International Journal of Statistics and Applications 2018, 8(4): 167-172
DOI: 10.5923/j.statistics.20180804.02

Regularized Multiple Regression Methods to Deal with Severe Multicollinearity

N. Herawati*, K. Nisa, E. Setiawan, Nusyirwan, Tiryono

Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Bandar Lampung, Indonesia

Abstract This study aims to compare the performance of Ordinary Least Square (OLS), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression (RR) and Principal Component Regression (PCR) methods in handling severe multicollinearity among explanatory variables in multiple regression analysis using data simulation. In order to select the best method, a Monte Carlo experiment was carried out, it was set that the simulated data contain severe multicollinearity among all explanatory variables (ρ = 0.99) with different sample sizes (n = 25, 50, 75, 100, 200) and different levels of explanatory variables (p = 4, 6, 8, 10, 20). The performances of the four methods are compared using Average Mean Square Errors (AMSE) and Akaike Information Criterion (AIC). The result shows that PCR has the lowest AMSE among other methods. It indicates that PCR is the most accurate regression coefficients estimator in each sample size and various levels of explanatory variables studied. PCR also performs as the best estimation model since it gives the lowest AIC values compare to OLS, RR, and LASSO.

Keywords Multicollinearity, LASSO, Ridge Regression, Principal Component Regression

1. Introduction

Multicollinearity is a condition that arises in multiple regression analysis when there is a strong correlation or relationship between two or more explanatory variables. Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant, p-values, and degrade the predictability of the model [1, 2]. Since multicollinearity is a serious problem when we need to make inferences or looking for predictive models, it is very important to find a best suitable method to deal with multicollinearity [3].

There are several methods of detecting multicollinearity. Some of the common methods are by using pairwise scatter plots of the explanatory variables, looking at near-perfect relationships, examining the correlation matrix for high correlations and the variance inflation factors (VIF), using eigenvalues of the correlation matrix of the explanatory variables and checking the signs of the regression coefficients [4, 5].

Several solutions for handling multicollinearity problem have been developed depending on the sources of multicollinearity. If the multicollinearity has been created by the data collection, collect additional data over a wider X-subspace. If the choice of the linear model has increased the multicollinearity, simplify the model by using variable selection techniques. If an observation or two has induced the multicollinearity, remove those observations. When these steps are not possible, one might try ridge regression (RR) as an alternative procedure to the OLS method in regression analysis which suggested by [6].

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. By adding a degree of bias to the regression estimates, RR reduces the standard errors and obtains more accurate regression coefficients estimation than the OLS. Other techniques, such as LASSO and principal components regression (PCR), are also very common to overcome the multicollinearity. This study will explore LASSO, RR and PCR regression which performs best as a method for handling multicollinearity problem in multiple regression analysis.

2. Parameter Estimation in Multiple Regression

2.1. Ordinary Least Squares (OLS)

The multiple linear regression model and its estimation using OLS method allows to estimate the relation between a dependent variable and a set of explanatory variables. If data consists of n observations $\{y_i, x_i\}_{i=1}^{n}$ and each observation i

* Corresponding author:
netti.herawati@fmipa.unila.ac.id (N. Herawati)
Published online at http://journal.sapub.org/statistics

includes a scalar response y_i and a vector of p explanatory (regressors) x_{ij} for $j=1,...,p$, a multiple linear regression model can be written as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y}_{nx1}$ is the vector dependent variable, $\boldsymbol{X}_{nxp}$ represents the explanatory variables, $\boldsymbol{\beta}_{px1}$ is the regression coefficients to be estimated, and $\boldsymbol{\varepsilon}_{px1}$ represents the errors or residuals. $\widehat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ is estimated regression coefficients using OLS by minimizing the squared distances between the observed and the predicted dependent variable [1, 4]. To have unbiased OLS estimation in the model, some assumptions should be satisfied. Those assumptions are that the errors have an expected value of zero, that the explanatory variables are non-random, that the explanatory variables are lineary independent, that the disturbance are homoscedastic and not autocorrelated. Explanatory variables subject to multicollinearity produces imprecise estimate of regression coefficients in a multiple regression. There are some regularized methods to deal with such problems, some of them are RR, LASSO and PCR. Many studies on the three methods have been done for decades, however, investigation on RR, LASSO and PCR is still an interesting topic and attract some authors until recent years, see e.g. [7-12] for recent studies on the three methods.

