A study on group lasso for grouped variable selection in regression model

E Sunandi\textsuperscript{1,2}, K A Notodoputro\textsuperscript{2}, and B Sartono\textsuperscript{2*}

\textsuperscript{1}Statistics Program, Department of Mathematics, Universitas Bengkulu, Indonesia
\textsuperscript{2}Department of Statistics, Institut Pertanian Bogor (IPB), Indonesia

*Corresponding author: bagusco@apps.ipb.ac.id

Abstract. Estimation of regression parameters using the Least Squares (LS) method could not be performed when the number of explanatory variables exceeds the number of observations. An approach that can solve the problem is the LASSO (Least Absolute Shrinkage and Selection Operator) method. This method produces a stable model but with slight bias as the trade-off. Yuan and Lin [6] introduced the Group LASSO method which can be used when there are grouped structure in the variables. This current paper provided a study of the performance of the Group LASSO method through a simulation with several different scenarios. Furthermore, the Group LASSO method was applied to the Human Development Index (HDI) data of Bengkulu Province in 2019. The simulation yielded that the Group LASSO analysis was better than LASSO in term of its Mean Squared Error of Prediction (MSEP), False Negative Rate (FNR) and R-Squared. In the application of the approach to the HDI data, our result was in line with the simulation results that the analysis of Group LASSO was better than LASSO with MSEP Group LASSO of 0.25 and R-Squared of 98%.

1. Introduction

Regression analysis is the most common method to examine the relationship between some explanatory variables (X) and a response variable (Y) [1]. The analysts frequently use least squares (LS) method which works to find model estimate by minimizing the sum of the squared errors. If X is the design matrix from the dataset, LS can be implemented when the X'X matrix is nonsingular [2]. For the case when the number of variables exceed the number of observations, it is impossible to have non-singular X'X so that LS could not work any longer. One way to solve this problem is the penalized regression method which estimates the parameters by minimizing of the sum of squared errors that are penalized [3]. The well-known penalized regression methods is the LASSO (Least Absolute Shrinkage and Selection Operator) method which was popularized by Tibshirani [4].

LASSO works by adding constraints to the LS method. The LASSO shrinks the solution of LS to zero even right to zero especially for the coefficients of less important variables. Therefore, LASSO could work as selection procedure to result simpler and easier model to interpret. However, this approach leads some bias in the estimation. Previous research has applied the LASSO method to data that has many explanatory variables [5]. As a result, the regression coefficients generated from the LASSO method are able to select explanatory variables than the common regression method. Apart from that, LASSO can also solve multicollinearity problems in regression analysis.
The original version of LASSO method has a limitation in the precision when it is applied in data with grouped variable. Yuan and Lin [6] introduced a new method called Group LASSO to handle this situation. This study provided a study of the performance of the Group LASSO method through a simulation with three different scenarios. We also implemented the Group LASSO method to the Human Development Index (HDI) data of Bengkulu Province in 2019 which includes several groups of variables as explanatory variables. Those variables are Expenditure per Capita Per Month (ECM), Expected Years of Schooling (EYS), Gross Regional Domestic Product (GRDP), and Life Expectancy Rate (LER) in Bengkulu Province [7]. Goodness of fit measures used in this study are Mean Squared Error of Prediction (MSEP), R-squared ($R^2$), False Discovery Rate (FDR), and False Negative Rate (FNR).

2. Method and Data

2.1 Method

2.1.1 Standard Normal Variate (SNV) Method

Standard Normal Variate (SNV) is the most common method for normalize the data with the following formula [8]

$$\text{SNV}_k = \frac{x_k - \bar{x}_k}{S_k}; \quad k = 1, ..., p = \sum_{j=1}^{J} p_j,$$

(1)

where $\text{SN}_k$ is the k-th explanatory variable from the standardization, $x_k$ is the k-th explanatory variable before standardization, $\bar{x}_k$ is the mean of the k-th explanatory variables before standardization, and $S_k$ is the deviation of the explanatory variable before standardization. Normalization is needed to standardize the explanatory variables because the coefficient estimator obtained cannot be compared with the results if the original variable is not standardized. It is known as non-equivariant [9].

