A PROJECT REPORT

On:

EFFECTS OF THE MAJOR ECONOMICAL VARIABLES ON THE CHANGE IN GROSS NATIONAL INCOME(GNI) (AT CURRENT PRICES) OF INDIA (1980-2021)

A Regression Case Study

Submitted By;

Shiladitya Bose (211379)

Rajarshi Paul(211356)

Course Project of Regression Analysis : (MTH 416A)

Under Supervision of Dr. Sharmishtha Mitra

DEPARTMENT OF MATHEMATICS AND STATISTICS

Acknowledgement

Real learning comes from a practical work.We would like to thank our instructor of the course **Dr. Sharmishtha Mitra** (Department of Mathematics and Statistics, IIT KANPUR), for providing us constant guidance and motivation for this project, without which it would have been an impossible task to accomplish.We would like to thank our department professors for teaching all the necessary topics with immense care which was needed to make the project fruitful.

We would also like to thank our seniors for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

We also take this opportunity to thank the authors and publishers of the various books and journals we have consulted .Without those this work would not have been completed.

It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course lecture.

Shiladitya Bose
Rajarshi Paul

Contents

1 Introduction & Objective:

1.1 Introduction:

Gross National Income (GNI) is the total amount of money earned by a nation's people and businesses. It is used to measure and track a nation's wealth from year to year. The number includes the nation's gross domestic product (GDP) plus the income it receives from overseas sources. The more widely known term GDP is an estimate of the total value of all goods and services produced within a nation for a set period, usually a year. GNI is an alternative to gross domestic product (GDP) as a means of measuring and tracking a nation's wealth and is considered a more accurate indicator for some nations.

Gross national income (GNI) is an alternative to gross domestic product (GDP) as a measure of wealth. It calculates income instead of output. GNI can be calculated by adding income from foreign sources to gross domestic product. Nations that have substantial foreign direct investment, foreign corporate presence, or foreign aid will show a significant difference between GNI and GDP.

GNI calculates the total income earned by a nation's people and businesses, including investment income, regardless of where it was earned. It also covers money received from abroad such as foreign investment and economic development aid.

For nations, like the US, there is little difference between GDP and GNI, since the difference between income received versus payments made to the rest of the world does not tend to be significant. For some countries, however, the difference is significant. Conversely, it can be much lower if foreigners control a large proportion of a country's production, as is the case with Ireland, a low-tax jurisdiction where the European and U.S. subsidiaries of a number of multinational companies nominally reside.

$$\boldsymbol{GNI = C + I + G + X}$$

where : Personal consumption (C), Business investment (I), Government spending (G), Exports - imports (X)

1.2 Objective:

(a) Collected data on GNI (at current price) and on 20 other economic variables for past 41 years and performed Data Cleansing task.

(b) Before proceeding into theoretical analysis, we use a method to analyze the characteristics of the variables visually, named as Exploratory Data Analysis.

(c) Fitted an MLR model on the dataset and planning and working on checking for validation of basic assumptions i.e. Normality, Heteroscedasticity assumption of the errors and presence of Autocorrelation among the errors.

(d) Also working on to solve multicollinearity problems using VIF and Variance Decomposition Method.

(e) Also planning to see Ridge Trace to see the multicollinearity effect of the regressors and also Performed Ridge Regression for reduce variability of regressors which causes multicollinearity.

2 Description of Data:

2.1 Data Source:

We had collected the data on GNI(at current price) and on 20 other economic variables for past 41 years i.e 1980-2021 from Handbook of Statistics on Indian Economy available at http://www.rbi.org.in. and World Bank National Accounts Data.

2.2 Dataset Description:

Our data consists of 861observations with the information on the following variables:

Here,

ResponseVariable (Y): Gross National Income(in Current LCU) in Billion.

[Data Source: World Bank national accounts data and OECD National accounts data files.]

and the Explanatory Variables:

- **Agricultural production of food grains in Million Metric tonnes** (X_1)**:**

Agriculture plays an important role in the formation of the Indian economy. The production of food grains constitutes a major part of India's total agricultural production. The major food grains that are produced in India are Rice, Wheat, Coarse Cereals, and Pulse. Data were taken in Million tonnes units.

[Data Source: Ministry of Agriculture & Farmers Welfare, Government of India.]

