Journal of Al Azhar University-Gaza (Natural Sciences), 2011, 13 : 119-134

Remedy of multicollinearity using Ridge regression

Abdalla M. EL-Habil, Khaled I.A. Almghari
abdalla20022002@yahoo.com

Department of Applied Statistics
Al-Azhar University - Gaza

Received 17/04/2011 Accepted 15/9/2011

Abstract: One of the main problems in the model is two or more of the explanatory variables in the sample are overlapping, and this overlapping indicates a multicollinearity problem.

Ridge regression is one of the famous methods for remedy of the multicollinearity problem because it enables us to keep these explanatory variables, which violate the assumption of independency in the model. In this paper, we used real data and R package to find the multicollinearity problem in the established linear model between two explanatory variables. One of the multicollinearity problems solutions is to omit the explanatory variables, which cause the multicollinearity. We show that by using the ridge regression, we get the new estimates of the new model without omitting any of the explanatory variables.

Keywords: Ridge regression – multicollinearity - ordinary least square - singular value decomposition - generalized cross validation - Lawless and Wang method.

1. Introduction

Multicollinearity is the extreme problem for a regression models, because it violates the assumptions of the model that is the explanatory variables should be independent.

Multicollinearity means there is a relation between these explanatory variables, this relation makes overlapping between variables.

Multicollinearity can cause serious problem in estimation and prediction, increasing the variance of least squares of the regression coefficients and tending to produce least squares estimates that are too large in absolute value.

Theoretically, we have two types of multicollinearity; these types are partial multicollinearity and perfect multicollinearity or full multicollinearity.

In addition, multicollinearity has two cases: scalar case and matrix case.

Multicollinearity can be detected by examining one of two qualities: Variance Inflation Factor "VIF" and Tolerance.

One of the remedial methods of multicollinearity is a ridge regression, this regression enables us to obtain such an estimate by minimizing MSE risk, and the ridge regression enables us to inference on values of predictor variables that follow the same pattern of multicollinearity and this aspect is very important.

The paper is organized as follows. Section 2 recalls the technical background of multicollinearity. Section 3 ridge regression. Section 4 methods for selecting λ in ridge regression. Section 5 data analysis. Section 6 concludes.

2. Multicollinearity

Multicollinearity occurs when two or more of the explanatory variables in a sample overlap. Because of the overlap, methods of analysis cannot fully distinguish the explanatory factors from each other or isolate their independent influence.

Interpretation of the multiple regression equation depends on the assumption that the predictor variables are not strongly interrelated, therefore the multicollinearity violates this assumption, and this means there is no linear relationship among the predictor variables.

Mathematical aspects of multicollinearity

Scalar case

Any population model looks like the below model:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + ... + B_n X_n + e \qquad (2.1)$$

Where

y: The dependent variable

B_0: is the intercept

$B_1, B_2, ..., B_n$: are the slopes of coefficients for their respective explanatory variables.

$X_1, X_2, ..., X_n$: are the explanatory variables.

e : is the random error term.

Note: with absent of multicollinearity then Y is a linear function of the explanatory variables and a random error term.

If we suppose that, we have just only two explanatory variables X_1, X_2 and X_2 is a multiple of X_1 then $X_2 = dX_1$ for simplicity.

The regression would need to find the coefficient estimates b_1, b_2 that produce the best $\hat{Y}$.

Then $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$, we can substitute X_2 by dX_1 as:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 (dX_1)$$

$$\hat{Y} = b_0 + X_1 (b_1 + db_2) \tag{2.2}$$

The above result is true also for infinite number of coefficient pairs; also these pairs produce the same value of $\hat{Y}$.

Any small change in b_1 from one possible value to another (δb_1) is matched by corresponding change in b_2.

By compensation, we get the below result:

$$\delta b_2 = -\frac{\delta b_1}{d}$$

By repeating all the coefficient pairs and keeping the linearity we can minimize sum of square errors $[Y - \hat{Y}]^2$.

