The lasso binary logistic regression method for selecting variables that affect the recovery of Covid-19 patients in China

A Rusyana\textsuperscript{1,2}, K A Notodiputro\textsuperscript{*} and B Sartono\textsuperscript{2}

\textsuperscript{1} Department of Statistics, Universitas Syiah Kuala, Banda Aceh, Indonesia 23111
\textsuperscript{2} Department of Statistics, IPB University, Bogor, Indonesia 16680

\textsuperscript{*}Corresponding author: khairil@apps.ipb.ac.id

Abstract. The Lasso regression performs the least squares method with the $\ell_1$-constraint. This is a particular type of regularization which adds a penalty equal to the absolute value of the magnitude of coefficients that can result in sparse models with few coefficients in which some coefficients can become zero and be eliminated. Larger penalties result in coefficient values closer to zero, which is ideal for producing simpler models. In this study, the lasso method has been applied to select variables affecting the recovery of the Covid-19 patients. The data consisted of the number of patients treated in several hospitals or clinics in China, recovered patients, demographics, comorbidities, symptoms, and treatment. The performance of the lasso binary logistic regression was compared with the full model of logistic and the stepwise logistic regression. The results showed that the number of independent variables selected by the lasso method was larger than those selected by the stepwise method. It has also been showed that coefficient values of variables produced by the lasso method were smaller than those produced by the stepwise. The independent variables that affect the cure rate of covid-19 patients with the lasso are gender, comorbidities, diabetes, cardiovascular, cough, fatigue, diarrhea, platelet count and antibiotics.

1. Introduction

1.1. Background

Least Absolute Shrinkage and Selection Operator (lasso) regression uses the least squares method with constraint $\ell_1$ [1]. This is done by adding a penalty in the form of adding the absolute value of the regression coefficient which is multiplied by lambda to the Lagrange form to obtain a parameter estimator. The lasso makes coefficients of the regression shrink toward zero [2]. The lasso can select independent variables that explain the response variable. It is like the stepwise regression.

Coronavirus disease 2019 (Covid-19) was a disease that began to be discovered in Wuhan China in December 2019. The number of covid-19 cases, patient deaths and cases of recovery patient in the world are 43,405,696; 1,159,835; and 31,935,211, respectively on October 26, 2020 and their trends rise. However, the trend of them flats in China with cases 85,810; deaths 4,634; and recovery 80,911 [3].

Meta-analysis is a quantitative approach to systematically combine the results of previous research to reach certain conclusions [4]. In this study, 202 papers about Covid-19 in China were selected, then 26 papers were taken to be the observations of the study. The data were analyzed with the lasso binary regression model then the results of the lasso model was compared with the results of the full model and
the stepwise regression model. This was conducted to identify characteristics of the lasso which has smaller coefficient scores than the other methods.

1.2. Research purposes
There are two (2) objectives in this study, namely
(1) build a lasso logistic regression model to see the effect of demographics, disease history, symptoms and treatment on the recovery of Covid-19 patients and compare the model's performance with a full logistic regression model and a stepwise logistic regression model,
(2) find variables that explain the recovery of Covid-19 patients with lasso binary logistic regression.

2. Literature Review
2.1. Binary Logistic Regression
The formula of the binary logistic regression model is given below.

\[ \pi_i = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)} \]

or

\[ \logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij} \]

where:
- \( j = 1,2, \ldots, p \); \( p \) = the number of parameter \( \beta \)
- \( i = 1,2, \ldots, N \); \( N \) = the number of observations
- \( \pi_i \) = probability of the \( i \)-observation
- \( \beta_j \) = the \( j \)-parameter;
- \( x_{ij} \) = value of the \( i \)-observation and the \( j \)-independent variable \[5\].

2.2. Stepwise regression
The stepwise regression is a modification of the forward selection. After the forward process, all candidate variables in the model are checked again. The stepwise regression requires two significance levels: one for adding variables and one for removing variables \[6\].

2.3. The lasso binary logistic regression
The lasso is a technique to shrink the coefficients of predictors \[7\]. The lasso is a regression method which is penalized so the coefficient by shrinking toward zero \[8\]. General form for finding coefficient estimators of linear method is given below.

\[ \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - (X\beta))^2 \]

The lasso estimator is given by the following.

\[ \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \]

or
\[\hat{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^{N} (y_i - (X\beta)_i)^2 + \lambda P(\beta) \right)\]

Where: \(||\beta||_1 = \sum_{j=1}^{p} |\beta_j|; ||.||_2 = the\ usual\ Euclidian\ norm; P(\beta) \leq t [1].

