Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 159-168
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Choosing Appropriate Regression Model in the Presence of Multicolinearity

Maruf A. Raheem[1*], Nse S. Udoh[2], Aramide T. Gbolahan[3]

[1]Department of Engineering and Mathematics, Sheffield Hallam University, Sheffield, UK
[2]Department of Mathematics & Statistics, University of Uyo, Uyo, Nigeria
[3]Department of Computer Science, Sheffield Hallam University, Sheffield, UK
Email: *rahemarsac@yaoo.com, nseudoh07@yahoo.com, aramide.gbolahan@gmail.com

How to cite this paper: Raheem, M.A., Udoh, N.S. and Gbolahan, A.T. (2019) Choosing Appropriate Regression Model in the Presence of Multicolinearity. *Open Journal of Statistics*, **9**, 159-168.
https://doi.org/10.4236/ojs.2019.92012

Received: February 3, 2019
Accepted: March 29, 2019
Published: April 1, 2019

Abstract

This work is geared towards detecting and solving the problem of multicolinearity in regression analysis. As such, Variance Inflation Factor (VIF) and the Condition Index (CI) were used as measures of such detection. Ridge Regression (RR) and the Principal Component Regression (PCR) were the two other approaches used in modeling apart from the conventional simple linear regression. For the purpose of comparing the two methods, simulated data were used. Our task is to ascertain the effectiveness of each of the methods based on their respective mean square errors. From the result, we found that Ridge Regression (RR) method is better than principal component regression when multicollinearity exists among the predictors.

Keywords

Multicollinearity, Adequacy, Regression coefficients, Variance Inflation Factor (VIF), Mean Square Error

1. Introduction

Regression Analysis is a statistical tool used in studying if there is existence of relationship, of any forms, either linear or nonlinear between the two variables, subject to certain constraints, such that one of the two variables can serve well to predict for the other. Meanwhile, it is important to note that our focus in this study is on the linear form of such relationship. Thus, when we talk of regression, we only consider the linear regression, which may either be simple, multiple and or, multivariate in nature depending on the levels (or number) of variables on either side of the equation. When we compare a single dependent with a single independent variable, the regression is said to be simple, so we have

simple (linear) regression. But, if only one dependent variable is being compared with more than one independent variables, the regression is said to be multiple in form; and thus we have multiple (linear) regression. The multivariate regression (which is outside the scope of this research), only comes to play if we are comparing more than one level of the dependent variable with two or more levels of the independent variables.

1.1. Fundamental Principles

Consider the multiple regressions model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i \tag{1}$$

In matrix form, (1) becomes:

$$Y = X'\beta + \in \tag{2}$$

where in (1); $\beta_{j's}$ $\forall j = 1, 2, \cdots, p$ are the regression coefficients; $Y(n \times 1)$ matrix represents the outcome (response or dependent) variable and $X_{i's}$, $\forall i = 1, 2, \cdots, n$ are the explanatory (predictor or independent) variables, which are fixed and e_i is the error term, which with $Y_{i's}$ are assumed to be random. With assumptions that: $E(e_i) = 0$; $E(e_i e_j) = 0$ $\forall i \neq j$; $E(e_i e_i) = \sigma^2$, which implies randomness, independence and homoscedasticity of error terms respectively. Indicating $e_i \sim IID\ N(0, \sigma^2)$. Other interesting assumptions are Zero Covariance between e_i and each of the $X_{i's}$ variable; *i.e.*

$$Cov(e_i, X_1) = Cov(e_i, X_2) = \cdots = Cov(e_i, X_p) = 0;$$

no specification bias; the model should correctly be specified and no exact linear relationship between any two predictor variables. However, violating any of these assumptions brings about serious problem in regression analysis. Hence, causes of multicollinearity which constitutes a major problem sets in as a result of violation of the said assumptions [1], which constitute the pivot of our discussion in this section.

Multicollinearity is an important concept in regression analysis, given the serious threat it poses on the validity or the predicting strength of the regression model. It is usually regarded as a problem arising out of violation of the assumption that the explanatory variables are linearly independent. It is a phenomenon that plays its way in regression, especially multiple regressions when there is a high level of inter-correlation or inter-associations among the independent variables.

