Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:89

https://doi.org/10.28919/cmbn/8759

ISSN: 2052-2541

QUANTILE REGRESSION MODELING WITH GROUP LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR CLASSIFICATION ON TUBERCULOSIS DATA

IRWAN USMAN, ANNA ISLAMIYATI*, ERNA TRI HERDIANI

Department of Statistics, Hasanuddin University, Makassar 90245, Indonesia

Copyright © 2024 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Tuberculosis (TB) is one of the top 10 causes of death in the world and is a deadly infectious disease in Indonesia. One of Indonesia's provinces that contributed the most TB cases in 2018 was South Sulawesi, with 84 cases per 100,000 population. This study aims to identify variables that can explain the proportion of TB cases in South Sulawesi, potentially leading to more effective prevention and treatment strategies. The data used has many predictor variables, and there are outliers. Quantile regression can be used to overcome outlier data, but it cannot overcome multicollinearity problems. Multicollinearity causes the variance of the estimated parameters to be too large and reduces the accuracy of the estimates, thus requiring a different approach to data analysis. There are various methods for handling regression analysis on data that experiences multicollinearity problems. One of the most commonly known penalized regression methods is Group LASSO. Group LASSO can be used to select variables and overcome multicollinearity. In this study, six naturally formed sector group variables are thought to influence the proportion of TB cases. Quantile regression modeling with LASSO group penalties was carried out using 3 quantile levels, namely (0.25, 0.5, and 0.75). The results of the quantile regression analysis with the LASSO penalty group obtained a different model for each quantile. The best model that is able to explain the proportion of TB cases obtained at the 0.5 quantile level with an R^2 value of 0.99 is closer to 1 than the other quantile model levels.

*Corresponding author

E-mail address: annaislamiyati701@gmail.com

Received July 17, 2024

Keywords: group LASSO; multicollinearity; outliers; quantile regression; tuberculosis.

2020 AMS Subject Classification: 62P10.

1. INTRODUCTION

Regression analysis is a powerful method for evaluating the functional relationship between one dependent variable and one or a set of independent variables. It is widely used to describe, control, and estimate the values of response variables with the help of observations for explanatory variables. The conventional approach to regression modeling is based on sharp data and clear relationships between dependent and independent variables [1]. Regression models develop based on problems in the data, whether in the pattern of relationships between variables [2], types of quantitative and qualitative data [3], or violations of assumptions in the data. One condition often occurs in data is that outliers can cause violations of basic assumptions in regression analysis. To overcome this problem, researchers have developed a quantile regression model that can address the limitations of linear regression [4]. However, quantile regression cannot overcome multicollinearity problems.

Multicollinearity is a violation of the assumptions in regression, which will increase the variance of the regression coefficients, making them unstable and leading to problems in estimating the coefficients. [5-6]. There are various methods for handling regression analysis on data that experiences multicollinearity problems, including PCA [7], LASSO (Least Absolute Shrinkage and Selection Operator), and ridge regression [8]. The main difference between LASSO and ridge analysis lies in the penalty given, where in LASSO, the penalty given is multiplied by the absolute of the regression coefficient. In contrast, in ridge analysis, the penalty given is multiplied by the square of the regression coefficient. This results in LASSO shrinking the regression coefficient value to exactly zero, while Ridge only shrinks the regression coefficient value to close to zero. A study on handling multicollinearity problems found that the LASSO method gave the smallest mean square error value compared to the variable selection method, principal component regression, partial least squares, Ridge regression, and elastic net. [9].

However, there are cases where variable reduction is needed in group form, so the LASSO method was developed into group LASSO, which can also be used in variable selection and overcoming multicollinearity [10-11]. Previous studies found that Group LASSO provided better analysis results than Lasso [12-14]. Therefore, the study uses LASSO group classification in

quantile regression to overcome the problems of outliers and multicollinearity that coincide in the data.

