Comparing the Principal Regression Analysis Method with
Ridge Regression Analysis in Overcoming Multicollinearity
on Human Development Index (HDI) Data in Regency/City of
East Java in 2018

A F Yoantika and Susiswo
Mathematics Department, Universitas Negeri Malang, Indonesia
E-mail: feliciayoantika@gmail.com, susiswo.fmipa@um.ac.id

Abstract. East Java Province is one of the densely populated provinces. The population
density in the province of East Java does not hamper the welfare of its people. The welfare
of the people in the province of East Java can be seen through the Human Development Index
(HDI), which includes elements of life expectancy, education, and economy. According to the
Central Statistics Agency (BPS) of East Java, the HDI level from 2017 to 2018 has increased,
but this data experiences a case of multicollinearity. Researchers compared the Principal
Regression Analysis method with Ridge Regression Analysis in overcoming the
multicollinearity. The conditions taken are methods that have a small MSE value and a large
$R^2$ value. Based on the results and discussion, the Main Component Regression Analysis
method has an MSE value of 2.51 and an $R^2$ value of 91.0%, the Ridge Regression Analysis
Method has an MSE value of 0.00071 and an $R^2$ of 97.7%. The provisions obtained are the
value of MSE Regression Ridge is smaller than the value of MSE Regression Main
Components and the value of $R^2$ Regression Ridge is greater than the value of $R^2$ Main
Components. So, the best method in overcoming the multicollinearity problem in the Regency/
City Human Development Index (HDI) data in East Java in 2018 is the Ridge Regression
Analysis method.

1. Introduction
The number of problems or cases that occur in everyday life, do not escape from the factors that
influence the problem or case. Problems that occur can include economic, social, cultural, political,
religious, health problems, to population density. Talking about population density in an area,
synonymous with the welfare of the people who live in it. One area that has a dense population is in
the province of East Java. The level of welfare in the province of East Java can be seen through the
Human Development Index (HDI), which includes life expectancy, education, and the economy.

According to the East Java Central Statistics Agency (BPS), the level of HDI in East Java
Province from 2017 to 2018 increased by 0.72%. Surabaya City has the highest HDI of 81.74, while
the region with the lowest HDI is in the Sampang region, which is 61.00. However, the district/city
Human Development Index (HDI) data in East Java Province in 2018 experienced a case of
multicollinearity. Multicollinearity is a strong relationship between one independent variable with
another independent variable. Independent variables or factors that influence the level of district/city
Human Development Index (HDI) in East Java Province in 2018, include the percentage of health
complaints, percentage of poverty rate, percentage of old school expectations, percentage of the
average length of the school, and percentage of gross domestic product.

Referring to the problems above, a method of analyzing the Main Component Regression
Analysis with Ridge Regression Analysis will be used to overcome the multicollinearity. The context
in the research conducted by (Mangaraj, Aparajita [4]) discusses the analysis and building of a general Human Development Index (HDI) model. According to (Sen, Zhang, Nagarajaiah, Sun [6]) multicollinearity problems can be overcome by using the Principal Regression Analysis method. Multicollinearity can also be overcome by Ridge Regression Analysis (Gorski, Jakubowska [2]). The advantage of the Principal Component Analysis method is that high correlations can be reduced without eliminating the original variables. However, it can only show the characteristics of the independent variables. The strength of Ridge Regression Analysis is that the estimated value of the ridge regression is biased, more stable, and relatively small. This bias estimator can meet the ridge regression estimator coefficient.

The purpose of this study is to find out the best method for dealing with multicollinearity cases in the Human Development Index (HDI) data of regencies/cities in East Java Province in 2018. Through this article, a comparison of the methods of Main Component Regression Analysis with Ridge Regression Analysis with the method selection requirements is having a small Mean Square Error (MSE) and a large Determination Coefficient ($R^2$). This research expects to be able to see the best method to overcome multicollinearity in the district/city Human Development Index (HDI) data in East Java Province in 2018 and produce significant results.

2. Method

Quantitative research is a method that plays a role in the research of this article. The nature of this quantitative research is systematic, detailed, and structured. In addition, this study also uses tables to show the results of the data obtained. The data used in a quantitative approach is population data or sample data and using research instruments and data analysis processes that have statistical properties in the presence of a predetermined hypothesis test.

The Human Development Index data used in this study are secondary data from the Central Statistics Agency of East Java Province in 2018 [1]. The amount of data used is 38 data. Factors that influence Human Development Index (HDI) include life expectancy, education, and economy. There are aspects that can be described from each of these Human Development Index (HDI) factors, (Yektiningsih [10]).

