Journal of Engineering and Applied Sciences 13 (2): 286-290, 2018
ISSN: 1816-949X

Overcoming the Multicollinearity by Using Principal Component Regression in Economic Growth Model

[1]Julita Nahar, [1]Sri Purwani, [1]Sudradjat Supian and [2]Fatimah Khonsa Syahidah
[1]Department of Mathematics, [2]Department of Statistics,
Padjadjaran University, Bandung, Indonesia

Abstract: Development of a developing country is essentially aimed at improving the welfare and prosperity of its people. National or regional development puts more emphasis on development in the economic field. In the implementation of their economic development factors that influence it should be considered. One measure of the success can be seen from the economic growth. In modeling the economic growth we are often constrained by models that do not meet the assumptions, one of which is multicollinearity. This occurs because the data obtained is taken from uncontrollable circumstances. The existence of these cases can cause difficulty in separating the influence of each independent variable on the response variable, so, we need a method to solve it. One method that can be used is Principal Component Regression (PCR). PCR is one method that has been developed to overcome the problem of multicollinearity. PCR is a regression analysis of the variables in response to the principal components that are not correlated with each other, where each principal component is a linear combination of all predictor variables.

Key words: Principal Component Regression (PCR), economic growth, multicollinearity, prosperity, assumptions, emphasis

INTRODUCTION

Development of a country cannot be separated between national development and regional development which is essentially intended for the community to improve the welfare and prosperity of society. National or regional development emphasizes on development in the economic field. In the implementation of economic development, its high growth is the main target for every region. One measure of the success of economic development can be seen from the economic growth in every region.

In modelling the economic growth we are often constrained by models that do not meet the assumptions, one of which is multicollinearity. This occurs because the data obtained is taken from uncontrollable circumstances. These cases can cause difficulty in separating the influence of each independent variable (X) on the response variable (Y), so, we need a method to solve it. One method that can be used is Principal Component Regression (PCR).

Multicollinearity was initially introduced by Frisch which means that the linear relationship is "perfect" or "almost perfect" among some or all of the independent variables of the regression model. According to Dewi (2014), multicollinearity was caused by the method of data collection employed, the specification model and the overdetermined model, a situation where in a particular estimation model, the number of explanatory variables is more than the amount of data (observations).

Detection of multicollinearity: To detect the existence of multicollinearity in the regression model (Imam, 2013) are as.

Through value t, R^2 and test F: If coefficient of determination R^2 is large and test F shows are not significant result, however most or even all predictor variables individually are not insignificant, then there is possibility of multicollinearity in the data.

Analyze the correlation matrix: If between two or more predictor variables have a fairly high correlation, usually above 0.9 then, it indicates the occurrence of multicollinearity.

VIF (Variance Inflation Factor): Variance Inflation Factor (VIF) is one way to detect the presence of multicollinearity. VIF expressed by the equation:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (1)$$

Corresponding Author: Julita Nahar, Department of Mathematics, Padjadjaran University, Bandung, Indonesia

where, R_j^2 is the coefficient of determination of predictor variables regressed against other response variables. Regression multicollinearity can be seen when VIF = 10.

TOL (Tolerance): In addition to using the VIF, multicollinearity can be detected by looking at the value of Tolerance (TOL). The value of TOL can be found using the following equation:

$$TOL_j = \frac{1}{VIF_j} \quad (2)$$

In other words, TOL is the opposite of VIF. In general, values used to indicate the presence of multicollinearity are TOL = 0.01. If A is an n×n matrix, there is a nonzero scalar λ and vector, so that, it meets the following equation:

$$AV = \lambda V$$

λ is the Eigenvalues of A and V Eigenvectors associated with Eigenvalues λ. To obtain the Eigenvalues equation rewritten as:

$$AV = \lambda V; \text{ where } V \neq 0$$

$$AV = \lambda V = 0, AV = \lambda I = 0, (A-\lambda I)V = 0$$

In order λ be Eigenvalues, then there should be no zero solution of the equation. The equation will have a non zero solution if and only if:

$$\det (A-\lambda I) = 0 \text{ or } |A-\lambda I| = 0$$

There are several methods that can be used to overcome multicollinearity including Partial Least Squares (PLS), ridge regression and PCR. In this case, the method used is PCR conducted by using Statistical Software R.