2.1.2 Group LASSO

In many problems, regression is used to find important explanatory factors in predicting response variables, where each explanatory factor can be represented by a group of variables/factors. Several studies on the Group LASSO have been conducted before [10-13]. Consider common regression model with J groups

$$Y = \sum_{j=1}^{J} X_j \beta_j + \epsilon,$$

(2)

where $Y$ is an $n \times 1$ vector, $\epsilon \sim N(0, \sigma^2 I), X_j$ is an $n \times p_j$ matrix corresponding to the j-th group and $\beta_j$ is a coefficient vector of size $p_j; j = 1, ..., J$. Denoting $X = (X_1, X_2, ..., X_J)$ and $\beta = (\beta_1', ..., \beta_J')$, equation (2) can be written as $Y = X\beta + \epsilon$.

LASSO method is a special case of equation (2) when $p_1 = ... = p_J = 1$. According to Tibshirani [4], the LASSO estimator is defined as

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \frac{1}{2N} \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right\},$$

(3)

where $\lambda$ is a tuning parameter and $\|\|_1$ stands for the vector $l_1$-norm [4]. When $X$ is orthonormal, the Group LASSO estimator with tuning parameter $\lambda$ can be given as [6]
\[
\beta_j = \left( 1 - \frac{\lambda \sqrt{p_j}}{\|X'_jY\|} \right) X'_jY; \; j = 1, ..., J. 
\]

(4)

As \( \lambda \) descends from \( \infty \) to 0, Group LASSO follows a piecewise linear solution path with changepoints at \( \lambda = \frac{\|X'_jY\|}{\sqrt{p_j}}; \; j = 1, ..., J. \)

2.1.3 Leave-One-Out Cross-Validation (LOOCV)

Cross validation is used to determine the optimal \( \lambda \) value so that the best model is obtained. LOOCV is a method that deals with cross validation [14]. The k-fold CV is a form of LOOCV. In the k-fold CV method, observations are randomly divided into k folds of relatively the same size. The first fold is set as the validation data while the other as the training data. The value of the Mean Squared Error (MSE\(_k\)) is calculated based on the predetermined validation data. This step is repeated k times for each fold of observations that are different as validation data. Therefore, k estimated MSE values were obtained, MSE\(_1\), MSE\(_2\), ..., MSE\(_k\). The formula of k-fold CV, i.e.

\[
CV_k = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i. 
\]

(5)

In this study data divided into \( t \) training data sets and \( v \) validation data sets where \( n = t + v \) [9]. The training data set is used to determine the estimated \( \beta \) value. The validation data set is used to test the goodness of the model from the results of the predicted \( \beta \) value in the training data set. The cross-validation process used validation data.

2.1.4 Goodness of Fit Measures

The Goodness of fit measures that are used in this research are Mean Squared Error of Prediction (MSEP), R-squared \((R^2)\), and several measures that are measured from the confusion matrix. The main measure used to see the goodness of the resulting model is MSEP, according to references from Yuan and Lin [6], who used the error value as the main measure when comparing Group LASSO and LASSO models. The MSEP value is calculated based on validation data. Meanwhile, the R-Squared value is used to see how far the explanatory variables in the resulting model can explain the response variable. The R-Squared value is calculated based on training data. Through this measure, it can be concluded that the best model between the LASSO and Group LASSO analyzes, i.e. with a smaller MSEP value. The MSEP formula is as follows [13]

\[
\text{MSEP} = \frac{1}{v} \sum_{i=1}^{v} \frac{(y_i - \hat{y}_i)^2}{v}, 
\]

(6)

where \( v \) is the number of validation data, \( y_i \) is the response variable and \( \hat{y}_i \) is the predicted value of the response variable. Furthermore, \( R^2 \) is a percentage value that shows how much the explanatory variables in the model can explain the response variable. Its value is calculated based on training data. The formula of \( R^2 \) is [15]

\[
R^2 = \frac{\sum_{i=1}^{t} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{t} (y_i - \bar{y}_t)^2}, 
\]

(7)
| Predicted data | Actual data |      |      |
|----------------|-------------|------|------|
|                | Positive    | True Positive (TP) | False Positive (FP) |
| Positive       | True Positive (TP) |      |      |
| Negative       | False Negative (FN) | True Negative (TN) |      |
| TP + FN        | FP + TN     |      |      |