Mathematically, the standard errors for coefficients S_j equals[1]:

$$S_j = \sqrt{\frac{\sum_{1}^{n} (Y - \hat{Y})^2}{\sum (x_i - \bar{x}_i)^2 (1 - R_j^2)(n - k)}}$$, j=1,…,n, i=1,…,n , where

n: is the numbers of predictors.

$\bar{x}_i$: is the mean of x_i

R_j^2: is the square of the multiple correlation coefficients that result when the predictor variable X_j is regressed against the other entire predictor variable.

k : is the number of explanatory variables.

[1]D.Stephen Voss, (2004), Multicollinearity, Encyclopedia of social measurements, pages 11-13

At full multicollinearity R^2=1, then (1- R_j^2) = 0

Therefore S_j is undefined, then there is no linear regression.

Matrix case

In a matrix case the explanatory variable X can be take the following form:

$$X = \begin{bmatrix} 1 & X_{12} & X_{13} & ..X_{1k} \\ 1 & X_{21} & X_{22} & ..X_{2k} \\ . & & & \\ & . & . & . \\ . & & & \\ 1 & X_{n2} & X_{n3} & X_{nk} \end{bmatrix}$$

Where:
n: number of observations
k: number of explanatory variables.
The first column is intercept or constant.
Matrix linear model is:

$$Y = B_0 + \sum_{i=1}^{j} X_i^T B_i + e \qquad (2.3)$$

$$\hat{Y} = b_0 + \sum_{i=1}^{j} X_i^T b_i$$ Where Y is a vector $n \times 1$.

The coefficient b can be calculated as:

$X^T X\, b = X^T Y$

$\therefore\ b = (X^T X)^{-1} X^T Y$

Smith and Norman (1998).

At multicollinearity the determinant of $(X^T X)$ is equal zero, therefore the inverse will not existing.

Detection of multicollinearity[2]

We can detect the multicollinearity by examining a quality called

[2] Ali S. Hadi, Samprit Chatterjee (2006), Regression analysis by examples, Fourth edition, pages 337-339

Variance Inflation Factor (VIF).

$VIF_j = \frac{1}{1-R_j^2} = \text{diag}(\text{X}^{-1}X)^{-1}$, $j = 1,....., p$ where:

R_j^2: is the square of the multiple correlation coefficients that results when the predictor variable X_j is regressed against all the other predictor variable.

p : is the number of predictor variables.

X: in matrix case

At multicollinearity R^2 would be closed to 1 then VIF_j would be large,

When VIF_j greater than 10, the data have collinearity problems.

Also we detect the multicollinearity by examining a quality called tolerance which equals $\frac{1}{VIF} = (1-R_j)^2$ and the tolerance in multicollinearity should be small.

3. Ridge regression[3]

One of the goals of ridge regression is to produce a regression equation with stable coefficients, and these coefficients are stable and not affected by slight variations in the estimation data.

From equation (2.1), we can get

$$\hat{Y} = \text{b}_0 + \sum_{i=1}^{j} b_j X_j = b_0 + \text{X}^{\text{T}} b \qquad (3.1)$$

Where $\hat{Y}$ is the estimated output, B_0 , B_j are coefficients and we can estimate B as:

$\hat{B} = (\text{X}^{\text{T}} X)^{-1} \text{X}^{\text{T}} Y$

The out correlation $(\text{X}^{\text{T}}\text{X})$ could have some eigenvalues close to zero, that is in multicollinearity problems but in ridge regression it is one of the shrinking methods to penalize strong deviations of the

[3] Technical Notes on Linear Regression and Information Theory Hiroki Asariy, Septermber 22, 2005, pages 1-9

parameters from zero. Therefore the errors function to be minimized as:

$$E_{ridge}(\mathrm{B},\lambda) = (\mathrm{y}-\mathrm{XB})^{\mathrm{T}}(y-XB)+\lambda B^{T}B \quad (3.2)$$

Where $\lambda \geq 0$ which determines the strength of the ridge constraint. The solution of equation (3.2) is given by:

$$\hat{B}_{ridge} = (\mathrm{X}^{\mathrm{T}}X+\lambda\,\mathrm{I})^{-1}\,\mathrm{X}^{\mathrm{T}}y \quad (3.3)$$

Eric Doviak (2009).

