Available online at http://iorajournal.org/indx.php/orics/index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 3, No. 1, pp. 1-7, 2022

Comparison of Partial Least Squares Regression and Principal Component Regression for Overcoming Multicollinearity in Human Development Index Model

Ravika Dewi Samosir[1*], Deiby Tineke Salaki[2],Yohanes Langi[3]

[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Indonesia
**Corresponding author email: ravikasamosir103@student.unsrat.ac.id*

Abstract

One of the assumptions in Ordinary Least Squares (OLS) in estimating regression parameter is lack of multicollinearity. If the multicollinearity exists, Partial Least Square (PLS) and Principal Component Regression (PCR) can be used as alternative approaches to solve the problem. This research intends to compare those methods in modeling factors that influence the Human Development Index (HDI) of North Sumatra Province in 2019 obtained from the Central Bureau of Statistics. The result indicates that the PLS outperforms the PCR in term of the coefficient of determination and squared error.

Keywords: Human Development Index, Linear Regression, Multicollinearity, Partial Least Square, Principal Component Regression.

1. Introduction

Human development is defined as a process of expanding people's choices (enlarging people's choice). Human develoment index a country or region can be measured by a value named HDI. There are three basic dimensions of HDI namely a long and healthy life, knowledge, and a decent standard of living. The standard are usually interpreted in four indicators, namely life expectancy at birth, expected years of schooling, average length of schooling and per capita expenditure. Currently, the Indonesian government still struggles for various problems in the implementation of human development such as limited facilities and infrastructure, employment opportunities and poverty levels (Sagar & Najam, 1998; McGillivray& White, 1993).

HDI components or variables tend to have a fairly high correlation or contain multicollinearity (Ranis, et al., 2006; Dias, et al., 2006). The use of the least squares method to estimate multiple linear regression parameters whose data contains multicollinearity will produce estimates that are not unique and the resulting model is inaccurate (Khan, et. al., 2018; Yeniay&GÖKTAŞ, 2002; Wold, et. al., 1984).

One of the methods used to model the relationship between the dependent variable and the independent variable is called linear regression analysis. In linear regression, ordinary least squares is a method widely used to estimate the parameter model. OLS works by minimizing the sum of the squared errors. One of the assumptions in OLS in estimating regression parameter is lack of multicollinearity.There are several ways or solutions that can be used in overcoming the problem of multicollinearity, including the PLS and PCR methods.

This study utilizes PLS regression to develop a linear model for HDI data in North Sumatra in 2019 obtained from the Central Statistics Agency (BPS). PCR and OLS are also implemented as a comparison to PLS in a quiring the best model prediction for HDI. Some variables used as independent variables comprise of life expectancy at birth, expected years of schooling, mean years of school, expenditure per capita, number of residences, Puskesmas ratio, poverty resident, open unemployment rate, number of hospitals, Puskesmas accreditation percentage, number of Puskesmas that provide occupational health services and presentation of Puskesmas with standardized services.

2. Literature Review

2.1.Linear Regression

Multiple linear regression is a statistical analysis that generally used to determine the relationship between two or more independent variables and dependent variable. It mathematically can be expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{1}$$

Where,

Y_i	: dependent variable with the observation of$i$, for$i = 1, 2, \dots, n$; n: number of observations
$\beta_0, \beta_1, \beta_2 \dots, \beta_p$	: regression coefficient or regression parameter
$X_{i1}, X_{i2}, \dots, X_{ip}$	: independent variable
ε_i	: error-term

Ordinary Least Squares (OLS) regression is a generalized linear modelling technique that may be used to model a single response variable which has been recorded on at least an interval scale. OLS is a technique that is used to obtain$\hat{\beta}$. The OLS procedure minimizes:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \tag{2}$$

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{3}$$

In the OLS method there is one assumption that is used, namely there is no relationship between the independent variables which is called multicollinearity. Multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpretation (Cherchye&Abeele, 2005; Williamson, 2005).

