Proceeding of ICSA 2019, p: 18-28
ISBN 978-979-19256-3-1 (PDF)

Group Lasso for Identifying Tuberculosis Influenced Factors in West Java

Septian Rahardiantoro[1*], Anang Kurnia[1]

[1] *Department of Statistics, IPB University, Bogor, 16680, Indonesia*

[*)]Corresponding author: rahardiantoro_14@apps.ipb.ac.id

Abstract

Tuberculosis (TB) is a disease caused by bacteria that can affect serious infection in the lungs. It can spread from one person to another through tiny droplets released into the air via coughs and sneezes. West Java is one of the locations with the largest TB sufferer in Indonesia. This research used the number of TB cases data in West Java 2017 to identify group of factors that can influence TB infections. The group factors applied include health factor, environment factor, health facility factor, and demography factor. The method that used in this research was group lasso. After the group factors have been identified, the ordinary least square applied to determine significant factors to affect TB infections based on these group factors. The selected group factors of this research was health factor, environment factor, and health facility factor significant to influence TB cases. Furthermore, number of diarrhea sufferers, unemployment, and number of malnutrition sufferers were significant to be the variables that important to affect the TB cases in West Java based on ordinary least square.

***Keywords*:** group factor, group lasso, ordinary least square, tuberculosis

Introduction

Tuberculosis (TB) is a disease that attacks the lungs caused by bacteria. The general features of TB sufferers include chronic cough, sputum production, appetite loss, weight loss, fever, night sweats, and hemoptysis. The spread of this disease can be directly through the air when there is a sufferer who coughs (Zumla *et al*, 2013). Tuberculosis is a disease that often occurs in Indonesia. West Java is one of the provinces with the highest number of cases of tuberculosis. In 2017, number of TB cases in West Java was about 22,693 cases (BPS, 2018).

We are interested in finding important predictor variables in predicting the response variable, where each predictor variables may be represented by a group of factor of derived input variables. The factors studied were a group of variables suspected of influencing TB cases in West Java, namely groups of health, social and environmental factors, health facilities, and demographics. Within these groups of factors, there are several variables that represent them. In order to detect groups of factors that influence the number of TB cases, this study used the group lasso method.

The idea of group lasso is to apply some penalizations on the groups. It makes the model more interpretable and variables selection can be applied continuously. The algorithm found by Yuan and Lin (2007) to be very stable and usually reaches a reasonable convergence tolerance within a few iterations. However, the computational burden increases dramatically as the number of predictors increases (Yuan *et al*, 2007).

The research process, carried out in two schemes: analysis of data on unitization method of data standardization, and unitization with zero minimum method of data standardization (Jajuga *et all*, 2000). Group lasso modeling was applied to both schemes, and was selected with the smallest RMSE value. Furthermore, a significant factor group was chosen to determine the variables that tended to significantly influence the number of TB cases in West Java.

Materials

The data used are secondary data from BPS publications: Jawa Barat Province in Figures 2018 regarding the number of TB cases in 2017 consisting of 27 district or city observations. Predictor variables used were 12 units, which were divided into four groups of factors. Table 1 displays a list of groups of factors along with the unit variables and information in them.

Table 1. List of group factors and predictor variables

No	Group Factor	Predictor Variable Name	Scale
1	Health factor (Group 1)	Malnutrition	Person
2	Health factor (Group 1)	Diarrhea	Person
3	Social & environment factor (Group 2)	Poor & Miserable	Person
4	Social & environment factor (Group 2)	Level of Garbage Transferred	Percentage
5	Social & environment factor (Group 2)	Unemployment	Person
6	Social & environment factor (Group 2)	Prosperous Family	number of family
7	Health facility factor (Group 3)	Medical Personnel	Person
8	Health facility factor (Group 3)	Pharmacy Personnel	Person
9	Health facility factor (Group 3)	Hospital	Unit
10	Health facility factor (Group 3)	Public Health Center	Unit
11	Demography factor (Group 4)	Population Density	Density
12	Demography factor (Group 4)	Total Number of Registered Lands	Area