It is therefore a type of disturbance in the regression model which if allowed, the statistical inferences made about the model become misleading simply because the estimates of the regression coefficients are faulty or unreliable. Multicollinearity is a condition in multiple regression models whereby two or more covariates become redundant. The redundancy implies that what one independent variable (X) explains about the dependent variable (Y) is exactly what the other independent variable explains. In this case, the estimates of the regression coefficients for such redundant predictor variables would be completely erroneous.

1.2. Possible Causes of Multicollinearity

1) Multicollinearity generally occurs when two or more explanatory variables are directly and highly correlated to each other.

2) It may also set in when one or more of the predictors represent the multiples of or computed from some other predictor variables in the same equation.

3) It may also be experienced when repeating or including almost the same predictor variable in the same model.

4) It may as well occur when in situations of nominal variables; the dummy variables are not properly use.

However, the following have been identified as the primary sources of multicollinearity;

1) When a regression model is over defined; that is, including more than necessary predictor variables in the model;

2) The data collection method is faulty; or better still choosing in appropriate sampling scheme used for data collection or generation;

3) Placing a spurious or unnecessary constraint on the model or in the population;

4) When the regression model is wrongly specified.

1.3. How Is Multicollinearity Detected?

There are a number of ways by which multicollinearity may be detected in a multiple regression model, which include:

1) when the correlation coefficients in the correlation matrix of predictor variables become so high that is close to one, or the value of correlation coefficient between two highly correlated predictor variables is close to one.

2) when the coefficient of determination (R^2) value is so close to unity for a particular predictor variable that is regressed on other independent variables to such an extent that the variance inflation factor (VIF) becomes so large [2].

3) When one or more eigenvalues of the correlation matrix becomes so small that is close to zero then multicollinearity is at work.

4) Another rule of thumb to detecting the presence of multicollinearity is that while one or more eigenvalues of the predictor variables become so small, to the extent of getting so close to zero but the corresponding condition number (ϕ) becomes very large [3] [4] [5].

5) Comparing the decisions made using overall F-test and t-test might provide some indication of the presence of multicollinearity. For instance, when the overall significance of the model is good using F-test, but individually, the coefficients are not significant using t-test, then the model might suffer from multicollinearity

1.4. Possible Effects of Multicollinearity

The effects of existence of multicollinearity in regression are of concerns that it gives rise to circumstance whereby:

1) The partial contribution of each of the explanatory variable remains confounded leading to difficulties in interpreting the model;

2) The variances as well as the coefficients of the predictor variables becomes un duly bogus (or inflated), thereby making precise estimation of the parameters becomes impossible;

3) The presence of multicollinearity gives rise to considerably high mean square error, paving ways for committing type-1-error;

4) The ordinary least square (OLS) estimators as well as their standard errors may be sensitive to small changes in the data, in order words, the results will not be robust; and

5) Finally, [6] observes that as multicollinearity increases, it complicates the interpretation of the variable because it becomes more difficult to ascertain the effect of any single variable, due to the variable interrelationships.

Since obtaining robust estimates of regression coefficients possess significant level of difficulties with ordinary least square (OLS) method, we in this research, we hope to explore other possible regression models; principal component regression (PCR) and ridge regression (RR) methods as alternative techniques to estimating the model parameters, with a view to enhancing precision in estimating the parameters of the regression parameters when multicollinearity is suspected among the predictors without need for dropping any of the variables and that to determine which of the two methods perform better based on the mean square errors of the two models. Meanwhile, the use of condition number as well as variance inflation factor is endearing to us to check if multicollinearity exists among the covariates after estimating via OLS approach.

The remaining part of this paper is organized as follows; Section 2 has the brief discussion on the methodologies adopted while Section 3 presents the results and general discussion on the findings of the research.

2. Methodology

This section discusses statistical techniques which are applied and compared with Ordinary Least Square (OLS) method in multiple linear regressions. These methods are Principal component regression (PCR) and Ridge regression (RR), their formulations as well as the underlining assumptions governing each of them are discussed.

2.1. Principal Component Regression (PCR)

This is one of the methods for solving problem of multicollinearity such that better estimates of the model parameters and consequently, better and a more robust prediction could be made as compared to ordinary least squares. With this method, the original variables are transformed into a new set of orthogonal or uncorrelated variables called principal components of the correlated matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure then involves eliminating some of the principal com-

ponents to effect a reduction in variance. The major goal of PCR include; variable reduction, selection, classification as well as prediction. It would be recalled that this method is two procedural in application; first principal component analysis is applied, then the set of "*k*" uncorrelated or orthogonal component factors are used to replace the original *p* set of predictor variables. According to [7], PCR is a two-step procedure, in the first step, one computes principal components which are linear combinations of the explanatory variables while in the second step, the response variable is regressed on to the selected principal components. Combining both steps in a single method will maximize the relation to the response variables.