- **Agricultural production of commercial products in Million Metric tonnes** (X_2)**:**

Commercial products are also an important type of Agricultural production. Apart from food grains the products like Groundnut, Rapeseed & Mustard, Soyabean, Coffee, Cotton (Lint), Raw Jute & Mesta, Sugarcane, Tea, Tobacco. generally grown for commercial purposes. Data taken in Million tonnes unit.

[Data source : Ministry of Agriculture & Farmers Welfare, Government of India, Coffee Board of India, Tea Board of India.]

- **Production of crude oil and Petroleum in Million Metric tonnes** (X_3)**:**

Indian economy and Indian market are strongly affected by the prices of crude oil and petroleum. Therefore the production of these commodities are very much important in the Indian context. The overall economical cycle can be affected by the price of oil.

[Data source : Ministry of Petroleum and Natural Gas, Government of India, PPAC]

- **Import-of crude oil and Petroleum in Million Metric tonnes** (X_4)**:**

[Data source : Ministry of Petroleum and Natural Gas, Government of India, PPAC]

- **Spread of gold price in Rupees** (X_5)**:**

The gold reserve of a country affects the supply of currency within the economy. If the central bank imports gold then it can result to an inflation in the economy. Therefore, the price of gold affects the demand and supply of gold and alternatively it affects the economic cycle.

[Data source : Business Standard/ Business Line and Economic Times for Indian price(Mumbai) and LMBA for London price]

- **Direct and Indirect tax revenue in Billion Rupees (X_6):**

Direct and indirect tax revenue is a principal source of government's income. Direct tax includes Income Tax, commercial property tax, personal property tax, taxes on assets etc. Whereas indirect taxes are those taxes that are imposed on the goods and services like sales tax, consumption tax, Goods and Service tax (GST), tax collected by the intermediaries.

[Data source :Budget documents of the Government of India and the State Governments]

- **Total saving deposits in commercial banks in Billion Rupees (X_7):**

The savings account in a commercial bank includes the feature that only a pre-specified number of withdrawals can be taken within a specified period of time. This money plays an important role in building the Indian economy when the government invests this money for loan purposes.

[Data source :RBI]

- **Gross Fiscal Deficit in Billion Rupees (X_8):**

Fiscal deficit is the difference between the total income of the Government and its total expenditure. It is an important concept in the context of Indian Economy. The government needs to take necessary measures for financing this deficit and that in turn can lead to the changes of major aspects of Indian economic cycle.

[Data source :Budget documents of the Government of India]

- **Combined Net borrowing of Central and State Government in Billion Rupees (X_9):**

In many cases the government needs to raise money from the market to meet its necessary expenses. These expenses can include the financing of Fiscal deficit and repaying loans etc. The government borrowing affects the private investment of a country.

[Data source :RBI]

- **Currency with the public in Billion Rupees** (X_{10})**:**

The total currency with the public is the difference between the total value of the currency including the coins and the paper currency issued by the Reserve Bank of India and the amount of that currency withdrawn by the Reserve Bank of India. The currency with the public may affect the production of the total investment expenditure of the country.

[Data source :RBI]

- **Government's Developmental and Non-Developmental expenditure in Billion Rupees** (X_{11})**:**

The developer expenditure includes those expenditures that it helps in increasing the production and in turn the national income of the country. The expenditures incurred by the government that do not directly help in economic development or production can be termed as the non developmental expenditures. It includes the cost of tax collection, the cost of printing notes, the expenses for maintaining the law and order of a country, the expenditure on Defence etc.

[Data Source: Budget documents of the Government of India and the State Governments]

- **Net Bank Credited to Government in Billion Rupees** (X_{12})**:**

Net bank credit to Government comprise the RBI's net credit to Central and State Governments and commercial and co-operative banks' investments in Central and State Government securities. Bank credit to commercial sector include RBI's and other bank's credit to commercial sector.

[Data Source: RBI]

- **Investment by LIC in Billion Rupees** (X_{13})**:**

This factor plays an important role in increase in Indian Development.

[Data Source: Life Insurance Corporation of India]

- **Combined Liabilities of the Central and State Goverment in Billion Rupees** (X_{14})**:**

These include repayments of sovereign debt, budget expenditures for the current fiscal year, and longer-term expenditures for legally mandated obligations (such as civil service salaries and pensions and, in some countries, the overall social security system).