Methods

Suppose the data consists of an n response vector y, and an n by p matrix of features, X. In addition, in this part also explains about lasso method. For definition of lasso (Tibshirani, 1996), as in the usual regression set-up, assume that the observations are independent or that y_is are conditionally independent given the x_{ij}s; $i = 1, \dots, n; j = 1, \dots, p$, here also assume that the x_{ij} are standardized so that $\sum_i x_{ij}/N = 0, \sum_i x_{ij}^2/N = 1$. Letting $\hat{\beta} = \left(\hat{\beta}_1.\hat{\beta}_2, \dots, \hat{\beta}_p\right)'$, the lasso estimate $\left(\hat{\alpha}, \hat{\beta}\right)$ is defined by $\left(\hat{\alpha}, \hat{\beta}\right) = arg\, min\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\}$; subject to $\sum_j \left|\beta_j\right| \leq t$. Here $t \geq 0$ is a tuning parameter. The parameter $t \geq 0$ controls the amount of

shrinkage that is applied to estimates. Values of $t < t_0$ will cause shrinkage of the solutions towards 0, and some coefficients may be exactly equal to 0. In this research, the development method of lasso will be applied, that is group lasso. The stages of the method carried out in this study are as follows:

1. Exploration of data on the number of TB cases in West Java through mapping many cases for each district or city.
2. Apply the data in the two schemes of standardization process that carried out by the following formula (Jajuga *et all*, 2000).

   Scheme 1: $x_{ij}^{s1} = \frac{x_{ij}}{max(x_{.j}) - min(x_{.j})}$; $min(x_{.j})$: minimum value of j-th variable,

   $max(x_{.j})$: maximum value of j-th variable

   Scheme 2: $x_{ij}^{s2} = \frac{x_{ij} - min(x_{.j})}{max(x_{.j}) - min(x_{.j})}$; $min(x_{.j})$: minimum value of j-th variable,

   $max(x_{.j})$: maximum value of j-th variable
3. Apply the group lasso method for both schemes of data with the algorithm below.
   a. Identify the best λ by using 10-folds cross validation.
   b. Use the best λ above to estimate the group lasso parameter. It is to find

$$min_{\beta} \frac{1}{2} \left\| y - \sum_{l-1}^{m} X^{(l)} \beta^{(l)} \right\|_2^2 + \lambda \sum_{l-1}^{m} \left\| W_l \beta^{(l)} \right\|_2$$

   where $X^{(l)}$ is the submatrix of X with columns corresponding to the predictors in group l, $\beta^{(l)}$ is the coefficient vector of that group, and W_l is some penalty matrix (Simon *et al*, 2011).
   c. Find the prediction of the estimated model, and then calculate the Root Mean Square of Errors (RMSE) as follows.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

4. Variables are selected according to the results of the best group lasso method of the two schemes above then the selection of variables is done with the stepwise method.
5. Interpretation of results.

