Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Using ridge regression to analysis the meteorological data in sulaimani

Layla. A.Ahmed

Department of Mathematics, College of Education, University of Garmin, Kurdistan Region- Iraq

Article information

Article history:
Received may 15, 2020
Accepted June 21, 2020
Available online December 1, 2020

Keywords:
Ridge Regression,
Ordinary Least Squares,
Multicollinearity Problem,
Meteorological Variables

Correspondence:
Layla. A.Ahmed
laylaaziz1974@gmail.com

Abstract

Linear regression is one of the frequently used statistical methods that have applications in all field of daily life. In a statistical perspective, the regression analysis is used for studying the relationship between a dependent variable and a set of independent variables. The ridge regression is the most widely model in solving the multicolinearity problem, and it's an alternative to OLS. Multicollinearity is the most common problem in multiple regression models in which there exists a perfect relationship between two explanatory variables or more in the model. In this study, ridge regression model was used to estimate linear regression model. This result was compared with result obtained using ordinary least squares model in order to find the best regression model. We have used meteorological data in this study. The results showed that the ridge regression method can be used to resolve the multicollinearity problem, without deleting the independent correlated variables of the model and able to estimate parameters with lower standard error values.

DOI: https://doi.org/10.33899/iqjoss.2020.167390 , ©Authors, 2020, College of Education, University of Garmin, Kurdistan Region- Iraq

1. Introduction

Linear regression is one of the frequently used statistical methods that have applications in all field of daily life. In a statistical perspective, the regression analysis is used for studying the dependence relationship between a dependent (response) variable and a set of independent (predictor) variables (Rawlings et al, 1998). In general, the most popular method used for regression is ordinary least squares (OLS) for its ease and simplicity. The OLS method is claimed to be unbiased, efficient and consistent estimator as compared to other linear regression model are satisfied. If the assumption is violated, the OLS method will no longer produce the least variance, leading to the inefficiency in estimating a model. One of the assumptions is that there is no exact linear relationship between the explanatory variables (Zahari et al, 2014).

Multicollinearity refers to a situation in which or more predictor variables in a multiple regression model are highly correlated if multicolinearity is perfect, the regression coefficients are indeterminate and their standard errors are infinite, if it is less than perfect (Dereny etal, 2011). There are several techniques used for the reduction of multicolinearity problem. Some of these techniques can be listed as: obtaining more data, the removal of one or more independent variables from the model, clustering the independent variables, and biased estimation techniques (Tunah and Siklar, 2015).

The ridge regression is the most widely model in solving the multicolinearity problem, and it's an alternative to OLS. The main advantage of ridge regression method is to reduce the variance term of the slope parameters (Alibuhatto, 2016). The aims of this study are to study the ridge regression method, which resolves multicolinearity without removing independent variables from the model but provides biased estimator to study the effect of some meteorological factors on the rainfall.

2. Theoretical Part

2. 1. Regression Model

Linear regression model is the relationship between a dependent variable and a set of independent variables as (Olandrewaju et al, 2017).

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \varepsilon_i\,,\ i = 1, \dots, n \ \ and\ j = 1, \dots, p \qquad (1)$$

Where; Y_i is the i^{th} response variable, $X_1, X_{2,\dots,} X_p$ are explanatory variables, ε_i is error term, and $\beta_0, \beta_1, \dots\dots, \beta_{(p-1)}$ are the regression coefficients.

In matrix form, the model can be written as:

$$Y = X\beta + \varepsilon \qquad (2)$$

Where; Y is $(n \times 1)$ vector of observations on dependent variables, X is a $(n \times p)$ matrix, ε is $(n \times 1)$ vector of error term, and β is a $(p \times 1)$ vector of regression coefficients.

