Modeling the Proportion of Tuberculosis Cases in South Sulawesi using Sparse Least Trimmed Squares

by

EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol3.iss2.art6

Research Article

Modeling the Proportion of Tuberculosis Cases in South Sulawesi using Sparse Least Trimmed Squares

Trigarcia Maleachi Randa[1] , Georgina Maria Tinungki[1*], Nurtiti Sunusi[1]

[1] Department of Statistics, Faculty of Mathematics and Natural Science, Hasanuddin University, Makassar, Indonesia.

[*] *Corresponding author: georgina@unhas.ac.id*

Received: 8 July 2022; Accepted: 1 September 2022; Published: 7 September 2022

Abstract: The deadliest infectious disease in Indonesia is tuberculosis (TB), and South Sulawesi is one of the provinces that contributed the most tuberculosis cases in Indonesia in 2018 with 84 cases per 100,000 population. This study aims to identify variables that could explain the proportion of TB cases in South Sulawesi. The data used has many explanatory variables, and there are outliers. Sparse Least Trimmed Squares (LTS) analysis can be used to handle data that has many explanatory variables and outliers. The resulting sparse LTS model successfully selects and shrinks the variables to 14 variables only. In addition, based on the value of R^2 and RMSE for the model evaluation, the sparse LTS shows satisfying results rather than classical LASSO. The government can focus on these factors if they want to reduce the proportion of TB cases in South Sulawesi.

Keywords: LASSO, Outliers, Penalized regression, Tuberculosis, Robust regression

Introduction

Tuberculosis (TB) is still a public health problem both in Indonesia and internationally so that it has become one of the sustainable development goals (SDGs) in the health sector. Indonesia is ranked third with the highest TB sufferers in the world [1]. In Indonesia, one of the provinces with the most TB cases is in the South Sulawesi [2]. It is known that the proportion of TB cases in South Sulawesi Province in 2018 according to a diagnosis by health workers is 0.36% [3]

In 2020 the number of TB cases in South Sulawesi Province has decreased [4]. However, the decrease in the number of TB cases still has to be watched out for due to the disparity in the spread of TB disease between regions in South Sulawesi, for example, Barru and Sidrap districts in 2018 had the number of TB cases in 182 and 493 cases respectively, but in 2020 there was an increase the number of cases was 202 and 591 cases, respectively [5]. This is presumably because the data contains outliers. Therefore, it is necessary to conduct further research on the factors that influence the proportion of TB cases in South Sulawesi Province.

Regression analysis is one method that can be used to determine the factors that influence the proportion of TB cases. One way to get the regression coefficient is through Ordinary Least Squares (OLS). OLS cannot be applied to high-dimensional data because the design matrix becomes rank deficient. LASSO regression can be used to solve the problem of high dimensional data because it can perform the variable selection by shrinking the regression coefficient to zero [6].

However, LASSO can be adversely affected by the existence of outliers. In situations where there are outliers in the data, a robust approach is recommended. Therefore, a method that can handle these two problems simultaneously, in order to obtain a simpler model and one of them is called robust LASSO regression. Therefore, a method that combines LASSO regression and robust regression is needed. Based on some literature the recommended method is sparse LTS [7].

The explanatory variables used in this study are variables related to health, economy, human resources and the environment. These factors were chosen to be studied because according to research conducted by Sejati & Sofiana [8], overall research on TB involves these factors. The total data collected has 24 observations, namely districts/cities in South Sulawesi, with 25 explanatory variables related to health, economy, human resources and the environment. The response variable used is the proportion of tuberculosis cases in South Sulawesi in 2018 for each district/city. Based on the size ratio between the observations and the number of explanatory variables used, it can be said that this data is high dimensional data, because the number of explanatory variables is more than the number of observations.

Several studies have been conducted using the LASSO method, other studies have been carried out by [9] used to overcome the problem of multicollinearity and determine the inflation regression model of a number of major cities in Indonesia. The results obtained were compared with the OLS method which results in a regression model for inflation in a number of major cities in Indonesia which is better than the OLS method in this study. Another study was also conducted by [10] regarding a general family of trimmed estimators for robust high-dimensional data analysis. In this study, sparse LTS and several other methods were applied to the analysis of yeast genotype data and compared with the trimmed mean square error (T-MSE) computed from 10-folds cross validation for each method, the results obtained that Sparse LTS exhibits the smallest T -MSE.

