International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 6 Issue 4, April - 2022, Pages:350-353

Application Of Lasso And Lasso Quantile Regression In The Identification Of Factors Affecting Poverty Levels In Central Java

Trigarcia Maleachi Randa[1], Georgina Maria Tinungki[2], Nurtiti Sunusi[3]

[1,2,3]Department of Statistics, Hasanuddin University, Makassar, Indonesia
[1]trigarciaranda95@gmail.com
[2]georgina@unhas.ac.id
[3]nurtitisunusi@unhas.ac.id

Abstract: *The condition in which the unknown number parameter to be estimated, p, is much larger than the number of observations, n, is termed high-dimensional. Traditional statistical methods cannot solve high-dimensional problems because they assume many observations and few unknown variables. For high-dimensional modeling, multicollinearity is a frequent phenomenon, causing serious problems with parameter estimation and associated inference and interpretation.. As this reason, Belloni and Chernozhukov in 2011 developed combined methods from Quantile Regression (QR) that is useful for robust regression, and also LASSO that is popular choice for shrinkage estimation and variable selection, becoming LASSO QR. Extensive simulation studies demonstrate satisfactory using LASSO QR in high dimensional datasets that lies outliers better than using LASSO.*

Keywords—high dimensional data; LASSO quantile regression; outliers; robust regression

1. INTRODUCTION

The classical multiple linear regression problem follows the model $\boldsymbol{y}_i = x_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i;\ i = 1,2,\dots,n,$ with $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)'$ p-dimensional regression covariates, a response $\boldsymbol{y}_i$, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)'$ the associated regression coefficients where assumes errors $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$. Estimation of regression parameter, $\boldsymbol{\beta}$ could be using ordinary least squre (OLS) that minimize the sum square of error. The formula follows $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, implies assume $\boldsymbol{X}'\boldsymbol{X}$ is a nonsingular matrix, with matrix covariates $\boldsymbol{X}_{n\times p}$ and response vector $\boldsymbol{y}_{n\times 1}$.

Common issues in the certain background, there are a lot of regression cases in the condition number of predictor variables more than number of observations (p»n). When $\boldsymbol{X}$ is full rank ($p \le n$), the exploration of causal relationship could be accomplished using classical multiple regression above. But when the number of predictors is large compared to the number of observations, $\boldsymbol{X}$ is likely not full rank, that means $\boldsymbol{X}'\boldsymbol{X}$ become singular and the regression approach is no longer feasible (i.e., because of multicolinearity) [1]. LASSO (Least Absolute Shrinkage and Selection Operator) regression [2], is a penalized regression method that is so popular choice for handling this conditions. It is so useful for shrinkage estimation and variable selection.

The worst condition of datasets for regression problem is when they subject to heavy-tailed errors or outliers that may appear in the responses and/or the predictors. In such a situation, it is well known that the traditional OLS may fail to produce a reliable estimator, and the quantile regression (QR) estimator can be very useful. Belloni and Chernozhukov (2011) [3] developed the combined method from QR and LASSO regression. The basic idea is to combine the usual QR criterion and the LASSO-type penalty together to produce the LASSO QR method.

Simulation study have been developed to see the LASSO and LASSO QR processes for handling high-dimensional data containts outliers in a lot of scenarios. The simulation using R software and some of R packages.

2. METHODOLOGY

2.1 Linear Regression

Linear regression is an approach to model the relationship between a scalar response or dependent variable Y and one or more explanatory or independent variables denoted X. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. A linear regression model involving p independent variables can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad (1)$$

where $i = 1,2,\dots,n$. Y_i is the response variable on the i-th observation, $\beta_0, \beta_1, \dots, \beta_p$ are parameters, X_i is the value of the independent variable on the i-th observation, and ε_i is a normally distributed random variable. The error $\varepsilon_i \sim N(0, \sigma^2)$ is not mutually correlated [4].

The most commonly used regression method is the method of ordinary least squares (OLS). The OLS estimate is obtained as the solution of the problem

$$\min J = \min \sum_{i=1}^{n} \varepsilon_i^2 \qquad (2)$$

Taking the partial derivatives of J with respect to $\beta_j, j = 0,1,\cdots,p$ and setting them equal to zero yields the normal equations and obtains the estimated regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} \quad (3)$$

To judge how well the estimated regression model fits the data, we can look at the size of the residuals

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}\right) \quad (4)$$

A point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. The ordinary or simple residuals (observed - predicted values) are the most commonly used measures for detecting outliers. Standardized residuals are the residuals divided by the estimates of their standard errors. They have mean 0 and standard deviation 1.