[Data Source: Budget documents of the Government of India and the State Governments]

- **Export of principal commodities in Billion Rupees** (X_{15})**:**

India's major Exports are mainly the Petroleum products, Gems, Jewelleries, machineries, tea, coffee, tobacco, iron steel etc. The total income from exporting affects the Indian economy to a remarkable amount.

[Data Source: Directorate General of Commercial Intelligence and Statistics]

- **Import of principal commodities in Billion Rupees** (X_{16})**:**

The most important products that are imported to India are crude oil, gold, solid oil,diamonds etc. Not only that some major factors of production like machineries are imported so that a good quality production can be possible.

[Data Source: Directorate General of Commercial Intelligence and Statistics]

- **Foreign direct Investment Inflows in Billion Rupees** (X_{17})**:**

FDI net inflows are the value of inward direct investment made by non-resident investors in the reporting economy, including reinvested earnings and intra-company loans, net of repatriation of capital and repayment of loans.

[Data Source: RBI and World Bank]

- **Foreign Exchange Reserves in terms of Gold, Foreign Currency Assets, Reserve Tranche Position in Billion Rupees** (X_{18})**:**

Foreign exchange reserves are assets denominated in a foreign currency that are held by a central bank. These may include foreign currencies, bonds, treasury bills, and other Government Securities.

[Data Source: RBI]

- **Net Inflow of Aid in Crore Rupees** (X_{19})**:**

It is defined as foreign and as well as domestic aid designed to promote the economic development and welfare of developing countries. Loans and credits for military purposes are excluded.

[Data Source: Controller of Aid, Accounts and Audit, Ministry of Finance, Government of India.]

- **Currency in Circulation in Billion Rupees** (X_{20})**:**

Currency in circulation is all of the money that has been issued by a country's monetary authority, minus cash that has been removed from the system. Currency in circulation represents part of the overall money supply, with a portion of the overall supply being stored in checking and savings accounts.

[Data Source: RBI]

3 Description of Multiple Linear Regression:

3.1 Model:

Given a dataset of n observations having p regressors the MLR model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \ ; \forall i = 1(1)n$$

Where, ϵ be the error term in the model. with the following assumptions;

$$E(\epsilon_i) = 0 \ ; \forall i = 1(1)n$$

$$Var(\epsilon_i) = \sigma^2 \ ; \forall i = 1(1)n$$

$$Cov(\epsilon_i, \epsilon_j) = 0 \ ; \forall i \neq j$$

3.2 Normal equations:

We can write the above stated MLR equations in the matrix form as follows:

$$Y = X\beta + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{2p} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We want to find the estimate of β from the given data. We will apply the least squares technique to obtain the estimates.. The technique involves minimizing the Sum of Squares of errors with respect to β i.e. to minimize the following function:

$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^{'} \epsilon = (Y - X\beta)^{'}(Y - X\beta)$

Differentiating the above equations with respect to β, we get the Least Squares Normal Equations of out MLR model, given as:

$$X^{'}X\beta = X^{'}Y$$

Thus the least squares estimates of out MLR model is given by:

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$$

provided $(X^{'}X)^{-1}$ exists.

4 Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. In order to create a reliable dataset we need to adopt Data Cleaning method so that we can increase the quality of our data set. In our study we will go through following steps for cleaning our data:

i) Missing value Treatment

ii) Duplicate data Treatment

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

4.1 Loading Data and Treatment of Missing and Duplicate Values:

```
data = pd.read_csv("Project_dataset.csv")
data.head()
```

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
0	1474	129.59	867.52	35	22	1522.44	197.63	109.37	120.12	28.11	...	378.79	257.18	60.205	698.10	67.11	125.49	0.63	55.45	1281	143.07
1	1728	133.30	927.71	50	22	1719.17	240.36	129.95	127.28	35.57	...	444.79	306.33	68.944	804.33	78.06	136.18	0.78	40.25	964	154.11
2	1926	129.52	905.44	57	20	1722.54	271.38	150.55	156.13	41.99	...	520.57	352.57	78.350	978.16	88.03	142.93	0.66	47.82	1258	176.39
3	2241	152.37	887.90	61	21	1858.47	312.02	178.11	193.89	45.89	...	599.89	406.42	88.576	1111.22	97.71	158.31	0.10	59.72	1338	206.43
4	2508	145.54	1035.07	69	19	1983.92	357.65	217.27	256.15	48.72	...	716.54	503.43	99.542	1328.59	117.44	171.34	0.23	72.43	1452	238.75

