Addressing the problem of collinearity using regression methods

Sackineh Shamil Jasim, Nibras Talib Mohammed

sackineh.sh@uokerbala.edu.iq

nbrass.t@uokerbala.edu.iq

Department of statistics, University of Kerbala , Kerbala , Iraq.

Abstract: *In the field of statistics, building a model in multiple linear regression can be regarded one of the significant goals. However, when there is a problem of collinearity between explanatory variables in the research data, this situation can lead to inaccurate results in building the general linear regression model and estimating parameters. In order to overcome this problem, a method known as the "Lasso method" was used, and its results were compared with the Ridge regression method to evaluate its efficiency and accuracy in dealing with this type of problem. The results in the analysis of factors affecting thyroid function showed that the Lasso method for estimation and selection of variables also had better results in terms of the mean square error (MSE) criterion, which was less than the Ridge Regression method.*

Keywords: collinearity problem - inflation factor - Lasso method - Ridge regression method

Introduction

The construct of a model in multiple linear regression is considered as one of the crucial goals in the field of statistics, since this requires improving the data set that is used in the model through the process of selecting important variables. The goal is to build a model that adequately explains the response variable. To achieve this, the researcher must carefully consider the selection of variables. If there are multiple linear relationships between explanatory variables, it is inappropriate to use traditional methods such as the ordinary least squares method. Instead, advanced statistical methods such as the Lasso method [1] can be used and compared with the known regression method. These methods will be explained theoretically as well as applied to actual data, while presenting the most important recommendations and conclusions reached in this research.

1-Research hypotheses:

This research is based on the following statistical assumptions at the level of significance ($\alpha = 0.05$):

First hypothesis (H1): There is a significant statistical correlation between the level of the GOT enzyme in the blood and the level of thyroid hormone (T3).

Second hypothesis (H2): There is a significant statistical correlation between the level of the enzyme (GPT) in the blood and the level of thyroid hormone (T3).

Third hypothesis (H3): There is a significant statistical correlation between the level of ferritin in the blood and the level of thyroid hormone (T3).

Fourth hypothesis (H4): There is a significant statistical correlation between the level of calcium in the blood and the level of thyroid hormone (T3).

Fifth hypothesis (H5): There is a significant statistical correlation between the blood sugar level and the level of thyroid hormone (T3).

Sixth hypothesis (H6): There is a significant statistical correlation between the level of creatine in the blood and the level of thyroid hormone (T3).

Seventh hypothesis (H7): There is a significant statistical correlation between the level of urea in the blood and the level of thyroid hormone (T3).

Eighth hypothesis (H8): There is a significant statistical correlation between the level of the specific hormone (T3) and the percentage of blood in the body.

2-The problem of collinearity

One of the assumptions that the general linear regression model relies on is the absence of a perfect or almost perfect relationship between the explanatory variables.

One of the assumptions for the estimator of β ̂ for β in the multiple linear regression model for the matrix of observations on the independent variables is perfect rank (m), and one of the conditions in the regression model is the rank condition. In order to estimate the parameters of the model, it is necessary to find the inverse of the Fisher matrix, and this is not done. Unless this matrix is of perfect rank, that is:

rank $(X^{\prime} X)$ = m....(1)

Since:

:$X^{\prime} X$ Fisher degree matrix (m*n) for observations of explanatory variables

Multiple correlation occurs when there is a connection between the independent variables or all variables. This correlation is referred to as collinearity (multiple correlation). In other words, when a very strong linear correlation occurs between two or more variables, it is difficult to disentangle the effect of each variable independently of the response variable.

We say that there is a collinearity problem if there is a correlation between two variables 1 and $x2$.

This problem is one of the most common problems facing the researcher in analyzing data using a statistical method and regression analysis method. Especially in the fields of economic, social and medical sciences.

There are two types of collinearity:

1. **Complete collinearity**: The researcher faces this type of problem when there is a complete correlation between two or more explanatory variables in a linear relationship, and therefore their effect cannot be separated from the response variable, and this is called complete collinearity. Thus, it is not possible to find the inverse of the matrix $(x^{\prime} x)$ because the specific value of the Fisher matrix is equal to zero, and this case is called perfect collinearity.
2. **Semi perfect collinearity:** when there is a strong correlation between two or more independent variables, but it is not perfect, which makes it difficult to isolate their effect on the dependent variable. It occurs when the determinant of the Fisher matrix $(x^{\prime} x)$ is close to zero (very small), and this The condition is called perfect-semi collinearity.

3- Discovering the problem of collinearity using the variance inflation factor test

This test can be used to detect an association between variables or an independent variable that causes this strong association. As the value of the variance inflation factor increases, it indicates a problem with the correlation within the model, which may cause the model to not appear significant. Consequently, the value of the t-test statistic decreases, as this measure depends on the square of the multiple correlation coefficient (Coefficient of Determination), i.e. R2, and is known by the following equation:

VIP=1/(1-R_j^2) j=1,2,…q….(2)

Since:

R_j^2: represents the square of the correlation coefficient between the independent variable and other independent variables in the model

1-R_j^2: This amount is called the area (Tolerance).

