Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 116 (2019) 245-252 EISSN 2392-2192

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], Khafsah 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.

(Received 10 December 2018; Accepted 20 December 2018; Date of Publication 21 December 2018)

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 etal., 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 $C_p$statistics. The smallest Mallow's C_p value will be selected as best model. This paper will discuss the best model of LASSO with LARS algorithm using Mallow's C_p.

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):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

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 β parameter in multiple linear regression.

The estimated model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_j x_{ij} + \cdots + \hat{\beta}_p x_{ip} + \varepsilon_i$$

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

$$\hat{\beta}_{OLS} = (X^t X)^{-1} X^t y \tag{2}$$

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:

$$R^2 = \frac{\hat{\beta}^t X^t y - n\bar{y}^2}{y^t y - n\bar{y}^2} \tag{3}$$

2. 4. Data Standardization

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

$$x_{ij}^* = \frac{(x_{ij} - \bar{x}_j)}{s_{X_j}\sqrt{n-1}} \quad \text{where } s_{X_j} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2}{n-1}} \tag{4}$$

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).

$$\mathrm{V}IF_j = \frac{1}{TOL_j} = \frac{1}{1 - R_j^2} \quad \text{where } TOL_j = 1 - R_j^2 \tag{5}$$

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 regrression (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):

$$C_p = \frac{\|y - X\hat{\beta}\|_2^2}{\sigma^2} - n + 2(q+1) \text{ where } \sigma^2 = \frac{\|y - X\hat{\beta}_{OLS}\|_2^2}{n-p+1} \tag{6}$$

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 $Y^* = Y - \overline{Y}$ 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× 1, n is the amount of data, the value of $\hat{\mu}$ will change as the stage progresses. Suppose that $\hat{\mu}_A^{(i)}$ is the estimated value with active variable A and define $\hat{\beta} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & 0 \end{bmatrix}$ with the dimensions of n× p, p is the number of independent variables.

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

$$\hat{C}^{(i)} = max\{|\hat{c}_j{}^{(i)}|\}, \text{ so that } A = \{j \,|\, |\hat{c}_j{}^{(i)}||\hat{C}^{(i)}|\} \tag{7}$$

c) Calculating equiangular vector ($\mathbf{u}_\mathbf{A}^{(\mathbf{i})}$).
Defined $X_A^{(i)} = [\ldots s_j X_j^* \ldots]_{j \in A}$, $s_j = \text{sign}\{\hat{c}_j{}^{(i)}\}$, $j \in A$, $\omega_A^{(i)} = P_A^{(i)} G_A^{(1)^{-1}} 1_A$
while $G_A^{(i)} = X_A^{(i)^t} X_A^{(i)}$ and $P_A^{(i) 1_A^T G_A^{(i)^{-1}} 1_A)^{-\frac{1}{2}}}$. So that the equiangular vector value is obtained:

$$u_A^{(i)} = X_A^{(i)} \omega_A^{(i)} \tag{8}$$

d) Defined the inner product vector $a^{(i)} \equiv X^{*^t} u_A^{(i)}$
so that $\hat{\gamma}$ can be obtained by the following equation :

$$\hat{\gamma}^{(i)} = \min{}^+_{j \in A^c} \left\{ \frac{\hat{C}^{(i)} - \hat{c}_j^{(i)}}{P_A^{(i)} - a_j^{(i)}}, \frac{\hat{C}^{(i)} + \hat{c}_j^{(i)}}{P_A^{(i)} + a_j^{(i)}} \right\} \tag{9}$$

e) Calculating $\gamma_j^{(i)}$ with equation :

$$\gamma_j^{(i)} = \frac{-\hat{\beta}_j^{(i)}}{s_j \omega_{A_j}^{(i)}} \tag{10}$$

The LARS algorithm for LASSO must meet the following conditions: $\varphi^{(i)} = \underbrace{min}_{\gamma_j^{(i)}>0} \{\gamma_j^{(i)}\}$, 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

$$\hat{\beta}_j^{(i+1)} = \hat{\beta}_j^{(i)} + \hat{\gamma}^{(i)} \omega_{Aj}^{(i)} s_j \quad \text{and } \hat{\mu}_A^{(i+1)} \text{with } \hat{\mu}_A^{(i+1)} = \hat{\mu}_A^{(i)} + \hat{\gamma}^{(i)} u_A^{(i)} \tag{11}$$

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

$$\hat{\beta}^*_{LASSO_{(i+1)j}} = \frac{\hat{\beta}^{i+1}_{LASSO_j}}{Scale_j} \text{ with } Scale_j = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \overline{X_j})^2} \tag{12}$$

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

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

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.

Acknowledgement

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.

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

[1] M. Alauddin, H.S. Nghiemb, Do Instructional Attributes Pose Multicollinearity Problems ? An Empirical Exploration. *Economic Analysis and Policy* 40(3) (2010) 351-361.

[2] Z.Y. Algamal, M.H. Lee, Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications* 42 (2015) 9326–9332.

[3] S. Chand, S. Ahmad, M. Batool, Solution path ef fi ciency and oracle variable selection by Lasso-type methods. *Chemometrics and Intelligent Laboratory Systems* 183 (2018) 140-146.

[4] H. Chen, X. Yaoxin, The Study of Credit Scoring Model Based on Group Lasso. *Procedia Computer Science* 122 (2017) 677–684.