Results and Discussion

1. Exploration of TB cases in West Java

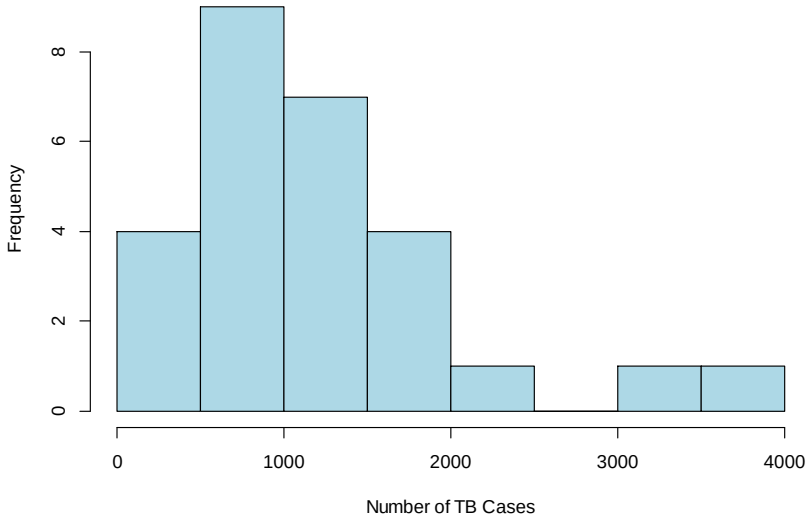

Figure 1. Histogram of TB cases in West Java in 2017

In this part, would be described the response variable of this study, number of TB cases in West Java in 2017. Figure 1 describes the distribution of TB cases in West Java in 2017. The pattern of this distribution seems to right skewed, that is there are several districts contain many cases of TB compared by others. The minimum value of this is 124 cases occurred in Banjar City, maximum value is 3831 occurred in Bogor District, also with the mean value 1179 cases.

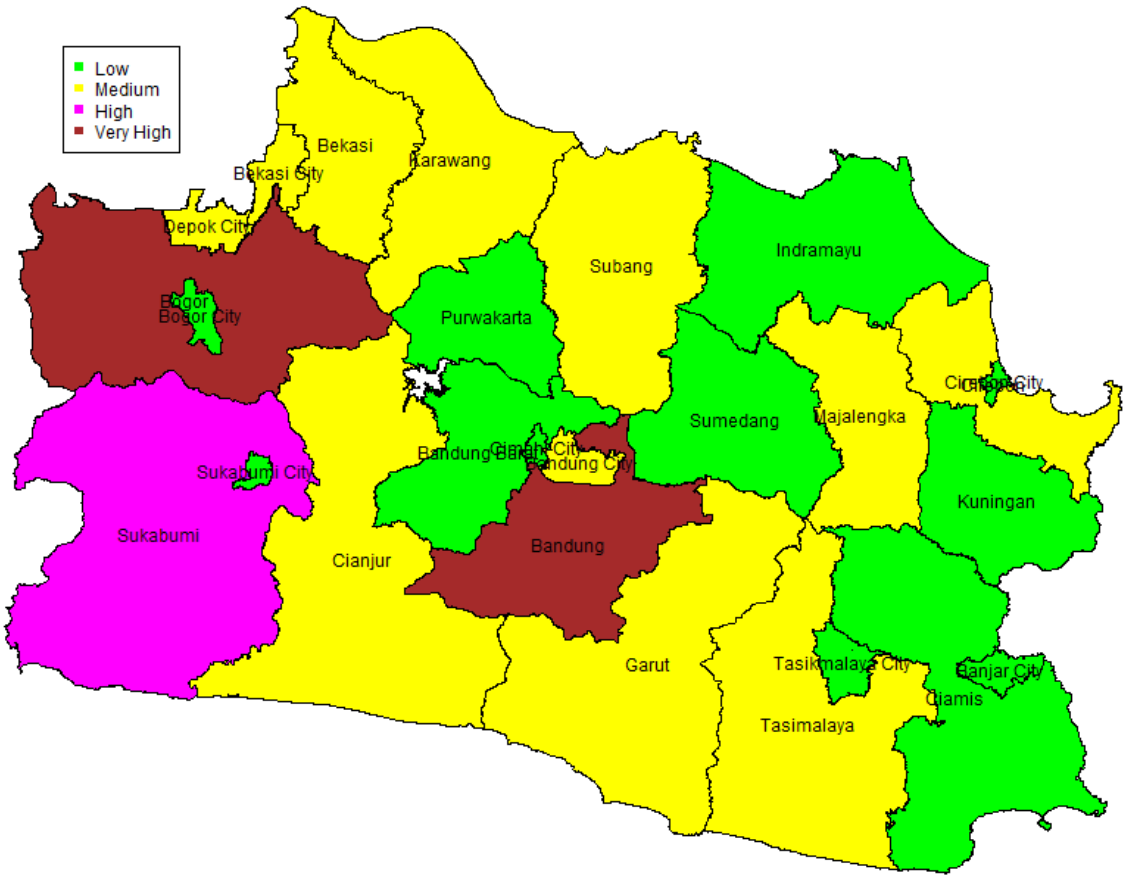

Figure 2. Mapping of distribution of TB cases in West Java in 2017 according four level categories

