Muthanna Journal of Administrative and Economics Sciences
مجلة المثنى للعلوم الادارية والاقتصادية

The selection of the most important factors affecting the infected with cholera using generalized Bayesian lasso regression

Meshal Harbi Odah*
* Al-Muthanna University/ College of Administration and Economic.

Abstract

Regression analysis is considered an important statistical tool in analyzing data that requires finding a statistical method it studies the relationship between the response variable and a group of explanatory variables, as it helps in estimating the regression equation to understand the relationship between variables, in addition to preparing statistical models, including the multiple linear regression model It is an essential tool for researchers in all fields, including the health field, to clarify the risks resulting from some diseases Detecting it helps prevent the development of the disease and treat the disease early through the necessary medical care To prevent deterioration of the condition. The successful researcher must choose the correct model and interpret the results accurately.

The Lasso regression model is one of the most important statistical topics for selecting explanatory variables by organizing and reducing important variables Also, choosing the wrong regression model to analyze and estimate the data will result in inconsistent estimates of the parameters of the model under study Which leads to unhelpful estimates, and therefore the explanatory variables included in the regression model do not explain the variables occurring in a clear and correct way. Therefore, Bayesian generalized Lasso regression it was used to obtain consistent and unbiased estimates and choose the most important explanatory variables that affect cholera disease. The issue of selecting variables in medical cases helps to determine the most important factors affecting the incidence of the disease, as well as ease of communication between the doctor and patients.

Information

Received: 1/3/2024
Revised: 20/3/2024
Accepted: 1/4/ 2024
Published: 6/7/2024

Keywords:
Variable selection, lasso method, Generalized Bayesian lasso, cholera

1. Introduction

In the field of regression analysis, the researcher must be careful in dealing with, identifying and choosing explanatory variables and the matter may get out of control when the results of the analysis impose a certain reality in the test. That is, choosing models carefully leads to improving the accuracy of the estimated model, as adding a large number of explanatory variables can a situation occurs over fitting. Therefore, the predictive performance of the model adopted for data analysis is poor. Over the past years, many methods and approaches have been developed for selecting variables and testing models, The Bayesian method was used to estimate the parameters of the multiple regression model (Mallick & Yi, 2014).

In addition, the subject of regression analysis is one of the important statistical tools in analyzing data that requires creating a model that studies the relationship between the response variable and a group of explanatory variables. It helps to know, identify and understand the risks that result from some diseases and detect them early, which helps prevent the development of the disease and treat it. This is through medical interventions.

From a statistical standpoint, the problem of the paper lies in solving the problem of misdiagnosis of the regression model, which will lead to finding inconsistent variables, and this affects the accuracy of the estimate and the accuracy of the prediction, so we resort to using Generalized Bayesian lasso regression (Denison & George, 2000)The Bayesian method is considered one of the best types of estimation methods, especially in small sample sizes , In (Kyung etal, 2010) he studied the regression of penalty functions and focused on linear regression multivariate and estimate the parameters of this type of model by assuming the Lasso penalty function according to the Bayesian method. In (Yuan & Lin, 2005) estimate the parameters of the multiple linear regression model according to the Bayesian method and assuming the Lasso penalty method. The aim of the research was to find an efficient method for testing variables according to the Bayesian method for the multiple linear Lasso regression model. Our goal was to identify the most important variables that affect the incidence of cholera.

1. Variable selection: Identifying the variables that best represent a subset of the overall variables is referred to as the variable selection procedure, meaning that the regression equation that describes the relationship between the response variable and the explanatory variables is tested to see if the function that was diagnosed for the regression equation is correct. We notice in many regression analysis applications that it is not specified which variables should be placed in the regression model, so it is the duty of the data analyst to use the variable selection method in order to obtain the best regression model. We also notice in many applied fields that theory in the study population entails selecting important variables in advance, and according to the theoretical basis of the case study. Therefore, in this case we do not need to resort to a method or procedure for selecting variables. However, in some studies and certain scientific fields, there is no basis for theory in determining variables in advance, so we rely on using the method of selecting variables in order to obtain the best estimate for the regression model. That is, the method of selecting variables and diagnosing the regression function are linked to each other. In (Wu & Martin, 2020) regarding the generalized Lasso method, that is, with the learning rate parameter, or in other words, the safe Lasso method, this research presented several proposed methods for choosing the learning rate parameter. These proposed methods aim to overcome the problem of misdiagnosis of the model fitting the data by choosing the best learning parameter that helps in making the prediction using generalized Bayesian.