Where: I is the identity matrix.

Singular Value Decomposition (SVD)[4]

Ridge regression is the developed methods of Least Square (LS) and Ordinary Least Square (OLS) and SVD is highly related to the Least Square solution of the below equation:

$$\hat{B}_{j} = (X^{T}X)^{-1}X^{T}y \quad (3.4)$$

And the ridge regression is the solution for the below equation:

$$\hat{B}_{ridge} = (X^{T}X+\lambda I)^{-1}X^{T}y$$

Then the SVD is of an $n\times p$ matrix X has the form

$$X = USV^{T} \quad (3.5)$$

where U is an $n\times p$ orthogonal matrix whose columns U_j span the column space of X, and V is a $p\times p$ orthogonal matrix whose columns span the column space of X^T, S is the $p\times p$ diagonal matrix of singular values $s_1 \geq s_2 \geq \ldots\ldots \geq s_p \geq 0$.

Using SVD and solving the equations (3.4) and (3.5) then the solving result as:

$$(X^{T}X)^{-1}X = (VS^{2}V^{T})^{-1}(USV^{T})^{T} = VS^{-1}U^{T}$$

Where $(\frac{1}{s_1},\frac{1}{s_2},\ldots\ldots\ldots,\frac{1}{s_p})$ are on the diagonal of S^{-1}.

The Least Square of equation (3.4) can be written as:

$$\hat{B} = VS^{-1}U^{T}y$$

The ridge regression solution for equation (3.3) is given as :

[4] Hernan R. , Luis F. (2002), The distribution of stochastic shrinkage parameters in ridge regression, Working paper of the central bank of Chile, pages 1-5

$\hat{B}_{ridge} = V(S^2 + \lambda I)^{-1} SU^T y$

Where the (i,i) elements of the diagonal matrix

$(S^2 + \lambda I)^{-1} S$ is $\frac{s_1}{(s_i^2 + \lambda)}$

Barak Weiss and Dimetry Kleinbock (2005).

From equations (3.3), (3.4), and (3.5) the estimated $\hat{y} = X\hat{B}$ for the Least Squares and the ridge regression can be expressed as:

$\hat{y} = X(X^T X)^{-1} X^T y = UU^T y$ [5]

4. Methods for selecting λ in ridge regression

The main step in the ridge regression analysis is to select a value of λ and to obtain the corresponding estimates of the regression coefficients.

If multicollinearity is a serious problem, the ridge regression estimators will vary dramatically as λ is slow increased from zero to 0.1.

Fixed point

Ali S. Hadi ets (2006) said that "Hoerl Kennard and Baldwine suggest estimating λ **"** by**:**

$$\lambda = \frac{\rho \hat{\sigma}^2(0)}{\sum_{1}^{p} [\hat{X}_i(0)]^2} \qquad (4.1)$$

Where:

$\hat{X}_i(0), ..., \hat{X}_p(0)$ are the least square estimates of $X_1, ..., X_p$.

$\hat{\sigma}^2(0)$ is the corresponding residual mean square.

Generalized Cross Validation (GCV)[6]

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to

[5] Barak Weiss and Dimetry Kleinbock (2005), friendly measures, homogenous flows and singular vectors pages 1-3

[6] Partrick B.(2009) R package"grpreg" page9

estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

$$GCV = \frac{2l}{\left(1 - \frac{df}{n}\right)^2} \tag{4.2}$$

Where:

l : Is the loss function (usually log likelihood estimator).

df : The effective number of model parameters at the chosen value of λ.

n : Is the sample size.