Let the random vector of the predictor variable be:

$X' = \left[X_1, X_2, X_3, \cdots, X_p \right]$, with the covariance matrix Σ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$. The linear combination:

$$Y_i = e_i' X \ ; \tag{3}$$

Subject to the constraints:

$$Var\left(Y_i\right) = e_i' \Sigma e_i; \ \ \forall i = 1, 2, \cdots, p \tag{4}$$

$$Cov\left(Y_i, Y_j\right) = e_i' \Sigma e_j; \ \ \forall i, j = 1, 2, \cdots, p; i \neq j \tag{5}$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, Y_3, \cdots, Y_p$ whose variances are as maximal as possible.

Looking closely at the model (2), suppose $X'X = \Sigma$ is rewritten as $P \Lambda P'$, with Λ being the $\left(p * p\right)$ diagonal matrix of the eigenvalues of the design matrix or variance covariance matrix, Σ and $P = \left[e_1, e_2, e_3, \cdots, e_p \right]$ representing associated normalized eigenvectors for the eigenvalues for Σ, such that:

$$PP' = P'P = I \quad \text{(Identity matrix)} \tag{6}$$

Rewriting the (2) by inserting (6) $Z = XPP'\beta + \varepsilon$; which then becomes:

$$Z = Y\Phi + \varepsilon \tag{7}$$

where $Y = X'P$ and $\Phi = P'\beta$; thus, $Y'Y = P'X'XP = P'\Sigma P = P'P\Lambda P'P = \Lambda$. The columns of *Y*, defined as the linearly uncorrelated components of the original random vector *X* are now the new set of orthogonal predictor variables; also known as the principal components. These now serve as the new covariates in the regression model, called principal component regression. Obtaining the estimates of the model using OLS, we have:

$$\hat{\Phi} = Y'Y^{-1}Y'Z = \Lambda^{-1}Y'Z \tag{8}$$

Such that the covariance of $\hat{\pi}$ is

$$V\left(\hat{\Phi}\right) = \sigma^2 \left(Y'Y\right)^{-1} = \Lambda^{-1}\sigma^2 = diag\left(\lambda_1^{-1} + \lambda_2^{-1} + \cdots + \lambda_k^{-1}\right) \tag{9}$$

This new variance estimate is now expected to produce a minimum variance, which eventually leads to having an improved and reliable estimate of the parameters, thus making a robust decision. According to [8], one of the simplest ways collinearity problem could be rectified in practice, is by the use of Principal Component Regression (PCR); claiming that from the experience, PCR usually

gives much better result than the least square for prediction purpose.

2.2. Ridge Regression Method

This method was originally suggested by [9] as a procedure for investigating the sensitivity of least-squares estimates based on data exhibiting near-extreme multicollinearity, where small perturbations in the data may produce large changes in the magnitude of the estimated coefficients.

The ridge regression estimate of the coefficients, β_j; $j=1,2,\cdots,k$ are given by

$$\hat{\beta}_R = \left(X'X + rI\right)^{-1} X'Y \tag{10}$$

where $r \geq 0$ is a constant called biasing factor, which needs to be set by the researcher; such that when $r = 0$, the ridge regression automatically reduces to ordinary least square. This is implies that ridge regression is an improved form of OLS, with minor transformation.

Thus:

$$E\left(\hat{\beta}_R\right) = \left(X'X + rI\right)^{-1} E\left(X'Y\right) = \left(X'X + rI\right)^{-1} \left(X'X\right) E\left(\beta\right) = P_R, \hat{\beta} \tag{11}$$

where $P_R = \left(X'X + rI\right)^{-1} X'X$. This results point to the fact that ridge regression is a biased estimator of $\hat{\beta}$, which is the necessary condition for getting away with the problem of estimating the model parameter.

Meanwhile the variance-covariance matrix of β_R is obtained as:

$$Var\left(\hat{\beta}_R\right) = \left(X'X + rI\right)^{-1} \left(X'X\right) \left(X'X + rI\right)^{-1} \sigma^2 \tag{12}$$

Giving rise to the mean square error:

$$MSE\left(\hat{\beta}_R\right) = Bias + variance;\ i.e.\ \left(Bias\hat{\beta}_R\right)^2 + Var\left(\hat{\beta}_R\right) \tag{13}$$

According to [10], ridge regressions are known to have favourable properties as shown by [9] β_R has smaller mean square error than the ordinary least square estimators $\hat{\beta}$, provided σ^2 is small enough so that the validity of the regression model holds. [11] [12] also pointed out that the ridge regressions are known as shrinkage estimator.

