A STUDY ON THE PERFORMANCE OF THE PARTIAL LEAST SQUARES REGRESSION IN HANDLING MULTICOLLINEARITY USING SIMULATED DATA

O. P. Babatoba[1], B. Ishaq[2*], Y. Zakari[1], A. S. Mohammed[1], A. Usman[1], J. Abdullahi[1] & I.A Sadiq[1]

[1]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria
[2]Department of Military Sciences, Nigerian Defence Academy, Kaduna, Nigeria

**Corresponding author: buhariishaqabu@gmail.com*

ABSTRACT

Multicollinearity is a common issue in regression analysis which occurs due to the violation of the assumptions of regression that there is no correlation between the explanatory variables of the least square estimator, and because of the violation, the estimate of the parameters tends to be less precise and unreliable, and this leads to unstable inflated variance. Thus, the biased regression techniques which stabilize the variance of the parameter estimate were employed. This study focused majorly on the Partial Least Square Regression, a biased regression technique for overcoming multicollinearity, the strength and limitations of the method, and also the performance of the method when compared with the Principal Component Regression (PCR) using the Root Mean Square Error (RMSE) as a performance metric. A simulation study of data that follows a normal distribution with varying levels of multicollinearity was conducted to evaluate the accuracy, interpretability, and robustness of PLSR models and also in comparison to the PCR using the root mean square error (RMSE) as a performance metric. Based on this study, it is observed that the PLSR is more robust to multicollinearity than PCR, as it is less likely to produce unstable parameter estimates in highly correlated datasets. Therefore, this technique can be applied to the same distribution used in this study by varying the sample sizes. It can also be used to look at the behaviors of distributions other than those used in this study.

Keywords: Regression; Multicollinearity; Partial Least Square Regression; Principal Component Regression; Simulation

INTRODUCTION

In regression, the objective is to explain the variation in one or more response variables, by associating this variation with proportional variation in one or more explanatory variables. This phenomenon called multicollinearity, is a common problem in regression analysis. Handling multicollinearity in Regression analysis is important because least squares estimations assume that predictor variables are not correlated with each other. Multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the explanatory variables (Gujarati, 2003). It is a problem that always occurs when two or more predictor (or explanatory) variables are correlated with each other or regressed on the other predictor variables in the applications of regression analysis. If it is regressed on the other explanatory variables, then the matrix of intercorrelations among the explanatory variables is singular and there exists no unique solution for the regression coefficients (Gordon, 1968). It is also a condition in a set of regression data that has two or more regressors that are redundant and have the same information. Redundant information means what one variable explains about the response (or dependent) variable is exactly what the other variable explains. In this case, the two or more redundant predictor variables would be completely unreliable since the regression coefficients would measure the same effect of the independent variables. The presence of multicollinearity in least squares regression can cause larger variances of parameter estimates which means that the estimates of the parameters tend to be less precise. As a result, the model will have insignificant tests and a wide confidence interval. Thus, the more the multicollinearity, the less interpretable the parameters. Several methods have been developed for detecting the presence of serious multicollinearity (Hair et al, 1998). One of the most commonly used methods is the variance inflation factor (VIF) which measures how much the variance of the estimated regression coefficients is inflated compared to when the independent variables are not linearly related Neter et al, (1990).

Depending on the goal of your regression analysis, you might not need to resolve the multicollinearity, but if you determine that you do need to fix multicollinearity, some of the common ways to resolve the problem of multicollinearity include: removing one or more of the highly correlated predictor variables and perform an analysis that is designed to account for highly correlated variables such as principal component analysis, partial least squares (PLS) regression, ridge regression and so on (Lukman et al, 2024).

METHODOLOGY

This chapter will discuss the methodology of the partial least squares regression as a tool for handling multicollinearity within simulated datasets and will also provide a structured approach to investigate the capability and evaluate the effectiveness of the partial least squares regression (PLSR) in handling multicollinearity within simulated datasets.

1. Principal Component Regression (PCR)

One of the simplest ways that the collinearity problem is solved in practice is by the use of principal component regression (PCR). Principal component regression (PCR) is a method that combines the advantages of principal component analysis with linear regression. It is a powerful tool for analyzing high-dimensional data when the number of observations is smaller than the number of predictor variables. PCR works by constructing a small set of principal components and then using them as predictors in a regression model.