Lawless and Wang (L_W) method[7]

G.R.Pasha,M.A.Shah(2004) said that "Lawless and Wang(1976)" they are used Bayesian prospective to estimate λ as:

$$\hat{\lambda} = \frac{\rho\hat{\sigma}^2}{\hat{\mathrm{B}}^{\mathrm{T}}\hat{B}} \tag{4.3}$$

Where:

$\hat{B} = (\mathrm{X}^{\mathrm{T}}X + \lambda I)^{-1}X^{-1}y$, B: It is a $p \times 1$ vector of unknown coefficients.

X : It is a $n \times p$ full rank matrix of explanatory variables.

y: It is a $n \times 1$ vector of observations of the dependent variable.

Notes: All the above three methods are omitted the intercept from the ridge model therefore the alternative names of ridge regression is the shrink regression.

5. Data

Our data set is a survey of National Crime Victimization Violent Crime Trends which published by FBI's (2008) as Crime report for USA since 1973 to 2008.

Our data set is the number of victimization per 1,000 population age 12 years or older.

[7] G.R.Pasha,M.A.Shah(2004).

Variables description

Year : year

Y : TOTAL VIOLENT CRIME

X_1: RAPE

X_2 : ROBBERY

X_3 : AGGRAVATED ASSAULT

X_4 : SIMPLE ASSAULT

Statistical Analyses

Estimate the general linear model by R outputs:

The R output shows the estimate of explanatory variables				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	**0.144421**	**0.05835**	**2.475**	**0.0192 ***
X_1	**1.015136**	**0.038091**	**26.65**	**<2e-16 *****
X_2	**1.066336**	**0.032445**	**32.866**	**<2e-16 *****
X_3	**0.947066**	**0.018587**	**50.954**	**<2e-16 *****
X_4	**1.003236**	**0.005739**	**174.795**	**<2e-16 *****

* The variable is significant at 0.05 level

*** The variable is significant at 0.01 level

From the table we can conclude the general linear model as:

$$\hat{Y} = 0.144421 + 1.015136\ X_1 + 1.066336\ X_2 + 0.947066\ X_3 + 1.003236\ X_4$$

The assumption of above model is: X_1, X_2, X_3, X_4 are independent.

To test this assumption we will:

- Examine the correlation between these variables.
- Investigate the scatter plot between each pair of variables.
- Test for multicollinearity if it exists.

The table below shows the correlation between the variables.

Table of correlations between variables				
	X_1	X_2	X_3	X_4
X_1	**1**	**0.902146**	**0.9014170**	**0.803882**
X_2	**0.902146**	**1**	**0.970647**	**0.919454**
X_3	**0.901417**	**0.970647**	**1**	**0.927912**
X_4	**0.803838**	**0.919454**	**0.927912**	**1**

From the above table we can conclude that the four variables are strongly correlated with each other.

Figure (1): Diagram below shows the scatter plot between RAPE and ROBBERY variables

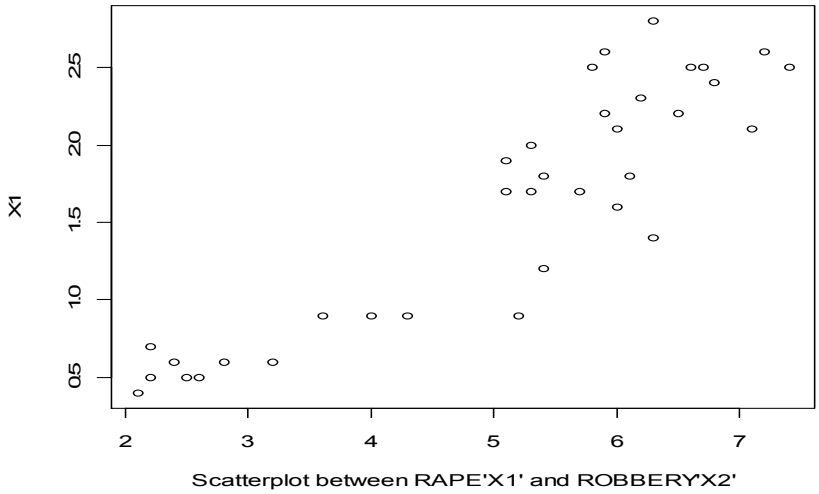


Figure (2): Diagram below shows the scatter plot between RAPE and AGGRAVATED ASSAULT variables

