# PRINCIPAL COMPONENT REGRESSION FOR SOLVING MULTICOLLINEARITY PROBLEM

**M.C. Alibuhtto[1], T.S.G. Peiris[2]**
[1]Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka
[2]Department of Mathematics, Faculty of Engineering, University of Moratuwa, Sri Lanka
mcabuhtto@seu.ac.lk and sarathp@mrt.ac.lk

**ABSTRACT:** Multicollinearity often causes a huge explanatory problem in multiple linear regression analysis. In presence of multicollinearity the ordinary least squares (OLS) estimators are inaccurately estimated. In this paper the multicollinearity was detected by using observing correlation matrix, variance influence factor (VIF), and eigenvalues of the correlation matrix. The simulation multicollinearity data were generated using MINITAB software and make comparison between methods of principal component regression (PCR) and the OLS methods. According to the results of this study, we found that PCR method facilitates to solve the multicollinearity problem.

**Keywords:** Linear Regression, Multicollinearity, Variance Influence Factor, Simulation.

## 1. INTRODUCTION

Multiple linear regressions is a widely used statistical technique that allows us to estimate models that describe the distribution of a response variable with the help of a two or more explanatory variables. The use of multiple regression mainly regards the interpretation of the regression coefficients. In case of independent coefficients the least-squares solution gives stable estimates and useful results.

Multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables. When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients. It will result in incorrect conclusions about the relationship between outcome variable and predictor variables. (Gujarat, 2004)

The presence of multicollinearity has several serious effects on the OLS estimates of regression coefficients such as high variance of coefficients may reduce the precision of estimation, it can result in coefficients appearing to have the wrong sign, the parameter estimates and their standard errors become extremely sensitive to slight changes in the data points and it tends to inflate the estimated variance of predicted values (Montgomery, 2001). Because multicollinearity is a serious problem when we are working for predictive models. So it is very important for us to find a better method to deal with multicollinearity.

A number of different techniques for solving the multicollinearity problem have been developed. These range from simple methods based on principal components to more specialized techniques for regularization (Næs and Indahl, 1998). The PCR method has been proposed as alternatives to the OLS estimators when the independent assumption has not been satisfied in the analysis. Through this study, we want to compare OLS and PCR methods by using Monte Carlo simulation data.

## 2. METHODOLOGY

### 2.2 Data

In this paper, the simulation data (50 observations) were generated using Minitab software, where the correlation coefficients between the predictor variables are large ($\rho = 0.95 \; and \; \rho = 0.99$) and the number of independent variables is five. The simulation procedure suggested by McDonaldand Galarneau (1975) and Gibbons (1981) was used to generate the explanatory variables:

$$X_{ij} = \left(1 - \rho^2\right)^{\frac{1}{2}} Z_{ij} + \rho Z_{ip} \; i = 1,2,...,n \quad and \quad j = 1,2,..., p \qquad (1)$$

Where $Z_{ij}$ are independent standard normal distribution, $\rho^2$ is the correlation between any two explanatory variables and p is the number of explanatory variables.

### 2.3 Detection of Multicollinearity

#### 2.3.1 Examination of Correlation Matrix

A simple method for detecting multicollinearity is to calculate the correlation coefficients between any two of the explanatory variables. A high value of the correlation between two variables may indicate that the variables are collinear. This method is easy, but it cannot produce a clear estimate of the degree of multicollinearity. (El-Dereny and Rashwan, 2011). The correlation coefficients are greater than 0.80 or 0.90 then this is an indication of multicollinearity.

#### 2.3.2 Variance Inflation Factor (VIF)

The VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. Let $R_j^2$ denote the coefficient of determination when $X_j$ is regressed on all other predictor variables in the model. The VIF is given by:

$$VIF = \frac{1}{1 - R_j^2} \qquad j = 1,2,3...p - 1 \qquad (2)$$

The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity. As per practical experience, if any of the VIF values exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery, 2001).

#### 2.3.3 Eigen Analysis of Correlation Matrix

The eigenvalues can also be used to measure the presence of multicollinearity. If multicollinearity is present in the predictor variables, one or more of the eigenvalues will be small (near to zero).

