©LUMEN Proceedings

Available online at: http://lumenpublishing.com/proceedings/published-volumes/lumen-proceedings/nashs2017/

3rd Central & Eastern European LUMEN International Conference
New Approaches in Social and Humanistic Sciences | NASHS 2017|
Chisinau, Republic of Moldova | June 8-10, 2017

New Approaches in Social and Humanistic Sciences

Ridge Regression for Addressing of the Multicollinearity Problem with Application in Cost of Production

Ali SADIG MOHOMMED BAGER,
Bahr KADHIM MOHAMMED, Meshal HARBI ODAH

https://doi.org/10.18662/lumproc.nashs2017.4

How to cite: Sadig Mohammed Bager, A., Badhim Mohammed, B., & Haribi. Odah, M. (2018). Ridge Regression for Addressing of the Multicollinearity Problem with Application in Cost of Production. In V. Manolachi, C.M. Rus, S. Rusnac (eds.), *New Approaches in Social and Humanistic Sciences* (pp. 56-63). Iasi, Romania: LUMEN Proceedings. https://doi.org/10.18662/lumproc.nashs2017.4

©LUMEN Proceedings

3rd Central & Eastern European LUMEN International Conference
New Approaches in Social and Humanistic Sciences |
NASHS 2017 | Chisinau, Republic of Moldova | June 8-10, 2017

Ridge Regression for Addressing of the Multicollinearity Problem with Application in Cost of Production

Ali Sadig Mohommed BAGER[1],
Bahr KADHIM MOHAMMED[2], Meshal HARBI ODAH[3]

Abstract

The regression analysis is statistical method of extensive use, which illustrates the relationship between the explanatory variables and the dependent variable in the form of a model useful in the interpretation of scientific phenomenon, bringing also benefits to society. In this paper we study the most important factors affecting the cost production of cement (Muthanna Factory) by using the ridge regression. The factors are described as follows: we consider the cost of production amount as response variable and factors that affect or may affect the explanatory variables are labor, Price per ton, Electric power, Quantity consumed. They all suffer from high correlation, indicating a problem of multicollinearity .The data analysis is included in the study of the ridge regression as the best approach in case of a multicollinearity problem in the context of financial and economic data being associated with each other often. We used R packages (MASS).

Keywords: *Ridge regression, Mullticollinearity problem, Production of cement.*

[1] The Bucharest University of Economic Studies, Department of Statistics and Econometrics, Iraq, Muthanna University, nader.ali62@yahoo.com.

[2] The Bucharest University of Economic Studies, Department of Statistics and Econometrics, Iraq, University of AL-Qadisiyah, baherm@yahoo.com.

[3] The Bucharest University of Economic Studies, Department of Statistics and Econometrics, Iraq, Muthanna University, m.algelidh@gmail.com.

https://doi.org/10.18662/lumproc.nashs2017.4
Corresponding Author: Ali Sadig Mohommed BAGER
Selection and peer-review under responsibility of the Organizing Committee of the conference

1. Introduction

The cement is used in many different purposes and consists of many materials for its manufacture at different rates for each type. It needs to be installed in renewable energy sources such as petroleum, solar energy, fossil fuels, coal as alternative fuel and other requirements. It is a simple industry compared to major industries of strategic industries, necessary in the case of availability of raw materials. Cement is the main element in the construction sector and it is one of the necessary requirements in the construction and reconstruction of the infrastructure, so we find a growing demand for cement. Studies have pointed the existence of 24 types of cement produced in the world and in Iraq in particular where there is Portland cement: a dry gray powder with formation and manufacture of calcium silicate, iron, aluminum and silicon. It is called "ordinary" because it is multi-tasking and uses in a variety of important works as it is used for general purposes which represent 80% of the total production. The work began at the Muthanna Cement Factory in 1984 with an actual production capacity of (1948000 tons per year). The energy consumption of electricity (185.2 thousand megawatts), while the actual production capacity in 2013 (448355 tons per year) and the consumed electricity was (69.1thousand megawatts). The actual production capacity of the factory for the period (1984-2013) showed decreased productivity, but higher than the actual production of 1984 without reaching the design capacity of the factory (2,500,000 tons per year). This study seeks to determine the most important factors affecting the cost production of cement (Muthanna Factory) by using Ridge regression.