3. Data and Methods

3.1. Data
The data in this study consisted of twenty six (26) observations and nine (9) independent variables. The data were selected from COVID Analytics web [9] which has 202 research results with more than one hundred variables where the research was conducted between December 2019 and April 2020. The observations in this study were groups of hospitals or clinics in certain areas in China.

The data consisted of response variables and nine independent variables. The response variable was formed by the population size of Covid-19 patients (Y1) and patients who recovered (Y2), while the independent variables consisted of demographic information, disease history, symptoms and treatment. Demographic information is male. Disease history includes diabetes and cardiovascular. Symptoms consisted of cough, fatigue, diarrhea, and platelets. The treatment is the percentage of antibiotics. The characteristics of the data can be seen in table 1.

| Variable | Label | Unit | Scale |
|----------|-------|------|-------|
| **Response Variables** | | | |
| Y1 | Covid-19 patient population | People | Numeric |
| Y2 | The number of people who are recovery. | People | Numeric |
| **Demography** | X1 | Male | Percent | Numeric |
| **Disease history** | X2 | Comorbidity | Percent | Numeric |
| | X3 | Diabetes | Percent | Numeric |
| | X4 | Cardiovascular | Percent | Numeric |
| **Symptoms** | X5 | Cough | Percent | Numeric |
| | X6 | Fatigue | Percent | Numeric |
| | X7 | Diarrhea | Percent | Numeric |
| | X8 | The number of platelets | $10^9/L$ | Numeric |
| **Treatments** | X9 | Antibiotics | Percent | Numeric |

3.2. Method

Steps of the research

The data were analyzed using R, SAS and Excel software. The steps for conducting the study are

(1) collecting data, the selected data consist of 26 observations and 9 independent variables,

(2) construct a lasso binary logistic regression model with minimum lambda, minimum the Akaike Information Criterion (AIC) and minimum Bayesian Information Criterion (BIC); the model is

\[\text{logit}(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \sum_{j=1}^{9} \beta_j x_{ij}, \quad i = 1, 2, \ldots, 26\]

(3) comparing the coefficients of result of the lasso regression with those of the full and stepwise logistic regression models. Proc hpgen select in SAS is applied to find results of the lasso binary logistic regression, the full model, and the stepwise model. Results of the lasso consisted of the
lambda, the number of significant independent variables, AIC, BIC and the coefficients. The smaller AIC and BIC is better.

\[
AIC = -2(\log\text{ likelihood}) + 2K \\
BIC = -2(\log\text{ likelihood}) + \log(N) \times K
\]

Where:
- \(K\) = the number of independent variables + intercept
- \(N\) = the number of samples in the training data

(4) identify independent variables that affect the recovery of Covid-19 patients with the lasso regression.

4. Results and Discussion

4.1. Data exploration

Description of response variables and independent variables can be seen from measures of location, measures of variability, the smallest and the biggest value. For patient population, the number of biggest patient in the observation is 1099 people and the smallest is 10 people whereas the mean, median and standard deviation are 202.6, 131.5 and 251.7, respectively. Independent variables can be divided four part, namely demography, disease history, symptom and treatment. Demography is percentage of male with mean 54.6%, median 53.3%, standard deviation 251.7%, minimum value 40.7% and maximum value 73%. Disease history consist of comorbidity, diabetes and cardiovascular. Comorbidity is percentage total of people who have disease in the original data. The mean and median of the variable is 36.2%. Symptoms consist of caugh, fatigue, diarrhea and average of platelets whereas treatment is antibiotics. For description of the variables detailly can be seen in tabel 2.

