

Comparison of Different Parametric Methods in Handling Critical Multicollinearity: Monte Carlo Simulation Study.

Abstract

In regression analysis, it is relatively necessary to have a correlation between the response and explanatory variables, but having correlations amongst explanatory variables is something undesired. This paper focuses on five methodologies for handling critical multicollinearity, they include: Partial Least Square Regression (PLSR), Ridge Regression (RR), Ordinary Least Square Regression (OLS), Least Absolute Shrinkage and Selector Operator (LASSO) Regression, and the Principal Component Analysis (PCA). Monte Carlo Simulations comparing the methods was carried out with the sample size greater than or equal to the levels ($n > p$) considered in most cases, the Average Mean Square Error (AMSE) and Akaike Information Criterion (AIC) values were computed. The result shows that PCR is the most superior and more efficient in handling critical multicollinearity problems, having the lowest AMSE and AIC values for all the sample sizes and different levels considered.

Keywords: Multicollinearity, Least Absolute Shrinkage and Selection Operator, Partial Least Square Regression, Akaike Information Criterion, Average Mean Square Error, Principal Component Analysis, Ordinary Least Square Regression, Ridge Regression.

1. Introduction

In regression analysis there are many assumptions about the model, namely, multicollinearity, non-consistent variance (non-homogeneity), linearity, and autocorrelation. If one or more assumption is violated, then the model in hand is no more reliable and also is not acceptable in estimating the population parameters [1]. Multicollinearity (or Collinearity) is a statistical phenomenon in multiple linear regression analysis where two (or more) independent or predictor variables are highly correlated with each other, or intercorrelated. Presence of multicollinearity violates one of the core assumptions of multiple linear regression analysis and as such it is problematic; the predicted regression coefficients are not reliable anymore [2]. This problem 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, non-significant, p-values, and degrade the predictability of the model [3]. There are a variety of informal and formal methods that have been developed for

detecting the presence of serious multicollinearity. Begin by studying pairwise scatter plots of pairs of independent variables, looking for near-perfect relationships [4]. Then glance at the correlation matrix for high correlations. Unfortunately, multicollinearity does not always show up when considering the variables two at a time. Thus, we consider the variance inflation factors (VIF), which measures how much the variances of the estimated regression coefficients are inflated compared to when the independent variables are not linearly related. VIFs over 10 indicate collinear variables. Also, Eigenvalues of the correlation matrix of the independent variables near zero indicate multicollinearity. Instead of looking at the numerical size of the eigenvalue, use the condition number. Large condition numbers indicate multicollinearity [5]. Investigating the signs of the regression coefficients could as well help in detecting the presence of multicollinearity as variables whose regression coefficients are opposite in sign from what you would expect may indicate multicollinearity. Correcting multicollinearity would depend on what the source of multicollinearity is, as the solutions will vary. 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. Above all, use care in selecting the variables at the outset. When these steps are not possible, you might try Ridge Regression or any other suitable approaches such as Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), and Ordinary Least Squares Regression (OLS) [6] etc. This paper looks at five different regression methods and explores which performs best as a method for handling multicollinearity problem using simulated data sets.

2. Materials and Method

Various methods have been developed to cope with multicollinearity problems. Among such methods are Ridge Regression, Principal Component Regression, Partial Least Squares Regression, Least Absolute Shrinkage and Ordinary Least Square.

In this study, we consider the true model as;

$$Y = X\beta + \varepsilon.$$

We simulate a set of data with sample size $n = 60, 100, 150, 200, 400, 1000$ contain severe multicollinearity among all explanatory variables ($\rho = 0.99$) using R package with 100 iterations. Following the explanatory variables are generated by

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

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 (Y) for each p explanatory variables is from $Y = X\beta + \varepsilon$ with β parameters vectors are chosen arbitrarily ($\beta_0 = 0$, and $\beta = 1$ otherwise) for $p = 2, 4, 6, 10, 20, 50, 100$ and $\varepsilon \sim N(0, 1)$. To measure the amount of multicollinearity in the data set, variance inflation factor (VIF) is examined. The performances of OLS, LASSO, Ridge Regression (RR), PLSR and PCR methods are compared based on the value of AMSE and AIC values. Cross-validation is used to find a value for the λ value for RR and LASSO.

2.1 Ridge Regression

Ridge Regression is developed by Hoerl and Kennard and this method is the modification of the least squares method that allows biased estimators of the regression coefficients [7]. Although the estimators are biased, the biases are small enough for these estimators to be substantially more precise than unbiased estimators. Therefore, these biased estimators are preferred over unbiased ones since they will have a larger probability of being close to the true parameter values. The ridge regression estimator of the coefficient β is found by solving for b_R in the equation [8]

$$(X'X + \delta I)b_R = X'Y$$

where $\delta \geq 0$ is often referred to as a shrinkage parameter. Thus, the solution for ridge estimator is given by

$$b_R = (X'X + \delta I)^{-1}X'Y$$

The matrix (X^*X^*) considered is replaced by $(X^*X^* \delta I)$, where δ is a small positive quantity. Since the V matrix diagonalizes (X^*X^*) , it also diagonalizes $(X^*X^* \delta I)$. Thus,

$$V'(X^*X^* + \delta I)V = \begin{bmatrix} \lambda_1 + \delta & 0 & 0 \\ 0 & \lambda_2 + \delta & 0 \\ 0 & 0 & \lambda_k + \delta \end{bmatrix}$$