2.2 Robust Regression

Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model. This method is an important tool for analyzing the data which is affected by outliers so that the resulting models are stout against outliers [5]. When researchers set of regression models and to test the common assumption that the regression assumptions are violated, the transformation seemed unlikely to eliminate or weaken the influence of outliers which eventually became biased predictions. Under these circumstances, robust regression is resistant to the influence of outliers is the best method. Robust regression is used to detect outliers and provide results that are resistant to the outliers [6].

2.3 LASSO Regression

Recent years, the LASSO has become one of the main practical and theoretical tools for sparse high-dimensional variable selection problems . LASSO is a penalized least squares technique which puts L1 constraint on the estimated regression coefficients The LASSO estimator, $\hat{\beta}$, for the linear regression model (1) is given as follows

$$\hat{\beta} \coloneqq \hat{\beta}(\lambda) = \underset{\beta \epsilon \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \quad (5)$$

where $\lambda \geq 0$ is the regularization parameter that controls the amount of shrinkage. Due to the geometry of the L1-norm penalty the LASSO shrinks some of the regression coefficients to exactly zero (to elaborate later). Thus it serves as a variable selection method also.

The lasso gives rise to a convex optimization problem and thus is computationally tractable even for moderately large problems. Indeed, the LARS (Least Angle Regression and Shrinkage) algorithm [7] can compute the entire solution path as a function of λ in $O(p^3 + np^2)$ operations

2.4 Quantile Regression

The setting of interest corresponds to a parametric quantile regression model, where the dimension p of the underlying model increases with the sample size n. Namely, we consider a response variable y and p-dimensional covariates x such that the u-th conditional quantile function of y given x is given by

$$F^{-1}_{y_i|x_i}(u|x_i) = x'\beta(u), \;\; \beta(u) \epsilon \mathbb{R}^p, \;\; \text{for all } u \epsilon \mathcal{U} \quad (6)$$

where $\mathcal{U} \subset (0,1)$ is a compact set of quantile indexes. The u-th conditional quantile $F^{-1}_{y_i|x_i}(u|x_i)$ is the inverse of the conditional distribution function $F_{y_i|x_i}(y|x_i)$ of y_i given x_i. We consider the case where the dimension p of the model is large, possibly much larger than the available sample size n, but the true model $\beta(u)$ has a sparse support

$$\begin{aligned} T_u &= \text{support}\big(\beta(u)\big) \\ &= \big\{j \in \{1, \dots, p\} : |\beta_j(u)| > 0\big\} \end{aligned} \quad (7)$$

having only $s_u \leq s \leq n/\log(n \vee p)$ non-zero components for all $u \epsilon \mathcal{U}$.

The population coefficient $\beta(u)$ is known to minimize the criterion function

$$Q_u(\beta) = \mathrm{E}[\rho_u(y - x'\beta)] \quad (8)$$

where $\rho_u(t) = (u - 1\{t \leq 0\})t$ is the asymmetric absolute deviation function [8]. Given a random sample $(y_1, x_1), \cdots, (y_n, x_n)$, the quantile regression estimator of $\beta(u)$ is defined as a minimizer of the empirical analog of (5):

$$\hat{Q}_u(\beta) = \mathbb{E}[\rho_u(y - x'\beta)] \quad (9)$$

2.5 LASSO Quantile Regression

In high-dimensional settings, particularly when $p \geq n$, ordinary quantile regression is generally inconsistent, which motivates the use of penalization in order to remove all, or at least nearly all, regressors whose population coefficients are zero, thereby possibly restoring consistency. A penalization that has proven quite useful in least squares settings is the L1-penalty leading to the Lasso estimator.

The L1-penalized quantile regression estimator $\hat{\beta}(u)$ is a solution to the following optimization problem:

$$\min_{\beta \epsilon \mathbb{R}^p} \hat{Q}_u(\beta) + \frac{\lambda\sqrt{u(1-u)}}{n} \sum_{j=1}^{p} \hat{\sigma}_j |\beta_j| \quad (10)$$

where $\hat{\sigma}_j^2 = \mathbb{E}\big[x_{ij}^2\big]$. The criterion function in (10) is the sum of the criterion function (9) and a penalty function given by a scaled L1-norm of the parameter vector. The overall penalty level $\lambda\sqrt{u(1-u)}$ depend on each quantile index u, while λ will depend on the set $\mathcal{U}$ of quantile indeces of interest. LASSO QR has been considered in [9] under small (fixed) p asymptotics. Therefore, the problem (10) can be solved in polynomial time, avoiding the computational curse of dimensionality. In order to describe our choice of the penalty level λ, we introduce the random variable

