0)=0 and is strictly increasing on [0,c]. Let $k = \int \rho(X) d\Phi(X)$, where Φ is the standard normal distribution. Introducedby [11], S-estimator is derived from a scale statistics corresponding to s(β). Let $(x_1, y_1), \dots, (x_n, y_n)$ be a sample of regression data with p-dimensional x_i. For each vector β, we obtain residuals $s\big(e_1(\beta), \dots, e_n(\beta)\big)$. The S-estimator $\hat{\beta}$ is defined by $\hat{\beta}_s = \min_{\beta} s\big(e_1(\beta), \dots, e_n(\beta)\big)$ with final scale estimator $\hat{\sigma}_s = s\big(e_1(\beta), \dots, e_n(\beta)\big)$. The objective function is by solving the equation of scale $\frac{1}{n}\sum_{i=1}^{n} \rho \frac{e_i}{\hat{\sigma}_s} = k$ where k is a constant defined as $E_\Phi[\rho(e)]$ and Φ is the standard normal distribution. ρ is taken Tukey's biweight function and hyperbolic tangent estimator given by

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & for\ |\mathrm{x}| \leq c \\ \dfrac{c^2}{6} & for |\mathrm{x}| > c \end{cases}$$

The parameter c is the tuning constant. This tuning parameter appears to be a very important part for asymptotic and breakdown properties of S-estimator for regression. [19] suggests c=1,548 and *k*=0,1995 for 50% breakdown and about 28% asymptotic efficiency.

M-Estimator

M-estimator was introduced by [13]. M-estimator is given by $\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho|e_i(\beta)|$. The M stands for maximum likelihood since ρ(.) is related to the likelihood function for a suitable assumed residual distribution. This estimator attemp to minimize the sum of a chosen function $\rho(e_i)$which is the residuals [1]. To obtain $\hat{\beta}_M$ we have to solve $\min_{\beta} \sum_{i=1}^{n} \rho(u_i) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right)$ with $\hat{\sigma} = \frac{MAD}{0.6745} = \frac{median|e_i - median(e_i)|}{0.6745}$. ρ function is Tukey's bisquare objective function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4} & , |u_i| \le c \\ \frac{c^2}{6} & , |u_i| > c \end{cases}$$

To minimize $\hat{\beta}_M$ is by taking partial derivatives with respect to β and setting them equal to 0, yielding $\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\hat{\sigma}}\right) = 0, j = 0,1,\dots,p$ where $\psi = \rho'$; x_{ij} is ith observation of jth independent variable and $x_{i0} = 1$. ψ function is selected with respect to the weight of assign outliers. A solution for ψ function by defining a weighted function $w(e_i) = \frac{\psi(e_i)}{e_i}$ and let $w_i = w(e_i)$. Because $u_i = \frac{e_i}{\hat{\sigma}}$, so that

$$w_i = w(u_i) = \frac{\psi(u_i)}{(u_i)} = \begin{cases} \frac{u_i\left(1 - \left(\frac{u_i}{c}\right)^2\right)^2}{u_i}, |u_i| \le c \\ 0, |u_i| > c \end{cases}$$

$$w_i = \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2 & , |u_i| \le c \\ 0 & , |u_i| > c \end{cases}$$

For Tukey‘s bisquare, take c=4,685, we get $\sum_{i=1}^{n} x_{ij} w_i \left(\frac{e_i}{\hat{\sigma}}\right) = 0, j = 0,1,\dots,p$. This equation can be solved by iteratively reweighted least squares (IRLS) method [20].

MM-Estimator

MM-estimator is a combination of high breakdown value estimation and efficient estimation which was introduced by [12]. This procedure estimates regression parameter using S-estimator which minimize the scale of the residual from M-estimator and then proceed with –estimator [21, 22]. The first stage is calculating an S-estimate with influence function $\rho(x) = 3\left(\frac{x}{c}\right)^2 - 3\left(\frac{x}{c}\right)^4 + \left(\frac{x}{c}\right)^6$ $if|x| \le c$, otherwise $\rho(x) = 1$. The value of tuning constant c=1.548. Then calculates the MM parameters which has minimum value of $\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}}\right)$ where $\rho(x)$ is the influence function used in the first stage with c=4.687 and $\hat{\sigma}$ is the estimate of scale form the first step (standard deviation of the residuals. The final step computes the MM estimate of scale as the solution to $\frac{1}{n-p} \sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{s}}\right) = 0.5$.