The variables in this study have two types of variables, namely the independent and dependent variables. The dependent variable used is the level of district/city Human Development Index (HDI) in East Java Province in 2018. Then for the independent variables used include the percentage of health complaints, percentage of poverty, percentage of old school expectations, percentage of the average length of the school, and the percentage of gross domestic product. This research was conducted with the help of Minitab 16 and Ms. Excel.

Before analyzing using methods Regression Analysis of Principal Components and Ridge Regression Analysis, the data needs to be tested using the assumption of multiple linear regression first. Things that must be done in the assumption of multiple linear regression include data that must be normally distributed, data not heteroscedasticity, data not autocorrelation and data not multicollinearity. after the data has passed the multiple linear regression assumption test and the data experiences multicollinearity, the data must be overcome by methods Regression Analysis of Principal Components and B.

The process of handling data using Regression Analysis of Principal Components method includes standardization of independent variables, determining eigenvalues and eigenvectors, regression of dependent variables with new independent variables (main component scores), returning the transformation results and ANOVA test of principal component analysis ([2], [3], [5], [9]). The second process of handling data using Ridge Regression Analysis method. The steps include transforming the Y and X variables through centering and rescaling, forming a matrix from the results of the transformation, determining the bias constant value c, determining the estimator coefficient $\beta^R_c$ of the selected VIF value c, the ridge regression ANOVA test and seeing the estimator relationship of the ridge regression parameters ([3], [6], [7], [8]).
The Data Analysis Process:

Start

Data Description

Normal Distribution

Yes

Yes

Heteroscedasticity

No

Yes

Autocorrelation

No

No

Multicollinearity

Yes

Interpretation of results

Principal Component

Independent Variable Standardization

Determination of Eigenvalues and Eigenvectors

Bound Variable Regression with Principal Component Scores

A

Ridge Regression

Centering and Rescaling

Forming a Matrix from the Transformation Results

Determination of Bias Constant Value $c$

A
Determination of the estimator coefficient $\beta_c R$ of the selected VIF value of $c$

Returns the Transformation Result

Principal Component ANOVA Test

Yes

No

Interpretation of results

Mean Square Error (MSE)

Determination of the Best Method

Finish

View Estimator Relationship of Ridge Regression Parameters

Ridge Regression ANOVA Test

Yes

No

Mean Square Error (MSE)
3. Results and Discussion

3.1 Result of Research

3.1.1 Assumption of Multiple Linear Regression. The results of the assumption of multiple linear regression in the Human Development Index (HDI) data of regencies/cities in East Java Province in 2018 are shown in Table 1.

| Assumption of Multiple Linear Regression | Information |
|----------------------------------------|-------------|
| Normal Distributed Data                | Yes         |
| Data Not Heteroscedasticity             | Yes         |
| Data Not Autocorrelated                | Yes         |
| Data Not Multicollinearity              | No          |

Based on Table 1 it can be seen that only the fourth assumption does not meet the assumption of Multiple Linear Regression, that is data, not multicollinearity. Assumptions that are not met will be overcome using two methods of Main Component Regression Analysis and Ridge Regression Analysis. Then, the results of the two methods will be compared to get the best method in overcoming the multicollinearity problem.

3.1.2 Regression Analysis of Principal Components. According to (Sriningsih, Hatidja, Prang [8]) the Principal Component Regression Analysis Process starts from the standardization of the independent variables, determines the eigenvalue and eigenvector vector, regresses the dependent variable with the new independent variable (the score of the main component), returns the transformation results and ANOVA Test of the Main Component Analysis. The process will bring up new independent variables that are no longer correlated.

| Predictor | Coef | SE Coef | T   | P  |
|-----------|------|---------|-----|----|
| Constant  | 70.9703 | 0.2568 | 276.33 | 0.000 |
| SK1       | 2.9835 | 0.1563 | 19.09 | 0.000 |

Based on Table 2 the variable regression model is tied to the scores of the main components, namely:

\[ Y = 71.0 + 2.98SK1 \] (1)

from the results of the Main Component Regression model, the transformation process will be returned to (Setiani [7]).

\[ Y = 33.39376 + 0.38442X_1 - 1.58834X_2 + 1.57642X_3 + 1.68966X_4 + 0.93572X_5 \] (2)

ANOVA test can show the MSE value and \( R^2 \) value of the Main Component Regression. The MSE and \( R^2 \) values of the Major Component Regression Analysis methods can be seen in Table 2.

| Criteria | Value |
|----------|-------|
| Mean Square Error (MSE) | 2.51 |
| KoefisienDeterminasi \( (R^2) \) | 91.0% |

