The LASSO (Least Absolute Shrinkage and Selection Operator) Method to Predict Indonesian Foreign Exchange Deposit Data

Trisha Magdalena Adelheid Januaviani
Master Program in Mathematics, Faculty of Mathematics and Natural Sciences
Universitas Padjadjaran
Bandung, Indonesia
trishadelheid@gmail.com

Nurul Gusriani, Khafsah Joebaedi, and Sukono
Department Program in Mathematics, Faculty of Mathematics and Natural Sciences
Universitas Padjadjaran
Bandung, Indonesia
gusriani99@gmail.com, khafsah.jbd@gmail.com, sukono@unpad.ac.id

Subiyanto
Department of Marine Science, Faculty of Fishery and Marine Science
Universitas Padjadjaran,
Bandung, Indonesia
subiyanto@unpad.ac.id

Abdul Talib Bon
Department of Production and Operations,
University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

Multicollinearity is the condition that there is a correlation between independent variables which is a problem. This event often occurs in regression analysis. LASSO (Least Absolute Shrinkage and Selection Operator) method regression can reduce multicollinearity and increase the accuracy of linear regression models. The lasso parameter estimator can be solved by the LARS (Least Angle Regression and Shrinkage) algorithm which calculates the correlation vector, the largest absolute correlation value, equiangular vector, inner product vector, and determines the LARS algorithm limiter for LASSO. LASSO method regression with a more detailed procedure and selecting the best model using the C_p Mallows statistics is discussed in this paper. LASSO method will be applied to Indonesia's foreign exchange deposit data.

Keywords
LASSO, LARS, Cp Mallows and Multicollinearity

1. Introduction

Correlation in regression analysis is correlation with the aim to predict the value of non-independent variables based on the independent variables. In fact, a 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. This independent variable that correlates with each other is called multicollinearity (Zhou and Huang, 2018). The LASSO (Least Absolute Shrinkage and Selection Operator) method regression can help to reduce multicollinearity and increase the accuracy of linear regression models (Sermpinis et al., 2018). 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).

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 algorithm in 2004. The solution paths of LAR are piecewise linear and thus can be computed very efficiently (Lee and Jun, 2017). Modification of LAR (Least Angle Regression) for LASSO produces a more efficient algorithm for estimating the LASSO coefficient estimator solution. 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 (Hastie et al., 2004). This paper will discuss the LASSO method (Least Absolute Shrinkage and Selection Operator) regression with a more detailed explanation and choose the best model using the Mallow's C_p statistics. This method will then be applied Indonesian foreign exchange deposit data that is affected by net foreign assets, net domestic assets, net bills to the central government and bills to other sectors. This model is expected to predict Indonesian foreign exchange based on factors.

2. Methods and Materials

2.1 Multiple Linear Regression Analysis

Correlation of two or more variables to non-independent variables, the regression model which is a multiple linear regression model (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 models :

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

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 coefficient of determination (R^2) to measure how far is the variation of the non-independent variable. Value coefficient of determination is between 0 (zero) to 1 (one). The coefficient of determination (R^2) defined as follows (Arthur et al., 2018):

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

2.4 Data Standardization

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

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

2.5 Multicollinearity

Multicollinearity is considered a weakness (black sign) in regression analysis because the parameter estimator is not BLUE (Best Linear Unbiased Estimator) and has a big error. Variance Inflation Factor (VIF) values are less than 10, so there is no multicollinearity (Alauddin and Nghiemb, 2010).

$$\text{VIF}_j = \frac{1}{TOL_j} = \frac{1}{1-R_j^2} \text{ where } TOL_j = 1 - R_j^2 \tag{5}$$

2.6 Mallow's C_p Statistic

Best Subset Regression starts with choosing the simplest model, a model with one variable. Furthermore, by entering other variables one by one until a model that meets the best criteria is obtained. Mallow's C_p is one way to evaluate the selection of the best models in best subset regrression (Miyashiro and Takano, 2015). Best models have small Mallow's C_p statistics value (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 Modification of LAR to LASSO