It can be noted that the correlation coefficients between the independent variables and the diagonal elements of the inverse matrix represent the variance inflation factors for the independent variables. When the value of the inflation factor is greater than 10, i.e. VIP >10, this indicates the presence of an independent variable xj that is linked in a linear relationship to some or all of the independent variables.

4-The Lasso method to solve the problem of collinearity

It is one of the penal least squares methods and represents a coefficient of least shrinkage, which performs two tasks in data analysis: organization and selection with high accuracy [2]. The Lasso method places restrictions on the sum of the absolute values of the model parameters, and the sum of the parameters must be less than the fixed value (The upper limit) and to achieve this matter[3]. the method of contraction (regularization) is applied and the best variable for the model is chosen by reducing the coefficients of the variables or deleting some of them to equate them to zero [4]. When performing this process, we notice that there are variables that are still non-zero, and the goal of performing this process (shrinkage) is to be partial. of the model is to reduce prediction error.

When performing the process of determining the parameter λ and to control the penalty power (Penalty), which is of great importance when using a value when λ is large, the coefficients are forced to be equal to zero, and in this way the dimensions will be small (low). That is, the parameters of the model are reduced and become equal. For zero. On the other hand, if...

λ=0, so we have a linear regression model with the sum of least squares (OLS).

There is a model mathematical formula used by researchers (Vande Geer & Buhlmann), which defines the use of the Lasso estimator to choose the best model, which is as follows:

Minimize $\left(\frac{\|y - x\beta\|_2^2}{n}\right)$ subject to $\sum_{j=1}^{k} \|\beta\|_1 < S$ ….(3)

whereas :

S: represents the upper limit of the total parameters.

The problem of choosing the best model is equivalent to estimating the parameter and is as follows:

β ̂(λ)=argmin ((‖y -xβ‖ _2^2)/n+λ ||β| |_1) …(4)

Since:

$\|y - x\beta\|_2^2 = \sum_{i=0}^{n} (y_i - x\beta))_i^2$, $||\beta||_1 = \sum_{j=1}^{k} |\beta_j|$, $\lambda \geq 0$ …(5)

As the value of λ increases, the parameters shrink more, and here lies the inverse relationship between λ, the penalty parameter and the highest limit of S. In this way, parameters with value equal to zero are excluded, according to Lasso's theorem. The constraint region in the Lasso model is a corner, meaning that if the initial point is close to the corner, then the parameter $\beta_j$is equal to zero.

5-Ridge Regression method

It is a way to improve the OLS method by adding a small positive quantity, which is a value between zero and one (1≤ J ≤ 0) to the diagonal elements of the coefficient matrix ($x' \ x$) before the inverse of the matrix[5], in order to estimate the parameter β of the regression model as follows:

$$b_{RR} = \begin{bmatrix} b_{1RR} \\ b_{2RR} \\ \vdots \\ b_{kRR} \end{bmatrix} = (\acute{x}x + KI_n)^{-1}\acute{x}y \quad ....(6)$$

Where K is the letter constant value.

When (J = 0), the OLS estimators are the same for estimating the parameter, and when J > 0, that is, its value is greater than zero, after adding a fixed quantity whose value is J, and it tends to be stable at a certain value, but it is biased in relation to the variables present in the data. Also, the mean square error of the letter regression estimators is less than the usual mean square error, that is, ($MSE_{\beta\ RR} < MSE_{\beta\ OLS}$) and when choosing the value of J, the test will be done using the graphical method assumed by the scientist (Hoerl). & kannand)

6- The practical aspect

Data were taken for some patients at Al-Hussein Teaching Hospital in Holy Karbala to analyze the relationship between the factors affecting the function of the thyroid gland (T3) using some regression methods and relying on the Lasso method and the Regression Ridge method. A comparison was made between the methods from Through the standard mean square error (MSE), all results were obtained using the statistical program (SPSS).