3.1.3 Ridge Regression Analysis. According to (Hasriani [3]) the Ridge Regression Analysis process starts from the transformation of variables Y and X through centering and rescaling, forming a matrix of the results of the transformation, determining the value of bias \( c \), determining the estimator coefficient \( \beta_c \) of the selected VIF value of \( c \), the ANOVA Ridge Regression test, and looking at the estimator relationship from the Ridge Regression parameter. Selection of a constant \( c \) from the VIF value, the matrix used is \( (Z^T Z + Ic)^{-1}Z^T Z(Z^T Z + Ic)^{-1} \). (Rohmah [5]). Based on the Ridge Regression Analysis process, the value of bias \( c \) obtained is \( c = 0.2 \).
Table 4. A constant $c$ from the VIF value

| C  | $\beta_1^c$ | $\beta_2^c$ | $\beta_3^c$ | $\beta_4^c$ | $\beta_5^c$ |
|----|------------|------------|------------|------------|------------|
| 0.01 | -0.00916  | 0.000971  | 0.157141  | 0.82673    | 0.066497   |
| 0.02 | -0.00631  | -0.01926  | 0.169196  | 0.791579   | 0.067598   |
| 0.03 | -0.00383  | -0.03679  | 0.179419  | 0.760846   | 0.068618   |
| 0.04 | -0.00166  | -0.05208  | 0.188155  | 0.733724   | 0.069569   |
| 0.05 | 0.000268  | -0.06554  | 0.195671  | 0.709589   | 0.070458   |
| 0.06 | 0.001985  | -0.07746  | 0.202175  | 0.687954   | 0.071292   |
| 0.07 | 0.003525  | -0.08807  | 0.20783   | 0.668433   | 0.072076   |
| 0.08 | 0.004915  | -0.09756  | 0.212767  | 0.650715   | 0.072815   |
| 0.09 | 0.006176  | -0.10609  | 0.217092  | 0.63455    | 0.073512   |
| 0.1  | 0.007326  | -0.11378  | 0.220892  | 0.619728   | 0.074161   |
| 0.2  | 0.014948  | -0.16182  | 0.241279  | 0.518204   | 0.079157   |
| 0.3  | 0.018979  | -0.18332  | 0.246484  | 0.459871   | 0.082169   |
| 0.4  | 0.021422  | -0.19361  | 0.245972  | 0.420386   | 0.083951   |
| 0.5  | 0.023003  | -0.19826  | 0.242926  | 0.390974   | 0.084906   |
| 0.6  | 0.024055  | -0.19979  | 0.238678  | 0.367673   | 0.08528    |
| 0.7  | 0.024756  | -0.19946  | 0.23386   | 0.348424   | 0.085238   |
| 0.8  | 0.025213  | -0.19798  | 0.228801  | 0.332038   | 0.084892   |
| 0.9  | 0.025492  | -0.19579  | 0.22368   | 0.317778   | 0.084323   |

from the selection of the bias constant $c = 0.2$, the regression equation can be written, i.e.

\[
\hat{Y}^* = 0.014948Z_1 - 0.16182Z_2 + 0.241279Z_3 + 0.518204Z_4 + 0.079157Z_5
\]

(3)

and according to (Wasilaine, Talakua, Lesnussa [9]) Ridge Regression model is taken from the return of the transformation equation

\[
\hat{Y}^* = 87.37961 - 0.00560X_1 - 0.02085X_2 + 0.40349X_3 + 1.41499X_4 + 0.07606X_5
\]

(4)

ANOVA test can show the MSE value and the $R^2$ value of the Ridge Regression. The MSE and $R^2$ values of the Ridge Regression Analysis method can be seen in Table 3.

Table 5. Result of ANOVA test

| Criteria                  | Value  |
|---------------------------|--------|
| Mean Square Error (MSE)   | 0.00071|
| KoefisienDeterminasi ($R^2$) | 97.7%  |

3.2 Discussion of Result

Based on the research results, the general model obtained from the Ridge Regression method in overcoming multicollinearity problems is

\[
\hat{Y}^* = 87.37961 - 0.00560X_1 - 0.02085X_2 + 0.40349X_3 + 1.41499X_4 + 0.07606X_5
\]

. This model interprets that the increase in Human Development Index (HDI) for East Java in 2018 is influenced by the level of health complaints and percentage of poverty that has decreased, then the percentage of old school expectations, percentage of the average length of the school and the percentage of gross domestic product increased.