3. METHODS

We simulate a set of data with sample size n=25, 50, 100 contain full multicollinearity (ρ=0.99) among all independent variables (p=10) and contain various number of outliers (10%, 15%, 20%) of the data using true model $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. R package is used with 100 iterations. The independent variables are generated by $x_{ij} = (1 - \rho^2)^{1/2} u_{ij} + \rho u_{ij}$, $i = 1,2,\dots,n$ $j = 1,2,\dots,p$, where u_{ij} are independent standard normal pseudo-random numbers and ρ is specified so that the theoretical correlation between any two explanatory variables is given by ρ^2. Dependent variable ($\boldsymbol{Y}$) for each p independent variables is from $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with β parameters vectors are chosen arbitrarily (β_0=0, and β=1 otherwise) for p= 10 and ε~N (0, 1). We considered contamination proportion in data 10%, 15%, 20%. To measure the amount of multicollinearity in the data set, variance inflation factor (VIF) is examined. The behaviour of RM, RMM and RS in estimating the regression coefficient is evaluated by standard error and mean square error (MSE) of the parameter estimates.

4. RESULTS AND DISCUSSION

Independent variables are designed to have full multicollinearity. To ensure that the condition occur as designed, it is examined using VIF values of the variables. As can be seen in Table 1, the initial VIF of the variables are greater than 10, it indicates the presence of full multicollinearity among all the independent variables being studied.

Tabel 1. VIF of Independent Variables

Variables	VIF
x1	47.1756
x2	54.9222
x3	38.7089
x4	43.8664
x5	49.1229
x6	53.3555
x7	30.625
x8	46.7618
x9	44.4623
x10	38.8778

After the ridge regression is applied to the data, the VIF is reexamined to see if the multicollinearity problem is resolved. The result shows that after applying ridge regression the VIF values reduce significantly to be close to one. It indicates that multicollinearity is handled very well by ridge regression. Further, RM, RMM and RS are used to handle the the presence of outliers in the data. The behavior of the RM, RMM and RS in estimating the regression coefficients is evaluated by standard error and mean square error (MSE) of the parameter estimates. RMM, RM and RS have small standard errors of parameter estimates for each sample size and number of outliers. The smallest standard errors of parameter estimates is produced by RMM. This shows that RMM gives better coefficient regression estimator than other methods being studied.

To see the best behavior of the robust ridge regression in the study, mean square error of each method is evaluated. The result is shown in Table 2 and Figure 1-3.

Tabel 2. MSE of RMM, Rm, RS for different sample sizes and number of outliers

Number of Outliers	Method	MSE		
		n=25	n=50	n=100
10%	RMM	2158.6	2375.6	15766.1
	RM	3162.6	2406.7	15785.7
	RS	8184.7	7945.4	35726.5
15%	RMM	1836.1	2107.8	13620.5
	RM	2349.3	2136.2	13639.5
	RS	6819.8	7839.7	28889.5
20%	RMM	1689.5	2488.7	13107.3
	RM	2189.3	2516.5	13120.5
	RS	3605.2	6780.7	25587.4