Let $\lambda_1, \lambda_2, ..., \lambda_p$ be the eigenvalues of correlation matrix. The condition number of correlation matrix is defined as:

$$K = \frac{\lambda_{\max}}{\lambda_{\min}} \text{ and } K_j = \frac{\lambda_{\max}}{\lambda_j} \qquad j = 1,2,..., p \qquad (3)$$

Where $\lambda_{\max}$ is the largest eigenvalue.

$\lambda_{\min}$ is the smallest eigenvalue

$\lambda_j$ is the eigenvalue of jth independent variable

If the condition number is less than 100, there is no serious problem with multicollinearity and if a condition number is between 100 and 1000 implies a moderate to strong multicollinearity. Also, if the condition number exceeds 1000, severe multicollinearity is indicated (Montgomery, 2001).

### 2.4 Principal Component Regression(PCR)

The PCR provides a unified way to handle multicollinearity which requires some calculations that are not usually included in standard regression analysis. The principle component analysis follows from the fact that every linear regression model can be restated in terms of a set of orthogonal explanatory variables. These new variables are obtained as linear combinations of the original explanatory variables. They are referred to as the principal components.

Consider the following model,

$$Y = X\beta + \varepsilon \qquad (4)$$

Where Y is an n x 1 matrix of response variable, X is an n x p matrix of the independent variables, $\beta$ is a p x 1 vector of unknown constants, and $\varepsilon$ is an n x 1 vector of random errors.

There exists a matrix C, satisfying

$$C'(X'X)C = \Lambda \;\; and \;\; C'C = CC' = 1 \qquad (5)$$

Where $\Lambda$ is a diagonal matrix with ordered characteristics roots of X′X on the diagonal. The characteristic roots are denoted by $\lambda_1 \geq \lambda_2 \geq ..., \geq \lambda_p$ C may be used to calculate a new set of explanatory variables, namely

$$\left(Z_{(1),} Z_{(2),} ..., Z_{(p),}\right) = Z = XC = \left(X_{(1),} X_{(2),} ..., X_{(p),}\right) \qquad (6)$$

That are linear functions of the original explanatory variables. The Z’s are referred to as principal components.

Thus the regression model can be restated in terms of the principal components as:

$$Y = Z\alpha + \varepsilon \text{ , where } Z = XC, \; \alpha = C\beta \qquad (7)$$

$$Z'Z = C'X'XC = C'C\Lambda C'C \qquad (8)$$

The least square estimator of $\alpha$ is

$\hat{\alpha} = (Z'Z)^{-1} Z'Y = \Lambda^{-1} Z'Y$ and the variance covariance matrix of $\hat{\alpha}$ is

$$V(\hat{\alpha}) = \sigma^2 (Z'Z)^{-1} = \sigma^2 \Lambda^{-1} \qquad (9)$$

Thus a small eigenvalue of X′X implies that the variance of the corresponding regression coefficient will be large.

Since $Z'Z = C'X'XC = C'C\Lambda C'C = \Lambda$, we often refer to the eigenvalues $\lambda_j$ as the variance of the j$^{th}$principal component. If all $\lambda_j$ equal to unity, the original regressors are orthogonal, while if a $\lambda_j$ is exactly equal to zero, then it implies a perfect linear relationship between the original regressors. One or more near to zero implies thatmulticollinearity is present.

The principal component regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimators, assume that the regressors are arranged in order of decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq ..., \geq \lambda_p > 0$. In principal components regression the principal components corresponding to near zero eigenvalues are removed from the analysis and least squares applied to the remaining components.

## 3. RESULTS AND DISCUSSIONS

### 3.1 Detection of Multicollinearity

The correlation matrix based on a set of simulated data are given in table 1.