Principal Component Regression (PCR): PCR is a regression analysis of the response variables in relation to the principal components that are not correlated with each other where each principal component is a linear combination of all predictor variables (Draper and Smith, 1992). It is a technique that combines regression analysis with the Principal Component Analysis (PCA).

Regression analysis is used to determine whether there is a relationship between the response and predictor variables (Sembiring, 2003) whereas PCA is basically aimed at simplifying the observed variables in a way to shrink (reduce) its dimensions. This is done by eliminating the correlation between the variables through the transformation of the initial variable to a new variable (a linear combination of the initial variables) that are not correlated. Of p initial variable can be formed p principal components, selected k principal components (k<p) which has been able to explain the diversity of the data which is high enough (between 80-90%) (Johnson and Wichern, 2002). The k principal components selected can replace the p initial variables without reducing the information.

There are two ways of forming regression of the principal component through principal component analysis that is the principal component formed based on the covariance matrix and based on the correlation matrix. The covariance matrix is used if all the observed variables have the same unit of measurement. Whiles correlation matrix of data that has been standardized (standard form Z) is used if the observed variables do not have the same unit of measurement. Regression of the principal component can be expressed as:

$$Y = w_0 + w_1K_1 + w_2K_2 + ... + \varepsilon \quad (3)$$

Where:
Y = Response variable
K = Principal components
w = Parameter of principal component regression

$K_1, K_2, K_3, \ldots, K_m$ shows the principal components that were included in the regression analysis of principal components where m is smaller than p. The principal component is a linear combination of raw variables Z, so that:

$$\begin{aligned} K_1 &= a_{11}Z_1 + a_{21}Z_2 + ... + a_{p1}Z_p \\ K_2 &= a_{12}Z_1 + a_{22}Z_2 + ... + a_{p2}Z_p \\ &\vdots \\ K_m &= a_{1m}Z_1 + a_{2m}Z_2 + ... + a_{pm}Z_p \end{aligned} \quad (4)$$

If $K_1, K_2, \ldots, K_m$ are substituted into Eq. 3 this gives:

$$\begin{aligned} \hat{Z}_Y &= w_0 + w_1(a_{11}Z_1 + a_{21}Z_2 + ... + a_{p1}Z_p) + \\ &w_2(a_{12}Z_1 + a_{22}Z_2 + ... + a_{p2}Z_p) + ... + \\ &w_m(a_{1m}Z_1 + a_{2m}Z_2 + ... + a_{pm}Z_p) + \varepsilon = \\ &w_0 + w_1a_{11}Z_1 + w_1a_{21}Z_2 + ... + w_1a_{p1}Z_p + \\ &w_2a_{12}Z_1 + w_2a_{22}Z_2 + ... + w_2a_{p2}Z_p + ... + \\ &w_ma_{1m}Z_1 + w_ma_{2m}Z_2 + ... + w_ma_{pm}Z_p + \varepsilon = \\ &w_0 + (w_1a_{11} + w_2a_{12} + ... + w_ma_{1m})Z_1 + \\ &w_1a_{21} + w_2a_{22} + ... + w_ma_{2m})Z_2 + ... + \\ &w_1a_{p1} + w_2a_{p2} + ... + w_ma_{pm})Z_p + \varepsilon \end{aligned} \quad (5)$$

Equation 5 is simplified as:

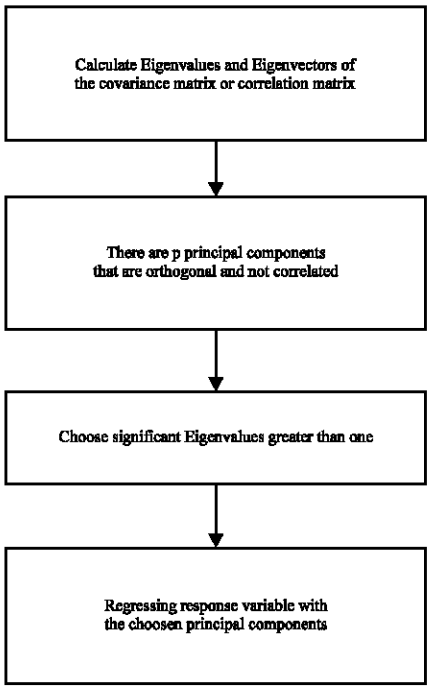

Fig. 1: The stages of PCR method

$$\hat{Z}_Y = b_0 + b_1Z_1 + b_2Z_2 + ... + b_pZ_p \quad (6)$$

Where:

$$b_0 = \hat{w}_0$$

$$b_1 = \hat{w}_1a_{11} + \hat{w}_2a_{12} + ... + \hat{w}_ma_{1m}$$

$$\vdots$$

$$b_p = \hat{w}_1a_{p1} + \hat{w}_2a_{p2} + ... + \hat{w}_ma_{pm}$$

MATERIALS AND METHODS

The data used is secondary data with Gross Regional Domestic Product (GRDP) as the response variable (Y) and the predictor variables are the contribution of the manufacturing industry (X_1), the number of workers manufacturing industry (X_2), labour productivity manufacturing industry (X_3), investment in the manufacturing industry (X_4). The stages in the implementation of PCR are shown in Fig. 1.

RESULTS AND DISCUSSION

Linear regression equation obtained by using software R is:

$$\hat{Y} = 41.65+2.35X_1-0.25X_2+2.05X_3+1.57X_4$$

The result of calculating the correlation value with software R shown in Table 1. The correlation matrix is the simplest measure to detect the presence of multicollinearity. The calculation of correlation show that almost all the variables close to 1, this indicate that there is a correlation between each independent variable (Ohyver, 2012).

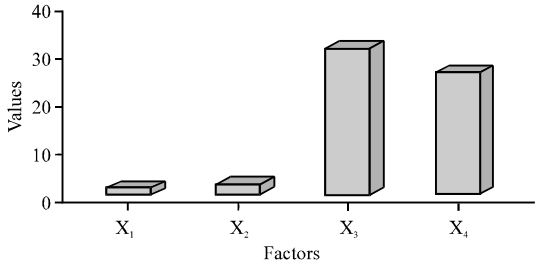

Fig. 2: The results of Variance Inflation Factor (VIF)

Table 1: The calculation of correlation matrix

Correlation	X_1	X_2	X_3	X_4
X_1	1	0.9092	0.9329	0.9690
X_2	0.9092	1	0.9519	0.8641
X_3	0.9329	0.9519	1	0.9110
X_4	0.9690	0.8641	0.9110	1

Table 2: The calculation of variance inflation factor

$VIFX_1$	X_2	X_3	X_4
1.635	2.546	31.187	25.935

Table 3: The calculation of eigen vector

	Component loadings			
Correlation	1	2	3	4
X_1	-0.50562	0.339848	0.356823	0.708190
X_2	-0.49405	-0.639270	0.506503	-0.301160
X_3	-0.50354	-0.317890	-0.781360	0.186735
X_4	-0.49670	0.612192	-0.074910	-0.610650

Table 4: The calculation of Eigenvalue

Component variances	Values
1	3.769371
2	0.162534
3	0.044236
4	0.023859

Table 5: The coefficient of determination

R^2	Adjusted R^2	SE
0.924	0.919	4.68

The initial results of implementing Variance Inflation Factor (VIF) with the software R are as Table 2 and Fig. 2. From VIF (Variance Inflation Factors) value above, there are variabels that have value greater than ten, so can be said that data have the multicollinearity. Whereas the outputs of Eigenvector and Eigenvalues are given, respectively in Table 3 and 4.

In this case, there is one principal component that has Eigenvalue greater than one, namely the first component of 3.7693 (Fig. 3). This means that the components one can explain the variance of 94.23% and

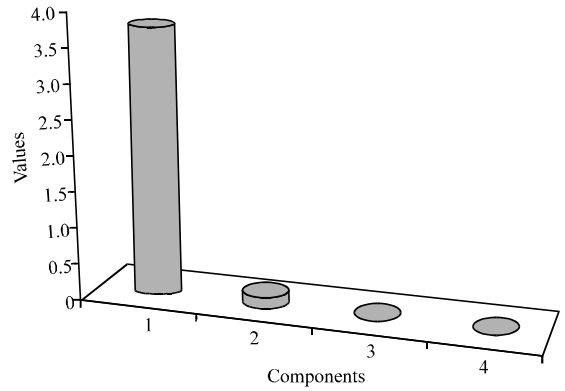

Fig. 3: Component variances

only parts 1 are selected to form the principal component regression. From the output value of Eigenvector above, obtained a model for component 1 is as:

$$K_1 = -0.5056Z_{x_1} - 0.4940Z_{x_2} - 0.5035Z_{x_3} - 0.4967Z_{x_4}$$

and the result for principle component regression model is:

$$Y = 76.3 - 3.7552\ K_1$$

If K_1 is submitted into Eq. 3, this gives:

$$\hat{Z}_Y = 76.3 - 3.76(-0.5056Z_1 - 0.4940Z_2 - 0.5035Z_3 - 0.4976Z_4)$$
$$\hat{Z}_Y = 76.3 + 1.9011Z_1 + 1.8576Z_2 + 1.8933Z_3 + 1.8712Z_4$$