3. Results and Discussion

In this work, we illustrate with an example to predicting gas productivity (Y) using density (X_1), volumetric temperature (X_2), sulphur content (X_3), feedback flow (X_4), output feedback temperature (X_5), catalyst temperature in regenerator system (X_6) and catalyst/feedback ratio (X_7) as the independent variables.

Now since some of the variables are significantly related as shown in **Table 1**, it then becomes impossible to determine which of the variables accounts for the variation in the dependent variable. This is because of high correlation among the predicting variables, resulting in less stability in the estimates of the regression parameters [13]. The results of the correlation matrix above showed a highly

Table 1. Correlation matrix between independent variables.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1						
X_2	−0.14	1					
X_3	0.96**	−0.24	1				
X_4	−0.25	0.05	−0.28	1			
X_5	0.10	−0.20	0.19	0.33	1		
X_6	−0.44	−0.49*	−0.46	−0.67	0.16	1	
X_7	0.67**	−0.17	0.66**	−0.13	−0.09	−0.30	1

*P-value < 0.05, significantly correlated at 5%, **P-value < 0.01, significantly correlated at 1%.

significant possible relationships between X_1 and X_3 (r-0.96, P-value = 0.004), X_3 and X_7 (r = 0.66, P-value = 0.004). These results showed that there is presence of multicollinearity among these independent variables.

3.1. Multicollinearity Diagnostic

The existence of multicollinearity was investigated using Variance Inflation Factor (VIF), variables proportion and condition index. The result obtained is shown in Table 2. It could be confirmed that X_1 and X_3, have VIF greater than 10 which shows that there is collinearity problem.

Variance Inflation Factor (VIF)

The following VIF values were obtained from each of the Independent Variables:

$$\begin{aligned}&\mathrm{VIF}(X_1)=20.74, \mathrm{VIF}(X_2)=3.639, \mathrm{VIF}(X_3)=38.95,\\ &\mathrm{VIF}(X_4)=2.58, \mathrm{VIF}(X_5)=2.82, \mathrm{VIF}(X_6)=5.58,\\ &\mathrm{VIF}(X_7)=2.24\end{aligned}$$

The result of VIF revealed presence of multicollinearity at VIF (1) and VIF (3) are greater than 10. This result confirmed a high level of multicollinearity among the independent variables.

The Condition Index (ϕ) = $\frac{\lambda_{\max}}{\lambda_{\min}}$; the ratio of maximum eigenvalue to minimum eigenvalue.

$\frac{7.492}{0.00001}=749200$. Since $\phi>1000$ (749,200 > 1000), the results also supported that obtained from VIF.

Mean Squared Error (MSE)

For any given regression model:

$$\mathrm{MSE}=\frac{\mathrm{SSE}}{n-p} \tag{14}$$

where: SSE: Sum of square error of the linear regression model; $n-p$: Degree of freedom; n: is the number of data point; p: Number of parameters in the model.

Table 2. Means square error for ordinary least, principal component and ridge regression.

OLS	PCR	RR
0.70554	0.70553	0.68624

3.2. Summary of the Regresion Model

Ordinary Least Square (OLS)

$$GP = 126.121 - 175.414\,DTY + 0.037\,VT - 2.121\,SC + 0.064\,FF - 0.042\,FT + 0.085\,CT + 1.718\,FR$$

Principal Component Regression (PCR)

$$GP = 51.400 - 2.336\,DTY + 0.336\,VT - 0.385\,SC + 0.552\,FF - 0.480\,FT + 0.674\,Ct - 0.049\,FR$$

Ridge Regression (RR)

$$GP = 54.1387 - 28.0014\,DTY + 0.0029\,VT + 0.5137\,SC + 0.0045\,FF + 0.0087\,FT + 0.1236\,CT - 0.62374\,FR$$

3.3. Comparison of OLS, PCR AND RR

Computing the mean Square Error for each of the model we obtain the following result. The result can be summarized in Table 2. From the table, it could be confirmed that among the three estimates, ridge regression seems to produce the least mean square error; followed by the PCR and OLS. Though the difference between the MSE for PCR and OLS seems insignificant, this may be due to the data size used for the analysis.