The eigenvalues of the new matrix $(X^*X^* \delta I)$ are $\lambda_i + \delta$ for $i = 1, 2, \dots, k$ where adding δ to the main diagonal effectively replaces λ_1 by $\lambda_1 + \delta$. From the properties of the ridge estimator, the role of δ are revealed in moderating the variance of the estimators. The impact of eigenvalues on the variances of the ridge regression coefficients can be illustrated as

$$\sum_i \frac{Var(b_{i,R})}{\sigma^2} = \sum_i \frac{\lambda_i}{(\lambda_i + \delta)^2}$$

Therefore, the δ in ridge regression moderates the damaging impact of the small eigenvalues that result from Collinearity. There are various procedures for choosing the shrinkage parameter δ . The ridge

trace is a very pragmatic procedure for choosing the shrinkage parameter where it allows δ increasing until stability is indicated in all coefficients. A plot of the coefficients against δ that pictorially displays the trace often helps the analyst to make a decision regarding the appropriate value of δ . However, stability does not imply that the regression coefficients have converged. As δ grows, variances reduce and the coefficients become more stable[9]. Therefore, the value of δ is chosen at the point for which the coefficients no longer change rapidly.

The C_p -like statistic that is based on the same variance-bias type trade-off is one of the proposed procedures. The use of C_δ statistic is by a simple plotting of C_δ against δ , with the use of δ -value for which C_δ is minimized. The statistic is given as [10]

$$C_\delta = \frac{SS_{Re\ S,\delta}}{\hat{\sigma}^2} - n + 2 + 2tr[H_\delta]$$

where $H_\delta = X^*(X^{*'}X^* \delta I) - X^{*'}$, $SS_{Re\ S,\delta}$ is the residual sum of squares using ridge regression and $tr[H_\delta]$ is the trace of H_δ . Notice that H_δ plays the same role as the HAT matrix in ordinary least squares. In ordinary least squares, residuals are helpful in identifying outliers which do not appear to be consistent with the rest of the data while the HAT matrix is used to identify “high leverage” points which are outliers among the independent variables. The HAT matrix H is given by $= X(X'X)^{-1} X'$. The trace is $tr\{H\} = p$, where p is the m vector of adjustable model parameters to be estimated from the available data set and for all diagonal elements, $0 < h_i < 1$. The statistic $\hat{\sigma}^2$ comes from the residual mean square from ordinary least squares estimation. The other criterion that represents a prediction approach is the generalized cross validation (GCV) that is given by [11]

$$\begin{aligned} GCV &= \sum_{i=1}^n \frac{e_{i,\delta}^2}{\{n - [1 + tr(H_\delta)]\}^2} \\ &= \frac{SS_{Re\ S,k}}{\{n - [1 + tr(H_\delta)]\}^2} \end{aligned}$$

where the value of 1 in $1 + tr(H_\delta)$ accounts for the fact that the role of the constant is not involved in H_δ . The use of this procedure is to choose δ so as to minimize GCV by a simple plotting of GCV against δ .

2.2 LASSO Regression

Lasso was originally introduced in the context of least squares, and it can be instructive to consider this case first, since it illustrates many of lasso’s properties in a straightforward setting.

Consider a sample consisting of N cases, each of which consists of p covariates and a single outcome. Let y_i be the outcome and $x_i := (x_1, x_2, \dots, x_p)^T$ be the covariate vector for the i^{th} case. Then the objective of lasso is to solve [12]

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} (y - \beta_0 - x_i^T \beta)^2 \right\}$$

$$\text{Subject to } \sum_{j=1}^p |\beta_j| \leq t.$$

Here t is a prespecified free parameter that determines the amount of regularisation. Letting X be the covariate matrix, so that $X_{ij} = (x_i)_j$ and x_i^T is the i^{th} row of X , the expression can be written more compactly as

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} \|y - \beta_0 \mathbf{1}_N - X\beta\|_2^2 \right\}$$

$$\text{Subject to } \|\beta\|_1 \leq t.$$

Where $\|\beta\|_p = (\sum_{i=1}^N |\beta_i|^p)^{1/p}$ is the standard ℓ^p norm, and $\mathbf{1}_N$ is an $N \times 1$ vector of ones.

Denoting the scalar mean of the data points x_i by \bar{x} and the mean of the response variables y_i by \bar{y} , the resulting estimate for β_0 will end up being $\hat{\beta}_0 = \bar{y} - \bar{x}^T \beta$, so that [13]

$$y_i - \hat{\beta}_0 - x_i^T \beta = y_i - (\bar{y} - \bar{x}^T \beta) - x_i^T \beta = (y_i - \bar{y})(x_i - \bar{x})^T \beta$$

and therefore it is standard to work with variables that have been centered (made zero-mean). Additionally, the covariates are typically standardized ($\sum_{i=1}^N x_i^2 = 1$) so that the solution does not depend on the measurement scale.

It can be helpful to rewrite [14]

$$\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\}$$

$$\text{Subject to } \|\beta\|_1 \leq t.$$

in the so-called Lagrangian form

$$\min_{\beta \in \mathcal{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where the exact relationship between t and λ is data dependent.

2.3 Principal Component Analysis

Let $V = [V_1, \dots, V_p]$ be the matrix of size $p \times p$ whose columns are the normalized eigenvectors of $X'X$, and let $\lambda_1, \dots, \lambda_p$ be the corresponding eigenvalues. Let $W = [W_1, \dots, W_p] = XV$. Then $W_j = XV_j$ is the j -th sample principal components of X . The regression model can be written as $Y = X\beta + \zeta = XVV'\beta + \zeta = W\gamma$ where $\gamma = V'\beta$. Under this formulation, the least estimator of γ is [15]

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

And hence, the principal component estimator of β is defined by

$$\hat{\beta} = V\hat{\gamma} = V\Lambda^{-1} W'Y.$$

Calculation of OLS estimates via principal component regression may be numerically more stable than direct calculation. Critical 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.

