Elastic-Net Method to Suppress the Presence of Multicollinearity in Dengue Fever Data in Indonesia

Adinda Sarianti[1], Netti Herawati[1*], Agus Sutrisno[1], Subian Saidi[1]
[1]Department of Mathematics, University of Lampung, Bandar Lampung, Indonesia, 35145
Email address: netti.herawati@fmipa.unila.ac.id

Abstract— *The purpose of this study is to determine the performance of the elastic-net method in overcoming multicollinearity problems and compare its estimated value with the Ordinary Least Squares (OLS) on dengue case data in Indonesia as well as to find out what factors affect dengue cases in Indonesia. The results show that the elastic-net method can supress multicollinearity problems and produce better regression coefficient estimates compared to OLS based on AIC measurements. The analysis showed that the factors that influence dengue cases in Indonesia are number of environmental health workers (X_2), the percentage of households with proper sanitation (X_3), the percentage of households with decent housing (X_4), the percentage of households with decent drinking water (X_5), the amount of rainfall (X_6), number of rainy days (X_7), and population density (X_8).*

Keywords—*Multicollinearity, Elastic-Net, AIC, Dengue Hemorrhagic Fever (DHF).*

I. Introduction

Regression analysis is a method for examining how a dependent variable is influenced by one or more independent variables [1]. However, in its application, there are many specific problems when the analysis process is carried out, one of which is the problem of multicollinearity. Multicollinearity is one of the problems that often arises when performing multiple regression analysis, which occurs when there is a strong correlation or relationship between two or more independent variables [2]. The existence of multicollinearity can cause the estimated coefficients in the regression model to be unstable so that the standard error produced is large. [3]. Several methods can be used to deal with multicollinearity problems, including adding new samples, replacing or removing variables that have high correlation, regression of major components, and several other methods [4]. According to [5], the problem of multicollinearity can also be overcome by using the regularization method, which can shrink the estimation coefficient to zero. The regularization methods that are often used are Regression Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic-Net.

The Elastic-Net Regression method is a method that combines the penalty values of Ridge Regression and LASSO, where this regression is able to overcome the shortcomings of Ridge Regression and LASSO which can shrink the exact regression coefficient of zero as well as select variables simultaneously [6].

Based on the explanation, the analysis will be carried out on the data of dengue cases in Indonesia to see the performance of the Elastic-Net Regression method in overcoming the problem of multicollinearity and compare the value of the estimated regression coefficient of Elastic-Net with Ordinary Least Squares (OLS) and what factors affect the dengue cases in Indonesia from the regression model produced. A comparison between the Elastic-Net and OLS methods was made to determine which method is better in addressing the multicollinearity problem based on the lowest AIC value of each regression model.

II. Literature Review

2.1 Regression Analysis

Regression analysis is a method for examining how a dependent variable is influenced by one or more independent variables [1]. There are two types of regression analysis, namely simple linear regression used for data with one dependent variable and one independent variable, and multiple linear regression used for data with one dependent variable and more than one independent variables. If there are a number of k independent variables, $X_1, X_2, \dots, X_k$ and dependent variable Y, then the linear regression model has the following form:

$$Y_i = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j + \varepsilon_i \qquad (1)$$

with, Y_i = dependent variable, X_{ij} = independent variable, β_0 = intercept, β_j = slope/coefficient, and ε_i = error.

2.2 Ordinary Least Squares (OLS)

The least squares method, also known as Ordinary Least Squares (OLS) is used to determine estimate β_0, β_1, β_2, …, β_k by minimizing the Sum of Squares Errors (SSE) [4].

