Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 2, June 2020, pp.668-672

Solving multicollinearity problem of gross domestic product using ridge regression method

Ahmed D. Ahmed[1], Ebtisam K. Abdulah[2], Baydaa I. Abdulwahhab[3], Noura O. Abed[4]
[1,2,3,4] Statistics Department, College of Administration & Economics, University of Baghdad

ABSTRACT

This study is dedicated to solving multicollinearity problem for the general linear model by using Ridge regression method. The basic formulation of this method and suggested forms for Ridge parameter is applied to the Gross Domestic Product data in Iraq. This data has normal distribution. The best linear regression model is obtained after solving multicollinearity problem with the suggesting of k_{10} value.

Keywords: Multicollinearity, Ridge regression, General linear model, Gross Domestic Product, Mean square error

Corresponding Author:

Ahmed D. Ahmed
Statistics Department, College of Administration & Economics
University of Baghdad
Address
E-mail: ahmedthieb19@gmail.com

1. Introduction

Countries of the world have witnessed many economic crises, and Iraq is one of these countries whose economy faced many economic crises, including high oil prices, increased foreign investment, and the lack or disappearance of local industries, and other economic activities. The Gross Domestic Product is one of the country's economic performance indicators. Its association with the population and inflation rates shows the efficiency of this country's performance [1]. Many of the statistical techniques have been used to treat and solve like these economic crises. Multiple linear regression model one of the tools has been used in the analysis of data economic, health, social, and others of applied sciences, besides forecasting the local product of the country in the future. This model has many properties depend on certain assumptions which make estimations accurate. During the applications, this model faces a few problems due to existing collinearity among its predictor variables. This problem called multicollinearity which makes the General Least Squares (GLS) method unreliable [2]. To solve this problem, different methods have been found. One way is to increase the size of the sample to reduce the correlation between explanatory variables. Another way is deleting some of the predictor variables that have high correlation or standard conversion of predictor variables. Of course, there are other solutions and one of them is the Ridge regression [3]. Consider the following multiple linear regression model as:

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + U_i \qquad i = 1,2,...,n \quad , \quad j = 0,1,...,p \tag{1}$$

Here, Y is a dependent variable, X is the predictor variables, B is the model parameters, and $U \sim N\left(\mu, \sigma^2\right)$ is a random error.
The GLS estimator for the regression parameters is given by:

$$b_{GLS} = \left(X'X\right)^{-1} X'Y \tag{2}$$

Many researchers addressed multicollinearity problems and tried to solve it with different methods as stated by Goktas and Sevinc [3], Fayose and Ayinde [4], Lukman et al. [5], and others.

This study aims to treat a multicollinearity problem by using Ridge regression methods with different suggestion formulas in order to obtain the best model for estimation of gross product domestic in Iraq.

2. Diagnosing multicollinearity

One of the scales that are used to measure the strength of multicollinearity problem is the variance inflation factor (VIF) [6], which depends on coefficient of determination (R^2) as follows:

$$VIF = \frac{1}{1-R_j^2} \tag{3}$$

So, if the coefficient of variance inflation exceeds (10) [7], there will be a problem with the multicollinearity and we can use the method of condition number formula to measure the effect of multicollinearity.

$$CN = \sqrt{\frac{\lambda_{Max\,eigenvalue}}{\lambda_{Min\,eigenvalue}}} \tag{4}$$

Where,
(λ) is the eigenvalue of the matrix $(X'X)$

So, in the case of the value of $(CN \succ 30)$, this points that there is a presence of influence by the multicollinearity [8]. We can also use matrix of correlation coefficient between the explanatory variables to detect multicollinearity. If the effect is high. Then, this will expose the model to suffer from multicollinearity problem but it is not enough evidence without verifying (VIF) and (CN).
If the effect is low, this does not fully conclude the absence of multicollinearity problem. So examining both methods is necessary to conclude.

2.1. Ridge regression method

It is one of the ways to detect multicollinearity problem for the model. This method is done by adding a certain amount $(k \succ 0)$ to the diagonal elements of matrix, to obtain more precision [9].

This removes any correlations between the independent variables, which was thought by Kennard and Hoerl in 1970 [10]. The formula estimates the variables in the following way.

$$\hat{\beta}_{R.R} = (x'x + kI)^{-1} x'y \tag{5}$$

Where

I is a unit matrix and (k) is a Ridge parameter and there are different ways of finding each one of them.

As for in this research, we will be taking a suggested formula for Hoerl [11] that adapts solving multicollinearity problem.

$$k = \frac{(p-1)\sigma^2}{b'_{Gls} b_{Gls}} \tag{6}$$

p represents the number of variables σ^2 represents the population variance and its estimated from the original data b_{Gls} represents estimated parameters by ordinary least squares.