Furthermore, in this part also presents the mapping of distribution of TB cases. For simplify the description, it was separated to be four level categories; Low: TB cases < 1000, Medium: 1000 ≤ TB cases < 2000, High: 2000 ≤ TB cases < 3000, Very High: TB cases ≥ 3000. Figure 2 shows the mapping of this case. Based on the map, Bogor District and Bandung District detected have very high number of TB cases in West Java. The following district that has high number of TB cases was Sukabumi District. It was to be the opposite with the Bogor City, Bandung City, and Sukabumi City that have lower level of TB cases in West Java; Low level category for Bogor City and Sukabumi City, while medium level category for Bandung City. It seems that higher number of TB cases occurred in the district of the similar city.

Proceeding of ICSA 2019, p: 18-28
ISBN 978-979-19256-3-1 (PDF)

2. Group Lasso Modeling

The modeling processes of this study designed in two schemes. These schemes were chosen because both of them have similar properties of transformed standard deviation and transformed range.

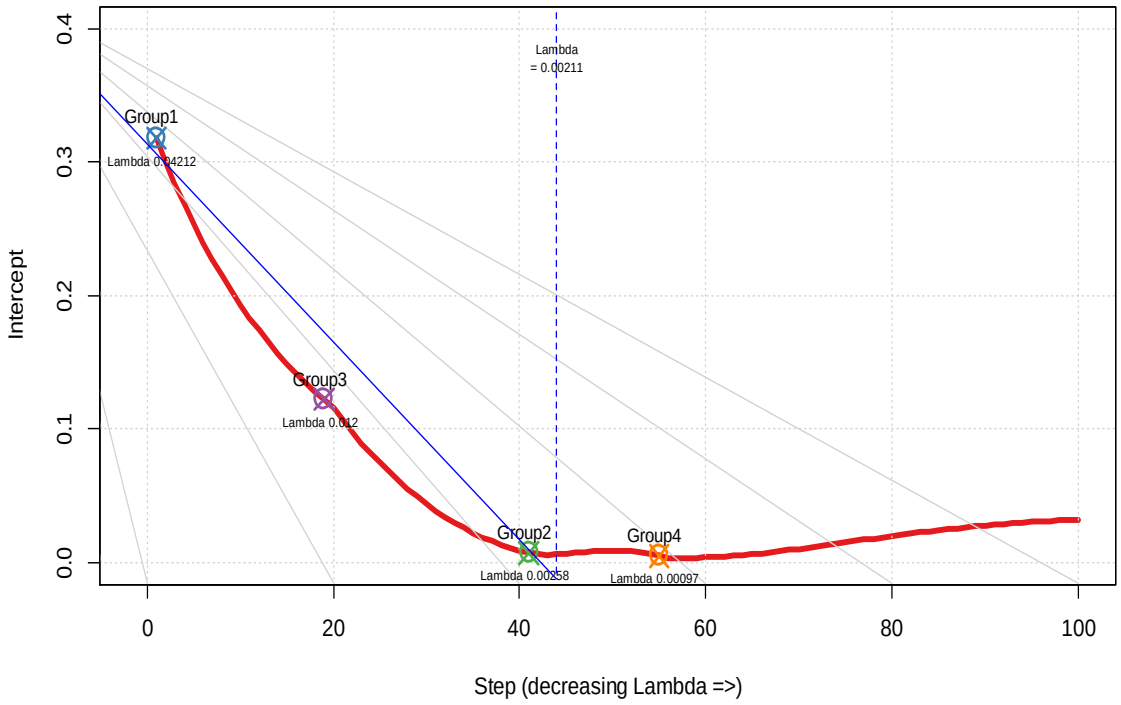

Figure 3. Plot of estimated coefficient of model versus the step of decreasing λ using scheme 1 of data standardization