2.2. Regularized Methods

a. Ridge regression (RR)

Regression coeficients $\widehat{\boldsymbol{\beta}}^{OLS}$ require X as a centered and scaled matrix, the cross product matrix ($X'X$) is nearly singular when X-columns are highly correlated. It is often the case that the matrix X'X is "close" to singular. This phenomenon is called multicollinearity. In this situation $\widehat{\boldsymbol{\beta}}^{OLS}$ still can be obtained, but it will lead to significant changes in the coefficients estimates [13]. One way to detect multicollinearity in the regression data is to use the use the variance inflation factors VIF. The formula of VIF is $(\text{VIF})_j = (VIF)_j = \frac{1}{1-R_j^2}$.

Ridge regression technique is based on adding a ridge parameter (λ) to the diagonal of X'X matrix forming a new matrix (X'X+λI). It's called ridge regression because the diagonal of ones in the correlation matrix can be described as a ridge [6]. The ridge formula to find the coefficients is $\hat{\beta}_\lambda = (\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X'Y}$, $\lambda \geq 0$. When λ =0, the ridge estimator become as the OLS. If all λ's are the same, the resulting estimators are called the ordinary ridge estimators [14, 15]. It is often convenient to rewrite ridge regression in Lagrangian form:

$$\hat{\beta}_\lambda = \underset{\beta_0,\beta}{argmin}\left\{\frac{1}{2n}\|y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}.$$

Ridge regression has the ability to overcome this multicollinearity by constraining the coefficient estimates, hence, it can reduce the estimator's variance but introduce some bias [16].

b. The LASSO

The LASSO regression estimates $\widehat{\boldsymbol{\beta}}^{OLS}$ by the optimazation problem:

$$\hat{\beta}_\lambda = \underset{\beta_0,\beta}{argmin}\left\{\frac{1}{2n}\|y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1\right\}$$

for some $\lambda \geq 0$. By Lagrangian duality, there is one-to-one correspondence between constrained problem $\|\beta\|_1 \leq t$ and the Lagrangian form. For each value of t in the range where the constraint $\|\beta\|_1 \leq t$ is active, there is a corresponding value of λ that yields the same solution form Lagrangian form. Conversely, the solution of $\hat{\beta}_\lambda$ to the problem solves the bound problem with $t = \left\|\hat{\beta}_\lambda\right\|_1$ [17, 18].

Like ridge regression, penalizing the absolute values of the coefficients introduces shrinkage towards zero. However, unlike ridge regression, some of the coefficients are shrunken all the way to zero; such solutions, with multiple values that are identically zero, are said to be sparse. The penalty thereby performs a sort of continuous variable selection.

c. Principal Component Regression (PCR)

Let $V=[V_1,...,V_p]$ be the matrix of size p x p whose columns are the normalized eigenvectors of $\boldsymbol{X'X}$, and let $\lambda_1,...,\lambda_p$ be the corresponding eigenvalues. Let $W=[W1,...,Wp]= XV$. Then $W_j = XV_j$ is the j-th sample principal components of X. The regression model can be written as $Y = \boldsymbol{X}\beta + \varsigma = \boldsymbol{XVV'}\beta + \varsigma = \boldsymbol{W}\gamma$ where $\gamma = \boldsymbol{V'}\beta$. Under this formulation, the least estimator of γ is

$$\hat{\gamma} = (\boldsymbol{W'W})^{-1}\boldsymbol{W'Y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{W'Y}.$$

And hence, the principal component estimator of β is defined by $\tilde{\beta} = \boldsymbol{V}_{\tilde{\gamma}} = \boldsymbol{V\Lambda}^{-1}\boldsymbol{W'Y}$ [19-21]. Calculation of OLS estimates via principal component regression may be numerically more stable than direct calculation [22]. Severe multicollinearity will be detected as very small eigenvalues. To rid the data of the multicollinearity, principal component omit the components associated with small eigen values.

