Modeling of Human Development Index Using Ridge Regression Method

F Yanuar[1*], M Tillah[2] dan D Devianto[3]

[*1]Mathematics Department, FMIPA, Universitas Andalas, Padang, Indonesia
[2,3] Mathematics Department, FMIPA, Universitas Andalas, Padang, Indonesia

*E-mail : ferrayanuar@sci.unand.ac.id

Abstract. This study aims to model factors affecting HDI (Human Development Index) in North Sumatra at 2015 using ridge regression method. We use ridge regression method because in the IPM data there is a multicolinearity problem. Ordinary least squares, as regression method commonly used in statistical modeling, cannot be applied in this case. This study makes the comparison between the use of OLS and the ridge regression method to the HDI data. This study proves that the ridge regression method produces better model and can solve the problem of multicolinearity case in data, while the OLS method can not. The significant factors that effecting the HDI at North Sumtera in 2015 are Total expenditure / capita / month (X_4) and Average school length (X_5). The indicator of the goodness of fit for the proposed model 81.81% which means that the model is good and could be accepted.

Keywords : Modelling, HDI, Index, Ridge Regression

1. Introduction

The basically target of development is human development. While the main goal of development is to create an environment that enables people in the nation to enjoy longevity, health and live a productive life. Human development places humans as the ultimate goal of development rather than a tool of development. The success of human development can be seen from how much fundamental human problems can be solved. These are the issues of health and education [1-3].

Human Development Index (HDI) is an indicator that is used to measure the important aspects related to the quality of economic development outcomes, ie., longevity, health and live a productive life[2-5]. These three elements are very important elements in determining the level of HDI. The three elements are interrelated and can not stand alone. In addition, HDI is also influenced by other factors, which is in the form of economic growth, infrastructure and government policies. So the HDI in an area will increase if there is an increase in the three elements. An area is said to have good economic development characterized by high HDI value. In other words, there is a positive correlation between the value of HDI with the degree of success of economic development.

Since the importance of HDI as an indicator of successful economic development of a region, it is important to identify the level of HDI in an area periodically. In determining the factors of HDI' model, it used the commonly modeling method is ordinary least squares (OLS) method. OLS can result acceptable model if all assumptions of classical model are fulfilled, such as error has normal distribution, no multicollinearity problems between indicator variables. All assumptions must be met in order to obtain predicted parameters that are BLUE (Best Linear Unbiased Estimator) [6].

One method to solve multicollinearity problem is ridge regression. The regression coefficient produced of this method is more stable and variance of the regression coefficient is smaller than classical method (OLS) [7,8]. Basically this method is a modification of the OLS method. The basically process in this method is the correlation matrix of the independent variable are converted using ridge regression method so that the value of the regression coefficient estimation is easy to be obtained.

Based on the above description, the problem in this research is the modeling of factors that affect HDI in North Sumatra in 2015 by using ridge regression method. In this article we will also compare the results of the OLS analysis and the ridge regression method.

2. Data and Methods

2.1. Data

The data used in this study are secondary data, ie Human Development Index (HDI) data and influencing factors for each district (33 districts) in North Sumatera at 2015. Factors assumed to affect the HDI are Number of poor (X_1), Population density (X_2), GRDP (Gross Regional Domestic Product) (X_3), Total expenditure / capita / month (X_4), Average school length (X_5) , Number of educational facilities (X_6) and Number of health facilities (X_7). Data obtained from Central Statistics Agency (BPS) of North Sumatra (BPS, 2016). Data has multicolinearity problems. Then the data is modeled using OLS dan ridge regression method.

2.2. Ridge Regression Analysis

Ridge regression method is bias estimator, but values for variance is small. Ridge regression is one method that can be used to solve the problem of multicollinearity that categorized as less perfect. This method do any modification to the OLS method [7], [8]. The modification is accomplished by adding the bias constant of *k* on the diagonal of the $Z'Z$ correlation matrix, so that the coefficient of the ridge estimator is influenced by the magnitude of the bias constant *k*. The *k* values for the ridge regression coefficients are generally between 0 and 1.