The study variables can be explained as follows:

Dependent variable (Yi): represents the level of thyroid hormone in the blood

The explanatory variables are represented by the following factors:

X1: represents the level of the enzyme (GOT) in the blood

X2: represents the level of the GPT enzyme in the blood

X3: represents the level of ferritin in the blood

X4: represents the level of calcium in the blood

X5: represents blood sugar level

X6: represents the level of creatinine in the blood

X7: represents the level of urea in the blood

X8: represents the blood level in the body

7- Data testing

The data was tested based on the variance inflation factor test, as shown in the table below

Table No. (1): shows the test to detect the problem of collinearity in the study variables based on the variance inflation factor (VIF) test.

variables	VIF	**Result**
X1	3.000	<10
X2	2·051	<10
X3	1·004	<10
X4	2·311	<10
X5	1·003	<10
X6	11·203	>10
X7	6·201	<10
X8	3·006	<10

We notice from Table No. (1) that there is an indication of the existence of a problem of double collinearity between the variable (X6) and the rest of the explanatory variables because the value of the (VIF) of the variable

8-Results analysis and interpretation

a) Data analysis with (Method – Lasso)

$$y\,i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_8 x_{i8} + e_i \quad ...(7)$$

Table (2) shows the values of the coefficient of determination$\boldsymbol{R^2}$**and** $\boldsymbol{R^2_{Adj}}$, and the mean square error (MSE) of the model (GIM) using the Lasso method.

Method	R^2	R^2_{adj}	MSE
Lasso	0.991	0.873	0.214

Table (3) shows estimates of the model parameters (GIM) and the standard deviation of the parameters using the Lasso method

Coefficients	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
$\hat{\beta}$	0.120	0.281	0.096	-0.399	-0.467	-0.173	0.672	0.067
$SD(\hat{\beta})$	0.489	0.188	0.017	0.353	0.436	0.298	0.533	0.392

b) Data analysis (Regression Ridge)

Table No. (4) shows the values of the coefficient of determination $\boldsymbol{R^2}$ **and** $\boldsymbol{R^2_{Adj}}$ and the mean square error (MSE) of the model (GIM) using the regression ridge method.

Method	R^2	R^2_{adj}	MSE
Rideg Regression	0.879	0.749	0.578

Table No. (5): shows estimates of the model parameters (GIM) and the standard deviation (SD) of the parameters using the letter regression method.

Coefficients	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
$\hat{\beta}$	0.062	-0.073	-0.082	0.171	-0.360	-0.079	0.193	-0.089
$SD(\hat{\beta})$	0.241	0.071	0.079	0.104	0.201	0.117	0.149	0.082

c) Analyzing the data using the OLS method.

The results of this method are shown in Tables No. (6) and (7) and as follows.

Table No. (6): Shows the values of the coefficient of determination $\boldsymbol{R^2}$ **and** $\boldsymbol{R^2_{Adj}}$ and the mean square error (MSE) of the model (GIM) using the OLS method

Method	R^2	R^2_{adj}	MSE
OLS	0.769	0.501	171.621

Table No. (7): shows estimates of the model parameters (GIM) and the standard deviation (SD) of the parameters using the OLS method**.**

coefficients	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
$\hat{\beta}$	221.743	-0.173	0.159	0.011	0.171	-0.008	-20.052	0.187	-12.39

$SD(\widehat{\boldsymbol{\beta}})$	46.507	0.241	0.421	0.010	0.920	0.062	22.131	0.049	4.815

d) Comparing the results of estimation methods

Table No. (8): shows the values of the mean square error (MSE) for all methods

Method	Lasoo	Ridge regression	OLS	Best
MSE	0.214	0.578	171.621	Lasso

According to the results of the above table, we notice that the Lasso method is superior in estimating the Generalized Linear Regression Model (GIM) compared to the LIS method since it gives the lowest value of the mean square error (MSE).

9 -Conclusions

Through the results obtained, the following conclusions were reached:

1- The results showed that the Lasso method for estimation and selection of variables has clear benefits compared to the Ridge Regression method, as it results in lower values of the regression square error (MSE)...

2- Based on the results of the optimal method (LASSO), the factors that significantly affect thyroid function were identified, which are the percentage of GOT and GPT enzymes in the blood, and the percentage of urea in the blood. The effect of these factors leads to an increase in the level of T3 hormone in the thyroid gland. The results indicate that the following factors have a negative impact on thyroid function, as their increase leads to a decrease in the level of T3 hormone.

3- The results showed that the level of urea in the blood represents the most influential variable on thyroid function compared to other variables, and this is based on the results of the Lasso method.

References

[1] Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of British Surgery*, *105*(10), 1348-1348.

[2] Wang, H., & Wang, G. (2021). Improving random forest algorithm by Lasso method. *Journal of Statistical Computation and Simulation*, *91*(2), 353-367.

[3] Sethi, J. K., & Mittal, M. (2021). An efficient correlation based adaptive LASSO regression method for air quality index prediction. *Earth Science Informatics*, *14*(4), 1777-1786.

[4] Huang, J. C., Tsai, Y. C., Wu, P. Y., Lien, Y. H., Chien, C. Y., Kuo, C. F., ... & Kuo, C. H. (2020). Predictive modeling of blood pressure during hemodialysis: A comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. *Computer methods and programs in biomedicine*, *195*, 105536.

[5] Mohammadi, S. (2022). A test of harmful multicollinearity: A generalized ridge regression approach. *Communications in Statistics-Theory and Methods*, *51*(3), 724-743.