2.3. Measurement of Performances

To evaluate the performances at the methods studied, Average Mean Square Error (AMSE) of regression coefficient $\widehat{\boldsymbol{\beta}}$ is measured. The AMSE is defined by

$$AMSE(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{l=1}^{m}\left\|\widehat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}\right\|^2$$

where $\hat{\beta}^{(l)}$ denotes the estimated parameter in the l-th simulation. AMSE value close to zero indicates that the slope and intercept are correctly estimated. In addition, Akaike Information Criterion (AIC) is also used as the performance criterion with formula: $AIC_C = 2k - 2\ln(\hat{L})$ where $\hat{L} = p(x|\hat{\theta}, M), \hat{\theta}$ are the parameter values that maximize the likelihood function, x = the observed data, n = the number of data points in x, and k = the number of parameters estimated by the model [23, 24]. The best model is indicated by the lowest values of AIC.

3. Methods

In this study, we consider the true model as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.

We simulate a set of data with sample size n= 25, 50, 75, 100, 200 contain severe multicolleniarity among all explanatory variables (ρ=0.99) using R package with 100 iterations. Following [25] the explanatory variables are generated by

$$x_{ij} = (1-\rho^2)^{1/2}u_{ij} + \rho u_{ij}, \quad i = 1,2,\dots,n \quad j = 1,2,\dots,p.$$

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 explanatory 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= 4, 6, 8, 10, 20 and ε~N (0, 1). To measure the amount of multicolleniarity in the data set, variance inflation factor (VIF) is examined. The performances of OLS, LASSO, RR, and PCR methods are compared based on the value of AMSE and AIC. Cross-validation is used to find a value for the λ value for RR and LASSO.

4. Results and Discussion

The existence of severe multicollinearity in explanatory variables for all given cases are examined by VIF values. The result of the analysis to simulated dataset with p = 4, 6, 8, 10, 20 with n = 25, 50, 75, 100, 200 gives the VIF values among all the explanatory variables are between 40-110. This indicates that severe multicollinearity among all explanatory variables is present in the simulated data generated from the specified model and that all the regression coefficients appear to be affected by collinearity. LASSO method is for choosing which covariates to include in the model. It is based on stepwise selection procedure. In this study, LASSO, cannot overcome severe multicollinearity among all explanatory variables since it can reduce the VIF in data set a little bit. Whereas in every cases of simulated data set studied, RR reduces the VIF values less than 10 and PCR reduce the VIF to 1. Using this data, we compute different estimation methods alternate to OLS. The experiment is repeated 100 times to get an accurate estimation and AMSE of the estimators are observed. The result of the simulations can be seen in Table 1.

In order to compare the four methods easily, the AMSE results in Table 1 are presented as graphs in Figure 1 - Figure 5. From those figures, it is seen that OLS has the highest AMSE value compared to the other three methods in every cases being studied followed by LASSO. Both OLS and LASSO are not able to resolve the severe multicollinearity problems. On the other hand, RR gives lower AMSE than OLS and LASSO but still high as compare to that in PCR. Ridge regression and PCR seem to improve prediction accuracy by shrinking large regression coefficients in order to reduce over fitting. The lowest AMSE is given by PCR in every case.

It clearly indicates that PCR is the most accurate estimator when severe multcollinearity presence. The result also show that sample size affects the value of AMSEs. The higher the sample size used, the lower the value of AMSE from each estimators. Number of explanatory variables does not seem to affect the accuracy of PCR.

Table 1. Average Mean Square Error of OLS, LASSO, RR, and PCR

p	*n*	AMSE			
		OLS	**LASSO**	**RR**	**PCR**
4	25	5.7238	3.2880	0.5484	0.0169
	50	3.2870	2.5210	0.3158	0.0035
	75	2.3645	2.0913	0.2630	0.0029
	100	1.7750	1.6150	0.2211	0.0017
	200	0.8488	0.8438	0.1512	0.0009
6	25	15.3381	6.5222	0.5235	0.0078
	50	5.3632	4.0902	0.4466	0.0051
	75	4.0399	3.4828	0.3431	0.0031
	100	2.8200	2.5032	0.2939	0.0020
	200	1.3882	1.3848	0.2044	0.0013
8	25	20.4787	8.7469	0.5395	0.0057
	50	8.2556	5.9925	0.4021	0.0037
	75	5.6282	4.7016	0.3923	0.0018
	100	3.8343	3.4771	0.3527	0.0017
	200	1.9906	1.9409	0.2356	0.0008
10	25	27.9236	12.3202	1.2100	0.0119
	50	12.1224	7.8290	0.5129	0.0089
	75	7.0177	5.8507	0.4293	0.0035
	100	4.7402	4.3165	0.3263	0.0022
	200	2.5177	2.4565	0.2655	0.0013
20	25	396.6900	33.6787	1.0773	0.0232
	50	33.8890	16.4445	0.7861	0.0065
	75	18.5859	13.1750	0.6927	0.0052
	100	12.1559	9.7563	0.5670	0.0032
	200	5.5153	5.2229	0.4099	0.0014