This study aims to overcome the problem of high-dimensional data and outliers with the sparse LTS approach in a case study of regression analysis of the proportion of TB cases in South Sulawesi Province in 2018. It is hoped that the results of this study can be a reference material for the government to pay attention to the factors that significantly influence the proportion of TB cases.

Sparse Least Trimmed Squares

Consider linear regression model that assumes a linear relationship between the explanatory variable $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response variable $\mathbf{y} \in \mathbb{R}^{n \times p}$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where the regression coefficients are $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ and $\boldsymbol{\varepsilon}$ is the error term that have zero expected value. For simplicity, let's assume that the matrix $\mathbf{X}$ is mean-centered and scaled to variance one, and vector $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is centered to mean zero. The method usually chosen in situations where p denotes the number of explanatory variables is smaller than n denotes the number of observations in the data set is ordinary least squares (OLS). However, in presence of multicollinearity among explanatory variables, OLS become estimator that have poor prediction performance, and if p exceeds n it can no longer be applied. Several alternatives have been proposed in this case; focus of this paper is on the penalty regression approach that shrinks the coefficients and reduces the variance by adding a penalty term on the parameter coefficients to the objective function. The LASSO is a penalized least squares technique which puts l_1 penalty on the estimated regression coefficients [6], while the ridge estimator takes an l_2 penalty instead [11]. Although using the l_1 penalty on LASSO has the advantage of shrinking some approximations to exactly zero and performing variable selection, the existence of an absolute value function means that LASSO is non-differentiable and therefore has no closed form solution. This is appropriate in particular for high dimensional low sample size (HDLSS) datasets that is, $n \ll p$, arising from applications in biometrics, econometrics, social sciences and many other areas, where the data include many non-informative variables that have no effect on the response variable or have very small contribution to the model.

The limitation of the LASSO regression is the lack of resistance to outlier data. The presence of outliers in a data set which is quite common in practice makes statistical analysis difficult, and thus robust alternative methods are often used, see, for example [12, 13]. In the linear regression setting, outliers may appear in the space of the the explanatory variables (so-called leverage points), or in the space of the response variable (vertical outliers) [14]. The Least Trimmed Squares (LTS) estimator has

E-ISSN: 2720-9326
P-ISSN: 2716-0459

been among the first proposals of a regression estimator being completely robust to both kinds of outliers [15]. Then it is defined as follows:

$$\widehat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} r_{(i)}^2(\boldsymbol{\beta}) \tag{2}$$

where the $r_{(i)}^2$ are the ordered statistics of the squared residuals $r_{(1)}^2 \leq r_{(2)}^2 \leq \cdots \leq r_{(n)}^2$, with $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$. The subset size h is chosen between $\lfloor (n+p+1)/2 \rfloor$ and n, where the largest integer $\leq a$ denote as $\lfloor a \rfloor$, and the robustness properties of the estimator determines by it [16]. Using a quick algorithm for its computation, namely FAST-LTS algorithm, thus making the LTS estimator become popular [17]. C-step or "concentration step" is the main feature of this algorithm, which is an efficient way to arrive at outlier-free data subsets where the OLS estimator can be applied. This only works for $n > p$, but recently the sparse and requlatized version of the LTS regression estimator has been proposed for high dimensional problems [7]:

$$\widehat{\boldsymbol{\beta}}_{sparseLTS} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{h} r_{(i)}^2(\boldsymbol{\beta}) + h\lambda\|\boldsymbol{\beta}\|_1 \right\} \tag{3}$$

This estimator adds an l_1 penalty with parameter λ to the objective function of the LTS estimator, and it can thus be seen as a robust counterpart of the lasso estimator. The sparse LTS estimator is robust against both leverage points and vertical outliers, and also a fast algorithm for its computation has been developed.

Key Performance Indicator

The Key Performance Indicator (KPI) used is based on the value of coefficient of determination/R^2, and Root Mean Square Error/RMSE [18]. This KPI used as comparison criteria to choose the best models, which are presented sequentially in Equation 4-5.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{5}$$

Where n number of observations, and. $\hat{Y}_i$ is the actual value of the response variable at location i, while Y_i is the predicted value of the response variable at location i and $\bar{Y}_i$ is the mean value of the response variable at location i

The model is better if it has a more significant coefficient of determination or R^2 because it means that the existing factor variables can explain the model more. Likewise, the smaller RMSE indicators show that the model is getting better.