Meanwhile, the regression parameters ($\beta$) are set with the classification of non-zero beta and zero beta. In this study, it was also investigated whether non-zero beta would be estimated by non-zero beta or conversely. The accuracy of the regression parameter estimator is measured by using confusion matrix. Confusion matrix can be interpreted as a tool that has a function to analyze whether the model is good at recognizing the estimator values from different classification classes (see Table 1) [16]. In this study, the focus is the beta classification (non-zero beta and zero beta). In a confusion matrix, there are four terms as a representation of the results of the classification process. They are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The True Negative (TN) value is the number of zero beta detected correctly. The False Positive (FP) is zero beta but is detected as non-zero beta. True Positive (TP) is the number of non-zero beta detected correctly. False Negative (FN) is number of non-zero beta, but detected as zero beta [17].

There are two values that are measured using the confusion matrix, i.e the False Discovery Rate (FDR) and the False Negative Rate (FNR). The FDR value describes the probability of the zero-beta error that is suspected as non-zero beta. The FDR value can be obtained by Equation 8. The FNR value shows the probability of non-zero beta error which is estimated to be zero beta. The FNR value is obtained by Equation 9. The smaller the FDR and FNR values describe the method used the better in the parameter estimation [18].

\[
FDR = \frac{FP}{FP + TP} \tag{8}
\]

\[
FNR = \frac{FN}{TP + FN} \tag{9}
\]

### 2.2 Data

The data used in this study are simulation data and empiric data. Simulation provided the performance of the LASSO and Group LASSO methods. The simulation scenarios can be seen in Table 2. The formula used to generate Y data is as follows.

- **Simulation 1:** three continuous groups are represented by a third-order polynomial and 3 categorical groups is trichotomized as 0, 1 or 2.
  \[
  Y = 0.5 + 0.940X_1 + 0.422X_1^2 + 0.822X_1^3 + 0.786X_2 + 0.361X_2^3 + 0.956X_3 + 0.473X_2^2 + 0.453X_3^3 + 0.359[Z_1 = 1] + 0.532[Z_1 = 2] + 0.002[Z_2 = 1] + 0.533[Z_2 = 2] + 0.824[Z_3 = 1] + 0.577[Z_3 = 2] + \text{epsilon}.
  \]

- **Simulation 2:** three continuous groups are represented by a third-order polynomial and 3 categorical groups is trichotomized as 0, 1 or 2; 3 continuous groups are represented by a third-order polynomial with zero beta.
  \[
  Y = 0.5 + 0.764X_1 + 0.889X_1^2 + 0.634X_1^3 + 0.922X_2 + 0.472X_2^2 + 0.251X_2^3 + 0.311X_3 + 0.419X_2^2 + 0.420X_3^2 + 0.746[Z_1 = 1] + 0.200[Z_1 = 2] + 0.856[Z_2 = 1] + 0.030[Z_2 = 2] + 0.901[Z_3 = 1] + 0.893[Z_3 = 2] + \text{epsilon}.
  \]

- **Simulation 3:** three continuous groups are represented by a third-order polynomial and 3 categorical groups is trichotomized as 0, 1 or 2; 3 continuous groups are represented by a third-order polynomial with zero beta.
\[ Y = 0.5 + 0.700X_1 + 0.462X_1^2 + 0.499X_1^3 + 0.490X_2 + 0.461X_2^2 + 0.155X_3^3 + 0.056X_3 + 0.819X_2^2 + 0.329X_3^3 + 0.179[Z_1 = 1] + 0.290[Z_3 = 2] + 0.839[Z_2 = 1] + 0.389[Z_2 = 2] + 0.541[Z_3 = 1] + 0.074[Z_3 = 2] + \text{epsilon}. \]

**Table 2. The Simulation Scenarios**

| Simulation | Number of observations (n) | Number of variables (p) | Number of groups (J) | Number of simulation samples (S) |
|------------|-----------------------------|-------------------------|----------------------|---------------------------------|
| Simulation 1 \(\left(\frac{p}{n} < 1\right)\) | 100                         | 18                      | 6                    | 1000                            |
| Simulation 2 \(\left(\frac{p}{n} < 1\right)\) | 50                          | 36                      | 12                   | 1000                            |
| Simulation 3 \(\left(\frac{p}{n} > 1\right)\) | 30                          | 36                      | 12                   | 1000                            |

Meanwhile, empiric data was obtained from the Statistics of Bengkulu Province in the book "Bengkulu Province in Figures 2020". As the unit of observation are districts/cities in Bengkulu Province. As for the variables used in this study, as a response variable, namely the Human Development Index (HDI) of Bengkulu Province in 2019. The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable, and have a decent standard of living. Based on the HDI compiler indicators, several groups of variables used are Expenditure Per Capita Per Month (ECM), Expected Years of Schooling (EYS), Gross Regional Domestic Product (GRDP), and Life Expectancy Rate (LER) in Bengkulu Province [7]. All groups are continuous variables. The definitions of the variables of each group can be seen in Table 3.

