International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 22(1), 2022, pp 213-218
© 2022 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

A Note on the LASSO Model for Predicting Vital Capacity

Md. Abu Shahin[1], Mohammad Zulficar Ali[2], Md. Abdul Khalek[3] and Md. Ayub Ali[4*]

[1]Agrani School & College, RUET Campus, Rajshahi 6204,
E-mail: shahin.hera10@gmail.com

[2]Department of Statistics, Patuakhali Science and Technology University,
Patuakhali 8602, Bangladesh, E-mail: zulficar_bd@yahoo.com

[3]Department of Statistics, University of Rajshahi, Rajshahi 6205,
E-mail: makstat09@gmail.com

[4]Department of Statistics, University of Rajshahi, Rajshahi 6205,
Email: ayubali67@gmail.com

[*]Correspondence should be addressed to Md. Ayub Ali
(Email: ayubali67@gmail.com)

[Received November 12, 2021; Accepted March 7, 2022]

Abstract

The purpose of the present study was to theoretical implementation of LASSO model for predicting vital capacity. There were many methods used for predicting and policy implicating of vital capacity. On the basis of accuracy of estimating and predicting vital capacity, the LASSO model was one for first choice. Therefore, LASSO model can be used for estimating and predicting vital capacity on the basis of its own related function in the world.

Keyword: Vital Capacity, Multicollinearity, LASSO Model.

AMS Subject Classification: 62F15, 62GO5,

1. Introduction

Lung is an important organ, not only for human being but also for animals. The capacity of lung is actually important for stability. There are several factors directly or indirectly related to lung. For example, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Pulse Rate (PR), Weight (Wt), Stature (St), Sitting Height (SHt), Oxygen Saturation (SpO_2), Continuity of Care Record (CCR) and so on. Several authors were investigate and predict the lung functions (Ahmadial et al., 2006; Chatterjee et al., 2011; Emerson and Green, 1921; and Lanza et al**.,** 2015). Lung characteristic assessments had been used to assess how nicely the lungs work, through figuring out how plenty air the lung can maintain and the way quick air actions inside and out of the lungs (Mohammad et al., 2015). Activities in lung in fitness is probably prejudiced through quite a number of things like systolic and diastolic blood pressure, oxygen saturation, pulse rate, stature, weight, sitting height, chest circumference and so on (Ali et al., 2018a and 2018b). Many researchers have investigated that for attaining uncommon event(s), you'll use the predicting equations the usage of without difficulty reachable variables (Ali and Ohtsuki, 2001; Ali et al. 2004; Rahman et al. 2004; Shahin et al. 2013, Ali et al., 2018a and 2018b). For example, to calculate vital capacity (VC), the quantity of air breathed out after the private inhalation, is steeply-priced in Bangladesh like approximately BD Tk1000 in step with test. Also, for measuring vital capacity the device Chestgraph HI-one hundred and one spirometer wishes a few anticipated equations of VC the ones had been advanced from special statistics base of various countries. While, those equations have to be built on the premise of its personal statistics base. These equations mainly developed vital capacity on the basis of age. No multicollinearity problem arises in this situation because there is only one regressor variable. But it is very important to consider other lung related variables for predicting vital capacity, more accurately. The problems is created when we consider others variables, because the multicollinearity problem arises in the regressors. Due to the problem of multicollinearity, the parameters of the equations cannot be estimated more precisely. In that case, we consider several types of biased estimators. Among them, advanced forms of regression analysis are LASSO (Tibshirani, 1996) and Ridge regression that can handle multicollinearity. Hoerl et al. (1975) studied the property of Ridge regression model using some simulation study. But, the ridge regression penalty ($\sum \beta_j^2$), although it helps with obtaining low values of the coefficients attached to the regressors, has two big shortcomings in this setting: (1) Heavy bias toward zero for large regression coefficients and (2) Interpretability: unimportant coefficients may be shrunken towards zero, but they're still in the model. For these reasons, this article suggests that the LASSO regression analysis is used for predicting vital capacity on the basis of precision, instead of ridge regression analysis.

2. Method

Consider the usual linear regression model with data (y_i, x_{ij}), i =1, 2, 3, …, n and j = 1, 2, 3, …, p, where y_i is the response variable of the i th observation and all other variables are regressors. The Ordinary Least Squares (OLS) regression method finds the linear combination of the x_{ij}, j = 1, 2, 3, …p and i =1,…,n that minimizes the residual sum of squares. However, if number of regressors is large or the regression coefficients are highly correlated (multicollinearity problem), the OLS may yield estimates with large variance which reduces the accuracy of the prediction. A widely-known method to solve this problem is the ridge regression and the method of selecting subset. As an alternative to these techniques, Tibshirani (1996) presented "LASSO" which minimized the residual sum of squares subject to the sum of absolute values of the coefficient being less than a constant.

$$\hat{\beta}^L = \arg\min\left[\sum_{i=1}^{n}\left\{y_i - \left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right\}^2\right]$$

subject to

$$\sum_{j=1}^{p}\left|\hat{\beta}_j^L\right| \le t \quad (\text{constant})$$

Let $\hat{\beta}_j^o$ be the full least squares estimate. If $t > \sum_{j=1}^{p}\left|\hat{\beta}_j^o\right|$, then the LASSO algorithm will yield the same estimate as OLS estimate. However, if $0 < t < \sum_{j=1}^{p}\left|\hat{\beta}_j^o\right|$, then the problem is equivalent to

$$\hat{\beta}^L = \arg\min\left[\sum_{i=1}^{n}\left\{y_i - \left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right\}^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]$$

$\lambda > 0$, the relationship between λ and LASSO parameter t is one-to-one. Due to the nature of the constraint, LASSO tends to produce some coefficients to be exactly zero. Compared to the OLS, whose predicted coefficient $\hat{\beta}^0$ is an unbiased estimator of β, both ridge regression and LASSO sacrifice a little bias to reduce the variance of the predicted values and improve the overall prediction accuracy.

