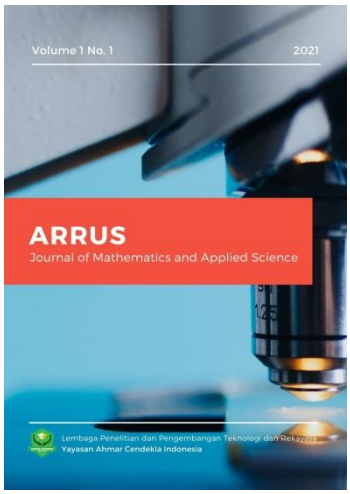
https://doi.org/10.35877/mathscience4813

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



*Corresponding author: Muh. Qodri Alfairus, Accounting Department, Faculty Economic and Bussines, Universitas Negeri Makassar, Makassar City, Indonesia

E-mail: qodri.alfairus@unm.ac.id

RESEARCH / REVIEW ARTICLE

Application of LASSO Regression for the Identification of Underdeveloped Regions in Central Sulawesi

Muh. Qodri Alfairus[1]*, Husnul Amira[2], Agung Tri Utomo[1], & Nur Abshari Abbas[1]

[1]Universitas Negeri Makassar, Makassar, Indonesia.
[2]IPB University, Bogor, Indonesia.

Abstract: This study aims to identify the main factors influencing regional underdevelopment in Central Sulawesi through Human Development Index (HDI) modeling and to develop a robust predictive model. To address the challenges of multicollinearity and the limited number of observations (13 districts/cities with 10 variables), this study employs LASSO (Least Absolute Shrinkage and Selection Operator) regression, which is capable of simultaneously shrinking coefficients and selecting variables. The data used are sourced from the 2019 publication of the Central Statistics Agency (BPS). The analysis was conducted using descriptive statistics, Ordinary Least Squares (OLS) modeling, VIF tests, and LASSO regression with cross-validation (leave-one-out cross-validation). The results indicate that very high multicollinearity (VIF > 10 for most variables) renders the OLS model unstable. Conversely, LASSO regression yielded better performance with superior RMSE (1.282), MAE (1.075), and R^2 (0.918) values compared to OLS (RMSE 21.67; MAE 9.85; R^2 0.78). Thus, LASSO is more suitable for limited data with high multicollinearity. The selected significant variables include the percentage of the poor population, the open unemployment rate, shopping facilities, the presence of hospitals, the population density ratio, and the number of elementary and secondary schools.

Keywords: Lasso, Regression, Multicollinearity, Human Development Index, Central Sulawesi.

1. INTRODUCTION

The issue of regional underdevelopment is a crucial concern in Indonesia's national development, particularly in achieving equitable development and social justice. Underdeveloped regions are areas with relatively low levels of development and community welfare compared to other regions. Identifying the main indicators that influence regional underdevelopment is an important step in formulating effective and targeted policies. Central Sulawesi Province was chosen as the focus of this study because of its complex geographical and socio-economic characteristics, as well as the continued presence of regencies and cities with high levels of development disparity. Indicators such as the Human Development Index (HDI), infrastructure, fiscal capacity, and accessibility often involve numerous variables that are highly correlated with one another (multicollinearity) (BPS, 2023)

Regression analysis is a statistical method used to model the relationship among variables and to make predictions (Neter et al., 1997). However, classical linear regression models have

https://doi.org/10.35877/mathscience4813

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

several assumptions, one of which is that the number of observations must be greater than the number of estimated parameters (Gujarati, 2009). In practice, particularly when dealing with data that involve many variables and limited observations, such as regional data, this assumption is often violated, leading to multicollinearity problems that can make the model unstable and difficult to interpret (Datta et al., 2007). Therefore, an alternative approach is needed to overcome these limitations.

One of the methods that can be used is Least Absolute Shrinkage and Selection Operator (LASSO) regression, introduced by Robert Tibshirani. This method combines shrinkage and variable selection through a penalty on the absolute values of regression coefficients, enabling it to produce a simpler and more interpretable model by eliminating insignificant variables (Tibshirani, 1996). The application of LASSO in this study is expected to generate a stable and efficient model, as well as to identify the dominant indicators influencing regional underdevelopment in Central Sulawesi, thereby providing a basis for more effective and data-driven development policy formulation.