In this research, a group of suggestions has been made to develop (k) and they are methods derived from method (6). Part of it seems similar to other researchers' researches such as [9]:

$$k_1 = \sqrt{\frac{(p-1)\sigma^2}{b'_{Gls} b_{Gls}}} \tag{7}$$

$$k_2 = \left(\frac{(p-1)\sigma^2}{b'_{Gls} b_{Gls}} \right)^{\frac{1}{p}} \tag{8}$$

$$k_3 = \frac{(p-1)\sigma^2}{Max(b_{Gls})} \tag{9}$$

$$k_4 = Max\left(\frac{(p-1)\sigma^2}{b_{Gls}}\right) \tag{10}$$

$$k_5 = Medain\left(\frac{(p-1)\sigma^2}{b_{Gls}}\right) \tag{11}$$

$$k_6 = \frac{1}{\sqrt{\frac{(p-1)\sigma^2}{b'_{Gls}b_{Gls}}}} \tag{12}$$

$$k_7 = Max\sqrt{\frac{(p-1)\sigma^2}{b_{Gls}}} \tag{13}$$

$$k_8 = Medain\sqrt{\frac{(p-1)\sigma^2}{b_{Gls}}} \tag{14}$$

$$k_9 = Max\left(\frac{1}{\sqrt{\frac{(p-1)\sigma^2}{b_{Gls}}}}\right) \tag{15}$$

$$k_{10} = Median\left(\frac{1}{\sqrt{\frac{(p-1)\sigma^2}{b_{Gls}}}}\right) \tag{16}$$

2.2. Describing of the data

The data represents gross domestic product at current prices in Iraq, which is represented by the variable Y. As for the other effecting variables, they are taken as X for the years 2002-2017 which has measured by a million Iraqi dinars as seen in Table 1 [1].

The data has been tested by using (easy fit) and it's distributed normal distribution with mean 1.5596E+8 and variance 8.2222E+7 and by using SPSS program the coefficient of determination value has been found which equals one. The model is significant according to F test and the value of VIF is greater than 10, which proves that there is a multicollinearity problem between the predictors. This is also clearly seen in the value of condition numbers for the seventh and eighth variable as seen in Table 2.

Table 1. Gross Domestic Product in Iraq for years 2002-2017

Imports of goods and services (X_8)	Exports of goods and services (X_7)	Gross fixed capital formation (X_6)	Government consumer spending (X_5)	Household consumer spending (X_4)	Subsidies (X_3)	Fixed capital consumption allocation (X_2)	Employs compensation (X1)	Gross domestic product at current prices (Y)
20179996.90	28949901.00	2199076.80	7919967.60	9956626.50	7304391.40	5641859.10	3394201.70	41022927.40
22734254.40	22897246.20	3151168.80	3631594.90	13616500.90	9210322.30	3889758.40	3654066.20	29585788.60
34050969.00	29956020.00	2857807.00	13608947.30	19538773.00	15422815.70	6388243.00	8371095.70	53235358.70
45145710.00	39963945.00	10182362.20	14683390.30	27593239.70	20924170.00	8824031.80	10790737.00	73533598.60
36914707.80	48780390.60	16911154.70	14984454.10	35526339.70	16388032.40	11470554.60	16573732.50	95587954.80
31422753.00	51158039.10	7530404.40	20871484.00	42963013.30	18799952.60	13062198.10	21371688.50	111455813.40
48249768.60	79028558.70	23240539.10	26139166.00	49091355.70	29571966.90	17780664.90	34400786.10	157026061.60

Imports of goods and services (X_8)	Exports of goods and services (X_7)	Gross fixed capital formation (X_6)	Government consumer spending (X_5)	Household consumer spending (X_4)	Subsidies (X_3)	Fixed capital consumption allocation (X_2)	Employs compensation (X1)	Gross domestic product at current prices (Y)
51326145.00	51473565.00	13471242.20	27517759.70	68256193.20	23193631.40	13835190.20	35228095.60	130643200.40
55232658.00	63880713.00	26252776.80	30660743.70	72026324.00	26182895.00	17473035.00	41560876.10	162064565.50
60316542.00	96531318.00	28234992.60	36999562.90	77412593.70	34232439.00	24854165.10	45919072.10	217327107.40
73980251.40	113151788.20	38139871.00	42158634.30	101299621.00	39238634.00	28238666.70	59476065.30	254225490.70
75910914.20	108514489.60	55036676.00	47755742.70	106171982.10	33822539.80	29479807.50	70278512.00	273587529.20
69948806.40	102738475.40	55837402.90	47946900.10	111317232.10	35297574.20	28247437.50	71610470.80	266420384.50
68289455.70	67192475.70	50650572.70	36339342.10	107245801.30	12616212.00	19037542.10	68422306.00	199715699.90
49267178.40	56312489.40	39634030.00	45872859.10	110514526.60	16222471.00	18413496.50	76880201.30	203869832.20