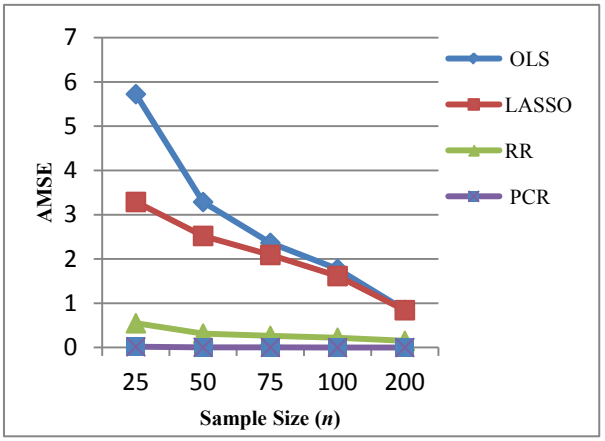

Figure 1. AMSE of OLS, LASSO, RR and PCR for p=4

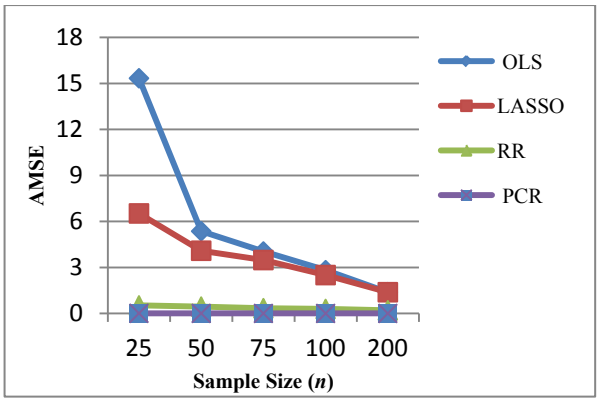

Figure 2. AMSE of OLS, LASSO, RR and PCR for p=6

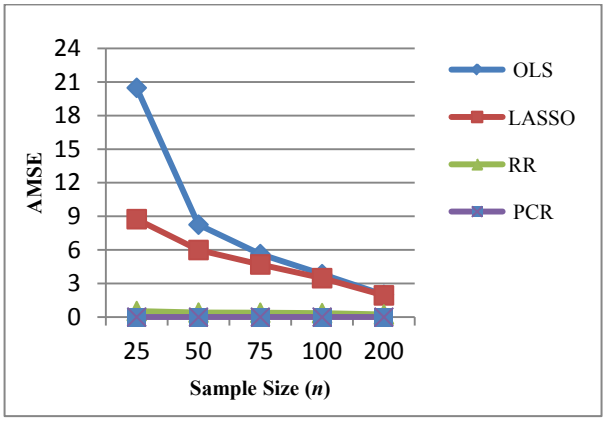

Figure 3. AMSE of OLS, LASSO, RR and PCR for p=8

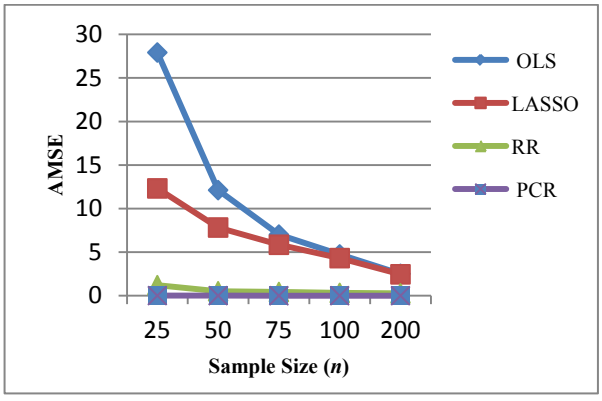

Figure 4. AMSE of OLS, LASSO, RR and PCR for p=10

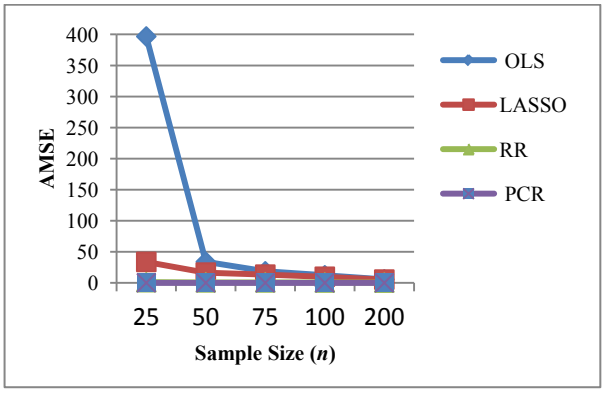

Figure 5. AMSE of OLS, LASSO, RR and PCR for p=20

To choose the best model, we use Akaike Information Criterion (AIC) of the models obtained using the four methods being studied. The AIC values for all methods with different number of explanatory variables and sample sizes is presented in Table 2 and displayed as bars-graphs in Figure 6 – Figure 10.

Figure 6 –Figure 10 show that the greater the sample sizes are the lower the values of AIC and in contrary to sample sizes, number of explanatory variables does not seem to affect the value of AIC. OLS has the highest AIC values in every level of explanatory variables and sample sizes. LASSO as one of the regularized method has the highest AIC values compare to RR and PCR. The differences of AIC values between the PCR performances from RR are small. PCR is the best methods among the selected methods