Least Angle Regression (LAR) is a variable selection method whose algorithm can be modified to obtain a LASSO solution (Hastie et al., 2004). Modification of LAR to LASSO produces an efficient algorithm for estimating computational LASSO parameters faster than quadratic programming. Modification of LAR to LASSO is called LARS or Least Angle Regression and Shrinkage. 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 $\hat{\mu}_A^{(i)}$ 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 $n \times p$, 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
$$\hat{C}^{(i)} = max\{|\hat{c}_j^{(i)}|\}, \text{ so that } A = \{j \mid |\hat{c}_j^{(i)}| | \hat{C}^{(i)}|\} \tag{7}$$

c. Calculating equiangular vector ($\mathbf{u}_A^{(i)}$). Defined $\mathbf{X}_A^{(i)} = [\dots s_j X_j^* \dots]_{j \in A}$, $s_j = sign\{\hat{c}_j^{(i)}\}, j \in A$
$$\boldsymbol{\omega}_A^{(i)} = \mathbf{P}_A^{(i)} \mathbf{G}_A^{(1)^{-1}} \mathbf{1}_A \text{ while } \mathbf{G}_A^{(i)} = \mathbf{X}_A^{(i)^t} \mathbf{X}_A^{(i)} \text{ and } \mathbf{P}_A^{(i)} \mathbf{1}_A^T \mathbf{G}_A^{(i)^{-1}} \mathbf{1}_A)^{-\frac{1}{2}}$$
So that the equiangular vector value is obtained $\mathbf{u}_A^{(i)} = \mathbf{X}_A^{(i)} \boldsymbol{\omega}_A^{(i)}$ \hfill (8)

d. Defined the inner product vector $\mathbf{a}^{(i)} \equiv \mathbf{X}^{*^t} \mathbf{u}_A^{(i)}$
so that $\hat{\gamma}$ can be obtained by the following equation :
$$\hat{\boldsymbol{\gamma}}^{(i)} = \min_{j \in A^c}^{+} \left\{ \frac{\hat{C}^{(i)} - \hat{c}_j^{(i)}}{P_A^{(i)} - a_j^{(i)}}, \frac{\hat{C}^{(i)} + \hat{c}_j^{(i)}}{P_A^{(i)} + a_j^{(i)}} \right\} \tag{9}$$

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: $\varphi^{(i)} = \underset{\gamma_j^{(i)}>0}{min} \{\gamma_j^{(i)}\}$, 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

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

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

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

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

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

j. Look for the best model with the smallest C_p value.

2.8 Indonesian Foreign Exchange Deposit Data

Least Absolute Shrinkage and Selection Operator method regression to overcome multicollinearity problems in Indonesian Foreign Exchange Deposit data. The data used are independent variables and non independent variable. The non-independent variable is foreign exchange deposit (Y), and the independent variables have four variables, namely net foreign assets (X_1), net domestic assets (X_2), net bills to the central government (X_3), and bills to other sectors (X_4).

3. Result

3.1 Ordinary Least Squares Method Parameter and Multicollinearity

Based on equation (2), the ordinary least squares meethod parameter is obtained from Indonesian foreign exchange data which has been transformed using equation (4) for X^* and $Y^* = Y - \overline{Y}$, which is shown in Table 1:

Table 1. OLS Parameter Value

	$\hat{\beta}_{OLS}$
$\hat{\beta}_{OLS_0}$	-2.21517E-14
$\hat{\beta}_{OLS_1}$	-3.725070268
$\hat{\beta}_{OLS_2}$	-25.33779207
$\hat{\beta}_{OLS_3}$	5.07550663
$\hat{\beta}_{OLS_4}$	32.39543664