The OLS estimate $\widehat{\beta}$ of β is obtained by minimizing the residual sum of squares (Salh, 2014).

$$\sum_{i=1}^{n} \varepsilon_i^2 = (Y - X\beta\,\acute{})(Y - X\beta) \qquad (3)$$

Then the best linear unbiased estimator of β is

$$\widehat{\beta} = (\acute{X}X)^{-1}\acute{X}Y \qquad (4)$$

With,

$$E(\widehat{\beta}) = \beta \qquad (5)$$

$$Var(\widehat{\beta}) = \sigma^2(\acute{X}X)^{-1} \qquad (6)$$

$$MSE(\widehat{\beta}) = \sigma^2 trace\,(\acute{X}X)^{-1} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad (7)$$

Assumptions made about the error and the variables:

1. ε is a random vector.
2. $E(\varepsilon_i) = 0$
3. $E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 I_n & , i = j \\ 0 & , i \neq j \end{cases}$
4. $\varepsilon \sim NID(0, \sigma^2 I_n)$
5. X is non-stochastic matrix.
6. There is no correlation between the non-stochastic x and the stochastic ε , i.e $E(\acute{X}\varepsilon_i) = 0$
7. The x variables are linearly independent, so $|\acute{X}X| \neq 0$

Thus, x matrix has rank $r = (p-1) < n$

2. 2. Multicollinearity

Multicollinearity is a statistical tool in which there exists a perfect relationship between the explanatory variables. When there is a perfect relationship between the explanatory variables, it is difficult to come up with reliable estimates of their individual coefficients. It will result in incorrect conclusions about the relationship between dependent variable and explanatory variables (Alibuhatto, 2016).

There are two types of multicollinearity (El-Sibakhi, 2016):

a. Perfect Multicollinearity

If exist perfect linear relationship among the explanatory variables then it is treated as exact multicollinearity. In case of perfect multicollinearity the design matrix as data matrix is not of full rank and consequently $(\acute{X}X)^{-1}$doesnot exist. In this case $|\acute{X}X| = 0$

b. Semi- Perfect Multicollineartity

If the explanatory variables are strongly as highly correlated but not perfectly then it is called semi- perfect mulitcolinearity. In this case $(\acute{X}X)^{-1}$ is exist but, with related large diagonal elements. Multicollineartity has several effects; these are described as follows (Dereny et al, 2011), (El-Sibakhi, 2016):

1. High variance of coefficients my reduced the precision of estimation.
2. Multicollineartity can result in coefficients appearing to have the wrong sign.
3. Estimates of coefficients may be sensitive to particular sets of sample data.
4. Some variables may be dropped from the model although they are important in the population.
5. The coefficients are sensitive of the presence of small number inaccurate data values.

2. 3. Detection of Multicollinearity

1. Correlation Matrix

Compute the correlation coefficients between any two of the explanatory variables. A high significant value of the correlation between two variables may indicate that the variables are collinear. This method is easy, but it cannot produce a clear estimate of the rate of multicollinearity (Alibuhatto, 2016).

2. The Variance **Inflation** Factor(VIF)

The VIF is computed from the correlation matrix of the independent variables (Rawlings et al, 1998), (Montgomery and Runger, 2002), (Raheem et al, 2019).

$$VIF = \frac{1}{1-R_j^2} \qquad ,j = 1,2,\dots\dots,p-1 \tag{8}$$

R_j^2 is coefficient of determination in the regression of explanatory variables on the remaining explanatory variables of the model.

3. Condition Number

The eigen values of the correlation matrix can also be used to measure the presence of multicollinearity. If multicollinearity is present in the predictor variables one or more of the egien values will be small. Let $\lambda_1, \lambda_2, \lambda_3, \dots\dots, \lambda_p$ be the egien values of correlation matrix. The condition number of correlation matrix is defined as:

$$k = \frac{\lambda_{max}}{\lambda_{min}} \tag{9}$$

If the condition number is less than 100, there is no serious problem with multicollinearity and if a condition number is between 100 and 1000 implies a moderate to strong multicollinearity. Also, if the condition number exceeds 1000, severe multicollinearity is indicated (Alibuhatto, 2016).