**Table 3. Details of the Variables in the Group**

| Group | Definition |
|-------|------------|
| HDI | Costs incurred (in rupiah) for the consumption of all household members for a month are divided by the number of household members. |
| 1. Expenditure Per Capita Per Month (ECM) | |
| 2. Expected Years of Schooling (EYS) | Length of schooling (in years) that the child is expected to experience at a certain age in the future. |
| 3. Gross Regional Domestic Product (GRDP) | Ability of an area to create the output (value added) at a certain time (in billions). |
| 4. Life Expectancy Rate (LER) | The average life year that will still be lived by someone who has reached the age of x, in a certain year, in a situation of mortality that prevails in his community. |
Based on a preliminary analysis that there were non-linear effects of both HDI and each explanatory variable, it is known that all groups of variables are represented by a third-order polynomial (alpha 5.5%). This can be seen based on the p-value of each effect in Table 4 [19]. So that for modeling with the Group LASSO in the application of HDI data, polynomial effects up to order 3 of each group are used as explanatory variables.

Table 4. The Significance of the polynomial effects Explanatory Variables

| Group | Linear | Quadratic | Cubic |
|-------|--------|-----------|-------|
| ECM   | 0.001  | 0.001     | 0.001 |
| EYS   | 0.000  | 0.000     | 0.000 |
| GRDP  | 0.000  | 0.000     | 0.000 |
| LER   | 0.054  | 0.001     | 0.001 |

2.2.1 Data Analysis Procedures

This research method is carried out in two stages. The first is to examine the performance of the LASSO and Group LASSO methods through simulation. The second stage is the application of the LASSO and Group LASSO methods to the 2019 Bengkulu Province HDI data. The simulation algorithm are as follows.
1. Determined the number of groups J.
2. Determined \( n = \sum_{j=1}^{J} p_j \) (number of variables), \( n \) (number of observations), \( S \) (number of simulation samples).
3. Generated dummy variables with size \( n \times \alpha \) for each of 3 categories.
4. Generated \( X \) \( n \times b \) (\( a + b = p \)) matrix from Normal Multivariate distribution \( (0, \sigma) \). set the correlation between \( X \) is high \((\rho = 0.9)\). Furthermore, the matrix \( X \) is standardized becomes \( X^* \).
5. Generated \( \beta \sim uniform(0,1) \).
6. Generated \( \epsilon \sim N(0,0.3) \).
7. Set \( \beta_0 = 0.5 \).
8. Calculated the generated \( Y \ n \times 1 \) vector with the formula \( Y = 0.51_n + X^*\beta + \epsilon \).
9. Analyzed Grup LASSO and LASSO:
   a) selected 80% training data and 20% testing data,
   b) determined the best lambda, and
   c) estimated the coefficient of each explanatory variable.
10. Determined the goodness of fit measures.
11. Interpreted results.

Analogous to the simulation flow, the real data analysis procedure uses the Group LASSO and LASSO method as follows.
1. Data exploration.
2. Data standardization.
3. Analyzed Grup LASSO and LASSO:
   a) selected 80% training data and 20% testing data,
   b) determined the best lambda, and
   c) estimated the coefficient of each explanatory variable.
4. Determined the the goodness of fit measures.
5. Interpreted results.
3. Results and Discussion

3.1 Simulation Results

The simulation was designed to assess the performance of the Group LASSO method with LASSO as a comparison method. In this study, a simulation was carried out with three scenarios, each generating 1000 times of simulation data. The detailed simulation results are summarized in Table 5. In simulation 1, all beta generated are non-zero beta at condition $p < n$. Based on the simulation results, it can be seen that FDR and FNR values in the LASSO and Group LASSO analysis are the same, i.e. 0%. Group LASSO and LASSO estimated all beta values that are not equal to zero. If seen from the R-Squared value, both analyzes have the same ability to explain the diversity of the response variables very well ($R^2 = 98\%$). When viewed from the MSEP value, the Group LASSO analysis has a smaller value (MSEP = 1.76) than compared to LASSO.