| Variable | Label | N  | Mean  | Median | StdDev | Min  | Max  |
|----------|-------|----|-------|--------|--------|------|------|
| Y1       | Patient population (people) | 26 | 202.6 | 131.5  | 251.7  | 10.0 | 1099.0 |
| Y2       | Recovery patient (people)  | 26 | 131.0 | 90.5   | 183.2  | 5.0  | 926.0 |
| X1       | Male (%)                     | 26 | 54.6  | 53.3   | 8.0    | 40.7 | 73.0  |
| X2       | Comorbidity (%)              | 26 | 36.2  | 36.2   | 11.6   | 1.0  | 64.3  |
| X3       | Diabetes (%)                 | 26 | 12.8  | 12.8   | 5.1    | 0.0  | 20.8  |
| X4       | Cardiovascular (%)           | 26 | 9.8   | 9.8    | 5.8    | 0.0  | 27.1  |
| X5       | Cough (%)                    | 26 | 65.5  | 65.9   | 14.1   | 10.2 | 81.6  |
| X6       | Fatigue (%)                  | 26 | 43.0  | 43.0   | 14.9   | 3.5  | 75.0  |
| X7       | Diarrhea (%)                 | 26 | 12.4  | 12.4   | 8.5    | 1.0  | 40.0  |
| X8       | Average Platelets            | 26 | 182.1 | 182.1  | 20.8   | 120.0| 227.1 |
| X9       | Antibiotics (%)              | 26 | 79.2  | 79.2   | 17.3   | 21.6 | 100.0 |

The correlation between the percentage of Covid-19 patients who recovered and the independent variables has a different pattern, see figure 1. This point pattern needs to be supported by looking at the results of formal tests using lasso regression in Table 3 so that the pattern becomes more visible. From figure 1 and table 3, Y (covid-19 patients who recovered) and X1 (percentage of male), and X8 (the number of platelets owned by the patients) had positive correlations. As an illustration of this positive correlation, the more male patients in the hospital, the more they recover. The Y and X2, X4, X5, X6, X7 and X9 have negative correlations. For an example of this negative correlation, the more people who suffer from cardiovascular (X4), the fewer Covid-19 patients who recover.
Figure 1. Scatter plot of recovery patient (Y) and independent variables (X).
4.2. The lasso binary logistic regression method

The making of the lasso binary logistic regression model involving the minimum lambda value, minimum AIC and minimum BIC is done through 20 stages, this is shown in table 3. Initially, the lasso model sets an intercept on the model using lambda equal to one, AIC = 1518.6 and BIC = 1519.9. Then step one enters X6, this is done by placing lambda = 0.8 and producing AIC = 1517.3 and BIC = 1519.8. Then the second step puts X8 into the model, this is done by placing lambda = 0.64 and producing AIC = 1509.0 and BIC = 1512.7. The information for stages 3 to 19 can be seen in table 3. Finally, stage 20 uses lambda minimum = 0.0115 with minimum AIC = 1290.1 and minimum BIC = 1301.4, this lambda produces 8 independent variables and one intercept that explains rate of recovery Covid-19 patients.

| Step | Description | Effects In Model | λ  | AIC   | BIC   |
|------|-------------|------------------|----|-------|-------|
| 0    | Initial Model | 1                | 1  | 1518.6| 1519.9|
| 1    | X6 entered  | 2                | 0.8| 1517.3| 1519.8|
| 2    | X8 entered  | 3                | 0.64| 1509.0| 1512.7|
| 3    | X9 entered  | 4                | 0.512| 1475.8| 1480.8|
| 4    | X9 entered  | 4                | 0.4096| 1435.5| 1440.5|
| 5    | X1 entered  | 5                | 0.3277| 1403.8| 1410.0|
| 6    | X7 entered  | 6                | 0.2621| 1376.8| 1384.3|
| 7    | X7 entered  | 6                | 0.2097| 1361.4| 1368.9|
| 8    | X7 entered  | 6                | 0.1678| 1344.8| 1352.4|
| 9    | X4 entered  | 7                | 0.1342| 1333.6| 1342.4|
| 10   | X4 entered  | 7                | 0.1074| 1323.3| 1332.1|

The most accurate coefficient of lasso binary logistic regression model are generated using a minimum lambda of 0.0115. Based on the results in Table 4, the lasso regression model can be written as follows:

\[
\ln \left( \frac{\pi_l(x)}{1 - \pi_l(x)} \right) = 0.607 + 0.023x_1 - 0.005x_2 - 0.021x_4 - 0.006x_5 - 0.014x_6 - 0.010x_7 + 0.012x_8 - 0.026x_9
\]

To see the characteristics of a lasso where the coefficient value shrinks towards zero, the coefficient is compared with the full model and the stepwise model. The full model produces four independent variables that can explain the recovery of Covid-19 patients. The variables are X4, X6, X7 and X9. The independent variables appear to have a p-value of less than 0.05. Likewise, in stepwise regression model, the stepwise has also four independent variables that can explain the cure for Covid-19 patients the same as full regression. Whereas the lasso regression has 8 independent variables that can explain the recovery of covid-19 patients. These variables can be divided into two, namely those that have a positive effect and those that have a negative effect. The variables that have a positive effect are X1 and X8 while the independent variables that have a negative effect are X2, X4, X5, X6, X7 and X9.
Table 4. Result for covid data.