2.4 Ordinary Least Square Regression

Ordinary least-squares (OLS) regression is a generalized linear modelling technique that may be used to model a single response variable which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been appropriately coded [16]. The OLS regression model can be extended to include multiple explanatory variables by simply adding additional variables to the equation. The form of the model is the same as above with a single response variable (Y), but this time Y is predicted by multiple explanatory variables (X_1 to X_3).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The interpretation of the parameters (α and β) from the above model is basically the same as for the simple regression model above, but the relationship cannot now be graphed on a single scatter plot. α indicates the value of Y when all vales of the explanatory variables are zero. Each β parameter indicates the average change in Y that is associated with a unit change in X , whilst controlling for the other explanatory variables in the model. Model-fit can be assessed through comparing deviance measures of nested models. For example, the effect of variable X_3 on Y in the model above can be calculated by comparing the nested models [17]

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

The change in deviance between these models indicates the effect that X_3 has on the prediction of Y when the effects of X_1 and X_2 have been accounted for (it is, therefore, the unique effect that X_3 has on Y after taking into account X_1 and X_2). The overall effect of all three explanatory variables on Y can be assessed by comparing the models

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Y = \alpha.$$

The significance of the change in the deviance scores can be assessed through the calculation of the F statistic using the equation provided above (these are, however, provided as a matter of course by most software packages). As with the simple OLS regression, it is a simple matter to compute the R -square statistics.

2.5 Partial Least Square Regression

Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data. PLS regression is especially useful when your predictors are highly collinear, or when you have more predictors than observations and ordinary least-squares regression either produces coefficients with high standard errors or fails completely [18]. PLS does not assume that the predictors are fixed, unlike multiple regression. This means that the predictors can be measured with error, making PLS more robust to measurement uncertainty.

PLS regression is primarily used in the chemical, drug, food, and plastic industries. A common application is to model the relationship between spectral measurements (NIR, IR, UV), which include many variables that are often correlated with each other, and chemical composition or other physio-chemical properties. In PLS regression, the emphasis is on developing predictive models. Therefore, it is not usually used to screen out variables that are not useful in explaining the response.

Unlike least squares regression, PLS can fit multiple response variables in a single model. PLS regression fits multiple response variables in a single model. Because PLS regression models the response variables in a multivariate way, the results can differ significantly from those calculated for the response variables individually. You should model multiple responses separately only if the responses are uncorrelated.

As in multiple linear regression, the main purpose of partial least squares regression is to build a linear model,

$$Y = XB + e,$$

where Y is an n cases by m variables response matrix, X is an n cases by p variables predictor matrix, B is p by m regression coefficient matrix, and E is a noise term for the model which has the same dimensions as Y . Usually, the variables in X and Y are centered by subtracting their means and scaled by dividing by their standard deviations.

Both principal components regression and partial least squares regression produce factor scores as linear combinations of the original predictor variables, so that there is no correlation between the factor score variables used in the predictive regression model. For example, suppose we have a data set with response variables Y (in matrix form) and a large number of predictor variables X (in matrix form), some of which are highly correlated. A regression using factor extraction for this type of data computes the factor score matrix $T=XW$ for an appropriate weight matrix W , and then considers the linear regression model $Y=TQ+E$, where Q is a matrix of regression coefficients (loadings) for T , and E is an error (noise) term. Once the loadings Q are computed, the above regression model is equivalent to $Y=XB+E$, where $B=WQ$, which can be used as a predictive regression model [19].

Principal components regression and partial least squares regression differ in the methods used in extracting factor scores. In short, principal components regression produces the weight matrix W reflecting the covariance structure between the predictor variables, while partial least squares regression produces the weight matrix W reflecting the covariance structure between the predictor and response variables.

For establishing the model, partial least squares regression produces a p by c weight matrix W for X such that $T=XW$, i.e., the columns of W are weight vectors for the X columns producing the corresponding n by c factor score matrix T . These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of Y on T are then performed to produce Q , the loadings for Y (or weights for Y) such that $Y=TQ+E$. Once Q is computed, we have $Y=XB+E$, where $B=WQ$, and the prediction model is complete.

One additional matrix necessary for a complete description of partial least squares regression procedures is the p by c factor loading matrix P which gives a factor model $X=TP+F$, where F is the unexplained part of the X scores [20]. We now can describe the algorithms for computing partial least squares regression.

3. Formal Method of Detecting Multicollinearity

3.1 Tolerance and Variance Inflation Factor (VIF)

Suppose we have a regression model with $p - 1$ regressors and an intercept, then the variance of the j th partial regression coefficient is given by [21]

$$\text{Var}(\beta_j) = \frac{\sigma^1}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right)$$

$$\text{Var}(\beta_j) = \frac{\sigma^1}{\sum x_j^2} \text{VIF}_j$$

Where β_j is the (partial) regression coefficient of the regressor X_j , R_j^2 is the R^2 in the (auxilliary) regression of X_j on the remaining $(p - 2)$ predictors. Rule of thumb:

The criterion for using VIF_i as a detection method is that the higher the VIF more certain we are that multicollinearity is present [22-24]. But how high VIF_i will be considered high? We can use the rule of thumb that if VIF_i is greater than 10 that variable is considered highly collinear [25]. Some statisticians also use the Tolerance to detect multicollinearity where

$$\text{TOL}_j = (1 - R_j^2) = \left(\frac{1}{\text{VIF}_i} \right)$$

3.2 Comparing the Performance

To evaluate the performances at the methods studied, Average Mean Square Error (AMSE) of regression coefficient $\hat{\beta}$ is measured. The AMSE is defined by [26]

$$AMSE(\hat{\beta}) = \frac{1}{n} \sum_{l=1}^m \|\hat{\beta}^{(l)} - \beta\|^2$$

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 [27]:

$$AIC_C = 2k - 2 \ln(\hat{L})$$

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. The best model is indicated by the lowest values of AIC and AMSE.