Multicollinearity shows the correlation between independent variables. indicators for knowing that there is multicollinearity in the data can use Variance Inflation Factor. If $VIF \geq 10$, there is a multicollinearity problem in Indonesian foreign exchange data. The VIF value in equation (5) using the SPSS Statistics 17.0 program can be seen in the following Table 2:

Table 2. Variance Inflation Factor Value

Model	Collinearity Statistics	
	Tolerance	VIF
X_1	0.121	8.248
X_2	0.002	419.722
X_3	0.070	14.224
X_4	0.003	369.514

According to Table 2 above can be seen the independent variables X_1, X_2 and X_3 have $VIF \geq 10$ values which means that there are multicollinearity problems between independent variables.

3.2 Analysis of Indonesian Foreign Exchange Deposit Data with LASSO Method

Calculation of LASSO method for data is done using R Studio software. So that the LASSO coefficient estimator can be seen in Table 3 below:

Table 3. LASSO Parameters

	$\widehat{\beta}_{LASSO}$			
i	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$
1	0	0	0	0
2	0	2.436433394	0	0
3	0	4.083652931	0	1.647219537
4	0	0	1.026443672	5.404790046
5	0	0	1.949173288	6.327519662
6	-1.807590278	0	1.332465393	8.620270647
7	-3.725070268	-25.33779207	5.07550663	32.39543664

Based on Table 3 it can be seen in the iteration of $i = 4$ variable X_2 shrinks to 0 because it violates the LARS algorithm limiter on LASSO. The $\widehat{\beta}_{LASSO}$ value used transformed data, the value of $\widehat{\beta}_{LASSO}$ is returned to the actual data with equation (12). The value of $\widehat{\beta}^*_{LASSO}$ in Table 4 is as follows:

Table 4. Real LASSO Parameters

	$\widehat{\beta}^*_{LASSO}$			
i	$\widehat{\beta}^*_1$	$\widehat{\beta}^*_2$	$\widehat{\beta}^*_3$	$\widehat{\beta}^*_4$
1	0	0	0	0
2	0	0.013498285	0	0
3	0	0.022624182	0	0.000674437
4	0	0	0.021035528	0.002212935
5	0	0	0.039945581	0.002590737
6	-0.037538261	0	0.027307015	0.00352948
7	-0.077358603	-0.14037599	0.104015412	0.013263973

Table 4 is the coefficient value of the LASS parameter using the RStudio program. Next is choosing the best model using the Mallow's C_p statistic in equation (6) presented in Table 5 as follows:

Table 5. Mallow's C_p Value

i	Cp
1	372.7238179
2	192.6608913
3	49.84910785
4	33.56544442
5	13.16500572
6	11.32731788
7	5

Based on Table 5 the model that will be used is a model with the smallest Mallow's C_p value that is found in the iteration to i $=$ 7.

3.3 Model Regresi dari Metode LASSO Regression

Based on Table 4 and Table 5 can be obtained the model of LASSO method as follows:

$$\hat{Y} = -0.077358603X_1 - 0.14037599X_2 + 0.104015412X_3 + 0.013263973X_4$$

The meaning of the model above is if net foreign assets increase by one trillion, the value of Indonesia's foreign exchange deposits will decrease by 0.077358603 or 7.7358603%. As with net domestic assets, if there is an increase of one trillion, the value of Indonesia's foreign exchange deposits will decrease by 0.14037599 or 1.4037599%. Different from the net bill to the central government, if it experiences an increase of one trillion, the value of Indonesia's foreign exchange deposit will increase by 0.104015412 or 1.04015412%. While bills to other sectors, if there is an increase of one trillion, the value of Indonesia's foreign exchange deposits will increase by 0.013263973 or 1.3263973%.