$$\Lambda = n \sup_{u\epsilon\mathcal{U}} \max_{1\le j\le p} \left| \mathbb{E}_n \left[\frac{x_{ij}(u - 1\{u_i \le u\})}{\hat{\sigma}_j \sqrt{u(1-u)}} \right] \right| \qquad (11)$$

where $u_i, \cdots, u_n$ are i.i.d uniform $(0,1)$ random variables, independently distributed from the regressors, variable $x_1, \cdots, x_n$. The random variable Λ has a known, that is, pivotal, distribution conditional on $X = [x_1, \cdots, x_n]'$. We then set

$$\lambda = c \cdot \Lambda(1 - \alpha|X) \qquad (12)$$

where $\Lambda(1-\alpha|X) :=$ $(1-\alpha)$-quantile of Λ conditional on X, and the constant $c > 1$ depends on the design [3].

3. Simulation Study

The simulation in this research using R software that would evaluate mean squared errors performance. It will be repeated 1000 times to get the average of mean squared errors. Using R package, glmnet and quantreg for LASSO and LASSO quantile regression, the simulation set in $= 200$, and vary from 500 to 3000 as shown in Table 1. The datasets generated independently with each row of the design matrix from a *p*-dimensional normal distribution ***N*** (5, 1). Then the response vector generated follows $y = 5x_{i1} + 4x_{i2} + 3.5x_{i3} + \varepsilon_i$, where ε is independently generated from ***N*** (0, 1). The scenario before generated without effects of outliers.

Next scenarios generated using the effects of outliers by replacing the distributions of errors that generated from vector of error components obtained from a combination of sampling data ***N*** (0, 1) as much 0.7 × *n* and data ***N*** (30, 1) as much as 0.3 × *n*. For comparison purpose, all of scenarios to be evaluated using LASSO and LASSO Quantile Regression.

Table 1. Average of Mean Squared Errors

Scenario 1 : Dataset without outliers				
Method	*p* = 500	*p* = 1000	*p = 1500*	*p* = 3000
LASSO	1.0526	1.1067	1.0638	1.2295
LASSO QR	0.0490	0.1845	0.0502	0.0931
Scenario 2 : Dataset with outliers				
Method	*p = 500*	*p = 1000*	*p = 1500*	*p = 3000*
LASSO	14.7597	15.3537	17.7238	14.7763
LASSO QR	0.0459	0.5184	0.2536	0.4928

Table 1 presents the average of mean squared errors for LASSO and LASSO QR estimator. It is shown that in the datasets without outliers, the better performance is from LASSO QR that have lowest average of mean squared errors compared to LASSO. It is also happen almost in the datasets with outliers, the performance of LASSO QR is better than LASSO This is indicated by the low the average mean squared errors value of LASO QR method.

4. Actual Data Analysis

This study also uses actual data as an application to examine the performance of the estimators. The actual data used is poverty levels data (in percent) as a response variable. The explanatory variables used are data on human development index in percent (X_1), data on the open unemployment rate in percent (X_2), data on the labor force participation rate in percent (X_3), and data on the district/city minimum wage in in million rupiahs (X_4), data on the mean years school in years (X_5), data on the gross regional domestic product in billion rupiahs (X_6), data on the expected years of schooling in years (X_7), data on the government consumption in million rupiahs (X_8), and data on the school enrollment rate in percent (X_9). The actual data used were obtained from the Jawa Tengah Province in Figures by the Badan Pusat Statistik in 2018 [10]. Here is the link for data download: https://jateng.bps.go.id. The data contains of 35 observations which are district/city poverty levels in Indonesia

4.1 Outliers detection

Outlier data detection using the difference of fits (DFFITS) method. Observations are declared as outliers if $|\text{DFFITS}| > 2\sqrt{p/n} = 1.0142$.

Table 2 Outliers detected by DFFITS

District/City	|DFFITS|
Demak	2.5726
Tegal	1.7639

Based on Table 2, it can be seen that there are two district/city declared as outliers, namely Demak District and Tegal District.