The value of β can be obtained by minimizing the square form:

$$Q(\hat{\beta}_j) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 \qquad (2)$$

In the form of a matrix the number of squares of error ε_i^2 can be written as follows:

$$\varepsilon_i^T \varepsilon_i = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n]\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 = \sum \varepsilon_i^2 \qquad (3)$$

Based on the general equation multiple linear regression with the matrix is obtained:

$$\varepsilon = Y - X\beta \qquad (4)$$

Therefore, the multiplication of the error matrix can be written as follows:

$$\varepsilon_i^T \varepsilon_i = (Y - X\beta^T)(Y - X\beta) \tag{5}$$

Then, a derivative of $\varepsilon_i^T \varepsilon_i$ partially to β and equated with zero.

$$\frac{\partial \varepsilon_i^T \varepsilon_i}{\partial \beta} = 0 - 2Y^T X\beta + \beta^T X^T X\beta = 0 \tag{6}$$

So that the following estimates for the OLS are obtained:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{7}$$

2.3 Multicollinearity

According to [7], multicollinearity describes a condition where some or all of the independent variables in a regression model are highly linearly correlated. There are several ways to detect the presence or absence of multicollinearity in a regression model, including, seeing a strong correlation between independent variables where, if there is a fairly high correlation between independent variables, say 0.8, then this indicates a multicollinearity problem. Then, looking at the Variance Inflation Factor (VIF) value in the regression model, the VIF value can be found using the following formula:

$$VIF_{(j)} = \frac{1}{1 - R_j^2} \tag{8}$$

with R_j^2 is the determination coefficient obtained from the X_j independent variable which is regressed with other regression variables. When the VIF value is above 10, it points to a potential multicollinearity problem.

2.4 Elastic-Net

Elastic-Net is a penalized regression with penalties from both Ridge and LASSO techniques, designed to manage problems related to multicollinearity. According to [6], the Elastic-Net method can reduce specific regression coefficients to zero and simultaneously performs variable selection, including the ability to choose groups of correlated variables. Elastic-Net was developed to overcome the shortcomings of LASSO Regression, where LASSO regression has the following drawbacks, when the number of variables (*p*) is greater than the number of observations (*n*), LASSO only selects *n* variables to be included in the model. Next, when a group of highly correlated variables exists, LASSO tends to randomly choose one variable from that group. Then, if the number of variables (p) is smaller than the number of observations (n), Ridge regression will have a greater impact on LASSO performance.

The Elastic-Net penalty is written as follows:

$$\sum_{j=1}^{p} [\alpha|\beta_j| + (1-\alpha)\beta_j^2]$$

The coefficient estimator in Elastic-Net can be written as follows:

$$\hat{\beta}^{net} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{9}$$

where $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$, $0 \le \alpha \le 1$.

The α value is a combination of shrinkage coefficients between Ridge Regression and LASSO Regression. When α = 0, the penalty applied is that of Ridge Regression, but if α = 1, the penalty is that of LASSO Regression.

2.5 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a method that is also used to select the best model found by Akaike and Schwarz. According to [8], the best model is determined by looking at the smallest AIC value. The calculation of the AIC value is carried out using the following formula:

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2p \tag{10}$$

with SSE = sum of squares error, *n* = number of data, anɑ *p* = number of parameters.

III. Methodology

The data used in this study is data on Dengue Hemorrhagic Fever (DHF) cases in Indonesia with the number of data (*n* = 34) and the number of independent variables (*p* = 9). The variables used in this data are number of hospitals (X_1), number of environmental health workers (X_2), percentage of households with proper sanitation (X_3), percentage of households with decent housing (X_4), percentage of households with decent drinking water (X_5), amount of rainfall (X_6), number of rainy days (X_7), population density (X_8), and number of poor people (X_9). This research began by looking at the VIF value on the data to find out the problem of multicollinearity. Next, Cross Validation (CV) is used to determine the optimal λ value for the Elastic-Net Regression. The best model determination is done by comparing OLS and Elastic-Net regression models based on the smallest AIC value.

IV. Result and Discussion

4.1 Identification of Multicollinearity

Before analyzing the data, multicollinearity is first checked by looking at the Variance Inflation Factor (VIF) value.

TABLE 1. VIF Value of Each Independent Variables

Variables	VIF
X_1	16,650395
X_2	7,356057
X_3	1,927731
X_4	2,713775
X_5	3,115674
X_6	1,765213
X_7	1,170997
X_8	2,559020
X_9	25,861271

Table 3 above shows that there are two variables with a VIF value greater than 10, namely in variables X_1 and X_9 so that it can be stated that the independent variables in the data used are detected multicollinearity problems.