Materials and Methods

Data

Data analysis was performed on proportion of TB cases in South Sulawesi in 2018. The dataset consist of a dependent variables and 25 independent variables with 24 observations. The data structure used includes data on proportion of tuberculosis, data related to health, data related to economy, data related to human resources and data related to environment. The dataset was obtained from the

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Indonesia Basic Health Research (Riskesdas) and Badan Pusat Statistik (BPS) with the following electronic address: https://pusdatin.kemkes.go.id/ and https://sulsel.bps.go.id/. Further explanation is shown in Table 1 as follows.

Table 1. Description of the Data

Variable	Explanation	Denomination
Y	Proportion of TB Cases	Percent
X_1	Proportion of Disability in Population Aged 18-59 Years	Percent
X_2	Number of Hospitals	Unit
X_3	Number of Clinics	Unit
X_4	Number of Public Health Centers	Unit
X_5	Number of Integrated Healthcare Center	Unit
X_6	Number of Doctors	People
X_7	Number of Nurses	People
X_8	Number of Pharmacists	People
X_9	Number of Nutritionists	People
X_{10}	Number of Undernourished Children	People
X_{11}	Number of Cases of HIV/AIDS	Case
X_{12}	Gross Regional Domestic Product (GRDP)	Billion Rupiah
X_{13}	Growth Rate GRDP	Index
X_{14}	Percentage of Poor People	Percent
X_{15}	Expected Years School	Year
X_{16}	Number of Employees	People
X_{17}	Number of Open Unemployment	People
X_{18}	Number of Labor Force	People
X_{19}	Human Development Index (HDI)	Indeks
X_{20}	Proportion of Smoking in the Population 10 Years and Over Who Smoke Every Day	Percent
X_{21}	Number of Districts	Unit
X_{22}	Number of Villages/Sub-districts	Unit
X_{23}	Proportion of Poor Household Waste Management	Percent
X_{24}	Population Density	People /Km^2
X_{25}	Percentage to Area	Percent

Methods

This study was started by implementing LASSO and then compared with the sparse LTS. The first step is data collection on the variable proportion of TB cases and 25 explanatory variable in South Sulawesi Province. R software is used for inferential analysis because of its convenience, power, and other resources [19]. The second stage is outlier detection in the data set using boxplot. Boxplot is a simple statistical method for exploratory data, which is a graphical display where the outliers appear tagged. The third step is LASSO regression analysis to see the estimation of the model parameters obtained from the data on the proportion of tuberculosis cases. The value of λ for LASSO regression will be chosen by 5-fold cross-validation. The fourth step is to find the estimated parameters of the sparse LTS model and determine the best model by selecting the lambda value for sparse LTS using 5-fold cross-validation. The last step is to calculate the R^2 and RMSE values from the LASSO regression and the sparse LTS estimator to compare the best model for analyzing case studies and concluding. The steps in this research can be described in the flow chart as follows:

E-ISSN: 2720-9326
P-ISSN: 2716-0459

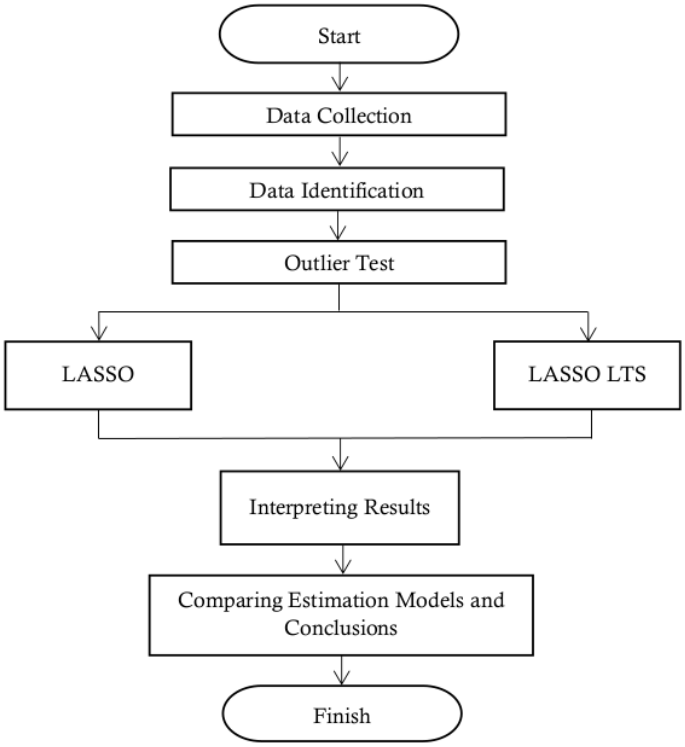

Figure 1. Flowchart of research methods

Results and Discussions

Descriptive Statistics

Table 2 reports descriptive statistics and Pearson correlation matrix for all variables used in this study. Where the highlighted values were known as real correlation values with *p*-values smaller than a significance level of 0.05. It can be seen that *p*-values less than 0.05 were found in the relationship between many explanatory variables. This means that the correlation that occurs between these variables is real. Thus, there is a multicollinearity problem in the data set used.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Table 2. Descriptive statistics and Pearson correlation