4.2 Multicollinearity detection

Multicollinearity refers to the significant correlation among the independent variables in the regression model. To measure the amount of multicollinearity in dataset, variance inflation factor (VIF) is examined. Table 3 shows the value of vif for each predictor variable Three predictor variables suffer from multicollinearity problem that is variables X_1, X_5, and X_9 with variance inflation factor (VIF) value higher than 10. Therefore, this dataset has both outlier and multicollinearity problems that is

Table 3 Variance Inflaion Factor of Poverty Level Datset

Variable	VIF	Variable	VIF
X_1	21.551387	X_6	4.581551
X_2	3.052960	X_7	15.206586
X_3	2.929930	X_8	5.121319
X_4	2.738633	X_9	2.338255
X_5	21.929398		

4.3 Regression Analysis

The results of the parameter estimates or regression coefficients estimates obtained from each of the methods used

in this study are presented in Table 4. These results indicate that the factors that can explain the poverty levels in Central Java Province according to the LASSO QR analysis are all explanatory variables. Meanwhile, from the LASSO model that has been obtained, it is known that the factors that can explain the poverty levels in Central Java Province according to the LASSO analysis are X_1, X_4 and X_8.

Table 4 Regression coefficients of each methods

Estimate	LASSO QR	LASSO
$\hat{\beta}_1$	-0.2313	-0.3549
$\hat{\beta}_2$	0.3675	0.0000
$\hat{\beta}_3$	-0.09745	0.0000
$\hat{\beta}_4$	-6.6546	-.3.6296
$\hat{\beta}_5$	0.1983	0.0000
$\hat{\beta}_6$	-0.00003	0.0000
$\hat{\beta}_7$	-0.4528	0.0000
$\hat{\beta}_8$	1.9878	0.1215
$\hat{\beta}_9$	-0.3832	0.0000

4.4 Evaluation Model Goodness of Fit

RMSE and R^2 values as the criteria for the goodness of the model are used in this study to examine the performance between methods in the actual data regression analysis. The method that produces the smallest values of RMSE and high R^2 is the best method. The RMSE and R^2 values of each method in estimating the poverty levels data regression model are presented in Table 5. The smallest RMSE and $R2$ values were produced by the LASSO QR method, namely 38.81 and 0.9978, respectively. Meanwhile, LASSO produced the largest RMSE value and the lowest R^2. Therefore, the LASSO QR method is the best method in high-dimensional datasets analysis to determine the effect of several explanatory variables used on poverty levels in Central Java Provinces in 2018.

Table 5 RMSE and R^2 value of actual data regression

Method	RMSE	R^2
LASSO QR	2.31254	61.5452%
LASSO	2.8039	39.68135%

5. CONCLUSION

LASSO QR has good performance on high-dimensional datasets without outliers based on the results shown from the simulation studies carried out. LASSO QR is also robust for high-dimensional datasets containing outliers, this is indicated by average the mean squared errors of the low values. The application of LASSO QR and LASSO on the actual data of poverty levels in Central Java Province in 2018 which has 35 data sizes and the conclusion that the LASSO QR method is the best method. This is indicated by the lowest RMSE values and the highest R^2. There are 2 observations that were detected as outliers based on the DFFITS method, namely the 21st observation, Demak District and the 28th observation, Tegal District. This means that the conclusions obtained from estimating parameters of the actual data on poverty levels in 2018 are in line with the results of estimating multiple regression parameters on high-dimensional data that containing outliers through simulation LASSO QR has best performance.

.

6. REFERENCES

[1] Young, M. (1990). *Classical and Modern Regression with Applications*. 2nd Edition. PWS-Kent Publishing Company. Boston.

[2] Tibshirani, R. (1996). *Regression Shrinkage and Selection via the LASSO*. Journal of the Royal Statistics Society Series: B, 58 (1), 267-288.

[3] Belloni, A. & Chernozhukov, V. (2011*). ℓ1-penalized quantile regression in high-dimensional sparse models*. The Annals of Statistics, 39(1), 82-130

[4] Montgomery, D. C. & Peck, E. A. (2006). *An Introduction to Linear Regression*. John Wiley Sons Inc., New York.

[5] Draper, N. R. & Smith, H. (1998). *Aplied Regression Analysis.* 3rd Edition. John Wiley Sons Inc., New York.

[6] Chen, C. (2002). Robust regression and outlier detection with the ROBUSTREG procedure. *Statistics and Data Analysis* Paper 265-27. SAS Instite Inc., Cary, NC.

[7] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). *Least angle regression*. The Annals of Statistics, 32(2), 407–499.

[8] Koenker, R. (1978). *Quantile regression*. Econometrics, 46(1), 33-50.

[9] Knight, K. & Fu, W. J. (2000). *Asymptotics for Lasso-type estimators*. The Annals of Statistics, 28(5), 1356–1378

[10] [BPS] Badan Pusat Statistik Provinsi Jawa Tengah. (2019). Jawa Tengah Province in Figures. BPS, Semarang.