4.2 Model Analysis with the Ordinary Least Squares (OLS)

In this study, a regression model of factors affecting dengue in Indonesia in 2021 will be formed through multiple linear regression analysis using OLS. The following is an estimate of the OLS parameters generated from the data used.

TABLE 2. Parameter Estimation with OLS

Intercept	-1,216
X_1	-2,669
X_2	-8,422
X_3	1,405
X_4	-1,452
X_5	1,145
X_6	1,517
X_7	-1,649
X_8	-1,024
X_9	5,375

Based on Table 2. then a multiple linear regression model with OLS is formed as follows:

$$\hat{Y} = -1{,}216 - 2{,}669\ X_1 - 8{,}422\ X_2 + 1{,}405\ X_3 - 1{,}452\ X_4 + \ 1{,}145\ X_5$$
$$+1{,}517\ X_6 - 1{,}649\ X_7 - 1{,}024\ X_8 + \ 5{,}375\ X_9$$

4.3 Model Analysis with Elastic-Net Regression

The optimal shrinkage parameter (λ) for Elastic-Net Regression is obtained by using Cross Validation (CV), namely by selecting a value λ that produces a minimum CVE value, where the λ value will produce the estimated value and the most optimal error value that minimizes MSE in Elastic-Net Regression. The CV was carried out with 10-fold cross validation. Figure 1. shows the CV value for each log (λ).

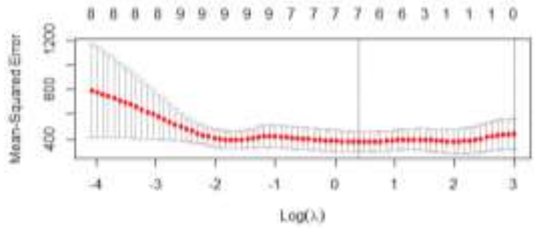

Fig. 1. Plot Cross Validation of the Shrinkage Parameter (λ) for Elastic-Net Regression

Based on Fig. 1, it can be seen that there are two vertical lines on the CV that represent the optimal λ values with the minimum CVE. The first vertical line shows the minimum log (λ) value (λ_{min}) and the second vertical line shows the largest log value (λ) in one standard error λ minimum (λ_{1se}). From the plot, the minimum CVE value was obtained which is 1.484 where log(1.484) = 0.3947411.

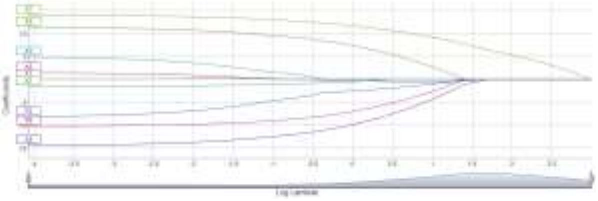

Fig. 2. Elastic-Net Regression Coefficient Plot

By using the optimal shrinkage parameter (λ) in the Elastic-Net Regression model, the Elastic-Net Regression coefficients for various λ values are obtained as shown in Fig. 2, where the coefficients for the variables $X_2, X_3, X_4, X_5, X_6, X_7, X_8$ shrink close to zero, while the coefficients for variables X_1 and X_9 shrink exactly to zero. This aligns with the principle of Elastic-Net Regression, which not only shrinks coefficients close to zero but also performs variable selection by shrinking some regression coefficients exactly to zero. Therefore, using the optimal λ value, the Elastic-Net Regression model is as follows:

$$\hat{Y} = -7{,}026044 \times 10^{-15} - 2{,}064105\ X_2 + 1{,}085229\ X_3 - 8{,}090547\ X_4$$
$$+ 5{,}838272\ X_5 - 4{,}457860\ X_6 - 1{,}258460\ X_7 - 5{,}628242\ X_8$$

4.4 Comparison of Regression Coefficients for OLS and Elastic-Net Models

The following is a table showing the regression coefficients for each of the independent variables in the OLS and Elastic-Net methods.