*Table 1. Correlation matrix of independent variables*

| $\rho = 0.95$ | | | | | |
|---|---|---|---|---|---|
| Variables | X1 | X2 | X3 | X4 | X5 |
| X1 | 1.0000 | 0.9509 | 0.9496 | 0.9599 | 0.9384 |
| X2 | 0.9509 | 1.0000 | 0.9379 | 0.9460 | 0.9367 |
| X3 | 0.9496 | 0.9379 | 1.0000 | 0.9452 | 0.9513 |
| X4 | 0.9599 | 0.9460 | 0.9452 | 1.0000 | 0.9302 |
| X5 | 0.9384 | 0.9367 | 0.9513 | 0.9302 | 1.0000 |
| $\rho = 0.99$ | | | | | |
| X1 | 1.0000 | 0.9876 | 0.9878 | 0.9914 | 0.9884 |
| X2 | 0.9876 | 1.0000 | 0.9882 | 0.9866 | 0.9821 |
| X3 | 0.9878 | 0.9882 | 1.0000 | 0.9871 | 0.9869 |
| X4 | 0.9914 | 0.9866 | 0.9871 | 1.0000 | 0.9844 |
| X5 | 0.9884 | 0.9821 | 0.9869 | 0.9844 | 1.0000 |

Table 1 shows the correlation between independent variables are highly correlated. This implies that the multicollinearity exits. This results further confirmed by VIF and Eigen values structure and the results are given in table 2 & 3.

*Table 2. VIF values of independent variables*

| Variables | VIF | |
|---|---|---|
| | $\rho = 0.95$ | $\rho = 0.99$ |
| X1 | 18.76 | 91.90 |
| X2 | 13.95 | 58.03 |
| X3 | 16.05 | 68.85 |
| X4 | 16.21 | 71.80 |
| X5 | 13.14 | 54.28 |

Table 2shows the VIF of each independent variables is greater than 10 in two different correlation coefficients which implies that the multicollinearity exist.

*Table 3. Results of Eigen analysis*

| Variables | $\rho = 0.95$ | | $\rho = 0.99$ | |
|---|---|---|---|---|
| | $\lambda_j$ | Kj | $\lambda_j$ | Kj |
| X1 | 4.7785 | 1.00 | 4.9482 | 1.00 |
| X2 | 0.0796 | 60.06 | 0.0183 | 270.72 |
| X3 | 0.0593 | 80.65 | 0.0150 | 328.94 |
| X4 | 0.0434 | 110.03 | 0.0108 | 456.77 |
| X5 | 0.0392 | 121.76 | 0.0076 | 649.00 |

From the table 3, the corresponding condition indices are large in two different data. This indicates that there is multicollinearity exist.

According to the above results, there is multicollinearity exist in the independent variables. The OLS estimates of two different types of multicollinearity data are given in table 4.

*Table 4. Results of multiple regression models*

| Variables | $\rho = 0.95$ | | | | $\rho = 0.99$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}$ | SE of $\hat{\beta}$ | t-values | p-values | $\hat{\beta}$ | SE of $\hat{\beta}$ | t-values | p-values |
| C | -0.012 | 0.050 | -0.24 | 0.815 | -0.005 | 0.024 | -0.19 | 0.849 |
| X1 | 0.361 | 0.166 | 2.17 | 0.035 | 0.435 | 0.170 | 2.55 | 0.014 |
| X2 | -0.090 | 0.154 | -0.58 | 0.513 | 0.196 | 0.138 | 1.42 | 0.163 |
| X3 | 0.358 | 0.153 | 2.34 | 0.024 | 0.069 | 0.144 | 0.48 | 0.632 |
| X4 | 0.325 | 0.155 | 2.10 | 0.042 | 0.374 | 0.147 | 2.54 | 0.015 |
| X5 | 0.024 | 0.129 | 0.19 | 0.853 | -0.084 | 0.147 | -0.64 | 0.527 |
| S = 0.3321 R-Sq(adj) = 92.9% F=75.80 (0.000) | | | | | S = 0.1541 R-Sq(adj) = 98.5% F=626.25(0.000) | | | |

Table 4 shows the overall models of both simulated data is significant at 5% significance level. However, only three independent (X1, X3, and X4) variables are statistically significant in the first model and two independent (X1 and X4) variables are statistically significant in the second model and the other variables are not statistically significant because of multicollinearity.

### 3.2 Principal Component Regression

The principal components technique can be used to reduce multicollinearity in the estimation data. The reduction is accomplished by using less than the full set of principal components to explain the variation in the response variable.