4. Eigen structure of $\acute{X}X$, Let $\lambda_1, \lambda_2, \lambda_3, \dots\dots, \lambda_p$ be the egien values of $\acute{X}X$ .when at least one eigen values is close to zero, then multicollineartity is exist (Dereny et al, 2011).
5. Checking the relationship between the F and T test might provide some indication of the presence of multicollinearity. If the overall significance of the model is good by using F- test but individually the coefficients are not significant by using T- test, then the model might suffer from multicollinearity (El-Sibakhi, 2016), (Raheem et al, 2019).

2.4. Ridge Regression

Ridge regression represents one of the methods which deal with multicollinearity problem (Kamel and Aboud, 2013). A possible remedy to this problem is the ridge estimator suggested by Hoerl and Kennard (Gullkey and Murrhy, 1975) represented it in 1970 (Kamel and Aboud, 2013). This reduces the variance of the estimates at the expense of introducing some degree of bias. This is accomplished by adding a small positive number, k, to each of diagonal elements of correlation matrix. The ridge estimator is shown as follow (Fitrianto and Yik, 2014).

$$\widehat{\beta}_R = (\acute{X}X + kI)^{-1}\acute{X}Y \tag{10}$$

Where, the I denote an identity matrix and k is ridge parameter.

The ridge regression estimator has several properties, which can be summarized as follow:

$$\begin{aligned} E(\widehat{\beta}_R) &= (\acute{X}X + kI)^{-1}E(\acute{X}Y) \\ &= (\acute{X}X + kI)^{-1}(\acute{X}X)E(\beta) \\ &= A_k\beta \end{aligned} \tag{11}$$

Where

$$A_k = \left(\mathrm{I} + \mathrm{k}(\acute{\mathrm{X}}\mathrm{X})^{-1}\right)^{-1} \tag{12}$$

$$\begin{aligned} V(\widehat{\beta}) &= (\acute{X}X + kI)^{-1}\acute{X}X(\acute{X}X + kI)^{-1}\sigma^2 \\ &= \sigma^2 A_k(\acute{X}X)^{-1}\acute{A}_k \end{aligned} \tag{13}$$

Where, $\widehat{\beta}_R$ is a biased estimator, but reduce the variance of the estimate, and $\widehat{\beta}_R$ is the coefficient vector with minimum length. The MSE of $\widehat{\beta}_R$ is given by:

$$\begin{aligned} MSE\,\widehat{\beta}_R &= E[(\widehat{\beta}_{(R)} - \beta)'(\widehat{\beta}_{(R)} - \beta) \\ &= \sigma^2 trace\,[A_k(\acute{X}X)^{-1}A_k'] + \acute{\widehat{\beta}}\,(I - A_k)'(I - A_k)\,\widehat{\beta} \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} k^2\ \acute{\widehat{\beta}}\,(\acute{X}X + kI)^{-2}\,\widehat{\beta} \end{aligned} \qquad (14)$$

3. Application Part

The data was obtained from the meteorological directorate of Sulaimani for the period (Jan. 2012- Aug. 2017) in order to reach an appropriate model, have been used NCSS19 and SPSS22.

The data that is including one response variable (Y) and seven explanatory variables (X_i):
$Y =$ Rainfall
$X_1 =$ Average Temperature
$X_2 =$ Relative Humidity
$X_3 =$ Wind Speed
$X_4 =$ Average Vapors
$X_5 =$ Sunshine
$X_6 =$ Station Pressure
$X_7 =$ Soil Temperature

Now since some of the variables are significantly related as shown in table (1).The results of the correlation matrix above, showed a highly significant possible relationships between variables. These results showed that there is presence of multicollinearity among these independent variables.