Next, the author applies quantile regression with LASSO group classification to data on tuberculosis cases in South Sulawesi Province. Tuberculosis is a dangerous public health problem throughout the world. This disease requires significant community and health system activity [15]. In 2020, according to data from the Indonesian Ministry of Health, the number of tuberculosis cases in South Sulawesi decreased. However, the decline in the number of Tuberculosis cases still needs to be watched out for because there are disparities in the spread of Tuberculosis between regions in South Sulawesi. For example, Barru Regency and Sidrap Regency in 2018 had 182 Tuberculosis cases each. and 493 cases, but in 2020, the number of cases increased to 202 and 591, respectively [16].

2. MATERIALS AND METHODS

Quantile Regression

Quantile regression is instrumental in various fields, including econometrics, biomedicine, finance, health, the environment, etc. The general equation of linear quantile regression is specifically for conditional quantiles, $Q(\theta|x_{1i}, x_{2i}, \dots x_{ki})$ from the dependent variable Y_i namely:

$$Y_i = \beta_0(\theta) + \beta_1(\theta)x_{1i} + \cdots + \beta_k(\theta)x_{ki} + \varepsilon_i(\theta),\ i = 1,2, \dots, n.$$

If the quantile regression model is presented in matrix form, the above equation can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0(\theta) \\ \beta_1(\theta) \\ \vdots \\ \beta_k(\theta) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(\theta) \\ \varepsilon_2(\theta) \\ \vdots \\ \varepsilon_n(\theta) \end{bmatrix}.$$

Furthermore, the equation above can be written in the following linear model form:

$$Y(\theta) = X\beta\ (\theta) + \ \varepsilon(\theta).$$

If the conditional function of the θ-th quantile has a certain independent variable X, then the conditional function is defined as follows:

$$\begin{aligned} Q(\theta|x_{1i}, x_{2i}, \dots x_{ki}) &= \ Q_Y(\theta|X) \\ &= X_i^T \beta\ (\theta),\ i = 1,2 \dots, n. \end{aligned}$$

So the optimization solution for quantile regression is as follows:

$$\underset{\beta\epsilon\mathbb{R}}{argmin} \sum_{i=1}^{n} \rho_\theta\,(y_1 - X_i^T\beta(\theta),\ \ i = 1,2,\dots,n\ \ \theta\,\epsilon\,(0,1).$$

Where $y_i = \{y_1, y_2, \dots, y_n\}$ is a random sample with dependent variable Y, and $x_i\,\epsilon\,R^\rho$ is a vector of covariates, while $\rho_\theta(u) = u\big(\theta - I(u < 0)\big), 0 < \theta < 1$ which is an asymmetric loss function, and u is the residual from the estimated parameters. The estimator is a general form with the aim of minimizing problems [17].

Group LASSO

Group LASSO is often used to select variables in independent variable data that form a group [18]. Group LASSO can be used in variable selection, overcoming multicollinearity, and categorical data [11]. LASSO Group is a development of LASSO. The group LASSO method was developed by adding group constraints to the LASSO method. A generalization of LASSO regression that may also be affected by coding strategy, but in a different way, is group LASSO regression. Unlike LASSO, group LASSO selects variables by selecting groups of variables, not individual variables [14].

For the vector $\eta\,\epsilon\,R^d,\ d\ \geq 1$ and K is a positive definite symmetric matrix with dimensions $d \times d$, it can be written as follows:

$$\|\eta\|K=(\eta' K\eta)1/2.$$

Written as $\|\eta\|=||\eta||_{I_d}$ for brevity. Given positive definite matrices K_1,..., K_j, the LASSO group estimation is defined as follows:

$$\hat{\beta}^{(g.LASSO)} = argmin\ \frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_j\,\beta_j\right\|^2\ +\ \lambda\sum_{j=1}^{J}\left\|\beta_j\right\|.$$

3. MAIN RESULTS

In the research data, multicollinearity problems occur. Therefore, the quantile regression estimate must be completed using Group LASSO. Before the group LASSO classification, each independent variable is standardized to avoid inequality.