This paper is organized in 7 sections as follows: In Section1 there is an introduction the topic, in Section2 we illustrate the problem statement, and in Section3 w present the aims of the research. In Section 4, we explain the methodology of multicollinearty and Ridge regression model. In section 5, we analysis the results, followed by discussions in section 6 and conclusion in section 7.

2. Problem Statement

In 1970, Hoerl &Kennard presented a study entitled "addressing multicollinearty by using a ridge regression " [5]. The objective of the study was to address the problem of linear multiplicity by providing regression coefficients whose errors were less than the standard errors of the least squares . Nieuwoudt used ridge regression for a production function of time series data in order to estimate the rate of return in the South African sugar industry. In 2014, Fitrianto &Yik reported that when a linear correlation between the explanatory variables of the multiple linear regression model is high, the variance is higher than the OLS method [3]. The researchers used ridge regression study to compare the performance of the ridge regression estimator and OLS. They found that Hoerl& Kennard's ridge regression method had a better performance than other methods.

The main objective of the study is to find out the issues emerging from significant determinants. The factors that affect the cement cost of production may suffer from the problem of multicollinearity, as is the case in many of the research and economic studies, where data encounters the problem of a linear relationship between the explanatory variables [10]. When there is a problem in the data, it usually means that the estimators method of least squares classic will fail for not achieving one of the basic method (OLS) hypothesis which states that the lack of a linear relationship between the explanatory variables and linear relationship will not get the estimator features characteristic (Best Linear Unbiased Estimator). In order to overcome the problem of linear multiplicity Ridge regression was used, considering it the best approach for multicollinearity in financial and economic data [11].

3. Research Questions/Aims of the research

The aim of this paper is study the most important factors affecting the cost of production cement in Iraq (Muthanna factory),The study focuses on and seeks to answer the question: What are the significant determinants cost of production cement And empirically prove their significance using Ridge regression. When variables experience the mullticolinearity problem this will be addressed in order to obtain more expressive estimators that affect the explanatory variables on the production in the regression equation.

4. Research Methods

Regression analysis is one of the important statistical methods which describes the relationship between variables in the form of a mathematical equation through trough which we can know the direction and strength of the relationship between the variables under study. We can depend on this equation to reach the variables through the availability of appropriate conditions under certain assumptions and on the accuracy of the estimated parameters of the regression equation. The hypotheses depend on this correctness and if the available data cannot support the scientific idea, problems usually arise. The empirical error in the data prevents good measurement as noted in many areas of scientific research. The variables of a study appear to be interconnected and their correlation has an impact on the results of the analysis. The multicollinearity problem is shown when there is an explanatory variable correlated with other explanatory variables or a set of explanatory variables in a linear trend. These Interrelated linear relationships are common in many data or economic information and Business management [1].

Most of the times it is difficult or impossible to isolate the individual effects from the dependent variable. In the case of this multicollinearty problem, the coefficients of the least squares may be statistically non-significant and may offer the wrong signal even though the value of the R-Squared (Coefficient of Determination) is high [9].

In order to overcome the multicollinearity problem, there a set of available methods, such as ridge regression. The philosophy of this method is to find a constant value (K) called a ridge parameter added to the elements of the information matrix ($X'X$). The advantage of this is to reduce the values of the inverted diagonal elements from the information matrix, which reduces the values of the variance of estimated parameter. When the explanatory variables diverge from independence, as is the case when the correlation strength between explanatory variables pairs increases, adding a constant (K) with small values makes a quick change in estimated parameter values. These values begin to stabilize gradually until they reach a point where the change is slight and the signal is fixed [2]. Whenever stability of parameters is faster, the explanatory variables are close to independence .The estimated values are obtained for ridge regression as showed below [6]:

$$\hat{\beta}_R = (X^{*'}X^* + kI_P)^{-1}X^{*'}Y \qquad (1)$$

Where:

$k > 0$: Ridge parameter.