Table 1: Correlation matrix of the variables

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	y
X_1	1							
X_2	-.893**	1						
X_3	.174	-.171	1					
X_4	.854**	-.624**	.201	1				
X_5	.846**	-.777**	.321**	.678**	1			
X_6	-.564**	.566**	.332**	-.347**	-.343**	1		
X_7	.932**	-.827**	.057	.804**	.748**	-.522**	1	
y	-.665**	.635**	-.159	-.526**	-.636**	.348**	-.596**	1

** Correlation is significant at the 0.01 level.

The existence of multicollinearity was investigated using variance inflation factor (VIF) and condition number. The VIF for all independent variables are as follow:
$VIF(X_1) = 36.854, VIF(X_2) = 7.781, VIF(X_3) = 1.70, VIF(X_4) = 6.56, VIF(X_5) = 4.533,$
$VIF(X_6) = 2.529,\ VIF(X_7) = 8.959$
The result of VIF revealed presence of multicollinearity at $VIF(X_1)$ is greater than 10. This result confirmed a high level of multicollinearity among the independent variables.
The eigenvalues of the correlation matrix as follow:
$\lambda_1 = 4.545,\ \lambda_2 = 1.344,\ \lambda_3 = 0.479,\ \lambda_4 = 0.311, \lambda_5 = 0.202,\ \lambda_6 = 0.097,\ \lambda_7 = 0.021$
The condition number $(\varphi) = \frac{\lambda_{max}}{\lambda_{min}} = 215.44 > 100$
The results also indicate the presence strong multicollinearity between variables. To estimate $\widehat{\beta}$ coefficients with the minimum variance it is need to resolve this multicollinearity. The parameter

estimations (β_{RR}) calculated with k in the range of [0, 1] in order to see the effects of multicollinearity, trying to resolve with ridge regression technique, on the coefficients $\widehat{\beta}$ are given in table (2).

Table 2: Standardized ridge regression coefficients and max VIF.

K	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Max VIF
0.000	-.539	0.151	0.024	0.052	-0.214	-0.038	0.129	36.854
0.001	-.521	0.155	0.023	0.047	-0.217	-0.035	0.123	33.601
0.002	-0.505	0.159	0.022	0.042	-0.219	-0.033	0.117	30.761
0.003	-0.489	0.162	0.021	0.038	-0.221	-0.031	0.112	28.267
0.004	-0.475	0.165	0.021	0.034	-0.223	-0.029	0.107	26.066
0.005	-0.463	0.168	0.020	0.030	-0.225	-0.028	0.102	24.112
0.006	-0.451	0.171	0.019	0.027	-0.227	-0.026	0.098	22.371
0.007	-0.439	0.173	0.018	0.024	-0.228	-0.025	0.094	20.812
0.008	-0.429	0.175	0.018	0.021	-0.229	-0.023	0.090	19.412
0.009	-0.419	0.177	0.017	0.018	-0.231	-0.022	0.086	18.148
0.010	-0.410	0.178	0.016	0.016	-0.232	-0.020	0.083	17.004
0.020	-0.344	0.190	0.011	-0.001	-0.239	-0.011	0.055	9.767
0.030	-0.304	0.194	0.006	-0.012	-0.241	-0.004	0.035	6.343
0.040	-0.276	0.196	0.003	-0.019	-0.241	0.000	0.020	4.456
0.050	-0.256	0.197	0.000	-0.024	-0.242	0.004	0.009	3.307
0.060	-0.241	0.196	-0.002	-0.028	-0.241	0.008	-0.000	2.860
0.070	-0.230	0.195	-0.005	-0.031	-0.240	0.010	-0.008	2.546
0.080	-0.220	0.194	-0.007	-0.034	-0.238	0.013	-0.015	2.285
0.090	-0.212	0.193	-0.009	-0.036	-0.236	0.015	-0.021	2.065
0.100	-0.206	0.192	-0.010	-0.038	-0.234	0.017	-0.027	1.878
0.200	-0.172	0.178	-0.021	-0.051	-0.232	0.029	-0.058	1.007
0.300	-0.158	0.168	-0.027	-0.058	-0.212	0.036	-0.072	0.680
0.400	-0.149	0.159	-0.030	-0.063	-0.196	0.040	-0.080	0.515
0.500	-0.143	0.152	-0.031	-0.067	-0.174	0.042	-0.084	0.421
0.600	-0.138	0.147	-0.032	-0.069	-0.166	0.044	-0.087	0.362
0.700	-0.134	0.142	-0.032	-0.071	-0.159	0.045	-0.088	0.316
0.800	-0.131	0.137	-0.032	-0.072	-0.153	0.046	-0.089	0.279
0.900	-0.128	0.133	-0.032	-0.072	-0.147	0.046	-0.090	0.248
1.000	-0.125	0.129	-0.032	-0.073	-0.142	0.046	-0.089	0.223

The regression coefficients and standard errors of these coefficients can be summarized in table (3), by using both OLS and RR methods to analyze the data, we get the following results.