In simulation 2 at the condition of $p < n$, the beta parameter was set as non-zero beta and zero beta, with a ratio of 50:50. The results of simulation 2 were different from simulation 1. For the Group LASSO analysis, the FDR and FNR values were similar, i.e 37.50%. It means, the Group LASSO was quite capable of classifying the beta estimator according to the true beta conditions. Meanwhile, LASSO was very good estimating the non-zero beta correctly. This was indicated by the FDR value of 0%. The FNR value was still relatively large (42.86%). This indicates that LASSO was weak in estimating zero beta correctly. From the MSEP value, it can be seen that the Group LASSO analysis has a much smaller MSEP value (MSEP = 18.49%) than compared to LASSO (MSEP = 28.75%) with the same R-squared value ($R^2 = 96\%$).

Simulation 3 is designed to be analogous to simulation 2, but the condition is $p > n$. In LASSO analysis, the probability of non-zero beta error estimated correctly is small (FDR = 6.25%). However, the probability of non-zero beta estimated incorrectly is quite large (53.13%). This indicates that LASSO has limitations in estimating zero beta. Furthermore, it can be seen from the R-Squared value, the Group LASSO analysis has a slightly larger value ($R^2 = 85\%$) than LASSO ($R^2 = 83\%$). Also, similar to simulations 1 and 2, the Group LASSO analysis has a much smaller MSEP value, even half of the LASSO MSEP value. Overall in 3 simulations, analysis Group LASSO has better performance than LASSO.

Table 5. The Goodness of Fit Measures

|                | Simulation 1 | Simulation 2 | Simulation 3 |
|----------------|--------------|--------------|--------------|
| **Group LASSO** |              |              |              |
| FDR            | 0\%          | 37.50\%      | 37.50\%      |
| FNR            | 0\%          | 37.50\%      | 41.18\%      |
| MSEP           | 1.76         | 18.49        | 84.46        |
| R-Squared      | 0.98         | 0.96         | 0.85         |
| **LASSO**      |              |              |              |
| FDR            | 0\%          | 0\%          | 6.25\%       |
| FNR            | 0\%          | 42.86\%      | 53.13\%      |
| MSEP           | 2.26         | 28.75        | 160.17       |
| R-Squared      | 0.98         | 0.96         | 0.83         |

3.2 Empirical Application to HDI Data

The application of the regression model using the LASSO and Group LASSO methods was carried out at the 2019 HDI of Bengkulu Province. This application was adjusted to the conditions in simulation 3, i.e $p > n$. There are 10 the city/district as units of observation ($n = 10$) with 12 explanatory variables ($p = 12$) which are divided into 4 groups. Next, Table 6 shows a summary statistical
description of the variables used. In the Table 6, it can be seen that the HDI scores in Bengkulu Province are in the range of 66.69-80.35.

Table 6. Statistics Description of the variables

|      | HDI   | ECM      | EYS     | GRDP   | LER   |
|------|-------|----------|---------|--------|-------|
| Mean | 69.29 | 10138.00 | 13.36   | 7208.70| 67.38 |
| Median| 67.90 | 9987.50  | 13.00   | 4883.16| 67.79 |
| Standard Deviation | 4.10 | 1627.26  | 1.00    | 5947.66| 1.80  |
| Minimum | 66.69 | 8209.00  | 12.56   | 3346.49| 63.12 |
| Maximum | 80.35 | 14030.00 | 16.01   | 23200.95| 70.04 |

Then the LASSO and Group LASSO analyzes were carried out. From the results of the LASSO analysis, several variables affect the HDI score in Bengkulu Province. They are Expenditure Per Capita Per Month (ECM), Expected Years of Schooling (EYS) and GRDP. This can be seen from the coefficient of variables. The variable that has a non-zero coefficient is the variable that affects the HDI Score in Bengkulu Province. All variables that explain the HDI score in Bengkulu Province have positive coefficients. This shows that the higher value of the explanatory variable reflects higher HDI score in Bengkulu Province.