Table 3 provides the results on the collinearity test obtained for the seven variables. In it we also have the VIF, Eigen-value and the conditional index.

4. Summary and Conclusion

Having fitted the respective regression models to the available data, we investigated the adequacies of the three models using the MSE square errors (see Table 2). It could be seen ridge regression gave the least error means square Error. Hence Ridge regression is considered superior.

Meanwhile, given the results obtained from the analyses, one may conclude there is no much difference in the error values, especially with the PCR and OLS; this may be due to the nature of the data used for the analysis. However, based on the results presented in Table 2, apparently the ridge regression performed better than principal components since it gave the smaller MSE value compared to PCR and OLS. The seeming insignificant difference between the MSE for PRC and OLS could be attributed to minimal level of multicollinearity between the covariates, except with X_1X_7; X_1X_3 and X_3X_7 (see Table 1). The fact that the RR's MSE is smaller compared to the other two is the confirmation of its efficiency power to perform well irrespective of what the hidden correlations among the covariates are.

Table 3. Computations on collinearity test.

VIF	Eigenvalue	Contn. Index	Intercept	X_1	X_2	X_3	X_4	X_5	X_6	X_7
-	7.492	1.0	0.00	0.000	0.000	0.001	0.000	0.000	0.000	0.0000
20.736	0.507	3.846	0.00	0.000	0.000	0.026	0.000	0.000	0.000	0.002
3.639	0.001	112.352	0.00	0.000	0.0003	0.0177	0.0137	0.0110	0.00	0.680
38.953	0.003	153.825	0.00	0.0004	0.0067	0.0005	0.3735	0.0001	0.000	0.030

Further, we observe that despite the little quantitative difference, there lie some advantages as well as disadvantages between the two methods. The advantages of RR method over PCR are that it is easier to compute and also provides a more stable way of moderating the model's degrees of freedom than dropping variables [14]. Meanwhile, it is important to note that PCR affords us the opportunity of testing a hypothesis to determine the most significant of the component variables.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mansfiled, E.R. and Billy, P.H. (1982) Detecting Multicollinearity. *The American Statisticians*, **36**, 158-160. https://doi.org/10.2307/2683167

[2] Hwang, J.T. and Netleton, D. (2000) Principal Component Regression with Data for Chosen Components and Related Methods. *Technometrics*, **45**, 70-79. https://doi.org/10.1198/004017002188618716

[3] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001) Introduction to Linear Regression Analysis. 3rd Edition, Willey, New York.

[4] Farrar, D.E. and Glanber, R.R. (1967) Multicollinearity in Regression Analysis. *Review of Economic and Statistic*, **49**, 92-107.

[5] Myers, R.H. (1986) Classical and Modern Regression with Applications. 2nd Edition, PWSKENT Publishing Company, USA.

[6] Hair, J.F., Anderson, R.E. and Black, W.C. (1998) Multivariate Data Analysis. 5th Edition, Prentice Hall, New Jersey.

[7] Filzmoser, P. and Groux, C. (2002) A Projection Algorithm for Regression with Collinearity. In: Jajuga, K., Sokołowski, A. and Bock, H.H., Eds., *Classification, Clustering, and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Berlin, Heidelberg, 227-234.

[8] Martens, H. and Naes, T. (1989) Assessment, Validation and Choice of Calibration Method. In: *Multivariate Calibration*, 1st Edition, John Wiley & Sons, New York, 237-266.

[9] Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[10] Bjokstrom, A. (2001) Ridge Regression and Inverse Problem. Research Report,

Stockholm University, Stockholm.

[11] Marquardt, D.W. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of Society of Industrial Applied Mathematics*, **11**, 431-441.

[12] Marquardt, D.W. (1970) Generalized Inverses, Ridge Regression. *Biased Linear Estimation*, *and Nonlinear Estimation*, *Techno Metrics*, **12**, 591-612. https://doi.org/10.2307/1267205

[13] Ijomah, M.A. and Nwakuya M.T. (2011) Ridge Regression Estimate. Unpublished M.Sc. Thesis, University of Port Harcourt, Port Harcourt.

[14] Pasha, G.R. and Shah, M.A. (2004) Application of Ridge Regression to Multicollinearity Data. *Journal of Research* (*Science*), **15**, 97-106.