Table 3: Regression coefficients and standard errors

Independent variable	Ridge Coefficient	Least Square Coefficient	Ridge Standard Error	Least Square Standard Error
intercept	204.995	428.476		
X_1	-2.544	-3.986	1.977	3.821
X_2	0.779	0.621	0.735	0.975
X_3	1.261	2.804	11.906	12.688

X_4	-0.053	1.604	4.843	6.701
X_5	-6.902	-6.195	4.843	5.239
X_6	-0.101	-0.346	1.091	1.234
X_8	0.538	1.266	1.909	2.489

In the study for (Jan. 2012- Aug. 2017) period, ridge parameter k was (0.02) and the ridge regression, which indicates the effects of independent variables to the rainfall in Sulaimani, is estimated as

$$\hat{y_i} = 204.995 - 2.544X_1 + 0.779X_2 + 1.261X_3 - 0.053X_4 - 6.902X_5 - 0.101X_6 + .538X_7$$

And ordinary least square model, is estimated as

$$\hat{y_i} = 228.476 - 3.986X_1 + 0.621X_2 + 2.804X_3 - 1.604X_4 - 6.195X_5 - 0.346X_6 + 1.266X_7$$

Table 4: Analysis of variance for k = 0.02

$O.S.V$	$D.F$	$Sum\ of\ Squares$	$Mean\ Square$	$F-Ratio$	$P-value$
Intercept	1	250937.5	250937.5		
Model	7	190049.7	27149.96	9.0330	0.00*
Error	73	219412.1	3005.645		
Total (Adjusted)	80	409461.8	5118.272		
$Mean\ of\ dependent\ variable$		55.659			
$Root\ mean\ square\ error$		54.824			
$R-Squared$		0.4641			
$Coefficient\ of\ variation$		0.985			
$**\ The\ result\ is\ significant\ at\ the$		0.01			

The root mean squares error of regression coefficients for RR and OLS methods are as follow:

$RMSE(\beta_{RR}) = 54.824,\ RMSE(\beta_{OLS}) = 55.543$

And the coefficient of determination (R^2) for each model, we obtain the following result:

$R^2(RR) = 0.464\ ,\ R^2(OLS) = 0.419$

We make a comparison between ridge regression and ordinary least squares. We noted that ridge regression model is better than ordinary least square model when the multicollinearity problem is exist because it has smaller mean square errors of estimators, smaller standard deviation for all estimators and has large coefficient of determination.

4. Conclusions

According to the results of this study the multicollinearity was detected, because variance inflation factor for X_1 equal (36.854) greater than 10 and condition number equal (215.44) greater than 100, this confirmed that the multicollinearty problem is existing. The most direct variables affecting the amount of rainfall are the average temperature which affects (-0.665), followed by sunshine that affects (-0.636), then relative humidity (0.635), then soil temperature (-0.596), and then other meteorological variables. The (k=0.02) value is the optimal value that resolves the multicollinearity problem. The ridge regression model is better than ordinary least square model when the multicollinearity problem is exist, because it has smaller mean square errors of estimators, smaller standard deviation for all estimators and has large coefficient of determination.