The first scheme was applied the group lasso in the unitization method of data standardization using the rescaled response variable. Based on 10-folds cross validation, the best λ was 0.002105952. Then, applied the group lasso modeling using that λ. Figure 3 describes the line plot of estimated coefficient of model versus the step of numerical processes by using decreasing λ. Based on the best λ obtained, the significant group of factor was group 1, group 3, group 2 sequentially (health factor, social & environment factor, and health facility factor), that described in the left position of selected λ in

Figure 3. The estimated coefficients displayed in the Table 2. This model had the RMSE value 0.08271039.

Table 2. The significant of estimated coefficients of group lasso in scheme 1 of data standardization

Variable	**Coefficient**
(Intercept)	0.0064
Malnutrition	0.2578
Diarrhea	0.6755
Poor & Miserable	-0.0044
Level of Garbage Transferred	-0.0027
Unemployment	-0.0024
Prosperous Family	0.0054
Medical Personnel	0.0509
Pharmacy Personnel	0.1162
Hospital	-0.1409
Public Health Center	0.0132

The second scheme was applied the group lasso in unitization with zero minimum standardized data. The result of 10-folds cross validation was reached the best λ=0.002105952, exactly same as the result in scheme 1. Then, the line plot of estimated coefficient of model versus the step of numerical processes by using decreasing λ showed in Figure 4. According this result; group 1, group 3, group 2 also sequentially significant in the modeling process. Therefore, the model result contained health factor, social & environment factor, and health facility factor. The estimated coefficients displayed in the Table 3. This model had the lower RMSE value 0.08268515.