	Mean	SD	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
X_1	29.5	11.0	1.0																							
X_2	3.3	5.3	0.4	1.0																						
X_3	12.1	28.4	0.5	0.9	1.0																					
X_4	19.1	8.6	0.4	0.6	0.7	1.0																				
X_5	408.3	224.9	0.5	0.5	0.6	0.9	1.0																			
X_6	171.9	356.2	0.4	1.0	1.0	0.7	0.6	1.0																		
X_7	682.1	872.3	0.5	0.9	1.0	0.7	0.6	1.0	1.0																	
X_8	89.0	106.3	0.4	1.0	1.0	0.6	0.6	1.0	1.0	1.0																
X_9	45.6	44.2	0.5	0.9	1.0	0.7	0.7	1.0	1.0	0.9	1.0															
X_{10}	1446.7	751.5	0.3	0.7	0.7	0.6	0.6	0.7	0.7	0.7	0.7	1.0														
X_{11}	72.8	232.5	0.4	1.0	1.0	0.7	0.6	1.0	1.0	1.0	0.9	0.8	1.0													
X_{12}	12948.2	21705.1	0.4	0.9	1.0	0.8	0.7	1.0	1.0	1.0	0.9	0.8	1.0	1.0												
X_{13}	6.5	2.0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	1.0											
X_{14}	9.8	3.2	0.0	-0.4	-0.4	0.0	-0.1	-0.4	-0.4	-0.5	-0.4	0.1	-0.4	-0.4	0.2	1.0										
X_{15}	13.1	0.9	0.2	0.7	0.6	0.2	0.1	0.6	0.6	0.6	0.5	0.3	0.6	0.5	0.0	-0.5	1.0									
X_{16}	157288.5	114625.8	0.4	0.8	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.7	0.8	0.9	0.2	-0.3	0.4	1.0								
X_{17}	8879.4	15928.8	0.5	1.0	1.0	0.7	0.7	1.0	0.9	1.0	0.9	0.8	1.0	1.0	0.2	-0.4	0.6	0.9	1.0							
X_{18}	166167.9	128858.9	0.4	0.8	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.7	0.8	0.9	0.2	-0.3	0.4	1.0	0.9	1.0						
X_{19}	69.6	4.2	0.1	0.7	0.6	0.1	0.0	0.6	0.6	0.7	0.5	0.4	0.6	0.6	-0.1	-0.6	0.9	0.3	0.6	0.4	1.0					
X_{20}	22.1	2.6	0.2	-0.2	-0.1	-0.1	0.0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0.0	-0.5	0.0	-0.1	0.0	-0.3	1.0				
X_{21}	12.8	5.3	0.3	0.1	0.2	0.7	0.6	0.1	0.1	0.1	0.2	0.2	0.1	0.2	0.2	0.3	0.0	0.4	0.2	0.4	-0.2	-0.1	1.0			
X_{22}	127.0	71.0	0.3	0.1	0.1	0.7	0.7	0.1	0.2	0.1	0.2	0.2	0.1	0.2	0.1	0.2	-0.1	0.5	0.1	0.4	-0.3	0.1	0.9	1.0		
X_{23}	70.6	21.9	-0.1	-0.7	-0.6	-0.1	0.0	-0.6	-0.6	-0.6	-0.5	-0.3	-0.6	-0.5	-0.1	0.6	-0.8	-0.3	-0.6	-0.3	-0.9	0.3	0.3	0.3	1.0	
X_{24}	645.7	1715.4	0.4	1.0	1.0	0.6	0.5	1.0	1.0	1.0	0.9	0.8	1.0	1.0	0.2	-0.4	0.6	0.8	1.0	0.8	0.7	-0.1	0.0	0.0	-0.7	1.0
X_{25}	4.2	4.2	-0.1	-0.2	-0.2	0.1	0.1	-0.2	-0.2	-0.2	-0.2	-0.1	-0.2	-0.1	-0.1	0.2	-0.3	0.0	-0.2	0.0	-0.1	0.3	0.3	0.5	0.4	-0.3	1.0

Outliers Detection

Boxplot analysis were performed to detect the outliers in the data. The boxplot in Figure 2 shows there is an outliers on the response variable, namely observation 9. The existance of these outliers contributes to the non-normalities of data.