Based on the data grouping carried out and analysis using quantile regression with LASSO group penalty, several groups of predictor variables were obtained, which were selected using optimal lambda. Optimal lambda is used because if the value λ is too large, it will result in parameter estimates being dominated by penalty elements, and as a result, the resulting model can tend to be underfitting and not optimal. Meanwhile, if the lambda value becomes smaller and closer to 0,

the resulting penalty value will be smaller. As a result, the resulting model may tend to be overfitting and not optimal [19]. The optimal lambda values obtained for each quantile point based on the cross-validation results are as follows:

Table 1. Optimal lambda value for each quantile

	Quantile		
	0.25	0.5	0.75
$\boldsymbol{\lambda}$	1.1884	3.0131	2.7454

Quantile regression modeling with LASSO group penalties in this study was carried out at quantiles 0.25, 0.5, and 0.75 using the optimal lambda values in Table 1. The results of quantile regression parameter estimation with LASSO group penalties for each quantile point. Based on Table 2, the group lasso quantile regression model for each quantile is as follows:

- Quantile 0.25

$$\begin{aligned} q_{0.25}(Y|X) = &-272.45192 - 0.00943x_1 + 0.62013x_2 + 5.37901x_3 - 0.17273x_4 \\ &+ 0.00029x_5 + 0.03471x_6 + 11.18028x_7 - 1.78582x_8 - 0.45372x_9 \\ &+ 0.23090x_{10} - 0.28305x_{11} - 0.29292x_{12} - 0.00026x_{13} + 0.00001x_{14} \\ &+ 0.01188x_{15} - 0.00300x_{16} + 0.00711x_{17} + 0.55221x_{18} + 0.16691x_{19} \\ &+ 0.00111x_{20} + 0.00429x_{21} \end{aligned}$$

- Quantile 0.5

$$\begin{aligned} q_{0.5}(Y|X) = &-288.19502 - 0.00950x_1 + 0.69314x_2 + 4.57778x_3 - 0.09341x_4 \\ &+ 0.00033x_5 + 0.03030x_6 + 15.71492x_7 - 1.56991x_8 + 0.32498x_9 \\ &+ 0.19601x_{10} - 0.24799x_{11} - 0.26168x_{12} + 0.00004x_{13} + 0.00020x_{14} \\ &+ 0.01011x_{15} - 0.00268x_{16} + 0.00680x_{17} + 0.55190x_{18} + 0.14977x_{19} \\ &+ 0.00141x_{20} + 0.00424x_{21} \end{aligned}$$

- Quantile 0.75

$$\begin{aligned} q_{0.75}(Y|X) = &-286.48254 - 0.00949x_1 + 0.68373x_2 + 4.69162x_3 - 0.10787x_4 \\ &+ 0.00032x_5 + 0.03097x_6 + 15.34787x_7 - 1.58565x_8 + 0.24775x_9 \\ &+ 0.19810x_{10} - 0.25070x_{11} - 0.26394x_{12} + 0.00003x_{13} + 0.00018x_{14} \\ &+ 0.01029x_{15} - 0.00272x_{16} + 0.00685x_{17} + 0.55741x_{18} + 0.14786x_{19} \\ &+ 0.00136x_{20} + 0.00422x_{21} \end{aligned}$$

Table 2. Results of the estimation of quantile regression parameters with the group LASSO