In the simplest form, the ridge regression procedure is as follows: Let *Z* be centrally and scaled of matrix *X* when the regression problem is in the form of correlation.

The ridge prediction value vector is obtained by minimizing the mean square error (MSE) for the regression formed model. Estimated values of ridge regression are as follows [7], [10]:

$$\hat{\boldsymbol{\beta}}_R^* = (\mathbf{Z}'\,\mathbf{Z} + k\,\mathbf{I})^{-1}\,\mathbf{Z}'\,\mathbf{Y}^* \qquad (1)$$

Where $\hat{\boldsymbol{\beta}}_R^*$ is ridge regression estimator, *k* is as ridge parameter, I is as identity matrix and ***Z*** is matrix X which is centering and scaling.

Mathematically there is a relationship between the ridge regression estimator, $\hat{\boldsymbol{\beta}}_R^*$ and the OLS estimator, $\hat{\beta}_{OLS}^*$ as follows [8] :

$$\hat{\beta}_R^* = \left[\mathbf{I} - k\left(\mathbf{Z}'\mathbf{Z} + k\,\mathbf{I}\right)^{-1}\right]\hat{\beta}_{OLS}^* \qquad (2)$$

The properties of ridge regression estimators are :

a. The expected value of the ridge estimator is biased :

$$E\,(\hat{\beta}^*_R) = E\Big[(\mathbf{I}-(\mathbf{Z}'\mathbf{Z}+k\mathbf{I})^{-1})\hat{\beta}^*_{OLS}\ \Big] \neq \beta^*$$

b. Minimum Variance, the variance matrix and the covariance of the ridge regression are given by

$$\mathbf{cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathbf{Z}'\mathbf{Z}+k\mathbf{I})^{-1}\mathbf{Z}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{Z}+k\mathbf{I})^{-1}$$

Variance for ridge regression is as follows:

$$\boldsymbol{var}(\hat{\boldsymbol{\beta}}^*_R) = \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} \qquad (3)$$

Variance for the OLS method is:

$$\mathrm{var}(\hat{\beta}^*_{OLS}) = \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j} \qquad (4)$$

When compared to the OLS variance in Equation (4) with the variance of the ridge regression in Equation (3) the variance of the ridge regression estimator is less than the OLS variance.

2.3 The Selection for Bias Constant, *k*

Selection of the magnitude for bias constant *k* has to be considered carefully. The desired bias constant, *k* will produce relatively small biases and produces a relatively stable parameter estimate. There are several ways to choose the magnitude of *k*, one of them is by minimizing MSE (Mean Square Error) of ridge regression [8] :

$$MSE(\hat{\beta}^*_R) = \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} + \sum_{j=1}^{p}\frac{\beta_j^{*2}k^2}{(\lambda_j+k)^2} \qquad (5)$$

So, it will be obtained :

$$k = \frac{p\sigma^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}} \qquad (6)$$

With p is number of parameters except β_0, while σ^2 and $\hat{\boldsymbol{\beta}}$ are obtained from OLS estimation method. Kibria & Banik (2016) did iteration procedure to determine the value of *k*, by following any steps below :

1. Determine the initial value, by making Equation (6) as the initial value that is $k_o = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ then substituted the value of k_o to the equation $\hat{\boldsymbol{\beta}}^*_R(k_0) = (\mathbf{Z}'\mathbf{Z}+k_0\,\mathbf{I})^{-1}\mathbf{Z}'\boldsymbol{Y}^*$
2. Define $k_1 = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^*_R(k_0)'\,\hat{\boldsymbol{\beta}}^*_R(k_0)}$ and then substitute the value of k_1 to the equation $\hat{\boldsymbol{\beta}}^*_R(k_1) = (\mathbf{Z}'\mathbf{Z}+k_1\,\mathbf{I})^{-1}\mathbf{Z}'\boldsymbol{Y}^*$

3. Define $k_2 = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}_R^*(k_1)'\,\hat{\boldsymbol{\beta}}_R^*(k_1)}$ and then substitute the value of k_2 to the $\hat{\boldsymbol{\beta}}_R^*(k_2) = (\mathbf{Z}'\mathbf{Z} + k_2\,\mathbf{I})^{-1}\mathbf{Z}'\mathbf{Y}^*$ and so on until it is satisfied $\frac{[k_{i-1} - k_i]}{k_i} > 20\,T^{-1.3}$ with $T = \frac{tr(\mathbf{Z}'\mathbf{Z})^{-1}}{p}$.