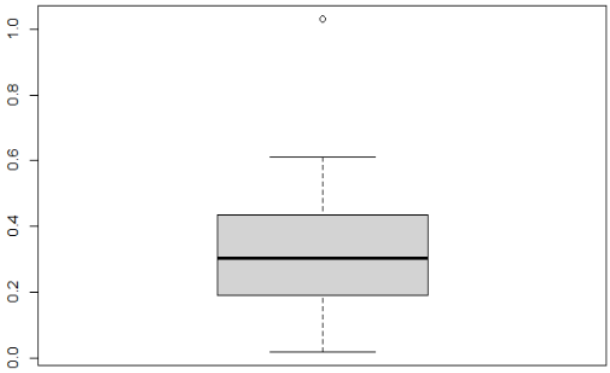

Figure 2. Boxplot for the response variabel (proportion of tuberculosis cases)

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Analysis of LASSO

LASSO regression shrinks the regression coefficient to zero and at the same time can be used for the selection of explanatory variables so that only important explanatory variables are included in the regression model. After the LASSO analysis, the results of several variables that affect the proportion of tuberculosis cases in South Sulawesi Province were obtained. This can be seen from the coefficient of the variable. The variable that has a non-zero coefficient is the variable that affects the proportion of tuberculosis cases in South Sulawesi Province. The coefficient values of each explanatory variable from LASSO analysis can be seen in Table 3.

Table 3. Regression coefficient of LASSO estimator result

Variable	Coefficient	Variable	Coefficient
Intercept	1.74698100	X_{13}	-0.00996912
X_1	-0.00024446	X_{15}	-0.09705172
X_4	0.00147205	X_{21}	-0.00203765
X_9	0.00019864	X_{23}	-0.00170448
X_{10}	0.00002454	X_{25}	-0.00070641
X_{12}	0.00000094		

Figure 3 shows the plots depict the mean squared error (MSE) against log(λ). 10-fold cross-validation as conducted to select λ with minimum mean squared error. The gray bars at each point show MSE plus and minus one standard error. One of the vertical dashed lines shows the location of the minimum of MSE which is equal to 0.02 for value of λ value was 0.02 or in log(λ) was -3.88. The second dashed line shows the point selected by the "one-standard-error" rule with value of λ was 0.04 or in log(λ) was -3.28

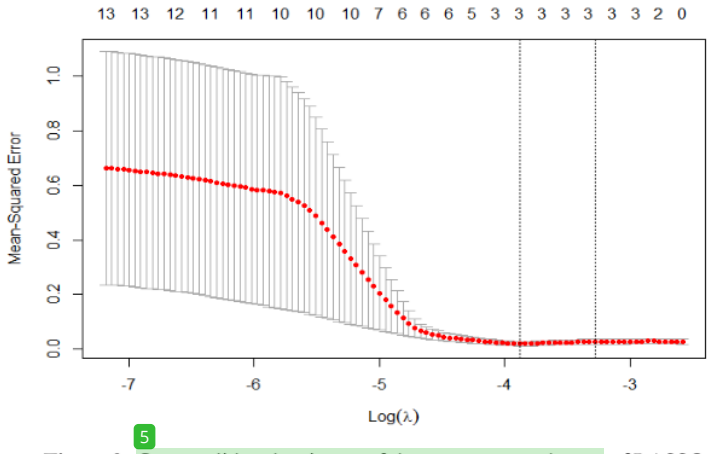

Figure 3. Cross-validated estimate of the mean squared error of LASSO

It can be seen from Figure 4 that the LASSO regression tends to "shrink" the regression coefficients to zero as l_1-norm decreases. l_1-norm is the total absolute of non-zero coefficients. The upper part of the plot shows the number of non-zero regression coefficients for a given value of l_1-norm.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