Thus, I can say that choosing the best subset of explanatory variables there is a standard for each purpose through which the work of the chosen set of variables can be tested and evaluated through a regression equation. It should be noted here that the method of selecting variables may lead to creating what are called over-fitting models. The model estimated for the data may not work on a data set, meaning that the model contains a large number of explanatory variables. Therefore, such a model may explain the changes that occur in the response variable to random errors instead of the effect of the explanatory variables.

2. Lasso method: To get acquainted with the Lasso method as a method for selecting variables, which is one of the penal least squares methods and represents the least shrinkage coefficient, as this method was formulated by the scientist (Tibshirani,1996) and it performs two tasks in data analysis, which are: organization and selection with high accuracy. The Lasso method places restrictions on the sum of the absolute values of the model parameters by reducing the coefficients of the variables or deleting some of them to equal them to zero. In addition, the Lasso method helps increase the ability to interpret the model by deleting variables that do not affect the response (dependent) variable.

When performing the process of determining the parameter of the variables and controlling the strength of the penalty (Penalty), as in the following equations:

$$\text{Min}(y - X\beta)'(y - X\beta)\ sub. \quad \sum_{j=1}^{k} \|\beta\|_1 \quad \dots\dots(1)$$

$$\hat{\beta}_{lasso}(\lambda) = argmin_{\beta}(|y - X\beta| + \lambda \|\beta\|_1) \quad \dots\dots(2)$$

So that

$$|y - X\beta|_2^2 = \sum_{i=0}^{n} \left(y_i - (x\beta)\right)_i^2, \|\beta\|_1 = \sum_{j=1}^{k} \|\beta_j\|, \lambda \geq 0 \quad \dots\dots(3)$$

$$\|\beta\|_r = \left(\sum_{j=1}^{k} |\beta_j^r|\right)^{1/r} \quad \dots\dots(4)$$

Whereas

r=1, we get $\|\beta\|_1$ It is the penalty function of Lasso, to provide zero solution information for the Lasso method.

3. Generalized Bayesian lasso regression & OLS: In order to reach the posterior distribution, the parameters of the regression model must be estimated to be close to the true distribution In order to make the subsequent distribution more consistent with the true distribution, and because consistency requires the provision of large data, working according to the Bayesian method has the ability to deal well and provide good estimates with small sample sizes. (Warahena etal, 2023)To choose the appropriate model, the best model approximating the data is selected based on the prediction performance among different models. According to the Bayesian method, the method was updated by adding a parameter α as in the equation below.

$$\pi(\beta|z^n,\alpha) \propto f(y^n|x^n,\beta)^{\alpha}\pi(\beta) \quad \dots\dots(5)$$

Whereas $\boldsymbol{\beta}$ it is a vector of features to be estimated, x & y It is the observation, $\boldsymbol{\alpha}$ it is a learning parameter.

Based on the definition of the generalized posterior distribution (Grunwald, 2007) until the regression model parameter estimator is represented according to the generalized Bayesian method for the studied parameter the multiple linear regression parameter was also estimated according to the Lasso method, as we explained in Equation (2).

We note from equation (2) it has the objective function and a constraint, and the objective function in it is to minimize the sum of the squares of the residuals according to the penalty function constraint of Lasso represented by $\boldsymbol{\lambda}\|\boldsymbol{\beta}\|_{\mathbf{1}}$ in addition, the equation (2) was assumed that the regression model under study is a multiple linear regression model and is defined in the following formula:

$$y = x\beta + e \quad \dots\dots(6)$$

Since

y: It is the response variable vector, x: Matrix of standardization values of observations, β: The model parameter vector to be estimated, e: A vector of independent errors that follow the same normal distribution.