After obtaining the expected value of the ridge regression parameter, *k*, the parameter value can be restored to the initial form before the variable is standardized by the formula [7] :

$$\beta_j^* = \left(\frac{s_j}{s_Y}\right)\beta_j,\ ,j = 1,2,\ldots,p \qquad (7)$$

Test of regression significance simultaneously can be done by F test or ANOVA test, and individual test with t test.

3. Results and Discussions

3.1 Multicollineary Test

The following is multicollinearity test on HDI data in North Sumatra 2015 using several ways :

1. VIF and Tolerance values
   The multicollinearity problem can be detected by using VIF and tolerance values for each independent variable. If the VIF value of the independent variables is greater than 10 or the tolerance value of the independent variables is less than 0.1 then the multicollinearity is less than perfect. Table 1 below presents the results of multicollinearity test on HDI data.

Table 1. VIF and Tolerance Values on HDI Data

Variable	VIF	Tolerance
X_1	10.123	0.0988
X_2	2.462	0.4062
X_3	20.467	0.0489
X_4	5.681	0.1760
X_5	6.365	0.1571
X_6	18.718	0.0534
X_7	11.533	0.0867

Table 1 shows that the VIF values of X_1, X_3, X_6 and X_7 are greater than 10 and also the tolerance values X_1, X_3, X_6 and X_7 are less than 0.1. Thus it can be concluded that there is an *imperfect* multicolinearity problem between the independent variables.

2. Determinant of Correlation Matrix.

Multicollinearity can be detected by using the determinant of correlation matrix. If the determinant value of the correlation matrix is close to 0, it indicates that there is an *incomplete* multicollinearity problem.

The following matrix R contains the correlation coefficient between the selected independent variables in the HDI data.

$$\mathbf{R} = \begin{bmatrix} 1 & 0.40174 & 0.894617 & 0.285885 & 0.088555 & 0.9235 & 0.851038 \\ 0.40174 & 1 & 0.586162 & 0.516978 & 0.560291 & 0.53648 & 0.656889 \\ 0.894617 & 0.586162 & 1 & 0.466106 & 0.323419 & 0.95877 & 0.943886 \\ 0.285885 & 0.516978 & 0.466106 & 1 & 0.871501 & 0.406748 & 0.498536 \\ 0.088555 & 0.560291 & 0.323419 & 0.871501 & 1 & 0.27968 & 0.360724 \\ 0.9235 & 0.53648 & 0.95877 & 0.406748 & 0.27968 & 1 & 0.907813 \\ 0.851038 & 0.656889 & 0.943886 & 0.498536 & 0.360724 & 0.907813 & 1 \\ 0.262791 & 0.467089 & 0.354746 & 0.308068 & 0.320811 & 0.402231 & 0.385334 \end{bmatrix}$$

The determinant of the correlation matrix R above is 0.000060719. Since the determinant value of the correlation matrix is close to 0, this means that the correlation matrix is almost singular, so it can be concluded that there is an *imperfect* multicolinearity problem between the independent variables.

3. Condition Value.

Multicolinearity can also be measured in terms of the ratio of the largest and smallest values of eigenvalues, obtained and expressed as the condition values of the correlation matrix. Eigen value is calculated using NCSS software. The condition value for this HDI data is :

$$\phi = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\lambda_1}{\lambda_7}$$

$$= \frac{4.629857}{0.029024}$$

$$= 159.52$$

The value of the obtained condition is larger than 100, so it can be concluded that *imperfect* multicolinearity due among independent variables.