4.1.1 Missing Value Treatment:

```
# Count no. of missing values per column

missing_values_count=data.isnull().sum()

missing_values_count
```

```
Y      0
X1     0
X2     0
X3     0
X4     0
X5     0
X6     0
X7     0
X8     0
X9     0
X10    0
X11    0
X12    0
X13    0
X14    0
X15    0
X16    0
X17    0
X18    0
X19    0
X20    0
dtype: int64
```

From the above analysis, we can see that our dataset does not contain any missing values.

4.1.2 Detection and Removal of Duplicate values:

```
#Detection of duplicate data

data.duplicated().sum()
```

Output: 0

Hence, our dataset also do not contain any duplicate values. So we are ready for further analysis using our cleaned data.

1 ORDINARY LEAST SQUARE FITTING

Test for significance of the regressors:

$H_0 : \beta_1 = ... = \beta_{20} = 0$ against
H_1: at least $\beta_j \neq 0$; at least one j

```
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 + X19 +
    X20, data = regression_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1112.10  -220.87   -18.69   146.74  1186.94

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.918e+03  2.466e+03  -1.183 0.250553
X1          -3.720e+00  1.280e+01  -0.291 0.774306
X2           5.602e+00  3.412e+00   1.642 0.116285
X3          -2.407e+01  1.585e+01  -1.519 0.144350
X4           1.225e+01  2.811e+01   0.436 0.667763
X5          -2.102e-01  2.400e-01  -0.876 0.391430
X6           1.164e+00  1.072e+00   1.086 0.290375
X7          -1.353e+00  5.169e-01  -2.618 0.016464 *
X8          -1.577e+00  1.192e+00  -1.323 0.200812
X9           6.993e-01  9.606e-01   0.728 0.475021
X10         -3.475e+00  1.118e+00  -3.108 0.005538 **
X11          3.754e+00  1.284e+00   2.923 0.008413 **
X12         -8.370e-01  8.409e-01  -0.995 0.331466
X13          3.256e+00  8.873e-01   3.670 0.001521 **
X14          2.040e-01  2.320e-01   0.879 0.389768
X15          1.986e+00  8.315e-01   2.388 0.026914 *
X16         -3.917e-01  3.605e-01  -1.087 0.290137
X17          1.444e-01  1.227e+00   0.118 0.907488
X18         -1.480e+00  3.319e-01  -4.458 0.000241 ***
X19         -2.943e-02  3.151e-02  -0.934 0.361443
X20          1.296e+00  7.053e-01   1.838 0.080986 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 659.3 on 20 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 1.71e+04 on 20 and 20 DF,  p-value: < 2.2e-16
```