I_P :Identity matrix .

The above represents the relationship between the Ridge regression model and ordinary least square when the value of $(k = 0)$. In this case, the Ridge regression method of the convert to the ordinary least square is the following [8]:

$$\hat{\beta}_{OLS} = (X^{*'}X^*)^{-1}X^{*'}Y \qquad (2)$$

The variance decreases when increasing the value of k, and the value of k should be chosen when it leads to a decrease in the value of the variance and the variance inflation factor (VIF) [4].Then it will be mean square error for ridge regression less than estimator for least square .To estimate the ridge parameter (k) we will follow the method of Hoerl and

Kennard and Baldwin (1975) of introducing new formula to determine ridge parameter optimal as estimator to (K) [6] [7]:

$$K_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}} \qquad (3)$$

Where

$\hat{\alpha}$and$\hat{\sigma}^2$ obtained from the ordinary least squares method .

p : number variables

The sample size was of (120) observation and data was taken from the records of the Muthanna Cement factory where the ridge regression method (RR) is applied to estimate the parameters of the multiple linear regression model over a set of data, where the (Y) response variable represents the monthly cost production of the factory for the period (1/1/2006 to 31/12/2015) and a set of explanatory variables explained as follows:

X_1: Labor.

X_2: Price (Measured in Iraqi dinar).

X_3 : Electric power. (Measured in MW)

X_4::Quantity consumed (Measured in tons).

For Farrah-Galaber , the value ($\chi_c = 546.123$) is greater than the value ($\chi_{table} = 12.59$) , indicating a multicollinearity problem. The K-M-O test is used to determine the sufficiency of the data with the condition that the minimum acceptable score is 0.5 for the sample size to be sufficient. The statistic value of the K-M-O is (0 783), which is more than 0.5, therefore the size of the calculated sample is sufficient.

5. Findings

5.1. Ridge regression analysis

This method is based on the estimation of model parameters when there is a multicollinearity problem between explanatory variables where the ridge regression coefficients are extracted by giving values to (k), thus solving the problem. The method of iteration was used to find the value of ridge parameter in accordance with formula (3) and we have done (50) iterations of this formula and reached ($k = 0.378969$).

Table (1) Represent value to the Variance Inflation Factors (VIF)

X_i	VIF
x1	113.3872

x2	8.3468
x3	6.3443
x4	110.9997

Table (2). Ridge vs. Least Squares Comparison for k = 0.378969

Independent Variable	Ridge Standard Error	O.L.S. Standard Error
x1	0.2944917	0.5341681
x2	1.9207825	4.456298
x3	0.0804115	0.671666
x4	6.062303	7.848474
R-Squared	0.8602	0.9792
Sigma	386398.08	149027.91

Table (3). Estimators Ridge regression

Independent Variable	RegularCoeff 's Ridge	Ridge Standard Error	t-value	Pr> |t|	VIF
Intercept	35337.96				
x1	1.722169	0.2944917	5.8479373	0.001	0.2304
x2	3.061636	1.9207825	1.5939525	0.501	0.3774
x3	0.4923529	0.0804115	6.1229165	0.003	0.4401
x4	10.75155	6.062303	1.7735092	0.341	0.2367

6. Discussions

For the represented value of the Variance Inflation Factors (VIF) :In Table (1). We note that the values of the VIF of the explanatory variables (X_1, X_4) are greater than (10). This means that these variables suffer from inflation in the variance of their parameters. The two variables are the cause of the multicollinearity problem.

Comparison Ridge vs. Least Squares at (k = 0.378969): Table (2) we note from the table the value of the standard error for the Ridge estimate for the variableX_1 $= 0.2944917$, while the standard error value for the Ridge estimate for variable $X_2 =$ 0.5207825 and $X_3 =$ 0.0804115, as for the value of the standard error to estimate Ridge to the variable $X_4 =$

1.2623303. Compared with the standard error values for OLS, estimation for all explanatory variables has been reached respectively: (0.4923529 , 3.061636 , 1.722169 , 10.75155)

It is clear that the standard error values when using the ridge regression estimation method are better and lower than the standard error values when using the OLS estimation method, which means that the ridge regression method reduces and removes the mullticolinearity problem between explanatory variables.