The general model obtained from the Principal Component Regression method in overcoming multicollinearity problems is

\[
Y = 33.39376 + 0.38442X_1 - 1.58834X_2 + 1.57642X_3 + 1.68966X_4 + 0.93572X_5
\]
Different from the Ridge Regression model, Human Development Index (HDI) will increase if the level of health complaints and poverty decreases, then the expectation level of school years, the average length of schooling, and gross domestic product increases. However, in the Principal Component Regression model, only the level of health complaints is still increasing. So, the best model of the two methods is the Ridge Regression method.

If you look at it from the calculation point of view, the reference here is the smallest MSE and the largest Coefficient of Determination. The value of MSE Regression Ridge = 0.0007 $< \text{the value of MSE Regression Main Components} = 2.51$. Then for the value of $R^2$ Ridge Regression = 97.7% $> \text{the value of } R^2 \text{ Regression Main Components} = 91.0\%$. The provisions that apply to the selection of the best method have been met, which have a small MSE value and a large $R^2$. So, it can be concluded that the best method for overcoming the multicollinearity problem in the Regency / City Human Development Index (HDI) data in East Java in 2018 is Ridge Regression Analysis.

4. Conclusion

Based on the results and discussion, the value of MSE Regression Ridge is less than the value of MSE Regression Main Components. Then for the value of $R^2$ Ridge Regression is more than the value of $R^2$ Regression Main Components. The provisions that apply to the selection of the best method have been met, which have a small MSE value and a large $R^2$. So, it can be concluded that the best method for overcoming the multicollinearity problem in the Regency / City Human Development Index (HDI) data in East Java in 2018 is Ridge Regression Analysis. The drawback of previous research is that the multicollinearity assumption test is not strong enough and uses only one method to overcome it. If more than one method, which one is more accurate and reliable can be compared. The author expects a more in-depth study of the Human Development Index (HDI), what other aspects are needed from each of the main Human Development Index (HDI) factors to determine the increase in Human Development Index (HDI) in an area.

5. References

[1] Badan Pusat Statistik (BPS) Provinsi Jawa Timur.2018.Indeks Pembangunan Manusia (IPM).Surabaya:Badan Pusat Statistik (BPS) Provinsi Jawa Timur.
[2] Gorski L and Jakubowska M 2019 Ridge Regression with Self - Paced Learning Algorithm in Interpretation of Voltammetric Signals Chemometrics and Intelligent Laboratory Systems 191 73–81
[3] Hasriani.2014.Perbandingan Regresi Ridge (Regresi Gulud) dan Principal Component Analysis (Analisis Komponen Utama) dalam Mengatasi Masalah Multikolinearitas.Skripsi.Makasar:Fakultas Sains dan Teknologi Universitas Islam Negeri Alauddin.
[4] Mangaraj B K and Aparajita U 2020 Constructing a Generalized Model of The Human Development Index Socio-Economic Planning Sciences 70
[5] Rohmah A 2019.Penerapan Analisis Regresi Ridge untuk Mengatasi Multikolinearitas pada Model Jumlah Pengangguran di Kota Malang.Skripsi.Malang:FMIPA Universitas Negeri Malang.
[6] Sen D, Erazo K, Zhang W, Nagarajahia S, Sun L 2019 On The Effectiveness of Principal Component Analysis for Decoupling Structural Damage and Environmental Effects inBridge Structures Journal of Sound and Vibration 457 280–298
[7] Setiani W 2019.Kombinasi Principal Component Analysis dan GeographicallyWeighted Regression untuk Mengatasi Masalah Multikolinearitas dan Heteroskedastisitas (Studi Kasus Indeks Pembangunan Manusia (IPM) di Kabupaten/Kotas di Jawa Timur Tahun 2017).Skripsi.Malang:FMIPA Universitas Negeri Malang.
[8] Sriningsih M, Hatidjia D, Prang J D 2018 Penanganan Multikolinearitas dengan Menggunakan Analisis Regresi Komponen Utama pada Kasus Impor Beras di Provinsi Sulut Jurnal Ilmiah Sains 18 1
[9] Wasilaine T L, Talakua M W, Lesnussa Y A 2014 Model Regresi Ridge untuk Mengatasi Model Regresi Linier Berganda
yang Mengandung Multikolinieritas (Studi Kasus: Data Pertumbuhan Bayi di Kelurahan Namaelo Rt 001, Kota Masohi) *Jurnal Barekeng* 8 31–37

[10] Yektiningsih E 2018 Analisis Indeks Pembangunan Manusia (IPM) Kabupaten Pacitan Tahun 2018 *ResearchGate* 18 2