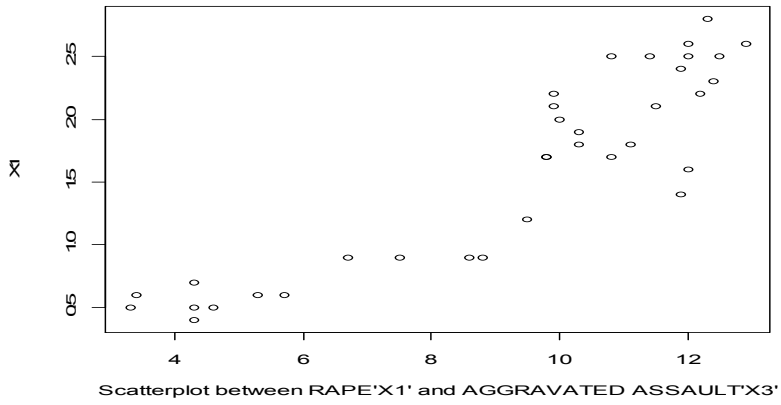


Figure (3): Diagram below shows the scatter plot between RAPE and SIMPLE ASSAULT variables

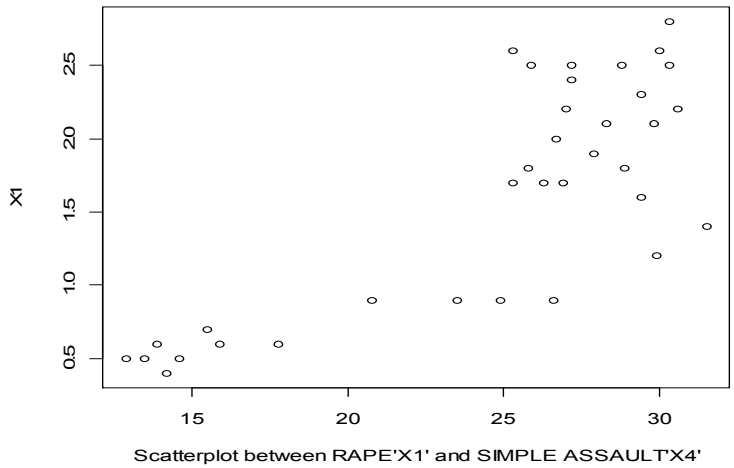


Figure (4): Diagram below shows the scatter plot between ROBBERY and AGGRAVATED ASSAULT variable