Table 7. Parameter Estimator in the LASSO and Group LASSO

|      | LASSO | GROUP LASSO |
|------|-------|-------------|
| Intercept | 69.29 | 69.29       |
| ECM | Linear | . | 0.56 |
|      | Quadratic | . | 0.53 |
|      | Cubic | 0.21 | 0.51 |
| EYS | Linear | . | 0.73 |
|      | Quadratic | . | 0.72 |
|      | Cubic | 1.17 | 0.72 |
| GRDP | Linear | 1.66 | . |
|      | Quadratic | . | . |
|      | Cubic | . | . |
| LER | Linear | . | 0.16 |
|      | Quadratic | . | 0.16 |
|      | Cubic | . | 0.16 |

The coefficient value of each explanatory variable in the LASSO and Group LASSO analysis can be seen in Table 7. In line with the results of the LASSO analysis, the Group LASSO results showed that the group of variables that can be indicator the HDI score in Bengkulu Province is all variables in the Expenditure Per Capita Per Month (ECM), Expected Years of Schooling (EYS), and Life Expectancy Rate (LER). These variables are polynomial effects up to order 3 of each group.
Chen et al. [11] conducted a recognized regression analysis through the Group LASSO and LASSO methods in identifying factors associated with tuberculosis cases in West Java. The results showed that the Group LASSO method is better than the LASSO [10]. This is indicated by the MSE value of the Group LASSO which is smaller than the LASSO. Furthermore, Yunus et al. [10] conducted a study on the characteristics of the Group LASSO in handling high correlated data. The simulation results showed the Group LASSO was better than LASSO and LS technique when group of is present included in a group [11].

In this further discussion, we see the significant differences between LASSO and Group LASSO. LASSO model uses far fewer variables which is not used in Group LASSO model. The measure of the goodness of these two models is presented in Table 8. Based on the results of the 10-fold cross validation, the lambda (λ) was 0.178. The model obtained by the Group LASSO analysis had an MSEP value of 0.25. This value was smaller than the MSEP LASSO value (1.40). Meanwhile, the Group LASSO R-squared value was 98%. This means that through Group LASSO analysis the variables used are able to explain HDI diversity very well. This is in line with the predicted vs actual comparison graph in Figure 1. In the figure, it can be seen that the predicted data pattern is almost the same as the actual data.

![Predicted vs Actual](image)

**Figure 1.** Distribution of Actual and Predicted HDI for Districts/Cities in Bengkulu Province

| Method      | MSEP | $R^2$ |
|-------------|------|-------|
| Group LASSO | 0.25 | 0.98  |
| LASSO       | 1.40 | 0.91  |

Table 8. Goodness of fit Measures of Group LASSO and LASSO Analysis

It can be seen from Table 8 that the Group LASSO model is better than LASSO with a lower MSEP value and a higher R-Squared. Referring to this value, it can be said that the Group LASSO is better than LASSO.
Figure 2. All possible lambda (\(\lambda\)) in the Group LASSO analysis

The Group LASSO analysis can also determine which group of variables has the most influence and does not affect the HDI score in Bengkulu Province. This can be seen in the graph presented in Figure 2. The red line in the graph above, shows all possible lambda. The horizontal axis (x) represents the step, that is, the index for each possible lambda, displayed from 0 to 100, where as the step index increases, the lambda value decreases further. The vertical axis is the intercept value in the model, during the x-step [8]. The greater the lambda value, the more explanatory variables are not included in the model. Based on Figure 2, if the lambda value is 1.134 \(< \lambda \leq 2.986\), then only the EYS variable group is included in the model. If the lambda value is 0.430 \(< \lambda \leq 1.134\), then the EYS and ECM variable groups will be included in the model. If the lambda value is 0.179 \(\leq \lambda \leq 0.430\), then the EYS, ECM, and LER variable groups will be entered in the model. Finally, if the lambda value is \(\lambda \geq 0.179\), then all of the grouped variables will be included in the model. Based on the results of the 10-fold cross validation, the best lambda value that minimizes the MSEP average value, and is used is 0.178. This resulted in the GRDP variable group not being included in the Group LASSO model. These results also indicate that the group order of the most important variables on HDI score in Bengkulu Province is Expenditure Per Capita Per Month (ECM), Expected Years of Schooling (EYS), and Life Expectancy Rate (LER).