2. Literature Review

2.1. Underdeveloped Regions

Underdeveloped regions are regencies whose territories and communities are less developed compared to other regions at the national level (Sari, 2014). As stated in Presidential Regulation (Perpres) Number 63 of 2020 concerning the designation of underdeveloped regions for 2020–2024, Article 1 paragraph (1) defines underdeveloped regions as regencies whose territories and communities are relatively less developed than other regions on a national scale. Furthermore, Article 2 paragraph (1) states that a region is classified as underdeveloped based on several criteria, including community economy, human resources, infrastructure, regional financial capacity, accessibility, and regional characteristics. These indicators are generally represented through macro measures such as the Human Development Index (HDI), which comprehensively reflects the quality of life of the population (Badan Pusat Statistik, 2023). Regional development disparities remain a significant issue in Indonesia; therefore, quantitative analytical approaches are needed to identify the dominant factors influencing regional underdevelopment.

2.2. Linear Regression Analysis

Linear regression analysis is a statistical method used to model the relationship between dependent and independent variables, as well as for prediction purposes. Classical linear regression models are generally estimated using the Ordinary Least Squares (OLS) method, which aims to minimize the sum of squared errors (Neter et al., 1997). However, the use of OLS requires several important assumptions to be satisfied, such as the absence of multicollinearity, homoscedasticity, and a number of observations greater than the number of estimated parameters (Gujarati, 2009). Violations of these assumptions can cause the model to become unstable and produce biased or inefficient estimates.

2.3. Multicollinearity

Multicollinearity is a condition in which there is a strong linear relationship among independent variables in a regression model. This problem causes the variance of regression coefficients to become large, making parameter estimates unstable and difficult to interpret. One commonly used method to detect multicollinearity is the Variance Inflation Factor (VIF), where a VIF value exceeding 10 indicates serious multicollinearity (Monthgomery et al., 2012). This issue frequently occurs in datasets with many variables or in regional data that exhibit similar characteristics across regions. (Yanke et al., 2022) explained that the Least Absolute Shrinkage and Selection Operator (LASSO) method is an effective regularization approach for addressing multicollinearity in regression models because it can improve the stability of parameter estimation while simultaneously performing variable selection.

https://doi.org/10.35877/mathscience4813

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

2.4. LASSO Regression (Least Absolute Shrinkage and Selection Operator)

LASSO regression is a penalized regression method introduced by Robert Tibshirani to address multicollinearity and overfitting problems by adding a penalty in the form of the absolute sum of regression coefficients (L1 penalty). This approach allows certain coefficients to be shrunk or even reduced to zero, thereby automatically performing variable selection (Tibshirani, 1996). This method has the advantage of producing a simpler and more interpretable model, particularly for data with many variables and limited observations. It has also been proven to be more stable than Ordinary Least Squares (OLS) regression under conditions of high multicollinearity (Hastie et al., 2009), making it highly relevant for identifying dominant factors in the analysis of regional underdevelopment.

3. Research Method and Materials

Before conducting further analysis, it is necessary to examine multicollinearity because it can cause regression model analysis using the least squares method to become insignificant and require further treatment. Therefore, multicollinearity testing is performed on the independent variables using the Variance Inflation Factor (VIF). The VIF test is conducted based on the following hypothesis (Hair et al. 2014):

(1). H_0 : There is no multicollinearity among independent variables.
(2). H_1 : There is multicollinearity among independent variables.
(3). Test Statistic :

$$VIF = \frac{1}{1-R_j^2} > 1 \ , \qquad j = 1,2,\dots,k \tag{1}$$

Where :

R_j : coefficient of determination
k : number of independent variables
Decisiom Rule :
Reject H_0 if any VIF value exceeds 10.

The Least Absolute Shrinkage and Selection Operator (LASSO) regression method was developed by Tibshirani (1996) to address problems related to estimation accuracy and interpretability. LASSO works by shrinking the coefficients of variables that are highly correlated with the error term, causing some coefficients to approach zero. Consequently, this method is also effective in overcoming multicollinearity (Sartika et al. 2020). regression minimizes the sum of squared errors according to the following equation :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2}$$

Where :

y_i = value of the dependent variable for observation i
β_j = regression coefficient for predictor j
x_{ij} = observation iii for predictor j
i = $1,2,\dots,n$; n number of observations
j = $1,2,\dots,k$; k number of independent variables

With the constrain $\sum_{j=1}^{k}|\beta_j| \leq t, t \geq 0$ and $t = \sum_{j=1}^{k}|\hat{\beta}_j|$, where $\hat{\beta}_j$ represents the LASSO estimator at each stage of the variable selection process. where ttt is a tuning parameter that controls the degree of coefficient shrinkage in the LASSO method (Sartika et al. 2020). In Lagrange form, the LASSO regression estimator can be written as:

https://doi.org/10.35877/mathscience4813

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \quad (3)$$

The term $\lambda \sum_{j=1}^{k} |\beta_j|$ is referred to as the L_1penalty $\sum_{j=1}^{k} |\beta_j| = ||\beta||_1$where $\lambda \geq 0$ is the tuning parameter controlling the magnitude of regularization..