Table 1: Factors and Level of the Simulated Dataset

Factors		
Number of Regressors (P)	of	2,4,6,10,20,50,100
Number of Observations in the dataset (n)	of	60,100,150,200,400,1000

4. Results and Discussion

Table 2: Average Mean Square Error of PLSR, RR, OLS, LASSO, and PCR

P	n	AMSE				
		PLSR	RR	OLS	LASSO	PCR
2	60	13.551	4.564	19.285	14.301	4.304
	100	13.052	4.058	19.017	11.987	4.008
	150	11.995	3.958	16.987	10.784	3.096

	200	9.795	2.147	14.218	9.741	2.107
	400	7.115	1.998	11.914	9.876	1.198
	1000	5.310	0.915	9.995	6.124	0.987
4	60	33.501	9.764	49.178	44.695	3.904
	100	31.059	8.098	45.709	41.784	2.608
	150	29.989	7.059	36.749	36.652	2.096
	200	25.825	6.487	29.098	26.741	1.975
	400	17.191	5.908	24.778	19.876	1.462
	1000	16.880	4.919	19.691	16.124	1.228
6	60	53.158	13.015	75.985	60.185	2.775
	100	51.852	10.495	72.745	58.254	1.658
	150	47.258	8.158	69.087	52.054	1.405
	200	44.895	7.958	65.954	47.951	1.207
	400	42.105	6.584	62.015	44.055	1.409
	1000	40.052	5.915	58.114	41.360	1.202
10	60	72.154	15.047	97.412	80.119	1.689
	100	61.518	14.020	92.415	77.020	1.299
	150	60.224	13.250	90.325	70.505	1.104
	200	50.256	12.132	87.451	65.512	0.925
	400	48.225	9.080	82.150	62.125	0.760
	1000	47.112	7.012	77.047	58.770	0.481
20	60	118.562	29.456	141.051	129.805	1.495
	100	107.521	28.009	138.705	119.852	1.091
	150	97.025	22.097	130.598	103.451	0.678
	200	92.097	17.963	122.215	97.856	0.475
	400	87.059	14.852	109.546	92.542	0.155
	1000	79.752	11.100	107.025	87.054	0.090
50	60	149.523	57.592	235.215	201.526	0.940
	100	147.001	48.526	231.954	198.751	0.759
	150	134.971	43.189	221.325	178.025	0.625

	200	126.594	35.054	211.259	154.978	0.452
	400	123.045	32.029	209.879	148.741	0.352
	1000	111.019	29.898	204.065	135.099	0.071
100	60	231.012	78.597	325.251	313.215	0.159
	100	217.942	69.457	321.985	303.124	0.090
	150	209.347	52.976	312.256	287.973	0.045
	200	198.245	48.237	307.152	269.125	0.033
	400	175.054	44.562	303.125	245.781	0.015
	1000	158.256	41.056	299.015	231.251	0.009