including based on the value of AIC. It is consistent with the result in Table 1 where PCR has the smallest AMSE value among all the methods applied in the study. PCR is approximately effective and efficient for a small and high number of regressors. This finding is in accordance with previous study [20].

Table 2. AIC values for OLS, RR, LASSO, and PCR with different number of explanatory variables and sample sizes

p	Methods	n				
		25	50	75	100	200
4	OLS	8.4889	8.2364	8.2069	8.1113	8.0590
	LASSO	8.4640	8.2320	8.2056	8.1108	8.0589
	RR	8.3581	8.1712	8.1609	8.0774	8.0439
	PCR	8.2854	8.1223	8.1173	8.0439	8.0239
6	OLS	8.7393	8.3541	8.2842	8.1457	8.0862
	LASSO	8.6640	8.3449	8.2806	8.1443	8.0861
	RR	8.4434	8.2333	8.1995	8.0868	8.0598
	PCR	8.3257	8.1521	8.1327	8.0355	8.0281
8	OLS	8.8324	8.3983	8.3323	8.2125	8.1060
	LASSO	8.7181	8.3816	8.3259	8.2104	8.1058
	RR	8.3931	8.2039	8.2062	8.1247	8.0660
	PCR	8.2488	8.1069	8.1162	8.0550	8.0254
10	OLS	9.0677	8.4906	8.3794	8.2595	8.1142
	LASSO	8.9011	8.4556	8.3711	8.2570	8.1140
	RR	8.4971	8.2275	8.2120	8.1446	8.0608
	PCR	8.2405	8.0969	8.1035	8.0674	8.0104
20	OLS	11.3154	9.1698	8.7443	8.5138	8.2652
	LASSO	9.8490	9.0324	8.7055	8.4968	8.2638
	RR	8.5775	8.4475	8.3195	8.2202	8.1390
	PCR	8.2628	8.2138	8.1375	8.0759	8.0535