The mathematical formula for PCA is:

$$Y = XB + \varepsilon \tag{1}$$

where Y is the response variable, X is the observed predictor matrix, B is the matrix of regression coefficients, and ε is the vector of residual errors.

The solution of multiple linear regression is:

$$\hat{B} = (X'X)^{-1}X'Y \tag{2}$$

In PCR, the collinearity that exists in the predictor variables can be eliminated by extracting a group of orthogonal predictors through the application of PCA on X and then performing regression on Y using a subset of the resulting components of X.

$$X = USV' + \varepsilon \tag{3}$$

$$U = XV \tag{4}$$

where U is the matrix of scores, S is the diagonal matrix of singular values, and V is the matrix of loadings.

The multiple linear regression can be written as the following:

$$Y = UB + \varepsilon \tag{5}$$

The solution of regression can be written as the following:

$$\hat{B} = (U'U)^{-1}U'Y \tag{6}$$

2. Partial Least Squares Regression (PLSR)

Partial least squares (PLS) is a method for modeling relationships between a dependent variable (Y) and explanatory variables (X) (Garthwaite, 1994). This method was first developed by Herman Wold (1966) in the social sciences, specifically in economics, but it gained popularity first in chemometrics through the work of his son, Svante Wold.

PLS is a predictive technique that can handle many independent variables, especially when these display multicollinearities. The goal of PLS regression is to predict Y from X and to describe their common structure when X is likely to be singular and the regression approach is no longer possible to be used because of multicollinearity problems. This method is similar to Principal Component Regression because components are extracted before they are regressed to predict Y. In contrast, PLS regression searches for a set of components called latent vectors, factors or components from X that are also relevant for Y that perform a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y (Abdi, 2003). In this method, the component is extracted from the rest of the components and the components are extracted in such a way that they are uncorrelated (orthogonal). How this algorithm functions will now be described to show.

Component is defined as:

$$t_i = W_{11}X_1 + W_{12}X_2 + \ldots + W_{ij}X_j \tag{7}$$

Where X_j are the explanatory variables, Y is the dependent variables.

The W_{ij} is the coefficient:

$$W_{ij} = \frac{\text{cov}(X_j, Y)}{\sqrt{\sum_j^p \text{cov}(X_j, Y)^2}}, j = 1, 2, 3, \ldots, p \quad (8)$$

From this, it can be deduced that to obtain the scalar product (X_j, Y) must be calculated for each j= 1, 2, …, p.

Calculating the second component is justified when the single-component model is inadequate i.e. when the explanatory power of regression is small and another component is necessary. The second component is denoted by t_2 and it will be a linear combination of the regression residues of X_j variables on components t_1 instead of the original variables. In this way, component orthogonality is assured. To do this, the residual for the single component regression must be calculated which will be,

$$\varepsilon_1 = Y - \hat{Y} = Y - \beta_1 t_1 \quad (9)$$

With,

$$\beta_1 = \frac{\text{cov}(y_i, t_i)}{\|t_1\|^2} \quad (10)$$

The second component is obtained as:

$$t_2 = W_{21}\varepsilon_{11} + W_{22}\varepsilon_{12} + \ldots + W_{2p}\varepsilon_{1p} \quad (11)$$

With,

$$W_{2j} = \frac{\text{cov}(\varepsilon_{ij}, \varepsilon_1)}{\sqrt{\sum_j^p \text{cov}^2(\varepsilon_{ij}, \varepsilon_1)}}, j = 1, 2, 3, \ldots, p \quad (12)$$

The residuals e_{ij} are calculated by computing the simple regression of x_j on t_1,

$$X_j^* = \alpha_j t_j, j = 1, 2, 3, \ldots, p$$

therefore,

$$\varepsilon_{ij} = X_j - X_j^* = X_j - \alpha_j t_j \quad (13)$$

where the estimators of the regression coefficients have been calculated thus:

$$\alpha_j = \frac{\text{cov}(x_j, t_1)}{\|t_1\|^2} \tag{14}$$

Now with e_i and e_{ij}, only the scalar products have to be computed cov(e_i, e_{ij}), for j = 1, …, p, to be able to compute t_2.

To construct subsequent components, the same steps are performed as for the two previous components. This iterative procedure is continued until the number of components to be retained is significant.

3. Performance Measures

The efficiency of the methods considered (PCR and PLSR) was evaluated using Root Mean Square Error (RMSE).