2.2.Partial Least Square

PLS is a method for constructing predictive models when the number of factors included is very large and highly correlated (Ramzan & Khan, 2010). The PLS regression model with m components can be written as follows:

$$Y = C_1 t_1 + C_2 t_2 + \cdots + C_H t_H + \epsilon \tag{4}$$

where Y is independent variable, C_H is the regression coefficient Y to t_H , and $t_h = X_{h-1} W_h / W'_h W_h$ is the principal component of h which is not correlated with each other, $(h = 1,2, \dots, m)$provided that the PLS component$t_H$is orthogonal. The first PLS component (t_1)represents the linear combination of the variable $X_i$that is most linearly correlated with the response variable Y.

$$t_1 = W_{11} X_1 + W_{12} X_2 + \cdots + W_{1p} X_p \tag{5}$$

$$W_{1j} = \frac{cov\left(X_j, Y\right)}{\sqrt{\sum_{j=1}^{p} cov^2\left(X_j, Y\right)}} \tag{6}$$

where:

$$cov\ (X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{7}$$

Simple regression between Y and component 1(t_1), can be expressed as follows,

$$Y = C_1 t_1 + Y_1 \tag{8}$$

where, $C_1$is a regression coefficient and $Y_1$is error term.

The second PLS component (t_2) which is also linear combination of the variables X_i, is not correlated with t_i (the contribution of new information) and best explains residual Y. It represents the most linearly correlated with response Y.

$$t_2 = W_{21} X_1 + W_{22} X_2 + \cdots + W_{2p} X_p \tag{9}$$

$$W_{2j} = \frac{cov\left(X_{1j}, Y_1\right)}{\sqrt{\sum_{j=1}^{p} cov^2\left(X_{1j}, Y_1\right)}} \tag{10}$$

where, $X_{1j}$is the residual regression of $Var\ X_j$on the first component(t_1).

Multiple regression between Y, component 1 (t_1) and component 2 (t_2), is expressed as follows:

$$Y = C_1 t_1 + C_2 t_2 + Y_2 \tag{11}$$

Where C_1, C_2: regression coefficient Y_2 is residual (error).

Furthermore, if the component is not enough, building component is done until reaches a maximum of p. The component is built from $t_1$up to t_{H-1}, as in the formation of the previous component where the significant variables Y in explaining Y at $t_1, t_2, \dots, t_{H-1}$.

$$t_H = W_{H1}X_1 + W_{H2}X_2 + \cdots + W_{Hp}X_p \tag{12}$$

The calculation of the PLS component stops when there are no more independent variables building the PLS component.

2.3.Principal Component Regression

PCR is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors (Donatos&Mergos, 1991; Roozbeh&Arashi, 2016; Wang, et al., 2015).

The main component is the technique of changing most of the original correlated variables with a set of independent variables. The principal component regression model is as follows:

$$Y = w_0 + w_1K_1 + w_2K_2 + \cdots + w_mK_m \tag{13}$$

where$K_1, K_2, \dots, K_m$ are the main components used in principal component regression analysis, where the number of quantities m is smaller than the number of variables p. w_0 is a constant, $w_1, w_2, \dots, w_p$is a regression parameter, and Y is the dependent variable.

The main component is a linear combination of the standard variable (Z), namely:

$$\begin{gathered} K_1 = a_{11}Z_1 + a_{21}Z_2 + \cdots + a_{P1}Z_p \\ K_2 = a_{12}Z_1 + a_{22}Z_2 + \cdots + a_{P2}Z_p \\ \vdots \\ K_m = a_{1m}Z_1 + a_{2m}Z_2 + \cdots + a_{Pm}Z_p \end{gathered} \tag{14}$$

with the estimated principal component regression equation is as follows:

$$Y = b_0 + b_1Z_1 + b_2Z_2 + \cdots + b_pZ_p \tag{15}$$

Whereb_0 is constant, $b_1, b_2, \dots, b_p$are regression parameter, and $Z_1, Z_2, \dots, Z_p$ are \ the standardized variable.