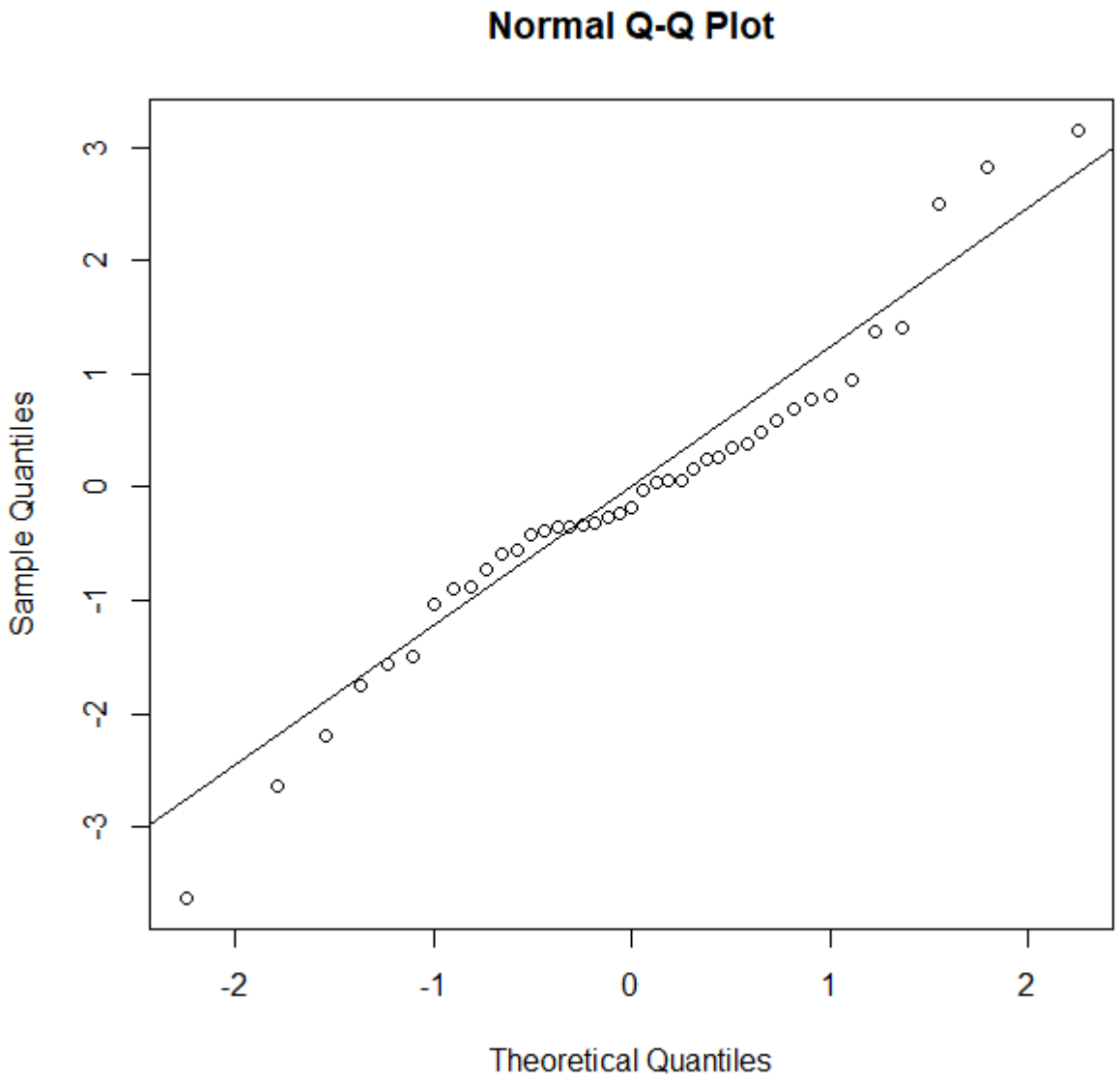
F statistics = 1.71e+04
p-value = 2.2e-16 < 0.05,
so, we reject the null hypothesis at 5% level of significance and conclude on the basis of the given data that all the variables are not insignificant in explaining the GNI data.