Prediction results are evaluated using Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2). RMSE measures the magnitude of prediction error; the smaller the RMSE value, the more accurate the prediction results. The RMSE value is calculated using the following equation:.

$$RMSE = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n}} \quad (4)$$

Where :

N = number of observations/data points.

4. Results and Discussion

4.1. Descriptive Analysis

Table 1: Descriptive Analysis

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Y	68,6	67,8	4,62	64,6	81,5
X1	13,4	14	3,14	6,8	17,4
X2	3,7	3	1,71	2,3	8,3
X3	5,7	2,2	12,05	0,6	45,6
X4	0,7	0,8	0,13	0,5	1
X5	3	1,5	5,64	0,56	21,7
X6	119,4	49,5	249,04	12	944,7
X7	70,6	66,7	11,14	55,7	93,6
X8	223,6	184	100,37	81	422
X9	65,1	63	24,4	36	107
X10	17,2	16	7,55	6	33

Overall, the descriptive statistics indicate that most variables, such as X1, X2, X3, X4, X5, and X10, exhibit relatively stable and homogeneous distributions across regencies/cities, as reflected by their low standard deviations and narrow value ranges. These characteristics suggest that they tend to contribute consistently to the regression model. However, variables X6, X7, and X9 show very high variability due to the presence of extreme outliers, particularly from Palu City, resulting in uneven data distributions that may reduce the quality of analysis when using ordinary linear regression. This condition also triggers multicollinearity problems, as indicated by high VIF values. Therefore, an alternative method such as LASSO regression is required to stabilize parameter estimation, perform automatic variable selection, and produce a more accurate and robust model against outliers and multicollinearity.

4.2. OLS Regression

Ordinary Least Squares (OLS) linear regression analysis was employed in the initial stage of this study to determine the extent to which the predictor variables (X1, X2, X3, …, X10) were able to explain the variation in the response variable Y. The use of OLS aimed to provide a basic overview of the linear relationship patterns among variables before

https://doi.org/10.35877/mathscience4813

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

conducting further advanced modeling. Considering the limited number of observations (n=13), this preliminary analysis was essential for identifying initial model performance as well as potential statistical issues such as multicollinearity and model instability.

Table 2 : OLS Regression Model Statistics

R-Square	Adj R-Square	F-Statistic	P-Value	Residual SE
0,983	0,899	11,76	0,08	1,46

The results in Table 2 indicate that the OLS model has a very high explanatory power, with an R-Square value of 0.983 and an Adjusted R-Square of 0.899, meaning that most of the variation in the response variable can be well explained by the predictor variables. The F-test produced a value of 11.76 with a p-value of 0.08, indicating that the model is significant at the 10% significance level but not at the 5% level. Meanwhile, the Residual Standard Error of 1.46 suggests a relatively low level of prediction error. However, this result still requires further examination through model validation.

4.3. Multicolinearity Test

The multicollinearity test was conducted to identify whether strong linear relationships existed among the predictor variables in the regression model. This assessment was performed using the Variance Inflation Factor (VIF), where VIF values exceeding 10 are generally considered to indicate serious multicollinearity.

Table 3: *Variance Inflation Factor* (VIF) Value

Variabel	VIF
X1	14,05
X2	28,75
X3	425,89
X4	4,87
X5	405,83
X6	508,79
X7	3,46
X8	81,06
X9	47,80
X10	48,57

The calculation results show that the majority of variables have very high VIF values, except for variables X4 and X7, which have VIF values below 10. This indicates the presence of serious multicollinearity among the variables used in the model. These conditions further strengthen the rationale for applying a penalized regression approach, specifically LASSO regression..

4.4. Lasso Regression

Based on the results of the multicollinearity test, significant multicollinearity was identified among the variables. To address this issue, LASSO regression was applied, using the lambda parameter as a reference for determining the variables that significantly influence the developed model. In constructing the LASSO regression model, the Leave-One-Out Cross Validation (LOOCV) method was employed to accommodate the limited number of observations.

Table 4 : LASSO Regression Model Results

R-Square	Lambda	RMSE	MAE
0,918	0,1325	1,282	1,075

The LASSO regression results revealed that variables X1, X2, X3, X4, X7, and X10 significantly influenced the modeling of the Human Development Index (HDI). Using these

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

six variables, the resulting LASSO regression model achieved an R-Square value of 0.918, an RMSE value of 1.282, and an MAE value of 1.075. The optimal lambda value was determined to be 0.1325, as it produced the smallest RMSE compared to other lambda values.