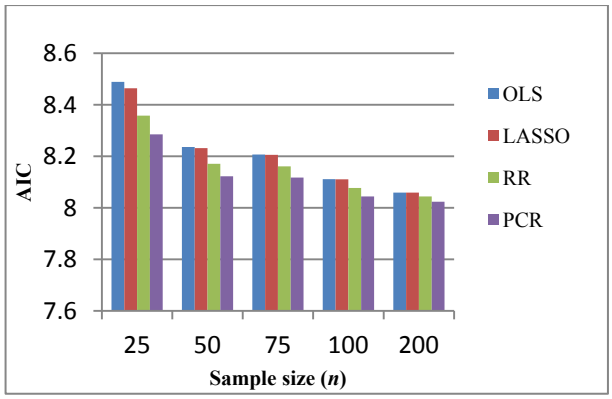

Figure 6. Bar-graph of AIC for p=4

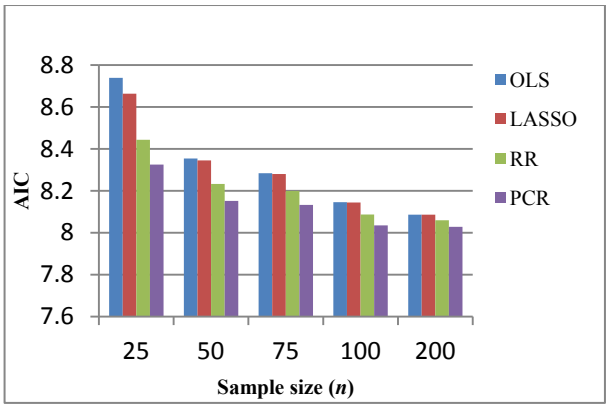

Figure 7. Bar-graph of AIC for p=6

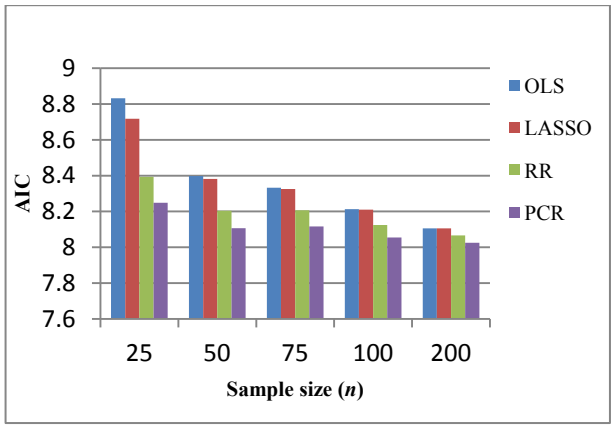

Figure 8. Bar-graph of AIC for p=8

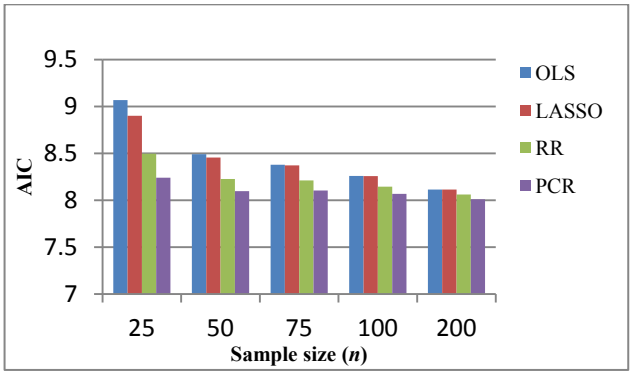

Figure 9. Bar-graph of AIC for p=10

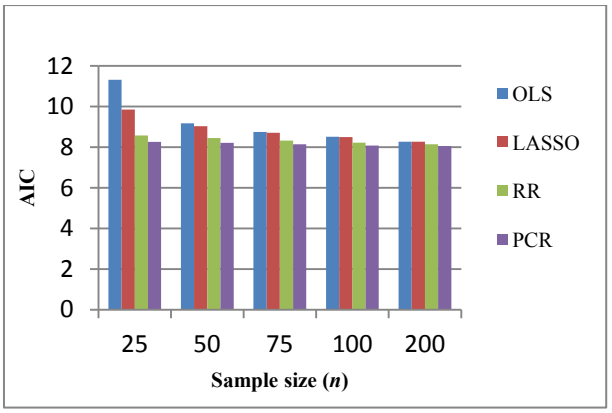

Figure 10. Bar-graph of AIC for p=20

5. Conclusions

Based on the simulation results at p = 4, 6, 8, 10, and 20 and the number of data n = 25, 50, 75, 100 and 200 containing severe multicollinearity among all explanatory variables, it can be concluded that RR and PCR method are capable of overcoming severe multicollinearity problem. In contrary, the LASSO method does not resolve the problem very well when all variables are severely correlated even though LASSO do better than OLS. In Overall PCR performs best to estimate the regression coefficients on data containing severe multicolinearity among all explanatory variables.