| Variables       | Full Model | Stepwise | Lasso  |
|-----------------|------------|----------|--------|
|                 | DF         | Coef     | SE     | χ²    | P-value | DF | Coef     | SE     | χ²       | P-value | DF | Coef     | SE     |
| Intercept       | 1          | 3.074    | 0.185  | 14.230| 0.000   | 1  | 3.553    | 0.216  | 271.169  | <.0001  | 1  | 0.607    |        |
| X1              | 1          | 0.001    | 0.010  | 0.012 | 0.913   | -  | -        | -      | -       | -       | -  | -        | -      |
| X2              | 1          | -0.004   | 0.005  | 0.769 | 0.381   | -  | -        | -      | -       | -       | -  | -        | -      |
| X3              | 1          | 0.011    | 0.013  | 0.732 | 0.392   | -  | -        | -      | -       | -       | -  | -        | -      |
| X4              | 1          | -0.028   | 0.011  | 6.526 | 0.011   | 1  | -0.201   | 0.007  | 8.083   | 0.005   | 1  | -0.201   |        |
| X5              | 1          | -0.002   | 0.004  | 0.167 | 0.683   | -  | -        | -      | -       | -       | -  | -        | -      |
| X6              | 1          | -0.019   | 0.003  | 37.880| <.0001  | 1  | -0.021   | 0.002  | 96.859  | <.0001  | 1  | -0.006   |        |
| X7              | 1          | -0.014   | 0.006  | 5.053 | 0.025   | 1  | -0.013   | 0.005  | 7.650   | 0.006   | 1  | -0.010   |        |
| X8              | 1          | 0.004    | 0.003  | 1.362 | 0.243   | -  | -        | -      | -       | -       | -  | 0.012    |        |
| X9              | 1          | -0.025   | 0.003  | 69.493| <.0001  | 1  | -0.023   | 0.002  | 97.599  | <.0001  | 1  | -0.026   |        |

Remarks:
X1 : Male (%)  X4 : Cardiovascular (%)  X7 : Diarrhea (%)
X2 : Comorbidity (%)  X5 : Cough (%)  X8 : Platelets (10^9/L)
X3 : Diabetes (%)  X6 : Fatigue (%)  X9 : Antibiotics (%)

5. Conclusions
The conclusions in the study are:
1) minimum lambda in the lasso is 0.0115 with nine selected variables, AIC of 1290.1 and BIC of 1301.4.
2) the lasso binary logistic regression model has more independent variables and smaller coefficient scores than those of the stepwise logistic regression, it is suitable with characteristics of the lasso,
3) the variables that affect recovery patient based on the binary lasso logistic regression are male, comorbidity, diabetes, cardiovascular, cough, fatigue, diarrhea, platelets and antibiotic treatment,
4) male and platelets have a positive effect, while comorbidity, diabetes, cardiovascular, cough, fatigue, diarrhea, platelets, and antibiotics have a negative effect.

References
[1] Hastie T, Tibshirani R and Wainwright M 2015 Statistical Learning with Sparsity: The Lasso and Generalizations (CRC Press, Taylor & Francis Group)
[2] Tibshirani R 1997 The Lasso Method for Variable Selection in the Cox Model Stat Med. 16 4 385–395
[3] Worldometer 2020 Covid-19 Coronavirus Pandemic https://www.worldometers.info/coronavirus/
[4] Purwani S W 2020 Meta Analysis. Lab Health and Environment Statistics (Surabaya: Department of Statistics - FSAD - ITS)
[5] Agresti A 2015 Foundations of Linear and Generalized Linear Models (USA: Wiley)
[6] NCSS Statistical Software 2020 Chapter 311: Stepwise Regression
[7] Muslim A, Hayati M, Sartono B and Notodiputro K A 2018 A Combined modeling of generalized linear mixed model and LASSO techniques for analizing monthly rainfall data in IOP Conference Series: Earth and Environmental Science 187 012044
[8] Rachmawati R N, Hidayat P N, Muslim A, Notodiputro K A and Sartono B 2017 Group LASSO for Rainfall Data Modeling in Indramayu District, West Java, Indonesia Procedia Comput. Sci. 116 p. 190–197
[9] Covid Analytics 2020 Dataset. https://www.covidanalytics.io/dataset.