4.5. Best Method Selection

After modeling with both OLS and LASSO, a comparison of the two models was conducted. This comparison aimed to identify the best final model and to demonstrate that handling multicollinearity using LASSO regression improves overall model performance. The comparison and evaluation of the best model were based on RMSE, MAE, and R^2.

Table 5 : Comparison of Regression Results

Model	RMSE	MAE	R^2
OLS Regression	21,67	9,85	0,78
LASSO Regression	1,28	1,07	0,91

Based on Table 5, the RMSE and MAE values of the OLS regression model are higher than those of the LASSO regression model. This indicates that the model produced by LASSO regression is able to provide better predictive performance compared to the model generated by OLS regression. Furthermore, when comparing the R-Square values of both models, the OLS regression model has a lower value than the LASSO regression model. Based on these results, the LASSO regression model is considered superior and more appropriate for predicting the Human Development Index (HDI) in Central Sulawesi Province than the OLS regression model. This finding is consistent with the results reported by Sunandi and Siswantining (2025), who stated that the LASSO regression model is capable of effectively and efficiently reducing variables while identifying the key factors influencing the Human Development Index.

5. Conclusion

This study demonstrates that high multicollinearity exists among the majority of the variables used, including the percentage of poor population, percentage of open unemployment, percentage of villages with shopping facilities, percentage of villages with hospitals, population density ratio, number of elementary schools, number of junior high schools, and number of senior high schools. This condition caused the OLS model to become unstable and less accurate during cross-validation. LASSO regression successfully addressed this issue through coefficient shrinkage and variable selection, where the LASSO model provided the best predictive performance with the lowest RMSE and MAE values, as well as the highest R-Square value. For future research, it is recommended to increase the number of observations, consider other penalized regression methods such as Ridge or Elastic Net, and explore robust approaches, data transformation techniques, as well as spatial or machine learning models to further improve model stability and predictive accuracy.

References

BPS. (2023). Statistika Indonesia. In badan pusat statistik: Vol. 53,2025. https://www.bps.go.id/publication/2020/04/29/e9011b3155d45d70823c141f/statistik-indonesia-2020.html

Datta, Susmita, Jennifer Le-Rademacher, and Somnath Datta. (2007). "Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO." In Biometrics, vol. 63. no. 1. Preprint, Blackwell Publishing Inc. https://doi.org/10.1111/j.1541-0420.2006.00660.x.

Gujarati, D. N., and D. C. Porter. (2009). The McGraw-Hill Series Economics.

Hair, J. F., W. C., Black, B. J., Babin, and R. E. Anderson. (2014). Multivariate Data Analysis_copy.

Monthgomery, D. C., Elizabeth A. Peck, & Vining, G. G. (2012). Introduction Linear Regression Analysis (Fifth). Wiley-Interscience Publication.

https://doi.org/10.35877/mathscience4813

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Neter, J., Wasserman, W., dan Kutner, M. H. (1997). Model Linier Terapan Buku I: Analisis Regresi Linear Sederhana. Penerjemah: Bambang Sumantri. Bogor: Jurusan Statistika FMIPA-IPB.

Sari, R. (2014). Dampak Kebijakan Desentralisasi Fiskal pada Daerah Tertinggal di Indonesia. Jurnal Ekonomi dan Kebijakan Publik, 5(1), 79–99.

Sartika, Imi, Naomi Nessyana Debataraja, and Nurfitri Imro'ah Intisari. 2020. "Analisis Regresi Dengan Metode Least Absolute Shrinkage And Selection Operator (Lasso) Dalam Mengatasi Multikolinearitas." In Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster), vol. 09. no. 1.

Sunandi, E., & Siswantining, T. (2025). Variable selection affecting Indonesian Human Development Index using LASSO. Inferensi Journal, 8(1), 1–12.

Tibshiranit, Robert. (1996). "Regression Shrinkage and Selection via the Lasso." In J. R. Statist. Soc. B, vol. 58. no. 1. https://academic.oup.com/jrsssb/article/58/1/267/7027929.

Yanke, A., Zendrato, N. E., & Soleh, A. M. (2022). Handling Multicollinearity Problems in Indonesia ' s Economic Growth Regression Modeling Based on Endogenous Economic Penanganan Masalah Multikolinieritas pada Pemodelan Pertumbuhan Ekonomi Indonesia Berdasarkan Teori Pertumbuhan Ekonomi Endogenous. 6(2), 228–244.