3. Research Methodology

This study utilizes data from BPS North Sumatra .The data consist of one dependent variable, namely the Human Development Index (HDI) and some independent variables in the form of:

1. Life expectancy at birth(X_1)
2. Expected years of schooling (X_2)
3. Mean years of school(X_3)
4. Per capita spending (X_4)
5. Number of residents (X_5)
6. Puskesmasratio(X_6)
7. Povertyresident(X_7)
8. Open Unemployment Rate(X_8)
9. Number of hospitals(X_9)
10. Puskesmas accreditation percentage(X_{10})
11. Number of Puskesmas that provide occupational health services(X_{11})
12. Presentation of Puskesmas with standardized services(X_{12})

The statistical analysis is conducted by employing software R where the procedure can be itemized as follows:

1. Preparing the data
2. Performing some statistical test dealing with assumption of linear regression
3. Implementing OLS, PLS and PCR successively to find a model prediction of HDI
4. Comparing the MSE and R^2 values to conclude the best method.

4. Results And Discussion

4.1. Statistical Test

There are several statistical tests used in this analysis, namely normality test, autocorelation test, heteroscedasticity test and multicollinearity test. The normality assumption needs to be considered for validation of data presented in the literature as it shows whether correct statistical tests have been used. Based on the R output, the value of the Kolmogorov-Smirnov test is 1 and the P value is 6.66×10^{-16}. This shows that the HDI data for North Sumatra Province in 2019 is normally distributed. Next is autocorelation test, based on the R output, the Durbin-Watson (D) value is 1.45. The value of $d_L$and $d_u$with a significance level of 0.05, with number of observations is 33, and K=12, equals to d_L = 0.66 and d_u = 2.48 . Since the value of D = 1.4 means, there is no negative autocorrelation. Afterwards heteroscedasticity test, heteroscedasticity is can be detected by looking at the pattern of dots on the regression plots. If the points spread with an unclear pattern above and below the number 0 on the Y axis, there is no heteroscedasticity problem. The following output plot from the R program is shown in Figure 1.

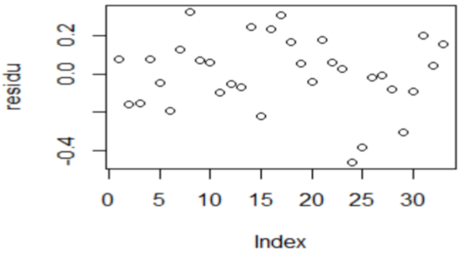

Figure1. Heteroskedasticity Plots

Based on Figure 1, it can be seen that the points spread with an unclear pattern above and below the number 0 on the Y axis, so there is no heteroscedasticity problem. The last is multicollinearity test, the multicollinearity between dependent variables can be detected by the value of VIF .The result shows the variableX_9, namely the number of hospitals contains multicollinearity as the VIF value equals to 18.81 which is greater than 10.

4.2. Multiple Regression Model

Regression analysis aimed to determine the relationship of the independent variables to the dependent variables.The linear regression equation obtained can be written as follows:

$$\begin{aligned}\hat{Y} = \ & 6.443 + 0.4283\,X_1 + 0.8153\,X_2 + 1.386\,X_3 + 1.013\,X_4 + 0.0000002142\,X_5 + 0.05401\,X_6 \\ & + 0.000003032\,X_7 + 0.004568\,X_8 - 0.03303\,X_9 + 0.009593\,X_{10} + 0.003522\,X_{11} \\ & - 0.0009773\,X_{12}\end{aligned} \tag{16}$$

Based on the equation (16), it can be seen that as independent variables $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{11}$increase the value of the HDI also increase. On the other hand, HDI will decrease as X_9 and X_{12} increase. Actually, based on the analysis, only variable X_{10} which is proved significantly influenced the HDI.