TABLE 3. Regression Coefficients in the OLS and Elastic-Net Models

Variables	**OLS**	***Elastic-Net***
X_1	-2,669	0,000000
X_2	-8,422	-2,064105
X_3	1,405	1,085229
X_4	-1,452	-8,090547
X_5	1,145	5,838272
X_6	1,517	-4,457860
X_7	-1,649	-1,258460
X_8	-1,024	-5,628242
X_9	5,375	0,000000

From Table 3. it can be seen that there is a shrinkage of regression coefficients in the Elastic-Net Regression model for variables that experience multicollinearity problems, specifically X_1 and X_9. The regression coefficient for X_1 with OLS is -2.669 and shrinks exactly to zero in Elastic-Net. Then, the regression coefficient for X_9 with OLS is 5.375 and shrinks exactly to zero on Elastic-Net. The independent variables that experience shrinkage to zero in the Elastic-Net will be excluded from the selected model. Therefore, the independent variables selected by the Elastic-Net model are those that do not have a significant effect on the dependent variable (Y).

4.5 Factors Affecting Dengue Hemorrhagic Fever (DHF) in Indonesia

Based on the model obtained, each factor that affects DHF cases will be studied. According to research conducted by [9], with so many health workers, it is important for them to support efforts to change community behavior, especially in terms of dengue prevention. Therefore, the role of health workers plays an important role in changing people's behavior to create an environment free from larvae, so the number of DHF cases depends on the number of health workers.

According to research by [10], material quality is one of the qualities that can describe the welfare of the house such as the type of roof, floor and wall that is the widest used, including other supporting facilities which include drinking water facilities, defecation facilities and fecal disposal sites with tanks. A house can be called livable if it has met several criteria for the quality of the house. So that the percentage of households with livable houses has a very important effect on the number of dengue cases, because if the percentage is small, this can increase the number of dengue cases, because uninhabitable houses can increase the transmission of diseases through mosquitoes [11].

According to research by [12], the incidence of DHF has a close relationship with home environmental sanitation that can create ideal conditions for *Aedes aegypti* mosquitoes to breed. The environmental sanitation factors that can affect DHF disease include water reservoirs, garbage disposal systems, lighting, and the presence of larvae. Thus, the number of dengue cases depends on what percentage of the population has proper sanitation.

According to [13], access to proper drinking water refers to the availability of water sources with good quality and maintained quality, so that it is safe for consumption by the community, where the percentage of drinking water facilities that meet the requirements has a significant effect on the number of DHF cases.

According to research by [14], rainfall is the amount of rainwater that falls during a certain period, and puddles arising from rain can be a breeding ground for *Aedes aegypti* mosquitoes. The thing that triggers the development of *Aedes aegypti* mosquitoes is not only due to rainfall but also depends on rainy days, where when the amount of rainfall and rainy days increases, the number of DHF cases will increase.

According to [15], population density is the ratio of population to area, where high density can cause dense settlements and accelerate the spread of DHF.

V. Conclusion

Based on the results of research and discussions that has been carried out to overcome the problem of multicollinearity in the data on dengue cases in Indonesia in 2021 using Elastic-Net Regression, it can be concluded that the Elastic-Net Regression model is better at estimating independent variables than the OLS model based on the AIC values obtained. Therefore, the best model with Elastic-Net Regression on dengue case data in Indonesia in 2021 is obtained as follows:

$$\hat{Y} = -7{,}026044 \times 10^{-15} - 2{,}064105\,X_2 + 1{,}085229\,X_3 - 8{,}090547\,X_4 \\ + 5{,}838272\,X_5 - 4{,}457860\,X_6 - 1{,}258460\,X_7 - 5{,}628242\,X_8$$

From this model, it can be seen that dengue cases are influenced by number of environmental health workers (X_2), the percentage of households with proper sanitation (X_3), the percentage of households with livable houses (X_4), the percentage of households with decent drinking water (X_5), the amount of rainfall (X_6), the number of rainy days (X_7), and population density (X_8).