Variable	Quantile		
	0.25	0.5	0.75
Intercept	-272.45192	-288.19502	-286.48254
x_1	-0.00943	-0.00950	-0.00949
x_2	0.62013	0.69314	0.68373
x_3	5.37901	4.57778	4.69162
x_4	-0.17273	-0.09341	-0.10787
x_5	0.00029	0.00033	0.00032
x_6	0.03471	0.03030	0.03097
x_7	11.18028	15.71492	15.34787
x_8	-1.78582	-1.56991	-1.58565
x_9	-0.45372	0.32498	0.24775
x_{10}	0.23090	0.19601	0.19810
x_{11}	-0.28305	-0.24799	-0.25070
x_{12}	-0.29292	-0.26168	-0.26394
x_{13}	-0.00026	0.00004	0.00003
x_{14}	0.00001	0.00020	0.00018
x_{15}	0.01188	0.01011	0.01029
x_{16}	-0.00300	-0.00268	-0.00272
x_{17}	0.00711	0.00680	0.00685
x_{18}	0.55221	0.55190	0.55741
x_{19}	0.16691	0.14977	0.14786
x_{20}	0.00111	0.00141	0.00136
x_{21}	0.00429	0.00424	0.00422

The model obtained shows that the coefficients of the predictor variables have positive and negative values and can explain the number of tuberculosis cases in South Sulawesi Province. The predictor variables in question come from several group sectors, namely the environmental group, population group, health group, economic group, human resources group, and education group. At all quantile points, the predictor variables that are considered to have a relationship to

the response variable are the same, namely those coming from the environmental group, population group, health group, economic group, human resources group, and education group, meaning that all predictor variables from all sector groups have a relationship to the response variable. However, there is a difference in the influence of the predictor variable on the response variable, namely, at the 0.25 quantile, the positive coefficients are ($x_2, x_3, x_5, x_6, x_7, x_{10}, x_{14}, x_{15}$, x_{17}, x_{18}, x_{19} , x_{20} and x_{21}) and the negative coefficients are ($x_1, x_4, x_8, x_9, x_{11}, x_{12}, x_{13}$ and x_{16}). while at quantiles 0.5 and 0.75, the coefficients have the same positive values, namely ($x_2, x_3, x_5, x_6, x_7, x_9, x_{10}, x_{13}, x_{14}, x_{15}$ x_{17}, x_{18}, x_{19} , x_{20} and x_{21}) and the coefficients that have negative values also equal ($x_1, x_4, x_8, x_{11}, x_{12}$ and x_{16}). A coefficient with a positive value means it has a positive effect on the number of TB cases, namely the higher the value of the predictor variable, the higher the number of TB cases, and a coefficient with a negative value means it has a negative effect on the number of TB cases, namely the higher the value of the predictor variable, the lower the number of TB cases. For example, at quantile 0.25, if the value of the predictor variable x_2 increases by 1 unit, the number of TB cases will increase by 0.62013 cases, and if the value of the predictor variable x_4 increases by 1 unit, the number of TB cases will decrease by 0.17273 cases.

The best regression model in this research is determined by looking at the R-Square (R^2) value, which is the percentage of diversity that measures the ability of the predictor variables in the model to explain the diversity of the response variable. The closer the R^2 value is to 1, or 100%, the better the model. Following are the R^2 values for each model based on quantile points:

Table 3. R^2 value at each quantile

	Quantile		
	0.25	0.5	0.75
$\boldsymbol{R^2}$	0.98	0.99	0.97

Based on Table 3, it can be concluded that the best model obtained is the model at the $\mathcal{Q}_{0.5}$ quantile, which has the lowest R^2 value among the other quantile models at 0.99, meaning this model is able to explain the relationship between the predictor variable and the response variable by 99%.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

[1] N. Chukhrova, A. Johannssen, Fuzzy regression analysis: Systematic review and bibliography, Appl. Soft Comput. 84 (2019), 105708. https://doi.org/10.1016/j.asoc.2019.105708.

[2] A. Islamiyati, N. Sunusi, A. Kalondeng, et al. Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model, J. Sci. Islam. Repub. Iran, 31 (2020), 175–183. https://doi.org/10.22059/jsciences.2020.286949.1007435.

[3] A. Islamiyati, Anisa, M. Zakir, et al. The use of the binary spline logistic regression model on the nutritional status data of children, Commun. Math. Biol. Neurosci. 2023 (2023), 37. https://doi.org/10.28919/cmbn/7935.