4.3. Partial Least Square

Before determining the PLS component the coefficients that will be used for the conversion of the PLS model to the linear regression model have to be found. The PLS component represents the linear combination of the X_i variables that is most linearly correlated with the Y response. Based on information criteria and Fit statistics in output R, it can be seen that R^2 started to stabilize at component 7. Therefore, it can be concluded that The PLS regression equation can be formed as follows:

$$\begin{aligned}t_1 = \ & 0.2348916X_1 + 0.2909979X_2 + 0.3715669X_3 + 0.3849528X_4 + 0.2972795X_5 \\ & + 0.2500166X_6 + 0.2543422X_7 + 0.2699512X_8 + 0.3516807X_9 \\ & + 0.2421589X_{10} + 0.3639095X_{11} + 0.1311052X_{12}\end{aligned} \tag{17}$$

$$\begin{aligned}t_2 = \ & 0.05221690X_1 + 0.15708498X_2 + 0.34658052\,X_3 + 00.12618007X_4 - 0.44088140X_5 \\ & + 0.07051519X_6 + 0.24912549X_7 - 0.12074225X_8 - 0.42772355X_9 \\ & + 0.38391608X_{10} - 0.23591743X_{11} - 0.14401048X_{12}\end{aligned} \tag{18}$$

$$\begin{aligned}t_3 = \ & 0.01948067X_1 - 0.35343849X_2 + 0.25986775\,X_3 + 0.39884393X_4 + 0.36596685X_5 \\ & - 0.55355103X_6 + 0.24912549X_7 - 0.31784596X_8 + 0.14414673X_9 \\ & - 0.23457759X_{10} - 0.30472312X_{11} + 0.13138327X_{12}\end{aligned} \tag{19}$$

$$\begin{aligned}t_4 = \ & 0.36157064X_1 - 0.09461747X_2 + 0.01478900\,X_3 + 0.02893688X_4 - 0.28652133X_5 \\ & + 0.42332707X_6 - 0.18973678X_7 + 0.34978154X_8 - 0.06368752X_9 \\ & - 0.64364959X_{10} - 0.30472312X_{11} - 0.18483523X_{12}\end{aligned} \tag{20}$$

$$\begin{aligned}t_5 = \ & 0.17281546X_1 + 0.84640303X_2 + 0.10054769\,X_3 - 0.30239309X_4 + 0.01597986X_5 \\ & - 0.47616792X_6 + 0.14576762X_7 - 0.48300065X_8 0.21247049X_9 \\ & - 0.13510227X_{10} + 0.13293838X_{11} - 0.69119663X_{12}\end{aligned} \tag{21}$$

$$\begin{aligned} t_6 = {} & -0.95771792X_1 + 0.04322347X_2 + 0.09849308\,X_3 + 0.42761618X_4 - 0.13345384X_5 \\ & + 0.39234738X_6 + 0.08214739X_7 + 0.23483427X_8 - 0.20951793X_9 \\ & - 0.28107294X_{10} - 0.16734217X_{11} + 0.54239406X_{12} \end{aligned} \tag{22}$$

$$\begin{aligned} t_7 = {} & 0.62980432X_1 - 0.35486553X_2 - 0.34639310\,X_3 + 0.07258252X_4 + 0.12001181X_5 \\ & + 0.23450197X_6 + 0.36546442X_7 - 0.09977548X_8 - 0.12547860X_9 \\ & + 0.44055794X_{10} - 0.26153821X_{11} - 0.54619219X_{12} \end{aligned} \tag{23}$$

Based on the equation (17)-(23), linear regression model is converted, as follows:

$$\hat{Y} = C_1t_1 + C_2t_2 + C_3t_3 + C_4t_4 + C_5t_5 + C_6t_6 + C_7t_7 \tag{24}$$

$$\begin{aligned} \hat{Y} = {} & 6.443168\,t_1 + 0.4283329\,t_2 + 0.8152602\,t_3 + 1.012529\,t_4 + 0.0000002142409\,t_5 \\ & + 0.05400745\,t_6 + 0.000003031815\,t_7 \end{aligned} \tag{25}$$

$$\begin{aligned} \hat{Y} = {} & 6.443168 + 0.484554\,X_1 - 0.33049\,X_2 + 0.822079\,X_3 + 0.872799X_4 + 0.02232X_5 \\ & - 0.15261X_6 - 0.19103\,X_7 - 0.05638X_8 - 0.07414X_9 - 0.57521X_{10} \\ & - 0.29623X_{11} - 0.03706X_{12} \end{aligned} \tag{26}$$

After obtaining the predicted linear regression results, it is necessary to check whether the data still contains multicollinearity or not. Because the values of $VIF = TOL = 1$then it means that there is no multicollinearityin the model.