The R-Squared value of the ridge regression method was 0.8602, as for the R-Squared value of OLS method was 0.9792, the ridge regression method was used by calculating the parameters model in ridge regression, and the results were better than OLS method.

Estimators Ridge regression: Table (3).We note the explanatory variables (labor (x1), price (x2), the electric power (x3), the quantity consumed (x4)).The relation between labor and cost production are positive then if the increase is one unit of the labor (X1), leading to an increase (1.722169) in cost production. The relation between the electric power consumed (X3) for the factory and cost of production is also positive, meaning that the increasing electric power is one line leading to an increase in an amount of (4.923529) in cost production. As for the rest of the variables (x2) (x4), the effect on the cost production is weak because it is not significant to the test. The results presented for the above two variables are consistent with the economic logic. We also note that the value of VIF has decreased the values of coefficients of amplification of variance for all four explanatory variables, noting the value (VIF) increases for all explanatory variables.

7. Conclusions

In this study, the main determinants of cost of production were identified with the procedures followed in order to arrive at the best-fit model. Using the ridge regression in cases where the explanatory variables suffer from multicollinearty problem is the best way to solve this problem and also note the decreasing value VIF. Through the results two determinants (Labor, Electric power) were identified to have positive effects for cost of production. Other two determinants (Price, Quantity consumed) have been found to be insignificant. We recommend using the ridge regression method in other studies because the estimator ridge regression of is better than the estimator of the ordinary least square method (OLS) in case the explanatory variables are related .The ridge regression has two important advantages over the linear regression: it penalizes the estimates and it doesn't penalize all the features of the estimate arbitrarily. In some domains, the number of independent variables is high. We are not sure which of the independent variables influences the dependent variable. In this kind of scenario, ridge regression plays a better role than linear regression. This study has also proved that most of the literature on cost of production applies adequately to factories in Iraq. The identified results can be used for further development of cement factories production.

References

[1] Al-Jubouri & Habib C. "Multiple Regression and Analysis of Differentiation", (translated), Iraq: Higher Education Printing Press. 1990

[2] Dorugade, A.V. New ridge parameters for ridge regression. Journal of the Association of Arab Universities for Basic and Applied Sciences. 2014 (15). pp. 94-99

[3] Anwar F. & Lee Ceng Y. "Performance of Ridge Regression Estimator Method on Small Sample size By Varying correlation coefficients: A simulation study", Journal of Mathematics and statistics 10 (1). 2014. pp. 25 – 29

[4] García C.B., García, J., López Martín, M.M., & Salmerón, R.. Collinearity: Revisiting the variance inflation factor in ridge regression. Journal of Applied Statistics, 42(3), 2015. pp. 648-661

[5] Hoer, A.E, & Kennard R.W. Ridge Regression. Advances, Algorithms and Applications 1981: American Sciences Press. 1980

[6] Hoerl A.E., Kannard R.W., & Baldwin K.F. Ridge regression: some esimulations. Communications in Statistics-Theory and Methods. 4 (2). 1975. pp. 105-123

[7] Hoerl A.E., & R W. Kennard. "Ridge regression: Iterative estimation of the biasing parameter," Commun. Statist.A5. 1976. pp. 77-88

[8] Kibria, B.G. Performance of some new ridge regression estimators. Communications in Statistics-Simulation and Computation. 2003. 32 (2). pp. 419-435

[9] Kraha A., Turner H., Nimon K., Zientek L. R., & Henson R. K. Tools to support interpreting multiple regression in the face of multicollinearity. Frontiers in psychology. 3. 2012

[10] Kazem, Armory H. & Shalibah, M.B. "Advanced Economic Measurement Theory and Practice", Baghdad: Dunia al-Amal Library. 2002

[11] McDonald G.C. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1). pp. 93-100. 2009