REFERENCES

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

[2] Gujarati, D. Basic Econometrics. 4th ed. New York: McGraw−Hill, 1995.

[3] Judge, G.G., Introduction to Theory and Practice of Econometrics. *New York:* John Willy and Sons, 1988.

[4] Montgomery, D.C. and Peck, E.A., Introduction to Linear Regression Analysis. New York: John Willy and Sons, 1992.

[5] Kutner, M.H. et al., Applied Linear Statistical Models. 5th Edition. New York: McGraw-Hill, 2005.

[6] Hoerl, A.E. and Kennard, R.W., 2000, Ridge Regression: Biased Estimation for nonorthogonal Problems. Technometrics, 42, 80-86.

[7] Melkumovaa, L.E. and Shatskikh, S.Ya. 2017. Comparing Ridge and LASSO estimators for data analysis. Procedia Engineering, 201, 746-755.

[8] Boulesteix, A-L., R. De Bin, X. Jiang and M. Fuchs. 2017. IPF-LASSO: Integrative-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. Computational and Mathematical Methods in Medicine, 2017, 14 p.

[9] Helton, K.H. and N.L. Hjort. 2018. Fridge: Focused fine-tuning of ridge regression for personalized predictions. Statistical Medicine, 37(8), 1290-1303.

[10] Abdel Bary, M.N. 2017. Robust Regression Diagnostic for Detecting and Solving Multicollinearity and Outlier Problems: Applied Study by Using Financial Data Applied Mathematical Sciences, 11 (13), 601-622.

[11] Usman, U., D. Y. Zakari, S. Suleman and F. Manu. 2017. A Comparison Analysis of Shrinkage Regression Methods of Handling Multicollinearity Problems Based on Lognormal and Exponential Distributions. MAYFEB Journal of Mathematics, 3, 45-52.

[12] Slawski, M. 2017. On Principal Components Regression, Random Projections, and Column Subsampling. Arxiv: 1709.08104v2 [Math-ST].

[13] Wethrill, H., 1986, Evaluation of ordinary Ridge Regression. Bulletin of Mathematical Statistics, 18, 1-35.

[14] Hoerl, A.E., 1962, Application of ridge analysis to regression problems. Chem. Eng. Prog., 58, 54-59.

[15] Hoerl, A.E., R.W. Kannard and K.F. Baldwin, 1975, Ridge regression: Some simulations. Commun. Stat., 4, 105-123.

[16] James, G., Witten D., Hastie T., Tibshirani R An Introduction to Statistical Learning: With Applications in R. New York: Springer Publishing Company, Inc., 2013.

[17] Tibshirani, R., 1996, Regression shrinkage and selection *via* the LASSO. J Royal Stat Soc, 58, 267-288.

[18] Hastie, T., Tibshirani, R., Mainwright, M., 2015, Statistical learning with Sparsity The LASSO and Generalization. USA: Chapman and Hall/CRC Press.

[19] Coxe, K.L., 1984, "Multicollinearity, principal component regression and selection rules for these components," ASA Proceed. Bus fj Econ sect'ion, 222-227.

[20] Jackson, J.E., A User's Guide To Principal Components. New York: Tiley, 1991.

[21] Jolliffe, LT, Principal Component Analysis. *New York*: Springer-Verlag, 2002.

[22] Flury, B. and Riedwyl, H., Multivariate Statistics. A Practical Approach, London: Chapman and Hall, 1988.

[23] Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B.N. Petrow and F. Csaki (eds), Second International symposium on information theory (pp.267-281). Budapest: Academiai Kiado.

[24] Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

[25] McDonald G.C. and Galarneau, D.I., 1975, A Monte Carlo evaluation of some ridge type estimators. J. Amer. Statist. Assoc., 20, 407-416.

[26] Zhang, M., Zhu, J., Djurdjanovic, D. and Ni, J. 2006, A comparative Study on the Classification of Engineering Surfaces with Dimension Reduction and Coefficient Shrinkage Methods. Journal of Manufacturing Systems, 25(3): 209-220.