4.4. Principal Component Regression

Principal component analysis was carried out to obtain principal components and principal component scores which were useful as independent variables in PCR. The first step is to look at the communality value which shows how much variance can be explained by the components formed. The results of the variance can be shown in the Table 1.

Tabel 1. Final Communality Estimates

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
0.826	0.573	0.031	0.271	0.147	0.837	0.058	0.738	0.101	0.686	0.409	0.917

Table 1 shows that the value of the existing variables has a meaning, namely how many percent of the variance of the variable can be explained by the components formed.

The next step is the Eigenvalues of the Correlation Matrix. In factor analysis there are several components which are variables. Each factor represents the variables analyzed. The ability of each factor to represent the analyzed variable is indicated by the magnitude of the variance described, which is called the eigenvalue. Eigenvalue shows the relative importance of each factor in calculating the variance of all analyzed variables.

From the output R it can be seen that the eigenvalues below 1 cannot be used in calculating the number of factors formed, so the factoring process should stop at only three factors. Eigenvalue in variable X_1 (life expectancy at birth) of 5.18 means that this factor can explain 43.20% of the total cumulative, variable X_2 (expected years of schooling) of 1.86 which means this factor explains 58.72% of the total cumulative and variable X_3 (mean years of school) of 1.32 which means this factor explains 69.68% of the total cumulative.

The next step is to find the value of the component scores that are formed to form the principal component regression equation. The score of this component will determine the number of variables that will be formed to replace the variables, $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}\ and\ X_{12}$with new variables.

The main component regression equation formed is as follows:

$$\begin{aligned} K_1 = {} & -0.22X_1 - 0.26X_2 - 0.32\,X_3 - 0.35X_4 - 0.33X_5 - 0.23X_6 - 0.30X_7 - 0.27X_8 - 0.38X_9 \\ & - 0.19X_{10} - 0.37X_{11} - 0.14X_{12} \end{aligned} \tag{27}$$

$$\begin{aligned} K_2 = {} & -0.12X_1 - 0.27X_2 - 0.40\,X_3 - 0.18X_4 + 0.37\,X_5 - 0.23X_6 + 0.48\,X_7 - 0.02X_8 + 0.31\,X_9 \\ & - 0.44X_{10} + 0.06\,X_{11} + 0.11\,X_{12} \end{aligned} \tag{28}$$

$$\begin{aligned} K_3 = {} & -0.48X_1 - 0.23X_2 + 0.05\,X_3 + 0.23\,X_4 - 0.12X_5 + 0.32\,X_6 - 0.12X_7 + 0.32\,X_8 - 0.14\,X_9 \\ & - 0.22X_{10} - 0.02X_{11} + 0.60\,X_{12} \end{aligned} \tag{29}$$

After the component scores are formed, the next step is to determine new variables to replace the variables$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11} and X_{12}$. The formation of the components formed by standardizing all variables and multiplying by the main component scores.

The main component regression is found by regressing with a new variable, namely the main component variable. Thus, the principal component regression model built for this case is:

$$Y = W_0 + W_1K_1 + W_2K_2 + W_3K_3 + \varepsilon \tag{30}$$

Then the equation model formed in Table 2.

Table 2. Y Output of Principal Component

Variable	Estimation Parameters	TOL	VIF
(Intercept)	70.5730	.	.
Factor 1	-1.7311	1	1
Factor2	-1.5185	1	1
Factor3	-0.0536	1	1

$$\hat{Y} = 70.5730 - 1.7311K_1 - 1.5185K_2 - 0.0536K_3 \tag{31}$$

$$\begin{aligned}\hat{Y} = {} & 70.5730 + 0.5888X_1 + 0.8724\,X_2 + 1.1587\,X_3 - 0.8669X_4 + 0.0159X_5 + 0.7303X_6 \\ & - 0.2031X_7 + 0.4806X_8 + 0.1946X_9 + 1.0088X_{10} + 0.5505X_{11} + 0.0432X_{12}\end{aligned} \tag{32}$$

After obtaining the predicted linear regression results, it is necessary to check whether the data still contains multicollinearity. By looking that the values of $VIF = TOL = 1$ which means that there is no multicollinearity.