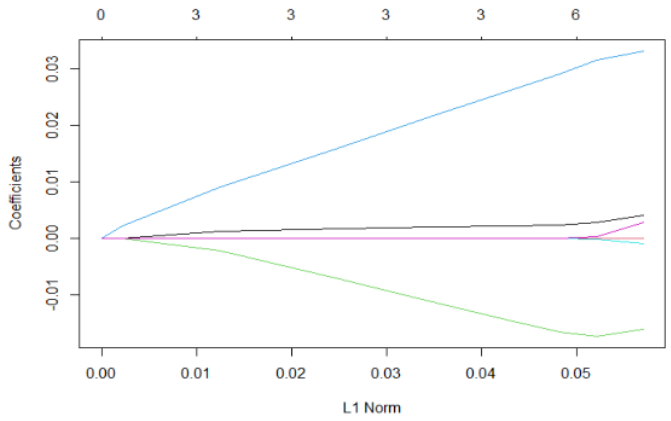

Figure 4. Regression coefficients of LASSO plot

Analysis of Sparse Least Trimmed Squares

To solve the problem of high dimensional data and outliers simultaneously, the LASSO regression is combined with one of the most popular robust regression estimators, namely the least trimmed squared (LTS) estimator to form a robust LASSO estimator, namely sparse LTS estimator. The coefficient values of each explanatory variable from sparse LTS analysis can be seen in Table 4. There are some of the linear regression coefficients shrink to zero. The explanatory variable that has a value of zero is a variable that has no significant effect on the response variable. From the sparse LTS model that has been obtained, it is known that there are 14 variables that can explain the proportion of tuberculosis cases in South Sulawesi Province.

Table 4. Regression coefficient of sparse LTS estimator result

Variable	Coefficient	Variable	Coefficient
Intercept	1.30315500	X_{10}	0.00001609
X_1	0.00061216	X_{11}	0.00340861
X_2	-0.10800260	X_{17}	0.00000822
X_3	-0.00311908	X_{19}	-0.00864150
X_4	-0.00464726	X_{21}	-0.01444924
X_5	0.00007002	X_{23}	-0.00065964
X_8	0.00209661	X_{24}	-0.00015877
X_9	-0.00253790		

Figure 5 shows selection of λ value in the sparse LTS model via 5-fold cross-validation. 10 values of λ were evaluated to select the optimal one that gives the minimum prediction error. The value 0.00006 was chosen for λ by 5-fold cross-validation with the minimum criteria. In this case the minimum value of prediction error was 0.14.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

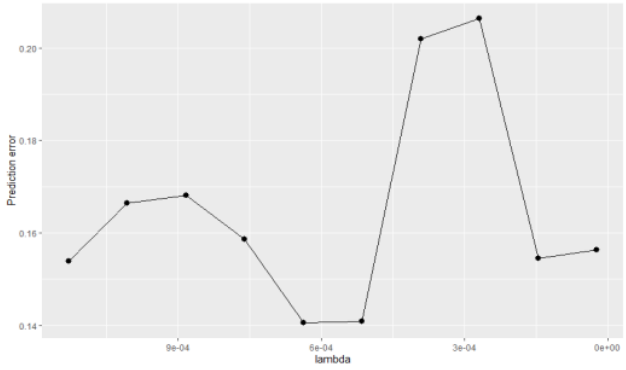

Figure 5. Cross-validated estimate of the prediction error of sparse LTS

Figure 6 shows the standardized residuals versus fitted values plot of the computed sparse LTS model. As observed from that plot, observations 4, 9, 10, 11, and 22 were identified as potential outliers.

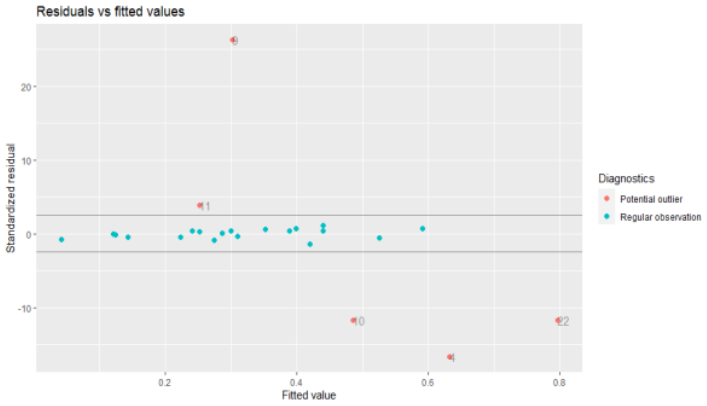

Figure 6. Plot of sparse LTS standardized residuals vs fitted values

Evaluation model goodnes of fit

In selecting the best model, KPI values are used in the form of R^2 and RMSE values from LASSO and sparse LTS models. KPIs from the LASSO model and sparse LTS model are presented in Table 5.