References

[1] W. H. David & L. Stanley, Applied Logistic Regression, John Wiley and Sons, pp. 1-2, 2000.

[2] A. M. Yeremia, T. S. Delby, & A. H. K. Hanny, “Kajian Model Prediksi Metode Least Absolute Shrinkage and Selection Operator (LASSO) pada Data Mengandung Multikolinearitas,” *Jurnal Matematika dan Aplikasi*, vol.10, issue 2, pp. 69-75, 2021.

[3] A. Shady, “Evaluation of Ridge, Elastic-Net, and LASSO Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study,” *Journal of Applied Economics and Business Studies*, vol.5, issue 1, pp. 131-142, 2021.

[4] C. M. Douglas, A. P. Elizabeth, & V. Geoffrey, Introduction to Linear Regression Analysis, John Wiley and Sons, pp. 70-319, 2012.

[5] F. F. Alin, A. Winalia, & A. Dian, “Performa Teknik Regularisasi dalam Penanganan Masalah Multikolinearitas,” *DJMA: Diophantine Journal of Mathematics and Its Applications*, vol.2, issue 1, pp. 46-51, 2023.

[6] Z. Hui & H. Trevor, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, issue 2, pp. 301-320, 2005.

[7] N. G. Damodar & C. P. Dawn, Basic Econometrics, McGraw-Hill Irwin, pp. 321-341, 2009.

[8] H. Netti, W. Ameliana, S. Agus, Nusyirwan, & Misgiyati, “The Performance of Ridge Regression, LASSO, and Elastic-Net in Controlling Multicollinearity: A Simulation and Application,” *Journal of Modern Applied Statistical Methods*, vol. 23, issue 2, pp. 3-13, 2024.

[9] S. Agung, D. Wawan, A. Hairil, H. Julius, & Fibrianti, “Faktor yang Mempengaruhi Pemberantasan Sarang Nyamuk (PSN) Melalui 3M Plus dalam Upaya Pencegahan Demam Berdarah Dengue (DBD),” *Jurnal Ilmu Kesehatan Masyarakat*, vol.11, issue 1, pp. 23-32, 2022.

[10] A. Farha, “Gambaran Sarana Sanitasi di Kawasan Pesisir (Studi di RT 03 Dusun Hurnala 1 Desa Tulehu Kecamatan Salahutu,” *Global Health Science*, vol.6, issue 3, pp. 118-123, 2021.

[11] H. S. Nurul & R. H. S. Jeffry, “Penyusunan Indeks Kerawanan Sosial Demam Berdarah Dengue Provinsi-Provinsi di Indonesia Tahun 2019, in *Seminar Nasional Official Statistics*, Jakarta Timur, pp. 373-382, 2021.

[12] M. Fatin, P. Suci, & A. T. T. Agustina, “Analisis Hubungan Kondisi Sanitasi Lingkungan dan Perilaku Keluarga dengan Kejadian Demam Berdarah Dengue di Kota Pontianak,” *Jurnal Teknologi Lingkungan Lahan Basah*, vol.10, issue 2, pp. 215-228, 2022.

[13] G. I. P. Rafdi, “Sistem Informasi Geografis pada Kasus Demam Berdarah Dengue di Kabupaten Sidoarjo Tahun 2019,” *Media Gizi Kesmas*, vol.12, issue 1, pp. 367-373, 2023.

[14] A. Kiki, O.A. Latifa, & F. Lasmi, “Pemodelan Incidance Rate Dema Berdarah Dengue di Indonesia yang Berkaitan dengan Faktor Lingkungan Menggunakan Metode Geographically Weighted Regression (GWR),” *Ekologia: Jurnal Ilmiah Ilmu Dasar dan Lingkungan Hidup*, vol.20, issue 2, pp.64-73, 2020.

[15] A. S. Esa, I. R. Hanum, F. M Ridwan, W. Wahyu, I. Yansi, & N. Rani, “Perbandingan Regresi OLS dan Robust MM-Estimation dalam Kasus DBD di Indonesia 2018,” *Jurnal Education and Development*, vol.8, issue 2, pp.68-74, 2020.