PRINCIPAL COMPONENT REGRESSION FOR SOLVING MULTICOLLINEARITY PROBLEM

M.C. Alibuhtto[1] and 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, 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 themulticollinearity problem.

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

INTRODUCTION

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. The objective of this paper is to compare OLS and PCR methods to solve multicollinearity problems using the Monte Carlo simulation data.

METHODOLOGY

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.

Detection of Multicollinearity

The following methods have been used to detect the multicollinearity.

Observing correlation matrix

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).

Variance influence 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:

$$VIF = \frac{1}{1-R_j^2} \qquad j = 1,2,3 \ldots p-1$$ (Montgomery, 2001)

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).

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.

RESULTS AND DISCUSSIONS

Detection of Multicollinearity

The correlation matrix based on a set of simulated data were given in table1.

Table1: 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 2 shows the VIF each independent variables is greater than 10 in two different correlation coefficients which implies that the multicollinearity exist.

Table3: Results of Eigen analysis

Variables	$\rho = 0.95$		$\rho = 0.99$	
	λ_j	Kj	λ_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 between independent variables.

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.

Table4: 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	-0.0118	0.0502	-0.24	0.815	-0.0045	0.0237	-0.19	0.849
X1	0.3610	0.1664	2.17	0.035	0.4345	0.1702	2.55	0.014
X2	-0.0896	0.1538	-0.58	0.513	0.1959	0.1380	1.42	0.163
X3	0.3579	0.1528	2.34	0.024	0.0695	0.1440	0.48	0.632
X4	0.3253	0.1550	2.10	0.042	0.3743	0.1473	2.54	0.015
X5	0.0241	0.1293	0.19	0.853	-0.0836	0.1473	-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 modelsof 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 (X1and X4) variables are statistically significant in the second model and other variables are not statistically significant because of multicollinearity.

Principal Component Regression

The principal components technique can be used to reduce multicollinearity in the estimation data.

Table 5: Eigen values and eigenvectors

Variables	Eigen values of the Correlation Matrix $(\rho = 0.95)$			Eigen values of the Correlation Matrix $(\rho = 0.99)$		
	Eigen value	Proportion	Cumulative	Eigen value	Proportion	Cumulative
X1	4.7785	0.9557	0.9557	4.9482	0.9896	0.9896
X2	0.0796	0.0159	0.9716	0.0183	0.0037	0.9933
X3	0.0593	0.0119	0.9835	0.0150	0.0030	0.9963
X4	0.0434	0.0087	0.9922	0.0108	0.0022	0.9985
X5	0.0392	0.0078	1.0000	0.0076	0.0015	1.0000

Variables	Eigenvectors					Eigenvectors				
	Z_1	Z_2	Z_3	Z_4	Z_5	Z_1	Z_2	Z_3	Z_4	Z_5
X1	0.449	-0.311	-0.183	0.194	-0.794	0.448	0.066	-0.411	0.212	-0.763
X2	0.447	-0.276	0.782	-0.321	0.103	0.447	-0.603	0.357	0.531	0.165
X3	0.448	0.352	-0.425	-0.702	0.041	0.447	-0.046	0.508	-0.704	-0.210
X4	0.448	-0.462	-0.369	0.312	0.595	0.447	-0.189	-0.641	-0.317	0.504
X5	0.445	0.700	0.198	0.519	0.059	0.447	0.771	0.187	0.270	0.305

From the table 5, the first principal components of the explanatory variables of both simulated data are given below.

$$Z_1 = 0.449X_1 + 0.447X_2 + 0.448X_3 + 0.448X_4 + 0.445X_5$$
$$Z_1 = 0.448X_1 + 0.447X_2 + 0.447X_3 + 0.447X_4 + 0.447X_5$$

Also table 5 indicates that the first component accounts for 95.57% variance by the first model and 98.96% of the variance accounts by the second model. All remaining components are not significant. Hence, the first components have been chosen in two models. Then the linear regression of Y against Z_1 is given by.

$Y = \alpha_1 Z_1 + \varepsilon$ (a)

The estimated value of α can be obtaining by the equation (a) and the results are given in table 6.

Table6: Results of principal component regressions

Variables	$\rho = 0.95$				$\rho = 0.99$				Both
	$\hat{\alpha}$	SE of $\hat{\alpha}$	t-values	p-values	$\hat{\alpha}$	SE of $\hat{\alpha}$	t-values	p-values	VIF
C	-0.024	0.049	-0.49	0.624	0.005	0.023	0.19	0.847	-
Z1	0.442	0.018	24.46	0.000	0.444	0.008	53.26	0.000	1.000
S = 0.3425 R-Sq(adj) = 92.4% F=598.09(0.000)					S = 0.1617 R-Sq(adj) = 98.3% F=2836.87(0.000)				

According to the table 6, selecting a model based on first principal component Z1 has removed the multicollinearity in both models.

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 goalis 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 thevariance inflation factor (VIF), Eigen value analysis and the remedial measures of principal component analysis helps to solve theproblem of multicollinearity.

REFERENCES

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

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, 407-416.