Table 5. Regression coefficient of sparse LTS estimator result

KPI	LASSO	Sparse LTS
R^2	-51.4731%	59.1952%
RMSE	0.3186	0.2014

Based on Table 5, it is found that the sparse LTS model has a coefficient of determination or R^2 greater than the LASSO model, which is 59.1952%. Thus, from looking at the value of the coefficient of determination or R^2, the sparse LTS model is better used than the LASSO model. Furthermore, the model with smaller RMSE values is the sparse LTS model with RMSE value of 0.2014.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Based on this, the sparse LTS model is the best model used in modeling the propotion of tuberculosis cases in South Sulawesi Province in 2018.

Conclusion

This study has resulted in modeling the proportion of TB cases in South Sulawesi Province in 2018 using sparse LTS. Sparse LTS is intended to form a more robust and simpler model, so that it can provide predictions efficiently by involving a minimum of explanatory. Sparse LTS model succeeded in selecting 14 out of 25 variables so that the model only needed fewer explanatory variables but was still able to explain the model better than the classical LASSO estimator according to R^2 and RMSE values. While there are 11 variables namely number of doctors (X_6), number of nurses (X_7), GRDP (X_{12}), growth rate GRDP (X_{13}), percentage of poor people (X_{14}), expected years school (X_{15}), number of employees (X_{16}), number of labor force (X_{18}), proportion of smoking in the population 10 years and over who smoke every day (X_{20}), number of villages (X_{22}), and percentage to area (X_{25}), these variables have a high Pearson correlation value with other variables making this variables unselected into the model. Based on these variables can be a reference for the government in reducing the proportion of TB cases in South Sulawesi Province.

References

[1] World Health Organization, Global tuberculosis report 2018, 2018.
[2] Badan Pusat Statistik Sulawesi Selatan, Provinsi Sulawesi Selatan Dalam Angka 2018, 2019.
[3] Badan Penelitian dan Pengembangan Kesehatan, Laporan Provinsi Sulawesi Selatan Riskesdas 2018, 2019.
[4] Kementerian Kesehatan Republik Indonesia, Profil Kesehatan Indonesia 2020, 2021.
[5] Badan Pusat Statistik Sulawesi Selatan, Provinsi Sulawesi Selatan Dalam Angka 2020, 2021.
[6] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series B., 58(1) (1996) 267-288.
[7] A. Alfons, C. Croux, S. Gelper, Sparse least trimmed squares of analyzing high-dimensional large data sets, Ann. Appl. Stat., 7(1) (2013) 226-248.
[8] A. Sejati, L. Sofiana, Faktor-faktor terjadinya tuberkulosis, KEMAS: Jurnal Kesehatan Masyarakat, 10(2) (2015) 122-128.
[9] Y. A. Mait, D. T. Salaki, H. A. Komalig, Kajian model prediksi metode least absolute shrinkage and selection operator (lasso) pada data mengandung multikolinearitas, d'CARTESIAN: Jurnal Matematika dan Aplikasi, 10(2) (2021) 69-75.
[10] E. Yang, A. C. Lozano, A. Aravkin, A general family of trimmed estimators for robust high-dimensional data analysis, Electron. J. Stat., 12(2) (2018) 3519-3553.
[11] A. Hoerl and R. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12(1) (1970) 55-67.
[12] Y. Z. Liang, O. M. Kvalheim, Robust methods for multivariate analysis-a tutorial review, Chemom. Intell. Lab. Syst., 32(1) (1996) 1-10.
[13] Y. Z. Liang, K T. Fang, Robust multivariate calibration algorithm based on least median of squares and sequential number theory optimization method, Analyst, 121(8) (1996) 1025-1029.
[14] R. Maronna, R. Martin, V. Yohai, Robust Statistics: Theory and Methods, John Wiley & Sons, New York, 2006.
[15] P. Rousseeuw and A. Leroy, Robust Regression and Outlier Detection, 2nd edition, John Wiley & Sons, New York, 2003.
[16] P. J. Rousseeuw, Least median of squares regression, J. Am. Stat. Assoc., 79(388) (1984) 871–880.
[17] P. J. Rousseeuw and K. Van Driessen, Computing LTS regression for large data sets, Data Min. Knowl. Discov., 12(1) (2006) 29–45.
[18] C. Sammut and G. I. Webb, Eds., Mean Squared Error, in Encyclopedia of Machine Learning, Boston, MA: Springer US, 2010, pp. 653.
[19] D. Rosadi, Analisis statistika dengan R, 1st ed., Gadjah Mada University Press, Yogyakarta, 2016.