(Tibshirani, 1996) The regression parameter in the model can be estimated (6) According to the Bayesian method, assuming that the regression parameter β it represents a variable that follows the Laplace distribution as follows:

$$f(\beta|b) = \frac{1}{2}exp\left[-\frac{|\beta|}{b}\right] \quad \dots\dots(7)$$

(Park & Casella, 2008) Bayesian estimation of the multiple regression parameter according to the Lasso method. The prior distribution of the parameter assumed β as follows:

$$\prod(\beta|\sigma^2) = \prod\frac{\lambda}{2\sqrt{\sigma^2}}exp\left\{\frac{-\lambda|\beta|}{\sqrt{\sigma^2}}\right\} \quad \dots\dots(8)$$

The above equation follows the Laplace distribution to obtain a single value to estimate the regression parameter β Use the mixed representation of the Laplace distribution, which consists of mixing the normal distribution with the exponential distribution to work with the mixed distribution.

The equation below will be raised to the force represented by the parameter, which results in the estimation work according to the generalized Bayesian Lasso method.

$$y/x,\beta,\sigma^2 \sim [N(x\beta,\sigma^2 I_n)]^{\alpha}$$

$$w/\lambda \sim \prod\frac{\lambda^2}{\sqrt{2}}w^{2-1}e^{-\lambda w}\,dw$$

$$\sigma^2 \sim \pi(\sigma^2) \quad \dots\dots(9)$$

Estimation by ordinary least squares method (OLS) according to the study variables, the equation will be as follows:

$$\hat{y} = \hat{\beta}_0\sum\hat{\beta}_i\,x_i \quad \dots\dots(10)$$

Table (1) of estimates of regression coefficients, standard error, calculated t value, and probability value at a significant level (0.05)

4. Applied aspect and research results: The data sample was collected from Al-Hussein Teaching Hospital in Samawah, Iraq. The sample size included 60 cases of both sexes. Those infected with cholera were selected for identifying the important factors that led to infection with this disease.
Regression model variables: The model includes the following variables, Dependent variable Cholera It is the bacteria that causes cholera in the human body.

The remaining explanatory variables are described as in the table:

The variables (X1: Contaminated water, X2: Anemia, X3: Low stomach acid, X4: The age, X5: Drinks sold by street sellers, X6: The weight, X7: Some foods, X8: Some fish and seafood, X9: Residence area, X10: Chronic diseases, X11: Blood group (O).

Estimation by ordinary least squares method (OLS) according to the study variables, the Estimate table will be as follows:

Var.	Estimates	S.E	t	P-Value
X1	-0.6151	0.6479	-0.8252	0.4262
X2	2.5378	0.8311	3.0701	0.0028
X3	0.0529	0.0913	0.7715	0.5120
X4	-0.0626	0.0261	-2.3631	0.0200
X5	-0.0187	0.1846	-0.1026	0.9294
X6	0.3619	0.0735	5.6413	0.0000
X7	1.3063	0.3164	4.1579	0.0001
X8	-2.4302	0.6201	-3.8517	0.0002
X9	-0.6961	0.4265	-1.6784	0.0997
X10	1.5441	0.5821	2.6746	0.0090
X11	-0.0827	0.2631	-0.3207	0.7581

We find in table (1) that the variables (X2: Anemia, X4: The age, X6: The weight, X7: Some foods, X8: Some fish and seafood, X10: Chronic diseases). It had a statistically significant value because P-Value Less than level 0.05.

As for the other variables (X1: Contaminated water, X3: Low stomach acid, X5: Drinks sold by street sellers, X9: Residence area, X11: Blood group (O)). The variables were not significant because the P-Value was greater than the level of significance 0.05.