3.2 Modeling the HDI Data Using OLS

In this section we will estimate the HDI model using OLS method. The regression analysis was performed with the help of SPSS software, parameter estimated are presenting in Table 2.

Table 2. Parameter Estimated Using OLS

Variables	Parameter	Estimated	Standard Error
	Constant	43.16153	1.91100
X_1	$\hat{\beta}_1$	0.00543	0.01897
X_2	$\hat{\beta}_2$	0.00007	0.00019
X_3	$\hat{\beta}_3$	0.00006	0.00003
X_4	$\hat{\beta}_4$	0.00764	0.00338
X_5	$\hat{\beta}_5$	2.15107	0.38667
X_6	$\hat{\beta}_6$	0.00008	0.00004
X_7	$\hat{\beta}_7$	0.00035	0.00043

Furthermore, the test of model significance simultaneously or together for all (β) with the test statistics F. The results are presented in Table 3. The hypothesis for this test is as follows:
H_0 : The variable X simultaneously has no effect on the predicted value of Y
H_1 : There is at least one variable X simultaneously affecting the predicted value of Y

Table 3. Variance Analysis for HDI Data Using OLS Method

Model	Sum of Square	Degree of Fredom	Mean Square	F_{Count}	F_{Table}
Regresi	709.648	7	101.378		
Error	43.099	25	1.724	58.805	2.4
Total	752.748	32			

Based on Table 3 it can be seen that the value of F_{Count} more than F_{Table}. This means that there is at least one of the independent variable affects the predicted value of Y significantly. Thus, t test should be done to determine the significant independent variables. Following are the hypothesis:

H_0: The independent variable has no significant effect on the predicted value of Y
H_1: Individually independent variables significantly influence the value of predictor Y

The t test statistic are presented in Table 4.

Table 4. T Test on HDI Data

Parameter	Estimate	Standard Error	t_{count}	t_{table}
Intercept	43.16153	1.911	22.586	2.05954
$\hat{\beta}_1$	0.00543	0.01897	0.287	

$\hat{\beta}_2$	0.00007	0.00019	0.383
$\hat{\beta}_3$	-0.00006	0.00003	-1.893
$\hat{\beta}_4$	0.00764	0.00338	2.259*
$\hat{\beta}_5$	2.15107	0.38667	5.563*
$\hat{\beta}_6$	0.00008	0.00004	1.853
$\hat{\beta}_7$	0.00035	0.00043	0.813

(* significant at α = 0,05)

Based on Table 4 it is known that the t_{count} value of the variables X_4 and X_5 is greater than t_{table}, meaning independent variables X_4 and X_5 individually significance to influence the value of estimated Y.
The value of determination coefficient or R^2 is 94.3 percent which means that the variability of the dependent variable can be explained by the regression model is 94.3 percent.

The HDI data has muticolinearity problem. Following are the consequences of multicolinearty if apply OLS method. *First*, in the partial test, it resulted that only few variables are statistically significant, as presented at Table 4. *Second*, the coefficient of the regression estimated is not suitable, look at the variable number of poor people (X_1) should have a negative sign, because based on previous research the number of poor people have a negative relationship with HDI. *Third*, the standard errors from OLS method are also quite large. Thus, OLS method cannot be applied to the data which has multicolinearity problem.

3.3 Modeling with Ridge Regression Method

To solve the multicollinearity problem that occurred in HDI data in North Sumatera in 2015, we used ridge regression method [8]. In the process of estimating the ridge regression parameters, the first step is doing centralization and scaling of the data [12], [13]. By using iteration procedure then we obtain the best value for *k* is 0.300719. At *k* = 0.300719 are obtained the expected value for all model parameters as presented in Table 5 below.