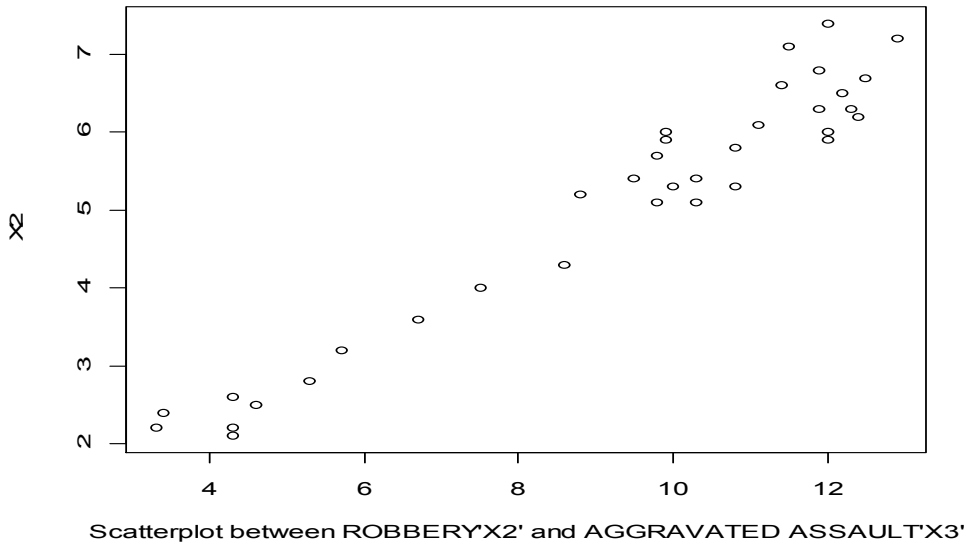


Figure (5): Diagram below shows the scatter plot between ROBBERY and SIMPLE ASSAULT variables

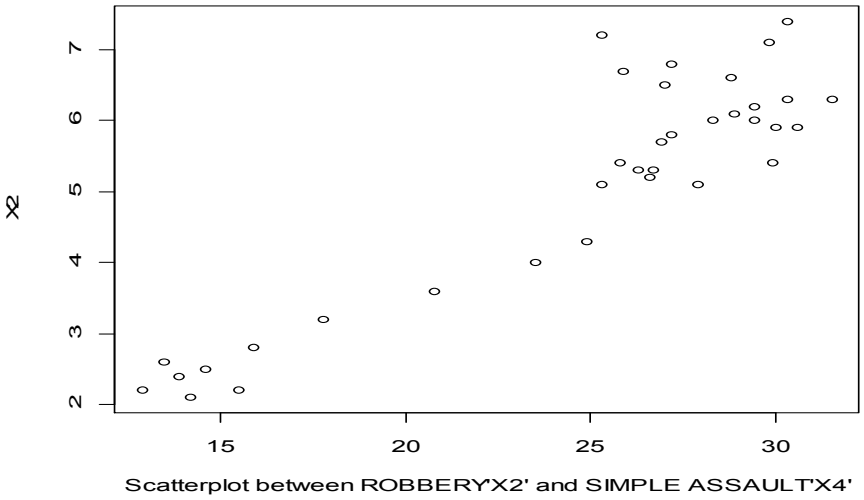
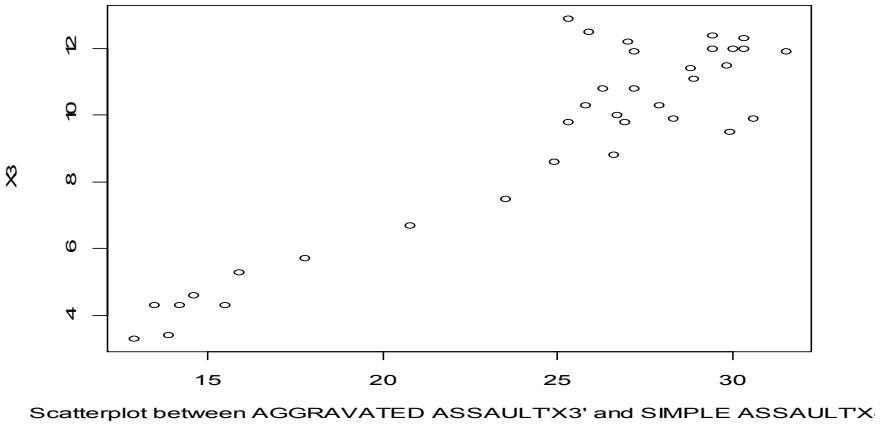


Figure (6): Diagram below shows the scatter plot between AGGRAVATED ASSAULT and SIMPLE ASSAULT variables



From the above all scatter plots; we can notice clearly that there is a linear combination between X_2 (ROBBERY) and X_3 (AGGRAVATED ASSAULT) variables.

The table below shows the R output of Variance Inflation Factor and Tolerance results for our variables.