References

1. Alibuhatto, M. C. (2016). Relationship between Ridge Regression Estimator and Sample Size When Multicollinearity Present among Regresses, world scientific news, 59, pp12-23.
2. Dereny, M. E. and Rashwan, N. I. (2011). Solving Multicollinearity Problem Using Ridge Regression Models, International Journal of Contempt Mathematic Sciences, Vol.6, No.12, pp585-600.
3. El- Sibakhi, R. A. (2016). A Comparative Study of Ridge Regression and Staged Logistic Regression as a Remedy of Multicollinearity Problem a Case study of Armenia Demographic and Health Survey 2010, M.Sc. Thesis, Al-Azhar University- Gaza.
4. Fitrianto, A. and Yik, L. C. (2014). Performance of Ridge Regression Estimator Methods on Small Sample Size by Varying Coefficients: A Simulation Study, Journal of mathematics and Statistics, Vol.10, No.1, pp25-29, doi:10.3844/jmssp.2014.25.29.
5. Gullkey, D. K. and Murrhy, J. L. (1975). Directed Ridge Regression Techniques in Casas of Multicollinearity, Journal of American Statistical Association, Vol.7, No.352, pp769-775.
6. Kamel, M. M., and Aboud, S. F. (2013). Ridge Regression Estimators with the Problem of Multicollinearity, Applied Mathematical Sciences, Vol.7, No, 50, pp2469-2480.
7. Montgomery, D. C. and Runger, G. C. (2002). Applied Statistics and Probability for Engineers, 3rd Edition, John Wiley and Sons, Inc, USA.
8. Olandrewaju, S. O., Yahaya, H. U., and Nasirru, M. O. (2017). Effects of Multicollinearity and Autocorrelation on Some Estimators in a System of Regression Equation, European Journal of Statistics and Probability , Vol.5, No,3, pp1-15.
9. Raheem, M. R., Udoh, N. S. and Gbolahan, A. T. (2019). Choosing Appropriate Regression Model in the presence of Multicolinearity, Open Journal of Statistics, 9, pp 159-168.
10. Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). Applied Regression Analysis: A Research Tool, 2nd edition, Springer- Verlag, Inc, New York.
11. Salh, S. M. (2014). Using Ridge Regression Model to Solving Multicollinearity Problem, International Journal of Scientific and Engineering Research, Vol.5, Issue 10, pp992-998.
12. Tunah, D. and Siklar, E. (2015). Analysis of Inflation in Turkey via Ridge Regression, International Research Journal of Applied Finance, Vol.6, No.12, pp811-820.
13. Zahari, S.M., Ramli, N. M., and Mokhtar,B. (2014). Bootstrapped Parameter Estimation in Ridge Regression with Multcollinearity and Multiple Outliers, Journal of Applied Environmental and Biological Sciences, Vol.4, No,7, pp150-156.

استخدام انحدار الحرف لتحليل بيانات الأرصاد الجوية في السليمانية

ليلى عزيز احمد

الخلاصة

الانحدار الخطي هو أحد الأساليب الإحصائية المستخدمة بشكل متكرر والتي لها تطبيقات في جميع مجالات الحياة اليومية. في المنظور الإحصائي ، يتم استخدام تحليل الانحدار لدراسة الارتباط الخطي المتعدد بين متغير تابع ومجموعة من المتغيرات المستقلة. انحدار الحرف هو النموذج الأكثر انتشارًا في حل مشكلة الارتباط الخطي المتعدد ، وهو بديل للمربعات الصغرى الاعتيادية (OLS).

تعد الارتباط الخطي المتعدد المشكلة الأكثر شيوعًا في نماذج الانحدار المتعددة التي توجد فيها علاقة التامة بين متغيرين توضيحيين أو أكثر في النموذج . في هذه الدراسة تم استخدام نموذج انحدار الحرف لتقدير نموذج الانحدار الخطي. تمت مقارنة هذه النتيجة مع النتيجة التي تم الحصول عليها باستخدام نموذج المربعات الصغرى الاعتيادية من أجل إيجاد أفضل نموذج انحدار. لقد استخدمنا بيانات الأرصاد الجوية في هذه الدراسة. أظهرت النتائج أنه يمكن استخدام طريقة انحدار الحرف لحل مشكلة الارتباط الخطي المتعدد ، دون حذف المتغيرات المترابطة المستقلة للنموذج وقادرة على تقدير أفضل المعلمة باقل قيمة الخطأ المعياري.