The tuning parameter $\sum_{j=1}^{p}\left|\hat{\beta}_j^L\right| = t$ is called LASSO parameter, which is also recognized as the absolute bound. Here we define another parameter, s, as the relative bound.

$$s = \frac{\sum_{i=1}^{p}\left|\hat{\beta}_j^L\right|}{\sum_{i=1}^{p}\left|\hat{\beta}_j^o\right|}, \quad s \in [0, 1]$$

The relative bound can be seen as a normalized version of LASSO parameter. There are two algorithms mentioned in Tibshirani (1996) to compute the best *s:* (i) *n*-fold cross-validation and (ii) generalized cross-validation (GCV).

Cross-validation is a general procedure that can be applied to estimate tuning parameters in a wide variety of problems. The bias in RSS is a result of using the same data for model fitting and model evaluation. Cross validity can reduce the bias of RSS by splitting the whole data into two subsamples: a training (calibration) sample for model fitting and a test (validation) sample for model evaluation. The idea behind the cross-validation is to recycle data by switching the roles of training and test samples.

The optimal s can be denoted by $\hat{s}$. Prediction error can be estimated for the LASSO procedure by ten-fold cross-validation (Tibshirani, 1996). The LASSO is indexed in terms of s, and the prediction error is estimated over a grid of values of s from 0 to 1 inclusive. We wish to predict with small variance, thus we wish to choose the constraint s as small as we can. The value $\hat{s}$ which achieves the minimum predicted error is selected (Tibshirani, 1996).

3. Conclusion

Regression is a well-known technique that is used for estimating and predicting data for particular purpose. This LASSO regression model helps in medical statistics for predicting vital capacity with greatest precisions. This article suggested that the LASSO regression analysis must be used for predicting vital capacity on the basis of precision, instead of ridge regression analysis.

Reference

[1] Ahmadial, N., Khamnei, S., Abedinzadeh, M., Najafi, H. and Mohammadi, M. (2006). Lung `function reference values in Iranian adolescents, Eastern Mediterranean Health Journal, 12(6), pp-834-839.

[2] Ali, M.A., and Ohtsuki, F. (2001). Prediction of adult stature for Japaness population: A stepwise regression approach, Ameican Jounal o Human Biology, 13, pp-316-322.

[3] Ali, M.A., Rahman, J. A. M. S., Ashizawa, K. and Ohtsuki, F (2004). Stepwise regression for predicting final stature of Japanese children. International Journal of Statistical Sciences, 3: 269–280.

[4] Ali, M.Z., Khalek, M. A. and Ali, M. A. (2018a). Lung variables in the students of Rajshahi University, International Journal of Advanced Research, 6(5), pp-901-912.

[5] Ali, M.Z., Khalek, M. A. and Ali, M. A. (2018b). Predicting equations for measuring vital capacity of the Rajshahi University students, International Journal of Advanced Research, 6(5), pp-1108-1115.

[6] Chatterjee, P., Banerjee, A. K, Das, P. (2011). A prediction equation for the estimation of vital capacity in nepalese young females, Journal Of Human Sport & Exercise, 6(1).

[7] Emerson, P. W. and Green, H. (1921). Vital capacity of lungs of Children, American Journal of Diseases of Children, 22, pp-20-22.

[8] Hoerl A. E, Kennard R. W. and Baldwin, K. F. (1975). Ridge regression: some simulation. Communications in Statistics, 4:105–123.

[9] Lanza, F. C., Santos, M.L.d.M, Selman, J. P. R., Silva, J. C., Marcolin, N.,Santos, J., Oliveira, C. M. G. Lago, P. D., and Corso, S. D. (2015). Reference Equation for Respiratory Pressures in Pediatric Population: A Multicenter Study, PLoS One, 10(8), PP-1-9.

[10] Mohammad, J., Maiwada, S. A. and Sumaila, F. G. (2015). Relationship between anthropometric variables and lung function parameters among primary school children, Annals of Nigerian Medicine, 9(1), pp-20-25.

[11] Rahman, J. A. M. S, Ali, M. A., Ashizawa, K. and Ohtsuki, F. (2004). Prediction of adult stature for Japanese population: an improvement of Ali-Ohtsuki equations. Anthropological science, 112, pp-61–66.

[12] Shahin, M. A., Ali, M. A. and Ali, A. B. M. S. (2013). An Extension of Generalized Triphasic Logistic Human Growth Model. Journal of Biometrics & Biostatistics 4(2): 162. doi:10.4172/2155-6180.1000162

[13] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B (Methodological), 58(1):267–288.