4. Conclusions

In this paper we have analyzed the LASSO method parameter can be specified with LARS algorithm which calculates correlation vector, the largest absolute correlation value, equiangular vector, inner product vector, and determines the LARS algorithm limiter for LASSO. The model of LASSO method as follows:

$$\hat{Y} = -0.077358603X_1 - 0.14037599X_2 + 0.104015412X_3 + 0.013263973X_4$$

Base on model if net foreign assets increase by one trillion, the value of Indonesia's foreign exchange deposits will decrease by 0.077358603 or 7.7358603%. As with net domestic assets, if there is an increase of one trillion, the value of Indonesia's foreign exchange deposits will decrease by 0.14037599 or 1.4037599%. Different from the net bill to the central government, if it experiences an increase of one trillion, the value of Indonesia's foreign exchange deposit will increase by 0.104015412 or 1.04015412%. While bills to other sectors, if there is an increase of one trillion, the value of Indonesia's foreign exchange deposits will increase by 0.013263973 or 1.3263973%.

Acknowledgements

References

Alauddin, Mohammad, and Hong Son Nghiemb. 2010. "Do Instructional Attributes Pose Multicollinearity Problems ? An Empirical Exploration." Economic Analysis and Policy 40(3): 351–61.

Arthur, David Ebuka et al. 2018. "Journal of King Saud University – Science Activity and Toxicity Modelling of Some NCI Selected Compounds against Leukemia P388ADR Cell Line Using Genetic Algorithm-Multiple Linear Regressions." *Journal of King Saud University - Science*.

Chen, Hongmei, and Yaoxin Xiang. 2017. "ScienceDirect The Study of Credit Scoring Model Based on Group Lasso." *Procedia Computer Science* 122: 677–84.

Dyar, M D et al. 2012. "Spectrochimica Acta Part B 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: 51–67

Gauthier, Philippe-aubert, William Scullion, and Alain Berry. 2017. "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: 134–53.

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

Jansen, Maarten. 2015. "Journal of Statistical Planning and Inference Generalized Cross Validation in Variable Selection with and without Shrinkage." *Journal of Statistical Planning and Inference*

Kazemi, A, A Mohamed, H Shareef, and H Zayandehroodi. 2013. "Electrical Power and Energy Systems Optimal Power Quality Monitor Placement Using Genetic Algorithm and Mallow ' s Cp." *International Journal of Electrical Power and Energy Systems* 53: 564–75.

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

Lee, Sudong, and Chi-hyuck Jun. 2017. "PT US CR." *Expert Systems With Applications*.

Lorchirachoonkul, Vichit, and Jirawan Jitthavech. 2012. "Journal of Statistical Planning and Inference A Modified C p Statistic in a System-of-Equations Model." *Journal of Statistical Planning and Inference* 142(8): 2386–94.

Miyashiro, Ryuhei, and Yuichi Takano. 2015. "Expert Systems with Applications Subset Selection by Mallow's Cp : A Mixed Integer Programming Approach." EXPERT SYSTEMS WITH APPLICATIONS 42(1): 325–31.

Permai, Syarifah Diana, and Heruna Tanty. 2018. "ScienceDirect ScienceDirect 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: 671–77.

Sermpinis, Georgios, Serafeim Tsoukas, and Ping Zhang. 2018. "Modelling Market Implied Ratings Using LASSO Variable Selection Techniques." *Journal of Empirical Finance* 48(May): 19–35.

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

Wang, Shengzheng et al. 2017. "Predicting Ship Fuel Consumption Based on LASSO Regression." *Transportation Research Part D.*

Zhang, Long, and Kang Li. 2015. "Automatica Forward and Backward Least Angle Regression for Nonlinear Zhou, Xiao-ping, and Xiao-cheng Huang. 2018. "Reliability Analysis of Slopes Using UD-Based Response Surface Methods Combined with LASSO." *Engineering Geology* 233(June 2017): 111–23.System." *Automatica* 53: 94–102.

Biographies

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.

Sukono is a lecturer in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently serves as Head of Master's Program in Mathematics, the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

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.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.
.