4. Conclusion

From the simulation results in 3 scenarios (when \(p < n \) and \(p > n\)) showed that the Group LASSO was better than LASSO. This can be seen from the value of MSEP and FNR Group LASSO which were smaller than LASSO and the value of R-Squared Group LASSO which was greater than LASSO. In the application of HDI data, it is in line with the simulation results that the analysis of Group LASSO was better than LASSO with MSEP Group LASSO of 0.25 and R-Squared of 98%. Based on the best lambda value (0.178), the order of the most important to insignificant group of variables, which was an indicator of the HDI score in Bengkulu Province was Expenditure Per Capita Per Month (ECM), Expected Years of Schooling (EYS), and Life Expectancy Rate (LER).

References

[1] Draper N R and Smith H 1998 Applied Regression Analysis New York: John Wiley Sons
[2] Setiawan and Sutikno 2010 Bayes Wavelet Regression Approach to Solve Problems in Multivariable Calibration Modeling JPTek: The Journal for Technology and Science 21 (2) 1-8
[3] Hastie T, Tibshirani R and Wainwright M 2015 Statistical Learning with Sparsity: The Lasso and Generalizations London: CRC Press

[4] Tibshirani R 1996 Regression Shrinkage and Selection Via the Lasso Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) 267-288

[5] Rajaratnam B, Roberts S, Sparks D, Yu H 2019 Influence Diagnostics for High-Dimensional Lasso Regression Journal of Computational and Graphical Statistics 28 (4) 877-890

[6] Yuan M and Lin Y 2006 Model selection and estimation in regression with grouped variables J. R. Stat. Soc. Ser. B 68 (1) 49-67

[7] Kuswardani D A 2020 Bengkulu Province in Figures 2020 Bengkulu: Badan Pusat Statistik (BPS)

[8] Grisanti E, Totska M, Huber S, Calderon C K, Hohmann M, Lingenfelser D and Otto M 2018 Dynamic localized SNV, Peak SNV, and partial peak SNV: Novel standardization methods for preprocessing of spectroscopic data used in predictive modeling Journal of Spectroscopy 2018 5037572

[9] James G, Witten D, Hastie T and Tibshirani R 2013 An Introduction to Statistical Learning with Applications in R Berlin: Springer

[10] Chen S, Notodiputro K A, Rahardiantoro S 2020 Penerapan Analisis Lasso dan Group Lasso dalam Mengidentifikasi Faktor-Faktor yang Berhubungan dengan Tuberkulosis di Jawa Barat Indonesia Indonesian Journal of Statistics and Its Applications 4 (1) 39-54

[11] Yunus M, Saefudin A, Soleh A M 2017 Characteristics of group LASSO in handling high correlated data Applied Mathematical Sciences 11 (20) 953-961

[12] Rahmat H S, Wijayanto H, Afendi F M, Sartono B, Anisa R and Septianingsih D A 2018 Variable-Group Selection on Estimated Metabolites of Curcuma aeruginosa Related to Antioxidant Activity by Using Group Lasso Regression International Conference on Mathematics and Mathematics Education 2018, Advances in Social Science, Education and Humanities Research 285 128-130

[13] Rachmawati R N, Pusponegoro N H, Muslim A, Notodiputro K A, Sartono B 2017 Group LASSO for Rainfall Data Modeling in Indramayu District, West Java, Indonesia Procedia Computer Science 116 190-197

[14] Riau B P R, Afendi F M, Anisa R 2019 Selection of Compound Group to Identify the Authenticity One of Jamu Product Using the Group Lasso for Logistic Regression Journal of Physics: Conference Series 1341 092020

[15] Darlington R B and Hayes A F 2017 Regression Analysis and Linear Models: Concepts, Application and Implementation New York: The Guilford Press

[16] Han J, Kamber M, Pei J 2012 Data Mining: Concepts and Techniques Netherland: Elsevier Inc.

[17] Narkhede S 2018 Understanding Confusion Matrix Towards Data Science 180 (1) 1-12

[18] Trevethan R 2017 Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice Front Public Health 5 (307) 1-7

[19] Widhiarso W 2010 Catatan pada Uji Linieritas Hubungan Fakultas Psikologi Universitas Gadjah Mada (UGM) Yogyakarta: UGM