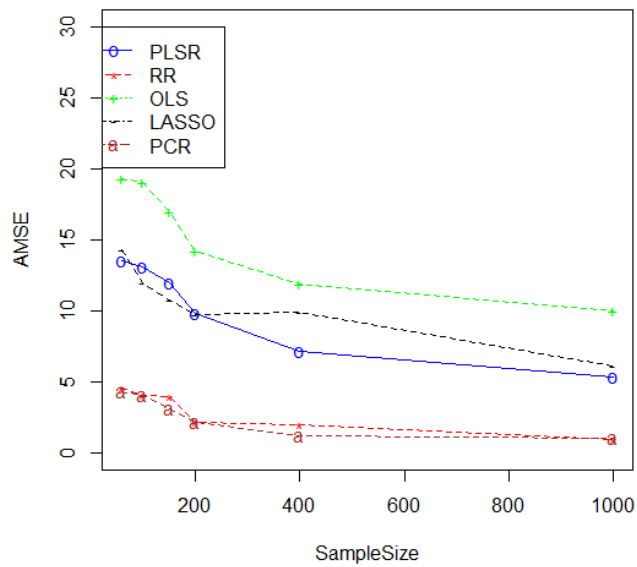


Figure 1: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=2$

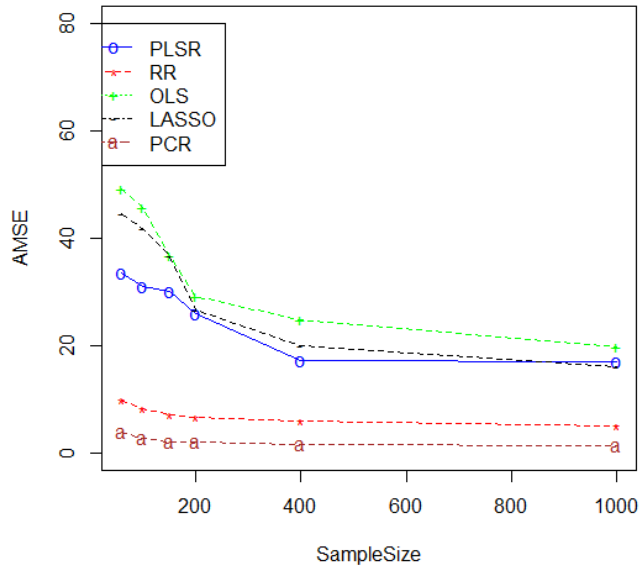


Figure 2: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=4$

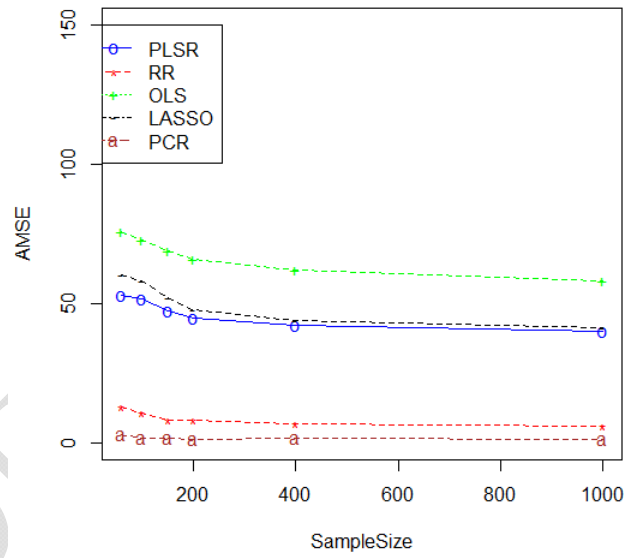


Figure 3: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=6$

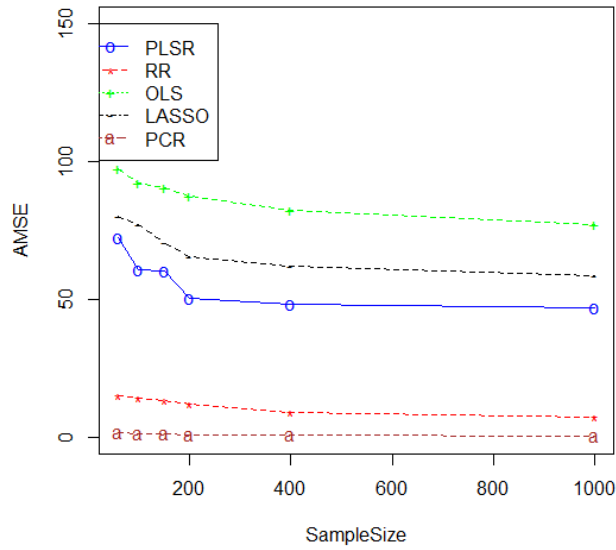


Figure 4: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=10$

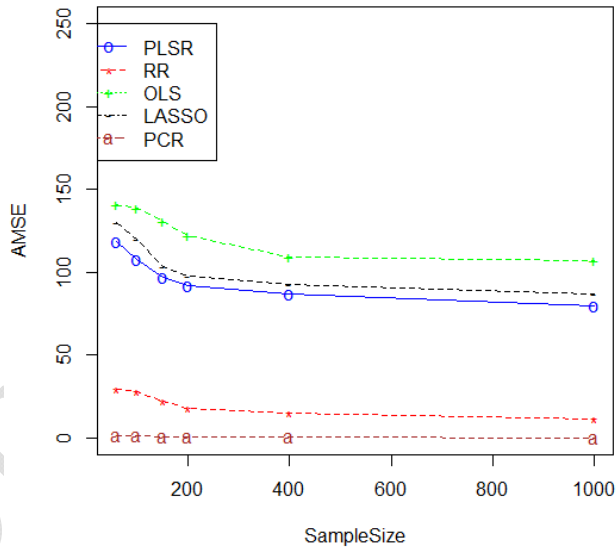


Figure 5: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=20$

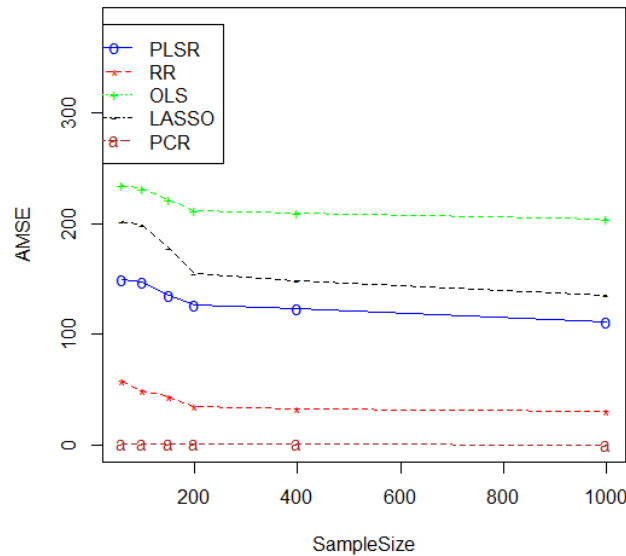


Figure 6: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=50$

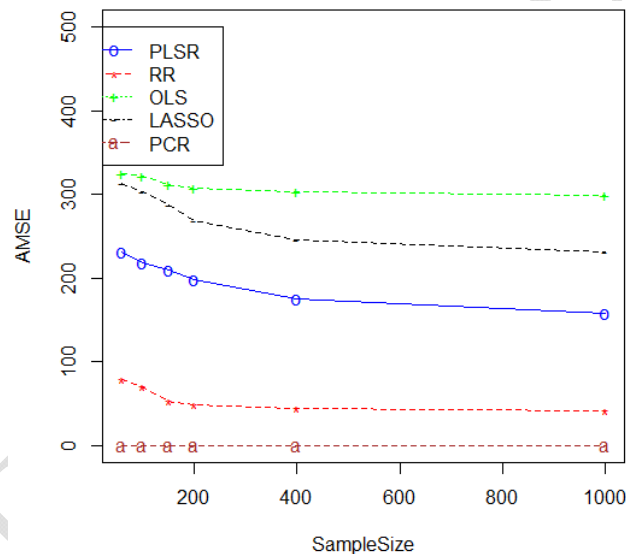


Figure 7: AMSE of PLSR, RR, OLS, LASSO, and PCR for $p=100$

From the simulation study, the AMSE values of the estimated regression parameters $\hat{\beta}$ for each specified p and n cases are calculated. These AMSE values indicate to what extent the slope and intercept are correctly estimated. So, the goal is to obtain an AMSE value close to zero. Figure 1-7 show the values for each method used. From Table 2 where $p = 2, 4, 6, 10, 20, 50$, and 100 , and the specified $n = 60, 100, 150, 200, 400, 1000$ observations, PCR performed best compared to the other methods, having the lowest AMSE values.

Table 3: AIC values for PLSR, RR, OLS, LASSO, and PCR with different number of explanatory variables and sample sizes

P	n	METHODS				
		PLSR	RR	OLS	LASSO	PCR