3.1 Root Mean Square Error (RMSE)

The RMSE is a measure of how well the model fits the data. It is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{15}$$

where $\hat{y}_i$ are the values of the predicted variable when all samples are include in the model formation, and n is the number of observations.

4. Simulation Study

In this section, the efficiencies of the PLSR and PCR methods were investigated via a simulation study. With the R Studio program, a great number of varying groups of datasets are generated from standard normal distribution with parameters mean (μ) = 0 and variance (σ) = 1 allowing for the inclusion of different degrees of collinearities for 50 replications.

The design of the study is based on simulation work that has been performed for three different correlation levels (0.2, 0.5, 0.8), indicating weak, moderate and strong relationship between the

predictor variables, five (5) number of variables and two different sample sizes (100 and 250). The two prediction regression methods were applied to the generated data.

The Root Mean Square Error (RMSE) value of the parameter estimates for each of these models was calculated to compare the performance of the regression methods employed in this study. Variance Inflation Factor (VIF) was also used to check the presence of multicollinearity in the data simulated. Sensitivity analysis was also performed to determine the number of the PLS components that are worth keeping so as to avoid over-fitting. Therefore the results of simulations are listed below:

Table 1: Evaluation of the effectiveness of PLSR using RMSE

Sample size	Number of Predictor variables	Multicollinearity level	Number of components	PLSR
100	5	Low correlation (0.2)	4	1.074944
		Moderate correlation (0.5)	3	0.9703781
		High correlation (0.8)	2	1.036041
250		Low correlation (0.2)	2	1.051946
		Moderate correlation (0.5)	1	1.027885
		High correlation (0.8)	2	1.152918

From the results of Table 1, it has been observed that partial least squares regression has high predictive abilities at moderate correlation of the two sample sizes considered, which means

that PLSR performed better when there is a moderate correlation between the predictor variables with for the sample sizes considered in this study.

Table 2: Comparison between PLSR and PCR using RMSE

Sample size	Number of Predictor variables	Multicollinearity level	Number of components for PLSR	PLSR	Number of components for PCR	PCR
100	5	Low correlation (0.2)	4	1.074944	5	1.074941
		Moderate correlation (0.5)	3	0.9703781	2	1.039347
		High correlation (0.8)	2	1.036041	1	1.052906
250		Low correlation (0.2)	2	1.051946	5	1.050641
		Moderate correlation (0.5)	1	1.027885	1	1.027903
		High correlation (0.8)	2	1.152918	2	1.149282

From the results of Table 2, it has been observed that partial least squares regression has high predictive abilities at the various levels of multicollinearity, and the sample sizes considered in this study, which means that PLSR performed better than the principal component regression (PCR).

CONCLUSION

In conclusion, our analysis demonstrates that partial least squares regression is a valuable technique for addressing multicollinearity in regression analysis. The findings contribute to a better understanding of multicollinearity mitigation strategies and have implications for improving the accuracy and reliability of regression models in various fields. The study also demonstrated that PLSR is a more effective method than PCR in handling multicollinearity in regression analysis. PLSR's ability to handle multiple dependent variables and its robustness to multicollinearity make it a suitable choice for datasets with highly correlated predictor variables. While PCR is a useful dimensionality reduction technique, its limitations in handling multicollinearity make PLSR a preferred choice in such scenarios.

REFERENCES

Abdi, H. (2003). Partial least squares (PLS) regression. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia of social science research methods* (Vol. 3, pp. 792-795).

Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, *89*(425), 122–127. https://doi.org/10.1080/01621459.1994.10476452

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology*, 73: 592-616.

Gujarati, D. N. (2003). Basic econometrics. *4th Ed., McGraw Hill, New York.*

Hair, J., Anderson, R., Tatham, R., & Black, W. (1998). Multivariate data analysis. *5th Ed, Prentice Hall, New Jersey*

Lukman, A. F., Adewuyi, E. T., Alqasem, O. A., Arashi, M., & Ayinde, K. (2024). Enhanced Model Predictions through Principal Components and Average Least Squares-Centered Penalized Regression. *Symmetry*, 16(4), 469.

Neter, J., Wasserman, W., & Kutner, M. (1990). Applied linear regression models. *3rd Ed. IRWIN Book Team USA.*

Wold, S. (1975). Partial least squares. *Encyclopedia of Statistical Sciences*, 6, 581-591.