This result does not agree with logic and cannot be contaminated water it is not a cause of cholera as for the other variables, their results appeared to be ineffective and cause cholera (Low stomach acid, Drinks sold by street sellers, Residence area, Blood group (O)).

As can be seen from the results of the least squares method, it gave inaccurate estimates of the regression coefficients, as it did not represent the reality of the studied phenomenon. The results of the method also showed that some variables are not important, but in fact, they are important and cause cholera.

Estimation by Bayesian lasso regression method according to the study variables, the Estimate table will be as follows:

Table (2) of estimates of regression coefficients, standard error, calculated t value, and probability value at a significant level (0.05)

Var.	Estimates	S.E	t	P-Value
X1	-0.246	0.016	-17.467	0.0000
X2	-0.087	0.0147	-6.807	0.0001
X3	-0.148	0.018	-4.853	0.0003
X4	0	-	-	-
X5	-0.066	0.015	-5.136	0.0000
X6	0	-	-	-
X7	0	-	-	-
X8	0.221	0.0264	4.772	0.0002
X9	0.828	0.1047	8.735	0.0000
X10	0	-	-	-
X11	-1.333	0.4021	-3.0441	0.0005

We find in table (2) that the variables (X1: Contaminated water, X2: Anemia, X3: Low stomach acid, X5: Drinks sold by street sellers, X8: Some fish and seafood, X9: Residence area, X11: Blood group (O)). It had a statistically significant value because P-Value Less than level 0.05.

As for the other variables (X4: The age, X6: The weight, X7: Some foods, X10: Chronic diseases). The penalty is imposed on the coefficients (variables), which reduces over fitting and performs variable selection by (shrinking) some coefficients to zero. This ensures that the model remains reliable in selecting the most important influential variables.

Table (3) compares the estimates of the two methods based on Mean square error and probability value:

Table (3) Mean square error and probability value for the two estimation methods

Method	MSE	P-Value
OLS	1.8784	0.0000
B Lasso regression	1.4728	0.0000

We note from the table that the method B Lasso regression had the least mean square error (1.8784), which is less than the error squares for the least squares method (1.4728), which indicates that the method B Lasso regression is better than the least squares method in determining the variables affecting cholera infection.

5. Conclusions: The most important conclusions we have reached will be presented in this paper, which will help researchers in determining the correct model to study a phenomenon and to know the factors affecting it and our study The least squares method was used, and Bayesian lasso regression method it was Bayesian lasso regression the best and most interpretable method, and the selection of variables was easier. The less important explanatory variables were also reduced and the estimates were equated to zero, meaning they have no significant effect. As well as helping medical specialists know the factors affecting the infection with cholera, as it appeared that Contaminated water, Anemia, Low stomach acid, Drinks sold by street sellers, Some fish and seafood, Residence area and Blood group (O) all of them have a direct impact on the infection with cholera, and this was consistent with medical logic, as you find that most studies indicate that the factors mentioned above affect the infection with cholera. We must also be careful that cholera thrives in situations where it is difficult to maintain a healthy and safe environment.

.

Reference

Ann Kirkland , L.. (2014). LASSO Simultaneous shrinkage and selection via L1 norm. A thesis in the Mathematical Statistics, University of Pretoria, Faculty of Natural and Agricultural Sciences.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos.

Denison, D., & George, E. (2000). Bayesian prediction using adaptive ridge estimators. *Dept. Math., Imperial College, London, UK, Tech. Rep*

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Mallick, H., & Yi, N. (2014). A new Bayesian lasso. *Statistics and its interface*, *7*(4), 571.

Chatterjee, S., & Hadi, A. S. (2006). Regression analysis by example. John Wiley & Sons

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the american statistical association*, *103*(482), 681-686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Warahena-Liyanage, G., Famoye, F., & Lee, C. (2023). A Generalization of LASSO Modeling via Bayesian Interpretation. *Austrian Journal of Statistics*, *52*(4), 15-45

Wu, P. S., & Martin, R. (2023). A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, *18*(1), 105-132.

Yuan, M., & Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, *100*(472), 1215-1225.