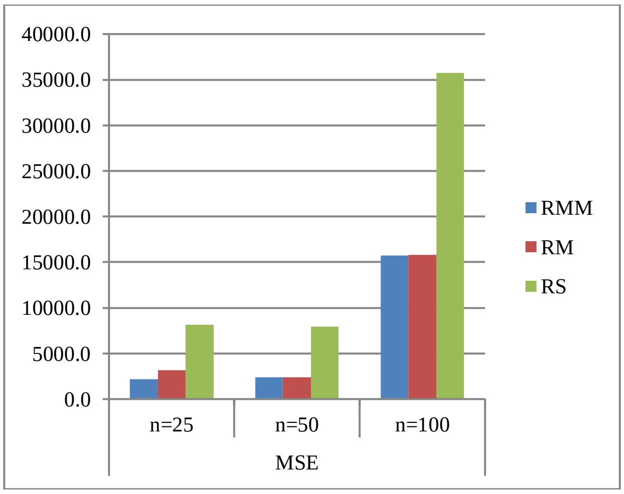

Fig.1. MSE of RMM, RM, RS for different sample sizes and 10% outliers

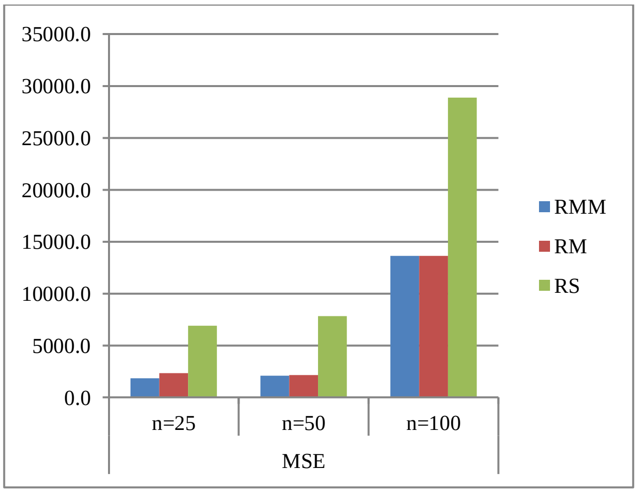

Fig.2. MSE of RMM, RM, RS for different sample sizes and 15% outliers

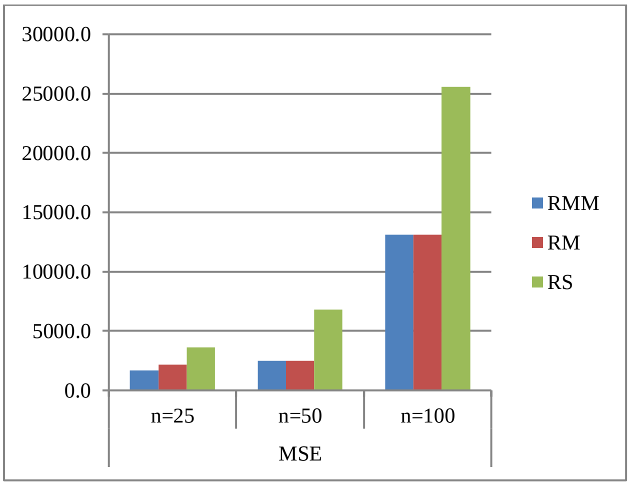

Fig.3. MSE of RMM, RM, RS for different sample sizes and 20% outliers

Table 2 and Fig. 1-3 show that RMM produces the smallest MSE value followed by RM and RS for each sample size and number of outliers. It denotes that RMM handles multicollinearity and number of outliers significantly compared to RM and RS in any number of sample sizes and number of outliers. This results are correspond to research by [18] who studied some robust ridge regression for handling multicollinearity and outliers for data the capital commodities and imported raw materials, in Iraq in the period from 1960 to 1990 shows that MSE of RMM is smaller than RM and RS.

5. CONCLUSION

Based on the results and discussion, it can be concluded that RMM is a better method in handling multicollinearity and outliers than RM and RS for small and large sample sizes.

REFERENCES

[1] Montgomery, D.C. and Peck, E.A. Introduction To Linear Regression Analysis. New York: Wiley and Sons, Inc., (2006).

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

[3] Hoerl, A. E. Application Of Ridge Analysis To Regression Problems. *Chemical Engineering Progress* **58**, 54-59, (1962).

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

[5] Dereny, M.El. and Rashwan, N.I., "Solving Multicollinearity Problem using Ridge Regression Models," *Int. J. Contemp. Math. Sciences*, **6**(12): 585600, (2011).

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

[7] 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).

[8] Herawati, N., Nisa, K., Azis, D., and Nabila, S.U., "Ridge Regression for Handling Different Levels of Multicollinearity," *Sci. Int. (Lahore)*, **30**(4): 597-600, (2018).

[9] Rousseeuw, P.J.and Hubert, M.,"Robust Statistics for Outlier Detection," *WIREs Data Mining Knowl Discov* **1**: 73–79, (2011).

[10] Chen, C., "Robust Regression ang Outlier Detection with the ROBUSTREG Procedure," *Statistics and Data Analysis.* SUGI Paper 265-27. SAS Institute, North Carolina, (2002).

[11] Rousseeuw, P.J. and Yohai, V., 'Robust Regression by Means of S estimators. In *Robust and Nonlinear Time Series Analysis.* Edited by J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics 26, New York: Springer Verlag, (1984).

[12] Yohai, V.J., "High Breakdown-point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics.* **15**:642-656, (1987).

[13] Huber, P.J., "Robust Estimation of a Location Parameter," *Ann.Math.Statist,* **35:** 73-101, (1964).

[14] Samkar, H. and Alpu, O., 'Ridge Regression Based on Some Robust Estimators,' *Journal of Modern Applied Statistical Methods,* ***9(*** 2), 495-501,(2010).

[15] Holland, P.W., "Weighted Ridge Regression: Combining Ridge and Robust Regression Methods," NBER Working Paper Series, Working Paper No.11, (1973).

[16] Askin, R.G. and D.C. Montgomery, D.C.,"Augmented robust estimators," *Technometrics*, **22**: 333-341, (1980).

[17] Midi, H. and Zahari, M., "A Simulation Study on Ridge Regression Estimators in the Presence of Outliers and Multicollinearity." J*urnal Teknologi*. **47**(C): 59-74, (2007).

[18] Lukman, A., Arowolo, O., and Ayinde, K., "Some Robust Ridge Regression for Handling Multicollinearity and Outlier," *International Journal of Sciences: Basic and Applied Research (IJSBAR),* **16**(2),192-202, (2014).

[19] Rousseeuw, P.J. and Leroy, A.M. Robust Regression and Outlier Detection, New York: Wiley-Interscience, (1987).

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

[21] Stromberg. A.J., "Computation Of High Breakdown Nonlinear Regression Parameters,*" J Am Stat Asso*c, **88**(421), 237-244, (1993).

[22] Alma, Ö.G., "Comparison Of Robust Regression Methods In Linear Regression," *Int. J. Contemp. Math. Sciences*, **6**(9), 409-421, (2011).