SHORT COMMUNICATION

The Best Model of LASSO With The LARS (Least Angle Regression and Shrinkage) Algorithm Using Mallow’s $C_p$

Trisha Magdalena Adelheid Januaviani$^1$, Nurul Gusriani$^2$, Khafsahe Joebaedi$^2$, Sudradjat Supian$^2$, Subiyanto$^3,*$

$^1$Master Program in Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia
$^2$Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia
$^3$Department of Marine Science, Faculty of Fishery and Marine Science, Universitas Padjadjaran, Indonesia

*E-mail address: subiyanto@unpad.ac.id

ABSTRACT

Multicollinearity often occurs in regression analysis. Multicollinearity is a condition of correlation between independent variables which is a problem. One method that can overcome multicollinearity is the LASSO (Least Absolute Shrinkage and Selection Operator) method. LASSO is able to help to shrink multicollinearity and improve the accuracy of linear regression models. Estimators of LASSO parameters can be solved by the LARS (Least Angle Regression and Shrinkage) algorithm by algorithm which calculates the correlation vector, the largest absolute correlation value, equiangular vector, inner product vector, and determines the LARS algorithm limiter for LASSO. Selecting the best model using the Mallow’s $C_p$ statistics. The smallest Mallow’s $C_p$ value will be selected as the best model. LASSO method with a more detailed procedure with LARS algorithm and selecting the best model using the Mallow’s $C_p$ statistics is discussed in this paper.
Keywords: LARS, LASSO, Cp Mallows, Multicollinearity

1. INTRODUCTION

Linear regression analysis is the correlation between two variables namely independent variable and independent variable (Tong and Ng, 2018; Chen et al., 2018). Correlation between variables does not only consist of two variables, but there can be a correlation between three or more variables called multiple linear regression (Permai and Tanty, 2018). Multiple linear regression analysis has many independent variables, so there is a correlation between two or more independent variables (Nakamura et al., 2017; Baskar et al., 2017). This correlated independent variable is called multicollinearity (Zhou and Huang, 2018; Katrutsa and Strijov, 2017). Reduce multicollinearity and increase the accuracy of linear regression models can use the LASSO Method (Least Absolute Shrinkage and Selection Operator) (Sermpinis et al., 2018; Melkumovaa and Shatskikhb, 2017). LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produce (Chen and Xiang, 2017; Kim et al., 2015; Algamal and Lee, 2015).

The lasso is the best-studied, most basic, shrinkage operator technique (Dyar et al., 2012). LASSO shrinks the coefficients (parameters) which correlate to zero or close to zero (Gauthier et al., 2017), resulting in estimators with smaller variants and a more representative final model (Tibshirani, 1996). The Lasso method became known after the LAR (Least Angle Regression) algorithm in 2004. The solution paths of LAR are piecewise linear and thus can be computed very efficiently (Lee and Jun, 2018; Iturbide et al., 2013). LARS (Least Angle Regression and Shrinkage) modification of LAR to LASSO. LARS is efficient algorithm for estimating computational LASSO parameters. The LASSO method can shrink the ordinary least squares method coefficient to zero so that it can select the fixed variable. The model produced by the LASSO method is simpler and indirectly free from multicollinearity (Efron et al., 2004). Selecting the best model using the Mallow’s Cp statistic. The smallest Mallow’s Cp value will be selected as best model. This paper will discuss the best model of LASSO with LARS algorithm using Mallow’s Cp.

2. METHODS AND MATERIALS

2.1. Multiple Linear Regression Analysis

Regression analysis is one of the data analysis techniques in statistics, regression is often used to examine the relationship between several variables and predict a variable. Variables consist of independent variables and non-independent variables. Multiple linear regression models examine the effect of two or more independent variables on non-independent variables (Kutner et al., 2004). The general form of multiple linear regression models (Kazemi et al., 2013; Miyashiro and Takano, 2015; Permai and Tanty, 2018):

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i \]  

(1)
2. 2. The Ordinary Least Squares Method

The Ordinary Least Squares Method (OLS) used to obtain a linear regression coefficient estimator. OLS is one method that can be used to estimate the \( \beta \) parameter in multiple linear regression.

The estimated model:

\[
y_\hat{i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i
\]

with \( \hat{\beta} \) parameter estimator on the Ordinary Least Squares Method (Kazemi et al., 2013):

\[
\hat{\beta}_{OLS} = (X^t X)^{-1} X^t y
\]

2. 3. The Coefficient of Determination

The Determination Coefficient \( (R^2) \) measures how far the model's ability to explain variations in non-independent variables (Hössjer, 2008; Renaud and Victoria-feser, 2010). The value of the determination coefficient ranges between 0 and 1. Value coefficient of determination close to one means that the independent variables provide almost all the information needed to predict variations in non-independent variables.