Table shows the Variance Inflation Factor & Tolerance results				
	X_1	X_2	X_3	X_4
VIF	**6.182179**	**20.0978**	**22.61593**	**8.126189**
Tolerance	**0.161755**	**0.049757**	**0.044217**	**0.123059**

From the above table we can conclude that:

The VIF of X_1 = 6.182179 < 10 and Tolerance = 0.161755 > 0.1 then there is no multicollinearity problem at this variable.

The VIF of X_2 = 20.0978 > 10 and Tolerance = .049757 < 0.1 then there is a multicollinearity problem at this variable.

The VIF of X_3 = 22.61593 > 10 and Tolerance = 0.044217 < 0.1 then there is a multicollinearity problem at this variable.

The VIF of X_4 = 8.126189 < 10 and Tolerance = 0.123059 > 0.1 then there is no multicollinearity problem at this variable.

From the above results, we can conclude that only X_2 (ROBBERY) and X_3 (AGGRAVATED ASSAULT) variables have the multicollinearity problems in our model.

One of the multicollinearity problems solutions is to omit the explanatory variables, which cause the multicollinearity. In our paper, we will use the ridge regression as the alternative method instead of omitting the explanatory variable.

The table bellow is R **outputs for methods, which select** λ

Methods for selecting λ		
modified HKB estimator	**modified L-W estimator**	**smallest value of GCV**
0.000210147	**8.13E-05**	**0.002**

From above table, we can conclude that the suitable method is GCV because it is the only value which is in the range between 0.001 and 0.1.

Therefore we are used R package to extract the model coefficients at λ=0.002 which they are at table below.

The table below is the R outputs of the extracting estimators of explanatory variables

Coefficients of ridge regression when λ=0.002			
X_1	X_2	X_3	X_4
0.7756401	**1.7259922**	**2.8359071**	**5.8305703**

So, the exact model of ridge regression to our data is:

$$\hat{Y} = 0.7756401\mathrm{X}_1 + 1.7259922\mathrm{X}_2 + 2.8359071\mathrm{X}_3 + 5.8305703\mathrm{X}_4$$

Here, we assessed how accurate the estimated model is by using the cross-validation method as we mentioned above.

6. Conclusion

We used R package for constructing the linear model between the dependent variable [TOTAL VIOLENT CRIME "Y"] and the explanatory variables: [RAPE "X_1"], [ROBERRY "X_2"], [AGGRAVATED ASSAULT "X_3"], [SIMPLE ASSAULT "X_4"].

In addition, we tested for the multicollinearity by extracting the VIF values and we found the multicollinearity problem between the explanatory variables X_2 and X_3 through the VIF quantities > 10, therefore we got a multicollinearity problem at our constructed model. We used a ridge regression as a remedial method for this problem. By using R package, we got the suitable λ for extracting the estimate values of B "coefficients" and constructed our new model with keeping all the explanatory variables.

References:

1. Ali S. Hadi, Samprit Chatterjee (2006), Regression analysis by examples, Fourth edition, Wiley series in probability and statistics.
2. Barak Weiss and Dimetry Kleinbock (2005), Friendly measures, Homogenous flows and singular vectors, Coruell university.
3. D.Stephen Voss, 2004, Multicollinearity, Encyclopedia of social measurements, University of Kentucky.
4. Draper Harry Smith and Norman R. (1998), Applied regression analysis, Third Edition.

5. Hernan R. , Luis F. (2002), The distribution of stochastic shrinkage parameters in ridge regression, working paper of the central bank of Chile, pages 1-5.
6. Hiroki Asariy (2005), Technical Notes on Linear Regression and Information Theory ,Cold spring harber NY11724, September 22, 2005, pages 1-9.
7. Partrick B.(2009), R package"grpreg", page 9.
8. G.R.Pasha,M.A.Shah (2004), Application of Ridge Regression to Multicollinear data, journal of research science vol. 15- June 2004, pp 97-106.
9. National Crime Victimization Violent Crime Trends which published by FBI's (2008) as Crime report for USA since 1973 to 2008.