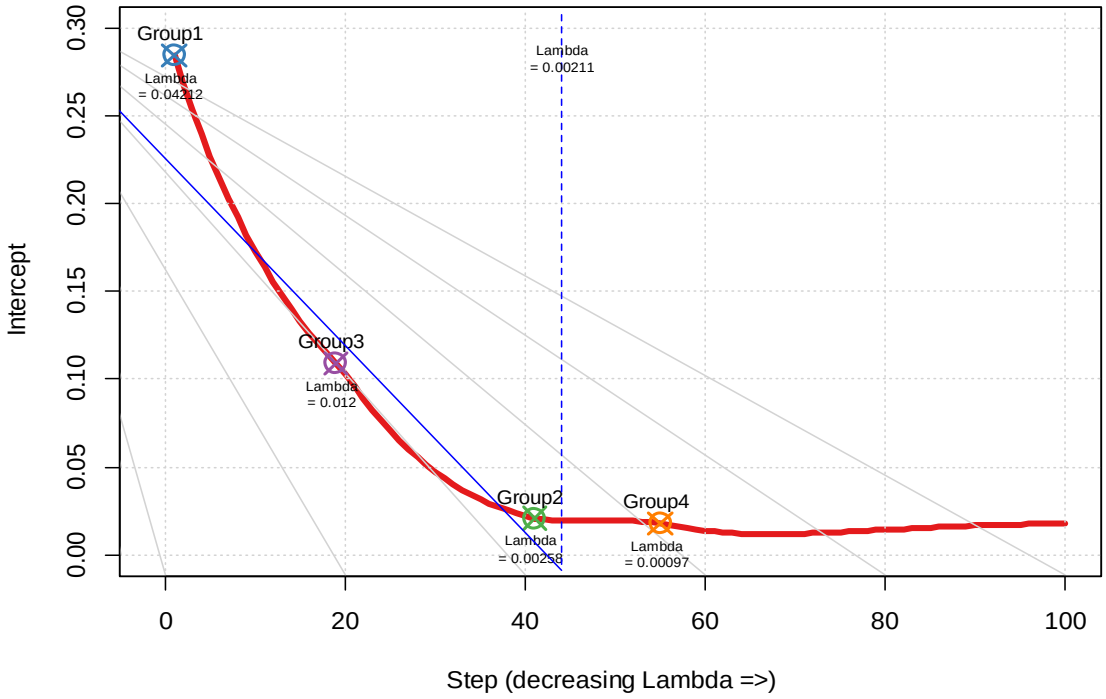

Figure 4. Plot of estimated coefficient of model versus the step of decreasing λ using in scheme 2 of data standardization

Table 3. The significant of estimated coefficients of group lasso in scheme 2 of data standardization

Variable	Coefficient
(Intercept)	0.0197
Malnutrition	0.2589
Diarrhea	0.6723
Poor & Miserable	-0.0062
Level of Garbage Transferred	-0.0037
Unemployment	-0.0035
Prosperous Family	0.0077
Medical Personnel	0.0511
Pharmacy Personnel	0.1162
Hospital	-0.1396
Public Health Center	0.0133

3. Interpretation of The Result

Based on both schemes above, the model that has better prediction was group lasso model in scheme 2 of data standardization because it has lower RMSE value. In addition, the study continued to reveal significant variables according three significant

groups in the second scheme model. The stepwise process was taken to handle this purpose. Figure 5 describes the each single step in the stepwise process.

Based on the result below, the best step in the stepwise process was step 3, with S = 256, R^2(adj) = 91.25, and Mallows Cp = 1.9, because in the step 3 has the closest Mallows Cp values with number of predictors, 3. Therefore, the significant predictor variables that affected the number of TB cases in West Java in 2017 was number of diarrhea, number of unemployment, and number of malnutrition.

Step	1	2	3	4
Constant	103.28	45.19	41.78	82.42
Diarrhea	0.0283	0.0548	0.0507	0.0513
T-Value	8.62	7.65	9.55	10.44
P-Value	0.000	0.000	0.000	0.000
Unemployment		-0.0140	-0.0155	-0.0161
T-Value		-3.98	-6.01	-6.70
P-Value		0.001	0.000	0.000
Malnutrition			0.395	0.476
T-Value			4.70	5.53
P-Value			0.000	0.000
Poor & Miserable				-0.00038
T-Value				-2.20
P-Value				0.039
S	443	351	256	237
R-Sq	74.82	84.83	92.26	93.66
R-Sq(adj)	73.82	83.57	91.25	92.50
Mallows Cp	44.9	19.9	1.9	0.1

Figure 5. Result of stepwise process using significant variable result of scheme 2

Conclusion

The group lasso applied in the unitization with zero minimum standardized data has better prediction value that indicated by the lowest RMSE value. Based on this result, health factor, environment factor, and health facility factor were significant to influence TB cases in West Java in 2017. Furthermore, by using stepwise process, the significant variable that affected TB cases was number of diarrhea, number of unemployment, and number of malnutrition.

References

BPS-Statistics of West Java Province. 2018. *Jawa Barat Province in Figure*. Adiatama, CV

Jajuga K, Walesiak M. 2000. *Standardisation of Data Set under Different Measurement Scales*. In: Decker R., Gaul W. (eds) Classification and Information Processing at the Turn of the Millennium. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg, pp 105-112

Simon N, Tibhsirani R. 2011. *Standardization and the Group Lasso Penalty*. [downloaded in http://faculty.washington.edu/nrsimon/standGL.pdf 14/07/19]

Tibshirani, R. 1996. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistics Society Series B*, 58, 267-288

Yuan M, Lin Y. 2007. Model selection and estimation in regression with grouped variables. *Journal Royal Statistical Society*. 68:1. pp 49-67.

Zumla A, Raviglione M, Hafner R, Reyn CFv. 2013. Current Concepts Tuberculosis. *The New England Journal of Medicine*. 368:8. pp 745-755.