2	60	14.12	12.91	31.98	35.71	12.09
	100	14.01	12.76	31.92	35.47	12.05
	150	13.92	12.57	31.84	35.33	12.01
	200	13.61	12.48	31.80	35.21	11.74
	400	13.43	12.30	31.74	35.19	11.37
	1000	13.30	12.07	31.72	35.09	11.08
4	60	14.85	12.96	31.65	35.97	12.31
	100	14.80	12.78	31.60	35.81	12.19
	150	14.69	12.65	31.54	35.79	12.07
	200	14.51	12.43	31.35	35.62	12.04
	400	14.43	12.31	31.30	35.59	12.03
	1000	14.28	12.19	31.27	35.47	11.89
6	60	14.65	12.71	31.40	35.70	12.64
	100	14.55	12.70	31.28	32.56	12.65
	150	14.39	12.68	31.22	35.45	12.54
	200	14.27	12.60	31.18	35.33	12.30
	400	14.21	12.59	31.10	35.30	12.14
	1000	14.09	12.30	31.04	35.21	12.11
10	60	14.93	12.97	31.90	35.98	12.90
	100	14.79	12.91	31.81	35.87	12.74
	150	14.69	12.78	31.74	35.74	12.62
	200	14.59	12.67	31.65	35.71	12.50
	400	14.53	12.59	31.50	35.49	12.38
	1000	14.49	12.40	31.40	35.33	12.27
20	60	14.79	12.69	31.48	35.62	12.54
	100	14.74	12.60	31.41	35.45	12.41

	150	14.71	12.49	31.37	35.41	12.34
	200	14.63	12.41	31.32	35.33	12.30
	400	14.59	12.37	31.30	35.23	12.22
	1000	14.41	12.25	31.27	35.09	12.09
50	60	14.38	12.99	31.89	35.82	12.71
	100	14.33	12.81	31.84	35.80	12.69
	150	14.26	12.77	31.82	35.74	12.62
	200	14.21	12.70	31.71	35.70	12.42
	400	14.13	12.60	31.70	35.65	12.31
	1000	14.09	12.49	31.65	35.60	12.19
100	60	14.70	12.74	31.81	35.53	12.40
	100	14.56	12.69	31.74	35.50	12.37
	150	14.41	12.47	31.65	35.47	12.30
	200	14.29	12.39	31.50	35.41	12.27
	400	14.19	12.31	31.43	35.23	12.25
	1000	14.01	12.28	31.40	35.12	12.22