[5] S. Chen, C.H.Q. Ding, B. Luo, Linear Regression Based Projections for Dimensionality Reduction. *Information Sciences* 467 (2018) 74-86.

[6] M.D. Dyar, M.L. Carmosino, E.A. Breves, M.V. Ozanne, S.M. Clegg, R.C. Wiens, Comparison of Partial Least Squares and Lasso Regression Techniques as Applied to Laser-Induced Breakdown Spectroscopy of Geological Samples. *Spectrochimica Acta Part B: Atomic Spectroscopy* 70 (2012) 51-67.

[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least Angle Regression. *Annals of Statistics* 32(2) (2004) 407-499.

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

Algorithms and Clustering Preprocessing. *Journal of Sound and Vibration* 400 (2017) 134-53.

[9] O. Hössjer. On the coefficient of determination for mixed regression models. *Journal of Statistical Planning and Inference* 138 (2008) 3022–3038.

[10] E. Iturbide, J. Cerda, M. Graff. A Comparison between LARS and LASSO for Initialising the Time-Series Forecasting Auto-Regressive Equations. *Procedia Technology* 7 (2013) 282–288.

[11] N.H. Jadhav, D.N. Kashid, S.R. Kulkarni, Subset selection in multiple linear regression in the presence of outlier and multicollinearity. *Statistical Methodology* 19 (2014) 44–59.

[12] M. Jansen, Generalized Cross Validation in Variable Selection with and without Shrinkage. *Journal of Statistical Planning and Inference* 159 (2015) 90-104.

[13] Kazemi, A. Mohamed, H. Shareef, H. Zayandehroodi, Optimal Power Quality Monitor Placement Using Genetic Algorithm and Mallow ' s Cp. *International Journal of Electrical Power and Energy Systems* 53 (2013) 564–575.

[14] Katrutsa, V. Strijov, Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems With Applications* 76 (2017) 1–11.

[15] Kim, J. Lee, H. Yang, W. Bae, Case influence diagnostics in the lasso regression. *Journal of the Korean Statistical Society* 44 (2015) 271–279.

[16] J. Kuan, Regression analysis estimation of stature from foot length. *Cognitive Systems Research* 52 (2018) 251–260.

[17] M.H. Kutner, C.J. Nachtsheim, J. Neter, Applied Linear Regression Models. 4th ed. New York: McGraw-Hill Companies, Inc (2004).

[18] S. Lee, C. Jun, Fast Incremental Learning of Logistic Model Tree Using Least Angle Regression. *Expert Systems With Applications* 97 (2018) 137-145.

[19] V. Lorchirachoonkul, J. Jitthavech, A Modified Cp Statistic in a System-of-Equations Model. *Journal of Statistical Planning and Inference* 142(8) (2012) 2386–2394

[20] L.E. Melkumovaa, S.Y. Shatskikhb, Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering* 201 (2017) 746-755.

[21] R. Miyashiro, Y. Takano, Subset Selection by Mallow's Cp: A Mixed Integer Programming Approach. *Expert Systems With Applications* 42(1) (2015) 325-331.

[22] K. Nakamura, T. Yasutaka, T. Kuwatani, T. Komai, Development of a predictive model for lead, cadmium and fluorine soil–water partition coefficients using sparse multiple linear regression analysis. *Chemosphere* 186 (2017) 501-509.

[23] H. Ogasawara, Accurate distributions of Mallows' Cp and its unbiased modifications with applications to shrinkage estimation. *Journal of Statistical Planning and Inference* 184 (2016) 105-116.

[24] S.D Permai, H. Tanty, Linear Regression Model Using Bayesian Approach for Energy Linear Regression Model Using Bayesian Approach for Energy Performance of Residential Building Performance of Residential Building. *Procedia Computer Science* 135 (2018) 671-677.

[25] O. Renaud, M. Victoria-Feser, A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference* 140 (2010) 1852-1862.

[26] G. Sermpinis, S. Tsoukas, P. Zhang, Modelling Market Implied Ratings Using LASSO Variable Selection Techniques. *Journal of Empirical Finance* 48 (2018) 19-35.

[27] X. Shi, Y. Huang, J. Huang, S. Ma, A Forward and Backward Stagewise algorithm for nonconvex loss functions with adaptive Lasso. *Computational Statistics and Data Analysis* 124 (2018) 235–252.

[28] H. Tong, M. Ng, Analysis of Regularized Least Squares for Functional Linear Regression Model. *Journal of Complexity* 49 (2018) 85-94.

[29] Torres-barr, C.M. Alaız, J.R. Dorronsoro, ν-SVM Solutions of Constrained Lasso and Elastic Net. *Neurocomputing* 275(31) (2017) 1921-1931.

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

[31] S. Wang, B. Ji, J. Zhao, W. Liu, T. Xu, Predicting Ship Fuel Consumption Based on LASSO Regression. *Transportation Research Part D: Transport and Environment* 65 (2017) 817-824.

[32] L. Zhang, K. Li, Forward and Backward Least Angle Regression for Nonlinear. *Automatica* 53 (2015) 94-102.

[33] X. Zhou, X. Huang, Reliability Analysis of Slopes Using UD-Based Response Surface Methods Combined with LASSO. *Engineering Geology* 233(31) (2018) 111-123.