*Table 5.Eigenvalues and eigenvectors* $(\rho = 0.95)$

| Variables | Eigenvalues of the Correlation Matrix | | | |
|---|---|---|---|---|
| | Eigen value | Difference | Proportion | Cumulative |
| X1 | 4.7785 | 4.6989 | 0.9557 | 0.9557 |

| | | | | | |
|---|---|---|---|---|---|
| X2 | 0.0796 | 0.0203 | 0.0159 | 0.9716 | |
| X3 | 0.0593 | 0.0158 | 0.0119 | 0.9835 | |
| X4 | 0.0434 | 0.0042 | 0.0087 | 0.9922 | |
| X5 | 0.0392 | | 0.0078 | 1.0000 | |
| Eigenvectors | | | | | |
| Variables | Z1 | Z2 | Z3 | Z4 | Z5 |
| X1 | 0.4491 | -0.3109 | -0.1826 | 0.1943 | -0.7942 |
| X2 | 0.4465 | -0.2756 | 0.7819 | -0.3206 | 0.1028 |
| X3 | 0.4477 | 0.3518 | -0.4251 | -0.7024 | 0.0411 |
| X4 | 0.4475 | -0.4616 | -0.3686 | 0.3119 | 0.5946 |
| X5 | 0.4451 | 0.7004 | 0.1975 | 0.5185 | 0.0592 |

From the table 5, the principal components of the explanatory variables are:

$$Z_1 = 0.4491X_1 + 0.4465\mathrm{X}_2 + 0.4477X_3 + 0.4475X_4 + 0.4451X_5$$

$$Z_2 = -0.3109X_1 - 0.2756\mathrm{X}_2 + 0.3518X_3 - 0.4616X_4 + 0.7004X_5$$

$$Z_3 = -0.1826X_1 + 0.7819\mathrm{X}_2 - 0.4251X_3 - 0.3686X_4 + 0.1975X_5$$

$$Z_4 = 0.1943X_1 - 0.3206\mathrm{X}_2 - 0.7024X_3 + 0.3119X_4 + 0.5185X_5$$

$$Z_5 = 0.4451X_1 + 0.7004\mathrm{X}_2 + 0.1975X_3 + 0.5185X_4 + 0.0592X_5$$

Then the model can be written in the form of principal components as:

$$Y = \alpha_1 Z_1 + \alpha_2 \mathrm{Z}_2 + \alpha_3 Z_3 + \alpha_4 Z_4 + \alpha_5 Z_5 + \varepsilon \quad \text{———————} \quad (10)$$

Also table 5 indicates that the first component accounts for 95.57 % of the variance. All remaining components are not significant. Hence, the first component has been chosen. Then the linear regression of Y against $Z_1$is given by.

$$Y = \alpha_1 Z_1 + \varepsilon \text{———————} \qquad (11)$$

The estimated value of $\alpha$ can be obtaining by the equation (11) and the results are given in table 6.

*Table 6. Results of principal component regression ( $\rho = 0.95$)*

| Variables | $\hat{\alpha}$ | SE of $\hat{\alpha}$ | t-values | p-values | VIF |
|---|---|---|---|---|---|
| C | -0.02397 | 0.0485 | -0.49 | 0.624 | - |
| Z1 | 0.4419 | 0.0181 | 24.46 | 0.000 | 1.000 |
| S = 0.3425R-Sq = 92.6% R-Sq(adj) = 92.4% F=598.09(0.000) | | | | | |

According to the table 6, selecting a model based on first principal component $Z_1$ has removed the multicollinearity.