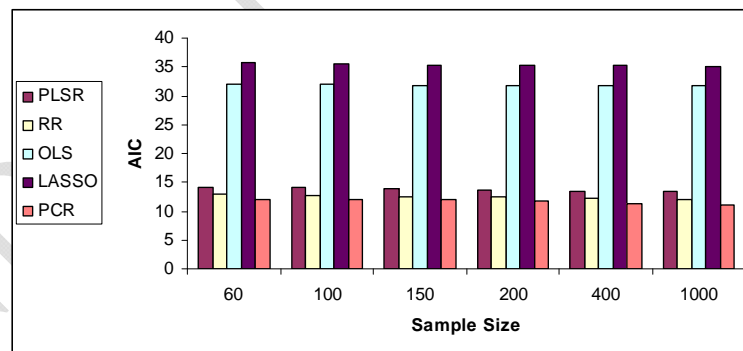


Figure 8: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=2$

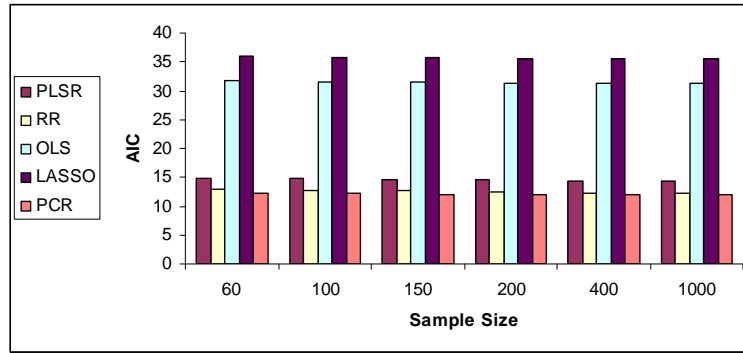


Figure 9: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=4$

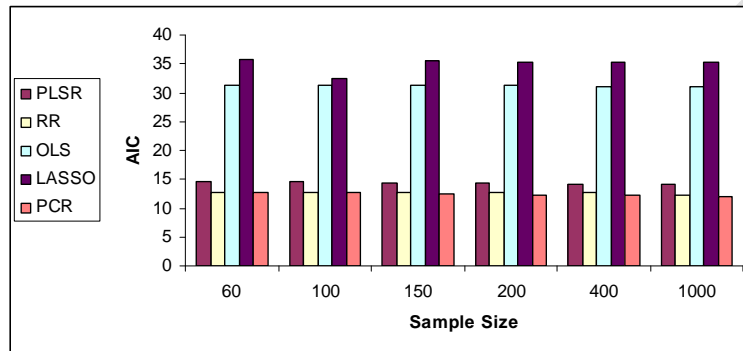


Figure 10: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=6$

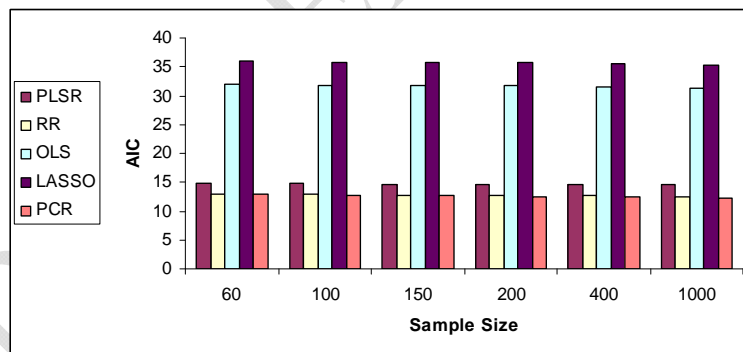


Figure 11: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=10$

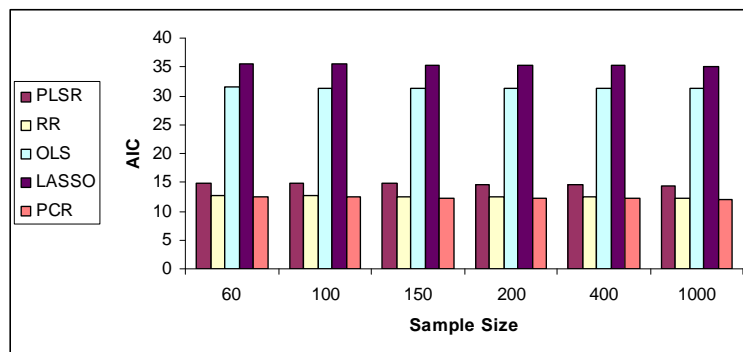


Figure 12: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=20$

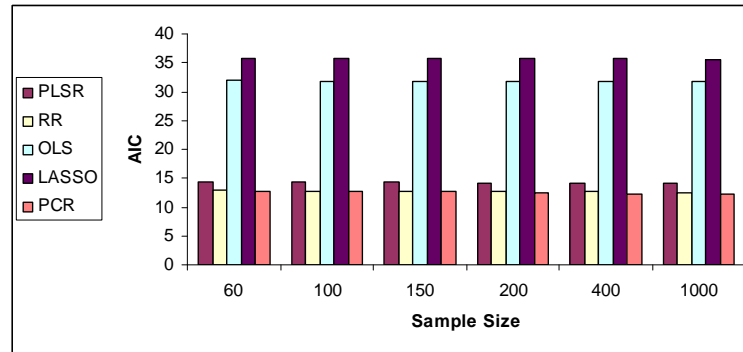


Figure 13: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=50$

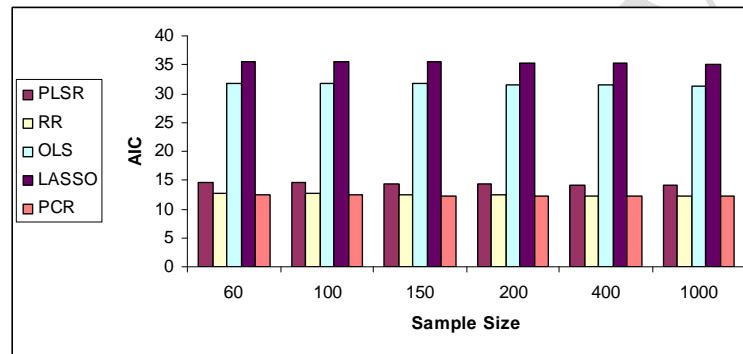


Figure 14: AIC of PLSR, RR, OLS, LASSO, and PCR for $p=100$

To choose the most ideal model, we use AIC of the models obtained using the five methods under review [28-29]. The Akaike Information Criterion values for all methods with different number of independent variables and sample sizes is presented in Table 3 and displayed as bars-graphs in Figure 8 – Figure 14. Figure 8 – Figure 14 show that the greater the sample sizes are the lower the values of Akaike Information Criterion and in contrary to sample sizes, number of explanatory variables does not seem to affect the value of Akaike Information Criterion. LASSO 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 most ideal methods among the selected methods including based on the value of Akaike Information Criterion. 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.

Conclusion

Based on the simulation results at $p = 2, 4, 8, 10, 20, 50, 100$ and the number of data $n = 60, 100, 150, 200, 400, \text{ and } 1000$ 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 multicollinearity among all explanatory variables.

Future Research

The performance of the five methods can also be done by comparing the use of all methods for high-dimensional regressors where $p > n$. It is known that the problem of multicollinearity is present in the data set where the number of variables is high compared to the number of observations.