4.5. Comparison between PLS and PCR

In selecting the best method, the coefficient of determination (R^2) and Mean Square Error (MSE) are compared with each of PCR and PLS as shown in Table 3. According to the table, in term of both R^2 and MSE, PLS outperforms PCR as the coefficient determination of PLS equals to 0.98 which is meant that as much as 98% variance in HDI can be explained by the factors included in this research. The lower MSE of PLS indicates the better performance of prediction model resulted by PLS than PCR

Table 3. R^2 and MSE Value

Methods	R^2	MSE
PCR	0.2462077	15.9193137
PLS	0.9814423	0.3919194

5. Conclusion

The estimated linear regression equation obtained from the application of the two methods in the case of the HDI of North Sumatra Province in 2019 is as follows:

$$\begin{aligned}\hat{Y}_{PLS} = {} & 6.443168 + 0.484554\,X_1 - 0.33049\,X_2 + 0.822079\,X_3 + 0.872799X_4 + 0.02232X_5 \\ & - 0.15261X_6 - 0.19103\,X_7 - 0.05638X_8 - 0.07414X_9 - 0.57521X_{10} - 0.29623X_{11} \\ & - 0.03706X_{12}\end{aligned}$$

$$\begin{aligned}\hat{Y}_{PCR} = {} & 70.5730 + 0.5888X_1 + 0.8724\,X_2 + 1.1587\,X_3 - 0.8669X_4 + 0.0159X_5 + 0.7303X_6 \\ & - 0.2031X_7 + 0.4806X_8 + 0.1946X_9 + 1.0088X_{10} + 0.5505X_{11} + 0.0432X_{12}\end{aligned}$$

By comparing PCR and PLS methods in term of the coefficient of determination and squared error it can be declared that PLS method is better than PCR method for overcoming the problem of multicollinearity and producing model prediction for HDI of North Sumatera in 2019.

References

Cherchye, L., & Abeele, P. V. (2005). On research efficiency: A micro-analysis of Dutch university research in Economics and Business Management. *Research policy*, *34*(4), 495-516.

Dias, R. A., Mattos, C. R., & Balestieri, J. A. (2006). The limits of human development and the use of energy and natural resources. *Energy Policy*, *34*(9), 1026-1031.

Donatos, G. S., & Mergos, G. J. (1991). Residential demand for electricity: the case of Greece. *Energy Economics*, *13*(1), 41-47.

Khan, N. H., Ju, Y., & Hassan, S. T. (2018). Modeling the impact of economic growth and terrorism on the human development index: collecting evidence from Pakistan. *Environmental Science and Pollution Research*, *25*(34), 34661-34673.

McGillivray, M., & White, H. (1993). Measuring development? The UNDP's human development index. *Journal of international development*, *5*(2), 183-192.

Ramzan, S., & Khan, I. M. (2010). Dimension reduction and remedy of multicollinearity using latent variable regression methods.

World Applied Science Journal, *8*(4), 404-410.

Ranis, G., Stewart, F., & Samman, E. (2006). Human development: beyond the human development index. *Journal of Human Development*, *7*(3), 323-358.

Roozbeh, M., & Arashi, M. (2016). Shrinkage ridge regression in partial linear models. *Communications in Statistics-Theory and Methods*, *45*(20), 6022-6044.

Sagar, A. D., & Najam, A. (1998). The human development index: a critical review. *Ecological economics*, *25*(3), 249-264.

Wang, G., Liu, Y., Li, Y., & Chen, Y. (2015). Dynamic trends and driving forces of land use intensification of cultivated land in China. *Journal of Geographical Sciences*, *25*(1), 45-57.

Williamson, O. E. (2005). Transaction cost economics and business administration. *Scandinavian journal of Management*, *21*(1), 19-40.

Wold, S., Ruhe, A., Wold, H., & Dunn, Iii, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735-743.

Yeniay, Ö., & GÖKTAŞ, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, *31*, 99-111.