The coefficient of determination \( (R^2) \) defined as follows:

\[
R^2 = \frac{\hat{\beta}^t X^t y - n \bar{y}^2}{y^t y - n \bar{y}^2}
\]

2. 4. Data Standardization

Data standardization means standardizing the independent variables in the ordinary least squares method equation as follows (Wang et al., 2017):

\[
x_{ij}^* = \frac{(x_{ij} - \bar{x}_j)}{s_{X_j} \sqrt{n-1}} \text{ where } s_{X_j} = \sqrt{\frac{\Sigma_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}}
\]

2. 5. Multicollinearity

Multiple linear regression analysis has many independent variables, so there is often a correlation between two or more independent variables (Jadhav et al., 2014). This correlated independent variable is called multicollinearity. Variance Inflation Factor (VIF) values are less than 10, so there is no multicollinearity (Alauddin and Nghiemb, 2010).

\[
VIF_j = \frac{1}{TOL_j} = \frac{1}{1 - R_j^2} \text{ where } TOL_j = 1 - R_j^2
\]

2. 6. Mallow’s Cp Statistic

Colin Mallow developed Mallow’s \( C_p \) statistic as a tool in estimating the number of independent variables in regression (Ogasawara, 2016) Mallow’s \( C_p \) is one way to evaluate the selection of the best models in best subset regression (Miyashiro and Takano, 2015). Mallow’s
\( C_p \) statistics value is the best model (Lorchirachoonkul and Jitthavech, 2012). The mathematical form of Mallow’s \( C_p \) Statistics is as follows (Kazemi et al., 2013; Miyashiro and Takano, 2015; Jansen, 2015):

\[
C_p = \frac{\|y-\hat{x}\|_2^2}{\sigma^2} n + 2(q + 1) \quad \text{where} \quad \sigma^2 = \frac{\|y-\hat{x}_{OLS}\|_2^2}{n - p + 1}
\]  

2.7. LARS for LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) regression method is a method to overcome multicollinearity. LASSO is one of the independent variable shrinkage regression techniques (Chand et al., 2018; Shi et al., 2018; Torres-barr et al., 2017). LARS (Least Angle Regression and Shrinkage) modification of LAR to LASSO. LARS is efficient algorithm for estimating computational LASSO parameters. Calculation of LASSO parameters using LARS can use the following steps (Zhang and Li, 2015):

a) The independent variable transformation \( X^* \) can be calculated by equation (4), while \( Y^* = Y - \bar{Y} \) then searches for \( \hat{\beta}^*_{OLS} \) with equation (2). Initially define \( i = 1 \) with \( \hat{\mu}^{(1)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \), in dimension \( n \times 1 \), \( n \) is the amount of data, the value of \( \hat{\mu} \) will change as the stage progresses. Suppose that \( \hat{\mu}^{(i)}_A \) is the estimated value with active variable \( A \) and define \( \hat{\beta} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \) with the dimensions of \( n \times p \), \( p \) is the number of independent variables.

b) Calculating the correlation vector \( \hat{c}^{(i)} = X^t (Y^* - \hat{\mu}^{(i)}_A) \) and the largest absolute correlation value

\[
\hat{c}^{(i)} = max\{ |\hat{c}_j^{(i)}| \}, \text{ so that } A = \{ j | |\hat{c}_j^{(i)}| |\hat{c}^{(i)}| \}
\]  

c) Calculating equiangular vector \( (u^{(i)}_A). \)

Defined \( X^{(i)}_A = [...sjX_j...]_{j \in A} \), \( s_j = \text{sign}\{\hat{c}_j^{(i)}\}, j \in A, \omega^{(i)}_A = P^{(i)}AG^{(i)-1}_A^{-\frac{1}{2}} \)

while \( G^{(i)}_A = X^{(i)}_A X^{(i)}_A^t \) and \( P^{(i)}_A \) are obtained:

\[
u^{(i)}_A = X^{(i)}_A \omega^{(i)}_A
\]  

d) Defined the inner product vector \( a^{(i)} \equiv X^t u^{(i)}_A \)

so that \( \hat{\gamma} \) can be obtained by the following equation :

\[
\hat{\gamma}^{(i)} = \min_{j \in A}^+ \{ \frac{\hat{c}^{(0)}_j - \hat{c}^{(i)}_j}{p^{(i)}_{A-a} - a_j}, \frac{\hat{c}^{(0)}_j + \hat{c}^{(i)}_j}{p^{(i)}_{A+a} + a_j} \}
\]
Calculating $\gamma_j^{(i)}$ with equation:

$$
\gamma_j^{(i)} = \frac{-\hat{\beta}_j^{(i)}}{s_j \omega_{\lambda_j}^{(i)}}
$$

(10)