REFERENCES

- [1] N. Herawati, K. Nisa, E. Setiawan, Nusyirwan, Tiryono (2018) "Regularized Multiple Regression Methods to Deal with Severe Multicollinearity" International Journal of Statistics and Applications Vol. 8, No 4 Pp: 167-172
- [2] Norliza Adnan, 1Maizah Hura Ahmad, Robiah Adnan (2006) "Study On Some Methods For Handling Multicollinearity Problems" MATEMATIKA, Volume 22, Number 2, pp. 109–119.
- [3] B. L. Bowerman and R. T. O' Connell, Linear Statistical Models an Applied Approach, 2nd Ed., Boston: PWS-KENT Publishing Co, 1990.
- [4] M. S. Younger, A Handbook for Linear Regression, USA: DUXBURY Press, 1979.
- [5] J. Neter, W. Wasserman and M. H. Kutner, Applied Linear Regression Models, 3rd Ed. USA: IRWIN Book Team, 1990.
- [6] A. E. Hoerl and R. W. Kennard, Ridge Regression: Biased Estimation to Nonorthogonal Problems, Technometrics, 12 (1970), 56 – 67.
- [7] R. H. Myers, Classical and Modern Regression With Applications, 2nd Ed. USA: PWSKENT Publishing Company, 1986.
- [8] W. F. Massy, Principal Components Regression in Exploratory Statistical Research, Journal of the American Statistical Association, 60 (1965), 234 – 246.

- [9] R. W. Farebrother, A Class of Statistical Estimators Related to Principal Components. *Linear Algebra and its Applications*, 289 (1999), 121 – 126.
- [10] P. Filzmoser and C. Croux, A Projection Algorithm for Regression with Collinearity, In K. Jajuga, A. Sokolowski, H. Bock (ed.): *Classification, Clustering and Data Analysis*, Berlin: Springer, 227 – 234, 2002.
- [11] R. R. Hocking, *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, USA: John Wiley & Sons, 1996.
- [12] J. T. Hwang and D. Nettleton, Principal Components Regression With Data Chosen Components and Related Methods, *Technometrics*, 45 (2000), 70 – 79.
- [13] Y. L. Xie and J. H. Kalivas, Local Prediction Models by Principal Component Regression, *Analytica Chimica Acta*, 348 (1997), 29 – 38
- [14] Slawski, M. 2017. On Principal Components Regression, Random Projections, and Column Subsampling. Arxiv: 1709.08104v2 [Math-ST].
- [15] Wethrill, H., 1986, Evaluation of ordinary Ridge Regression. *Bulletin of Mathematical Statistics*, 18, 1-35.
- [16] Hoerl, A.E., 1962, Application of ridge analysis to regression problems. *Chem. Eng. Prog.*, 58, 54-59.
- [17] Hoerl, A.E., R.W. Kannard and K.F. Baldwin, 1975, Ridge regression: Some simulations. *Commun. Stat.*, 4, 105-123.
- [18] James, G., Witten D., Hastie T., Tibshirani R *An Introduction to Statistical Learning: With Applications in R*. New York: Springer Publishing Company, Inc., 2013.
- [19] Tibshirani, R., 1996, Regression shrinkage and selection via the LASSO. *J Royal Stat Soc*, 58, 267-288.
- [20] Hastie, T., Tibshirani, R., Mainwright, M., 2015, *Statistical learning with Sparsity The LASSO and Generalization*. USA: Chapman and Hall/CRC Press.
- [21] Coxe, K.L., 1984, “Multicollinearity, principal component regression and selection rules for these components,” *ASA Proceed. Bus fj Econ sect'ion*, 222-227.
- [22] Jackson, J.E., *A User's Guide To Principal Components*. New York: Tiley, 1991.
- [23] Jolliffe, LT, *Principal Component Analysis*. *New York*: Springer-Verlag, 2002.
- [24] Flury, B. and Riedwyl, H., *Multivariate Statistics. A Practical Approach*, London: Chapman and Hall, 1988.
- [25] Osuji G.A., Obubu, M., Nwosu, C.A (2016) “Stock Investment Decision in Nigeria; A PC Approach” *World Journal of Multidisciplinary and Contemporary Research*, Vol. 2, No 1, pp. 1-11. <http://wjmc.com>

- [26] Osuji G.A., Obubu, M., Nwosu, C.A (2016) “Preconception sex selection using proper ovulation timing” World Journal of Probability and Statistics Research U.S.A, Vol. 2, No 1, pp. 1-12. <http://wjmc.com>
- [27] Osuji, G.A., Okoro, C.N., Obubu, M., Obiora-Ilouno H.O. (2016) “Effect of Akaike Information Criterion on Model Selection in Analyzing Auto-Crash Variables.” International Journal of Sciences: Basic and Applied Research (IJSBAR). Vol. 26, No 1, pp. 98-109.
- [28] Obubu, M., Okoye Valentine, Omoruyi Frederick, Ngonadi Lilian Oluebube (2017) “Infant Mortality; a continuing social problem in Northern Nigeria: Cox Regression Approach. American Journal of Innovative Research and Applied Sciences.2017; 5(5):1-5.
- [29] Maxwell, O., Happiness, O.-I., Alice, U.C. and Chinedu, I.U. (2018) An Empirical Assessment of the Impact of Nigerian all Share Index, Market Capitalization, and Number of Equities on Gross Domestic Product. Open Journal of Statistics, 8, 584-602. <https://doi.org/10.4236/ojs.2018.83038>.