Modeling the Proportion of Tuberculosis Cases in South Sulawesi using Sparse Least Trimmed Squares

ORIGINALITY REPORT

14%	8%	11%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mehmet Şamil Güneş, Doğan Yıldız, Fatma Sevinç Kurnaz. "A Comparative Assessment of Municipal Water Use in Turkey", Journal of Water Resources Planning and Management, 2022 Publication	1%
2	Irnawati Nur, Nur Nasry Noor, Andi Ummu Salmah, Anwar Mallongi, Hasnawati Amqam. "Risk Factors Analysis and Mapping of Pulmonary Tuberculosis in Community Health Centre Tamalatea of Jeneponto District", Open Access Macedonian Journal of Medical Sciences, 2020 Publication	1%
3	Khusnul Khotimah, Kusman Sadik, Anang Kurnia. "Robust multi-stage method (MM) and least median square (LMS) evaluation on handling outlier for multiple regression", Journal of Physics: Conference Series, 2021 Publication	1%

4	L.E. Melkumova, S.Ya. Shatskikh. "Comparing Ridge and LASSO estimators for data analysis", Procedia Engineering, 2017 Publication	1 %
5	www.researchgate.net Internet Source	1 %
6	projecteuclid.org Internet Source	1 %
7	Mayooran Thevaraja, Azizur Rahman. "Chapter 30 Assessing Robustness of Regularized Regression Models with Applications", Springer Science and Business Media LLC, 2020 Publication	1 %
8	bmcbioinformatics.biomedcentral.com Internet Source	1 %
9	patents.google.com Internet Source	1 %
10	Daniel M. McNeish. "Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences", Multivariate Behavioral Research, 2015 Publication	1 %
11	www.mdpi.com Internet Source	1 %

12	Submitted to KTH - The Royal Institute of Technology Student Paper	<1%
13	www.bus.miami.edu Internet Source	<1%
14	Viktoria Öllerer, Christophe Croux, Andreas Alfons. "The influence function of penalized regression estimators", Statistics, 2014 Publication	<1%
15	mdpi-res.com Internet Source	<1%
16	Submitted to Associatie K.U.Leuven Student Paper	<1%
17	Rendra Gustriansyah, Juhaini Alie, Nazori Suhandi. "Modeling the number of unemployed in South Sumatra Province using the exponential smoothing methods", Quality & Quantity, 2022 Publication	<1%
18	epdf.tips Internet Source	<1%
19	Submitted to (school name not available) Student Paper	<1%
20	Valdivieso Martinez Raul. "Validacion de la eficiencia y modelos de fijacion de precios en	<1%

el mercado mexicano de valores", TESIUNAM, 2004
Publication

21	www.springerprofessional.de Internet Source	<1 %
22	www.tandfonline.com Internet Source	<1 %
23	Chen, D.. "A strategy for enhancing the reliability of near-infrared spectral analysis", Vibrational Spectroscopy, 20080717 Publication	<1 %
24	J Iñigo, A Arce, JM Martín-Moreno, R Herruzo, E Palenque, F Chaves. "Recent transmission of tuberculosis in Madrid: application of capture–recapture analysis to conventional and molecular epidemiology", International Journal of Epidemiology, 2003 Publication	<1 %
25	Modern Nonparametric Robust and Multivariate Methods, 2015. Publication	<1 %
26	S. Safari, J.M. Londoño Monsalve. "Data-driven structural identification of nonlinear assemblies: Structures with bolted joints", Mechanical Systems and Signal Processing, 2023 Publication	<1 %

27	citeseerx.ist.psu.edu Internet Source	<1 %
28	mafiadoc.com Internet Source	<1 %
29	silo.pub Internet Source	<1 %
30	www.aimspress.com Internet Source	<1 %
31	www.ncbi.nlm.nih.gov Internet Source	<1 %
32	Eunho Yang, Aurélie C. Lozano, Aleksandr Aravkin. "A general family of trimmed estimators for robust high-dimensional data analysis", Electronic Journal of Statistics, 2018 Publication	<1 %
33	"Handbook of Applied Spatial Analysis", Springer Science and Business Media LLC, 2010 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		