The LARS algorithm for LASSO must meet the following conditions: $\varphi^{(i)} = \min \{\gamma_j^{(i)}\}$, $\gamma_j^{(i)} > 0$, if $\varphi^{(i)}$ does not have a value then $\varphi^{(i)} = \hat{\gamma}^{(i)}$, but if $\varphi^{(i)} < \hat{\gamma}^{(i)}$ stop the LARS process in this step, remove the variable $j$ from the calculation $\hat{\beta}^{(i+1)}$ then, the value $\varphi^{(i)}$ becomes equal to the value of $\hat{\gamma}^{(i)}$ and $A_+ = A - \{j\}$, but if all the independent variables have entered, ignore this step.

f) Renew value $\hat{\beta}^{(i+1)}$ and $\hat{\mu}_A^{(i+1)}$ with

$$
\hat{\beta}_j^{(i+1)} = \hat{\beta}_j^{(i)} + \hat{\gamma}^{(i)} \omega_{\lambda_j}^{(i)} s_j \\
and \hat{\mu}_A^{(i+1)} = \hat{\mu}_A^{(i)} + \hat{\gamma}^{(i)} u_A^{(i)}
$$

(11)

g) There is a data standardization so the $\hat{\beta}_{\text{LASSO}}$ value will be returned to the actual data with the equation:

$$
\hat{\beta}_{\text{LASSO}}^{(i+1)} = \frac{\hat{\beta}_{\text{LASSO}}^{(i+1)}}{\text{Scale}_j} \\
\text{with Scale}_j = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \bar{X}_j)^2}
$$

(12)

h) Calculate the statistical value of $C_p^{i+1}$ in equation (6)

i) If $\hat{\beta}_{\text{LASSO}}^{(i)} \neq \hat{\beta}_{\text{LSM}}^{*}$ returns to step (b) with $i = i + 1$. The iteration is carried out to a maximum of the amount of data, therefore the iteration $i = 1, 2, \ldots, n$ so that the value $\hat{\beta}_{\text{LASSO}}^{(i)} = \hat{\beta}_{\text{LSM}}^{*}$.

j) Look for the best model with the smallest Mallow’s $C_p$ value.

3. CONCLUSIONS

In this paper, The LASSO method can be determined by LARS algorithm. LARS is a more efficient algorithm to find lasso parameters. LARS for LASSO, that calculates vectors, the largest absolute value, equiangular vector, inner product vector, and determines LARS algorithm limiter for LASSO. The best model with the smallest Mallow’s $C_p$ value.

Acknowledgements

Acknowledgments are conveyed to the Rector, Director of Directorate of Research, Community Involvement and Innovation, and the Dean of Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran.

-249-
Biography

Trisha M. A. Januaviani is a graduate of Mathematics at Universitas Padjadjaran with honors in 2017. Miss Januaviani is currently continuing her studies in The Master Program in Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran.

Nurul Gusriani is a lecturer in the Department of Program in Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Mrs. Gusriani is a graduate of the Masters Program in Statistic, Bogor Agricultural University. Mrs Gusriani has published various journals in statistics especially in ridge regression, telbs regression, markov chains and others.

Khafsah Joebaedi is a lecturer in the Department of Program in Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Mrs. Khafsah has published various journals in mathematics especially in topology, space time autoregression (STAR) and banach space.

Sudradjat Supian a Professor of Operation Research in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently serves as dean of Faculty of Mathematics and Natural Sciences, the field of applied mathematics, with a field of operation research and modeling.

Subiyanto is a lecturer in the Department of Marine Science, Faculty of Fishery and Marine Science, Universitas Padjadjaran. He received his Ph.D in School of Ocean Engineering from Universiti Malaysia Terengganu (UMT), Malaysia in 2017. His research focuses on applied mathematics, numerical analysis and computational science.

References


[8] P. Gauthier, W. Scullion, A. Berry, Sound Quality Prediction Based on Systematic Metric Selection and Shrinkage: Comparison of Stepwise, Lasso, and Elastic-Net