[4] Anisa, A. Islamiyati, S. Sahriman, et al. Truncated spline quantile regression model on platelet changes in dengue fever patients based on body temperature, Commun. Math. Biol. Neurosci. 2024 (2024), 69. https://doi.org/10.28919/cmbn/7978.

[5] L. Peng, Quantile regression for survival data, Annu. Rev. Stat. Appl. 8 (2021), 413–437. https://doi.org/10.1146/annurev-statistics-042720-020233.

[6] N. Shrestha, Detecting multicollinearity in regression analysis, Amer. J. Appl. Math. Stat. 8 (2020), 39–42. https://doi.org/10.12691/ajams-8-2-1.

[7] A. Islamiyati, A. Kalondeng, N. Sunusi, et al. Biresponse nonparametric regression model in principal component analysis with truncated spline estimator, J. King Saud Univ. - Sci. 34 (2022), 101892. https://doi.org/10.1016/j.jksus.2022.101892.

[8] K. Enwere, E. Nduka, U. Ogoke, Comparative analysis of ridge, bridge and lasso regression models in the presence of multicollinearity, IPS Intell. Multidiscip. J. 3 (2023), 1–8. https://doi.org/10.54117/iimj.v3i1.5.

[9] A. Yanke, N.E. Zendrato, A.M. Soleh, Handling multicollinearity problems in indonesia's economic growth regression modeling based on endogenous economic growth theory, Indones. J. Stat. Appl. 6 (2022), 228–244. https://doi.org/10.29244/ijsa.v6i2p214-230.

[10] A.A. El Sheikh, S.L. Barakat, S.M. Mohamed, New aspects on the modified group LASSO using the least angle regression and shrinkage algorithm, Inf. Sci. Lett. 10 (2010), 527–536. https://doi.org/10.18576/isl/100317.

[11] X. Liu, P. Cao, J. Wang, et al. Fused group Lasso regularized multi-task feature learning and its application to the cognitive performance prediction of Alzheimer's disease, Neuroinform. 17 (2018), 271–294. https://doi.org/10.1007/s12021-018-9398-5.

[12] S.G. Feronato, M.L.M. Silva, R. Izbicki, et al. Selecting genetic variants and interactions associated with amyotrophic lateral sclerosis: A group LASSO approach, J. Pers. Med. 12 (2022), 1330. https://doi.org/10.3390/jpm12081330.

[13] F.J. Detmer, J. Cebral, M. Slawski, A note on coding and standardization of categorical variables in (sparse) group lasso regression, J. Stat. Plan. Inference 206 (2020), 1–11. https://doi.org/10.1016/j.jspi.2019.08.003.

[14] Y. Huang, T. Tibbe, A. Tang, et al. Lasso and group lasso with categorical predictors: Impact of coding strategy on variable selection and prediction, J. Behav. Data Sci. 3 (2024), 15–42. https://doi.org/10.35566/jbds/v3n2/montoya.

[15] V. Nafade, M. Nash, S. Huddart, et al. A bibliometric analysis of tuberculosis research, 2007–2016, PLoS ONE. 13 (2018), e0199706. https://doi.org/10.1371/journal.pone.0199706.

[16] Badan Pusat Statistik, Sulawesi Selatan dalam Angka 2020 (Sulawesi Selatan in Figures 2020), Makassar: Badan Pusat Statistik Provinsi Sulawesi Selatan, 2020.

[17] R. Koenker, Quantile regression, Econometric Society Monographs, Vol.38, Cambridge University Press, 2005.

[18] H. Chen, Y. Xiang, The study of credit scoring model based on group lasso, Procedia Computer Sci. 122 (2017), 677–684. https://doi.org/10.1016/j.procs.2017.11.423.

[19] L. Meier, S. Van De Geer, P. Bühlmann, The group lasso for logistic regression, J. R. Stat. Soc. Ser. B: Stat. Methodol. 70 (2008), 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x.