Table 5. Parameter Estimate, VIF Values on HDI Data Using Ridge Regression Method

Parameter	Estimate	Standard Error	VIF
$\hat{\beta}_1^*$	-0.0027	0.00731	0.473
$\hat{\beta}_2^*$	0.0918	0.00018	0.681
$\hat{\beta}_3^*$	0.0055	0.00001	0.331
$\hat{\beta}_4^*$	0.3170	0.00188	0.556

$\hat{\beta}_5^*$	0.4365	0.19566	0.513
$\hat{\beta}_6^*$	0.0857	0.00001	0.364
$\hat{\beta}_7^*$	0.0492	0.00015	0.462

The parameter estimated then be transformed to the original values before standardized, the values are presented in Table 6.

Table 6. Parameter Estimated Using Ridge Regression Method

Variable	Parameter	Estimate	Standard Error
X_1	$\hat{\beta}_1$	-0.00034	0.00731
X_2	$\hat{\beta}_2$	0.00023	0.00018
X_3	$\hat{\beta}_3$	0.00001	0.00001
X_4	$\hat{\beta}_4$	0.00940	0.00188
X_5	$\hat{\beta}_5$	1.39774	0.19566
X_6	$\hat{\beta}_6$	0.00001	0.00001
X_7	$\hat{\beta}_7$	0.00013	0.00015

3.4 Comparison of Parameter Estimation Model with Ridge Regression and Ordinary Least Square (OLS)

This section presents the results of ordinary least squares (OLS) method and ridge regression, presented in Table 7 below.

Table 7. Comparison of Parameter Estimate Using Ridge Regression and OLS.

Parameter	OLS 's Estimated	Ridge Regression's Estimated	OLS's VIF	Ridge's VIF	OLS's Standard Error	Ridge's Standard Error
$\hat{\beta}_1$	0.00543	-0.00034	10.124	0.473	0.01897	0.00731
$\hat{\beta}_2$	0.00007	0.00023	2.462	0.681	0.00019	0.00018
$\hat{\beta}_3$	0.00006	0.00001	20.468	0.331	0.00003	0.00001

$\hat{\beta}_4$	0.00764	0.00940	5.681	0.556	0.00338	0.00188
$\hat{\beta}_5$	2.15107	1.39774	6.366	0.513	0.38667	0.19566
$\hat{\beta}_6$	0.00008	0.00001	18.719	0.364	0.00004	0.00001
$\hat{\beta}_7$	0.00035	0.00013	11.533	0.462	0.00043	0.00015

Based on Table 7 it is known that the VIF value for all independent variables of the ridge regression results is less than 10. Thus the multicolinearity problem has been solved by using the ridge regression.

In addition, the parameter predictor mark for the variable of the poor population (X_1) has been in accordance with the theory that should be, that is negative value. Table 7 also shows that the all standard errors from ridge regression are smaller than the standard errors of the OLS method. Thus it can be concluded that the ridge regression method could produce better proposed model than the OLS method, as due to multicolinearity problem.

3.5 Test The Significance Of Model Parameters

Then will be tested the significance of the model parameters, to perform linear regression testing is done with the hypothesis as follows:

H_0 : The variable X simultaneously has no effect on the predicted value of Y

H_1 : There is at least one variable X simultaneously affecting the predicted value of Y

Following Table 8 is presenting ANOVA test using ridge regression method.

Tabel 8. *ANOVA* untuk Data IPM dengan Regresi *Ridge*

Model	Mean Square	Degree of Freedom	Mean Square	F_{count}	F_{table}
Regression	615.817	7	87.97385	16.0618	2.4
Error	136.930	25	5.47722		
Total	752.747	32	23.52336		

Based on Table 8 it can be seen that value of F_{count} is more than F_{table}, thus we should to reject H_0. Thus it can be stated that at least one of independent variables has a significant effect on the dependent variable. T test then has to done to determine the independent variables that have significant influence on HDI. The results of the T test are presented in Table 9.