The next step is testing the feasibility of a regression model to see the value of the coefficient of determination (R^2). The output of the coefficient of determination is shown in Table 5 as. From the Table 5, it can be seen adjusted R^2 of 0.919 which means that 91.9% of the variation $\hat{z}_Y$ can be explained by the principal component K_1.

Then, we re-examine whether the principal component regression contains multicollinearity or not. Based on the results obtained that the regression model output has a value of VIF<10, it can be concluded that the principal component regression model can cope the case of multicollinearity on the linear regression model. This model is still in the form of standardized estimates, so that to get the actual model it is required to be transformed back from standardized data into the initial data that is from $\hat{z}_Y$ returned back into $\hat{Y}$ with the result:

$$\hat{Z}_Y = 76.3 + 1.9011Z_1 + 1.8576Z_2 + 1.8933Z_3 + 1.8712Z_4$$

$$\hat{Y} = 76.3 + 1.9011\left(\frac{X_1 - \bar{X}}{S_{X_1}}\right) + 1.8576\left(\frac{X_2 - \bar{X}}{S_{X_2}}\right) + 1.8933\left(\frac{X_3 - \bar{X}}{S_{X_3}}\right) + 1.8712\left(\frac{X_4 - \bar{X}}{S_{X_4}}\right)$$

$$\hat{Y} = 44.0434 + 0.7530X_1 + 0.8282X_2 + 3.0861X_3 + 3.6449X_4$$

Summaries: Based on analysis above the result we get linear regression model is:

$$\hat{Y} = 41.65 + 2.35X_1 - 0.25X_2 + 2.05X_3 + 1.57X_4$$

and the result for principle component regression model is:

$$Y = 76.3 - 3.7552\ K_1$$

$$\hat{Z}_Y = 76.3 + 1.9011Z_1 + 1.8576Z_2 + 1.8933Z_3 + 1.8712Z_4$$

and if $\hat{z}_Y$ returned back into $\hat{Y}$ give the result:

$$\hat{Y} = 44.0434 + 0.7530X_1 + 0.8282X_2 + 3.861X_3 + 3.6449X_4$$

CONCLUSION

This method is useful to overcome problems of multicollinearity without need to remove independent variables with high multicollinearity. This is shown by the results of testing multicollinearity where VIF to less than ten and the result principal component regression model is not too different from linear regression model.

Using method of principal component regression as method to reduce the predictor variables, can be obtained new predictor variables that are not correlated, each independently of one another and can absorb most of the information contained in the initial variables or which could contribute to the variance of the variables.

RECOMMENDATIONS

PCR is a method in the factor analysis. This method would be better used if we first know about the factor analysis. Weakness in the factor analysis is the high subjectivity in determining factor. Therefore, in determining the number of factors it is required an additional theory to form factors appropriately. It is one of areas for our future work.

ACKNOWLEDGEMENT

We would like to thank the Academic Leadership Grant 1-1-6 (ALG) led by Prof. Dr. Sudradjat Supian for all support given.

REFERENCES

Dewi, A., 2014. Partial Least Square (PLS) and Principal Component Regression (PCR) for Linear Regression with Multicollinearity in the Case of the Human Development Index in Gunung Kidul. Yogyakarta State University, Yogyakarta, Indonesia,.

Draper, H. and H. Smith, 1992. Applied Regression Analysis. 2nd Edn., PT Gramedia Pustaka Utama, Central Jakarta, Indonesia,.

Imam, G., 2013. Multivariate Analysis Applications with IBM SPSS 21 Program Update Issue 7 PLS Regression. Diponegoro University, Semarang, Indonesia,.

Johnson, R.A. and D.W. Wichern, 2002. Applied Multivariate Statistical Analysis. Vol. 5, Prentice Hall, Upper Saddle River, New Jersey, USA.,.

Ohyver, M., 2012. Pemodelan Principal Component Regression dengan Software R. Comtech, 3: 177-185.

Sembiring, R., 2003. Regression Analysis. 2nd Edn., Bandung Institute of Technology, Bandung, Indonesia.