Table 2. Values of VIF and condition number for explanatory variables

Explanatory variables	VIF	Eigen value	Condition Number
X_1	184.732	89.689	1
X_2	704.942	8.049	3.338
X_3	20.456	1.015	9.400
X_4	148.063	.821	10.452
X_5	201.484	.235	19.536
X_6	30.349	.137	25.586
X_7	350.583	.053	41.137
X_8	24.592	.002	211.765

3. Analyzing the results of multicollinearity

The multicollinearity problem has been solved by using Ridge regression method and suggested formulas for Ridge parameter k. The comparisons between these formulas, has been accomplished by using the mean square error (MSE) statistical criteria with MATLB program for computation. In addition, the natural logarithms are taken for the data to get rid of some of the common problems such as singular matrix existence. Table 3 depicts the results.

Table 3. Values of Ridge parameter and MSE for the model

Ridge parameter	Ridge parameter value	MSE(y)
k	0.964203896814891	1.0e-003 * 0.816074
k_1	0.981938845761227	1.0e-003 * 0.816093
k_2	0.995453803439111	1.0e-003 * 0.816108
k_3	2.023108857023143	1.0e-003 * 0.816648
k_4	80.932384820246284	1.0e-003 * 0.817159
k_5	20.438944025799106	1.0e-003 * 0.817121
k_6	1.018393359542438	1.0e-003 * 0.816132
k_7	40.959298520326257	1.0e-003 * 0.817147
k_8	6.031370692880903	1.0e-003 * 0.816997
k_9	0.976676135837547	1.0e-003 * 0.816088
k_{10}	0.165799790946417	1.0e-003 * 0.810816

It's seen from Table 3 that the least value of the MSE for the model is based on Ridge parameter. The best linear regression model with Ridge regression is represented as follows:

$$\hat{Y} = -1.0886 - 0.1620X_1 - 0.2897X_2 - 0.5574X_3 + 0.4513X_4 + 0.4605X_5 + 0.0731X_6 + 0.8304X_7 - 0.2068X_8 \quad (17)$$

4. Conclusion

Here in this study, the authors suggested Ridge regression parameter, which has been better than the reported original one and this suggestion was successful in the case of MSE value is minimum at k_{10}. This can be applied to the Gross Domestic Product data in Iraq based on normal distribution. The best linear regression model is obtained after solving multicollinearity problem with the suggested value. On the other hand, the other values k_1 ,k_2,k_3,k_4,k_5,k_6,k_7, k_8 and k_9 were not better than the reported original one, which are not recommended.

References

[1] Ministry of Planning, Central Statistical Organization" Directorate of Publishing and Relations, "*Statistical Indicators on the Economic and Social Situation,* Iraq, 2007-2011.

[2] Al-Sabah S. and Muzhir Z., "Using the Ridge Regression Method in Addressing Linear Multiplicity Problem" *Journal of university of Kerbalah,* vol.16, no. 2, 2018 [in Arabic].

[3] Goktas, A. and Sevinc, V., "Two New Ridge Parameters and A Guide for Selecting an Appropriate Ridge Parameter in Linear Regression", *Journal of Science,* vol.29,no.1,pp.201-211, 2016.

[4] Fayose, T. and Ayinde, K., "Different Forms Biasing Parameter for Generalized Ridge Regression Estimator", *International Journal of Computer Applications,* vol.181, no.37, pp.2 –29, 2019.

[5] Lukman, A., "Some Improved Generalized Ridge Estimators and their comparison", *WSEAS Transactions on Mathematics,* vol.17, no.1, pp.369-376, 2018.

[6] Dereny, M. and Rashwan, N., "Solving Multicollinearity Problem Using Ridge Regression Models", *Int. Journal, Contemp, Mathematics, Sciences,* vol. 6, no.12, pp.585-600,2011.

[7] Khalaf, G. and Lguemane, M., "Multicollinearity and A Ridge Parameter Estimation Approach", *Journal of Modern Applied Statistical Methods*, vol.15, no.2, pp.400-410, 2016.

[8] Montgomery, D. and Peck, E., Introduction to Linear Regression Analysis, John Wiley and Sons, New York, 1992.

[9] Asar, Y. and Genc, A. "On Some New Modifications of Ridge Estimators", *Kuwait Journal of Science,* vol.44, no.3, pp.1-14, 2015.

[10] Hoerl, A. and Kennard, R., "Ridge Regression: Biased Estimation for Nonorthogonal problems", *Technometrics,* vol.12, no.1, pp.55-67, 1970.

[11] Hoerl, A. and Kennard, R., "Ridge Regression: Applications to Nonorthogonal problems", *Technometrics,* vol.12, no.1, pp.69-82, 1970.