Tabel 9. T test on HDI Data

Parameter	Estimate	Standard Error	T Count	T Table
$\hat{\beta}_2$	0.00023	0.00018	1.30455	2.05954
$\hat{\beta}_3$	0.00001	0.00001	0.11216	
$\hat{\beta}_4$	0.00940	0.00188	4.98349*	
$\hat{\beta}_5$	1.39774	0.19566	7.14364*	
$\hat{\beta}_6$	0.00002	0.00001	1.66473	
$\hat{\beta}_7$	0.00013	0.00015	0.84858	

* significant at level significant 0.05.

Based on Table 9 it can be seen that X_4 dan X_5 individually have significant influence on response varable, Y. Goodness of fit for this proposed model or R^2 is 81%, meaning the proposed model could be accepted.

4. Conclusion

The multicolinearity problem that occurs in multiple linear regression can be solved by the ridge regression method, which essentially seeks the variance to be smaller by adding a bias constant *k*. This causes the estimator of ridge regression parameters to be more stable even if biased.

This article modeled the factors of HDI in North Sumatera at 2015 by using ridge regression method because in the preliminary analysis it is known that in the data there are multicollinearity problems. This study conducted a comparison analysis between OLS and ridge regression method and it was proved that ridge regression method was able to overcome multicolinearity problem while OLS did not. This study found that that expenditure (X_4) dan mean of school length (X_5) individually have significant influence on response varable, Y. Goodness of fit for this proposed model or R^2 is 81%.

References

[1] Dumuid D, Maher C, Lewis LK, Stanford TE, Martin Fernandez JA, et al. 2018. Human development index, children's health-related quality of life and movement behaviors: a compositional data analysis. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 27:1473-82

[2] Freire RCJ, Pieruccini-Faria F, Montero-Odasso M. 2018. Are Human Development Index dimensions associated with gait performance in older adults? A systematic review. *Experimental gerontology* 102:59-68

[3] M. Jhingan, 2007, *Ekonomi Pembangunan dan Perencanaan*. Rajawali Press.

[4] Moustris K, Tsiros IX, Tseliou A, Nastos P. 2018. Development and application of artificial neural network models to estimate values of a complex human thermal comfort index associated with urban heat and cool island patterns using air temperature data from a standard meteorological station. *International journal of biometeorology* 62:1265-74

[5] Sabet Rohani H, Ahmadvand A, Garmaroudi G. 2018. The relationship between important reproductive health indices and human development index in Iran. *Medical journal of the Islamic Republic of Iran* 32:54

[6] F. Yanuar, L. Hasnah, and D. Devianto, 2017, The Simulation Study to Test the Performance of Quantile Regression Method With Heteroscedastic Error Variance, *Cauchy - J. Mat. Murni dan Apl.*, vol. 5, no. 1, pp. 36–41

[7] H. Duzan and N. S. Shariff, 2016, Solution to the Multicollinearity Problem by Adding some Constant to the Diagonal, *J. Mod. Appl. Stat. Methods*, vol. 15, no. 1, pp. 752–773.

[8] N. Sima, M. Shariff, and N. A. Ferdaos, 2017, A Robust Ridge Regression Approach in the Presence of Both Multicollinearity and Outliers in the Data," *AIP Conf. Proc.*, vol. 1870, No. 060003.

[9] B. G. Kibria and S. Banik, 2016, Some Ridge Regression Estimators and Their Performances," *J. Mod. Appl. Stat. Methods*, vol. 15, no. 1, pp. 206–238.

[10] M. El-Dereny and N. I. Rashwan, 2011, Solving Multicollinearity Problem Using Ridge Regression Models λ λ," *Int. J. Contemp. Math. Sci.*, vol. 6, no. 12, pp. 585–600.

[11] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, 2004, *Applied Linear Statistical Models Fifth Edition*. McGraw-Hill.

[12] S. Lipovetsky and W. M. Conklin, 2005, Ridge regression in two-parameter solution," *Appl. Stoch. Model. Bus. Ind.*, vol. 21, pp. 525–540.

[13] J. B. Forrester and J. H. Kalivas, 2004, Ridge regression optimization using a harmonious approach," *J. Chemom.*, vol. 18, pp. 372–384.