*Table 7. Eigenvalues and eigenvectors* $(\rho = 0.99)$

| Variables | Eigenvalues of the Correlation Matrix | | | |
|---|---|---|---|---|
| | Eigen value | Difference | Proportion | Cumulative |
| X1 | 4.9482 | 4.9299 | 0.9896 | 0.9896 |
| X2 | 0.0183 | 0.0032 | 0.0037 | 0.9933 |
| X3 | 0.0150 | 0.0042 | 0.0030 | 0.9963 |
| X4 | 0.0108 | 0.0032 | 0.0022 | 0.9985 |
| X5 | 0.0076 | | 0.0015 | 1.0000 |

| Eigenvectors | | | | | |
|---|---|---|---|---|---|
| Variables | Z1 | Z2 | Z3 | Z4 | Z5 |
| X1 | 0.4479 | 0.0659 | -0.4109 | 0.2117 | -0.7625 |
| X2 | 0.4469 | -0.6026 | 0.3569 | 0.5314 | 0.1654 |
| X3 | 0.4474 | -0.0455 | 0.5082 | -0.7038 | -0.2104 |
| X4 | 0.4473 | -0.1886 | -0.6405 | -0.3165 | 0.5038 |
| X5 | 0.4466 | 0.7713 | 0.1874 | 0.2701 | 0.3052 |

From the table 7, the principal components of the explanatory variables are:

$$Z_1 = 0.4479X_1 + 0.4469\mathrm{X}_2 + 0.4474X_3 + 0.4473X_4 + 0.4466X_5$$

$$Z_2 = 0.0659X_1 - 0.6026\mathrm{X}_2 - 0.0455X_3 - 0.1886X_4 + 0.7713X_5$$

$$Z_3 = -0.4109X_1 + 0.3569\mathrm{X}_2 + 0.5082X_3 - 0.6405X_4 + 0.1874X_5$$

$$Z_4 = 0.2117X_1 + 0.5314\mathrm{X}_2 - 0.7038X_3 - 0.3165X_4 + 0.2701X_5$$

$$Z_5 = -0.7625X_1 + 0.1654\mathrm{X}_2 - 0.2104X_3 + 0.5038X_4 + 0.3052X_5$$

Then the model can be written in the form of principal components as:

$$Y = \alpha_1 Z_1 + \alpha_2 \mathrm{Z}_2 + \alpha_3 Z_3 + \alpha_4 Z_4 + \alpha_5 Z_5 + \varepsilon \qquad (12)$$

Also table 7 indicates that the first component accounts for 98.96 % of the variance. All remaining components are not significant. Hence, the first component has been chosen. Then a linear regression of Y against $Z_1$ is given by.

$$Y = \alpha_1 Z_1 + \varepsilon \qquad (13)$$

The estimated value of $\alpha$ can be obtaining by the equation (13) and the results are given in table 8.

*Table 8. Results of principal component regression ( $\rho = 0.99$)*

| Variables | $\hat{\alpha}$ | SE of $\hat{\alpha}$ | t-values | p-values | VIF |
|---|---|---|---|---|---|
| C | 0.0045 | 0.0229 | 0.19 | 0.847 | - |
| Z1 | 0.4439 | 0.0083 | 53.26 | 0.000 | 1.000 |
| S = 0.1617 R-Sq = 98.3% R-Sq(adj) = 98.3% F= 2836.87(0.000) | | | | | |

According to the table 8, selecting a model based on first principal component $Z_1$ has removed the multicollinearity.

## 4. CONCLUSIONS

Multicollinearity often causes a huge explanatory problem in multiple linear regression analysis. When multicollinearity is present in the data, ordinary least square estimators are inaccurately estimated. If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. According to the results of this study the multicollinearity was detected using examination of correlation matrix, calculating the variance inflation factor (VIF), Eigen value analysis and the remedial measures of principal component analysis helps to solve the problem of multicollinearity.

## 5. REFERENCES

EL-DERENY, M. AND RASHWAN,N.I., (2011), Solving Multicollinearity Problem Using Ridge Regression Models, Int.J.Contemp. Math. Sciences, Vol.6, No.12:pp.585-600.

GUJRATI, D. N. (2004). Basic econometrics 4th edition, Tata McGraw-Hill, New Delhi.

MONTGOMERY, D. C., PECK, E. A., VINING, G. G. (2001). Introduction to linear regression analysis, 3rd edition, Wiley, New York.

MCDONALD, G. AND GALARNEAU,D.(1975). A Monte Carlo evaluation of some ridge type estimators, Journal of the American statistics association, 70, pp.407-416.

NÆS, T. AND INDAHL, U. (1998). A unified description of classical classification methods for multicollinear data. J. Chemometrics. 12, pp.205-220.