Standardized the Residuals and Inspection of the Normality Assumption of Errors

Inspection of the Normality Assumption of Errors

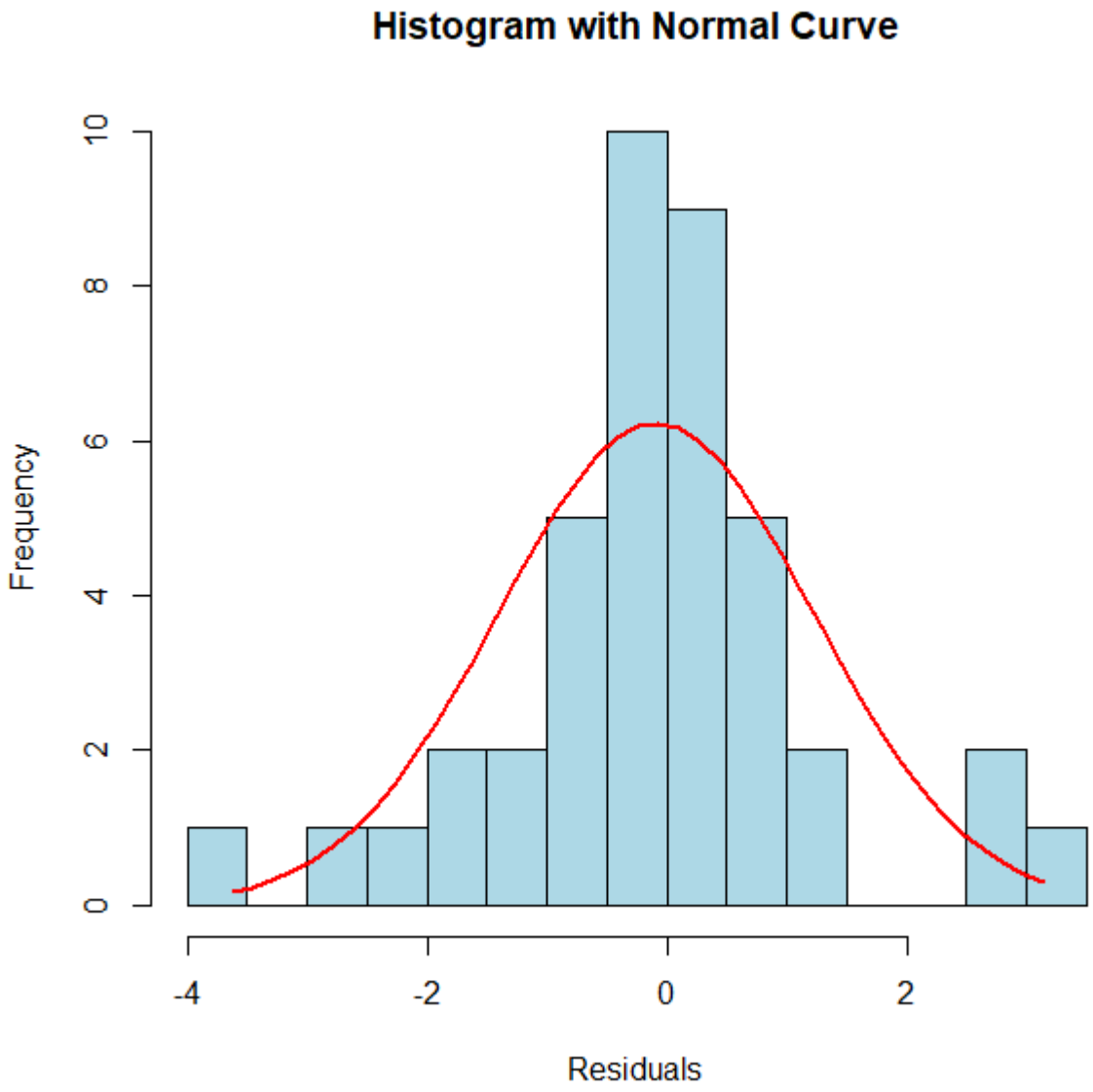
Q-Q Plot

In this method we would plot the the ordered residuals $e_{(i)}$ against $\phi^{-1}(\frac{i-0.5}{n})$, for i=1,2,...,n.If the errors are truely from Normal Distribution then the plot will be nearly a straight line.



The above Q-Q plot yields almost a straight line.So, it can be concluded that the residuals can be assumed to follow a Normal Distribution which supports our assumption. But we will use others methods to get check our assumption.

Histogram Approach



The histogram of Residuals is not significantly different from a Normal Curve. From here we could have concluded that our normality assumption for errors hold, but we will apply Shapiro-Wilk Test for Normality to get the final conclusion.

Shapiro-Wilk Test for Normality:

Here the null hypothesis is,
H_0:Errors are normally distributed
The Alternative hypothesis is
$H_1 : H_0$ is not true
The test Statistic is: $W = \frac{\sum_{i=1}^{n} a_i e_{\hat{(i)}}}{\sum_{i=1}^{n} (\hat{e_i} - \bar{e})^2}$
Here, $\hat{e_i}$ are the i^{th} fitted residual.
$e_{\hat{(i)}}$ is the ith order statistic.
$\bar{e}$ is the sample mean
$(a_1,a_2,...,a_n) = \frac{m^T V^{-1}}{C}$

And, $C = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$. Here m is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics. If p-value is greater than chosen level of significance, null hypothesis is accepted(i.e.distribution of error is not significantly different from a normal population).

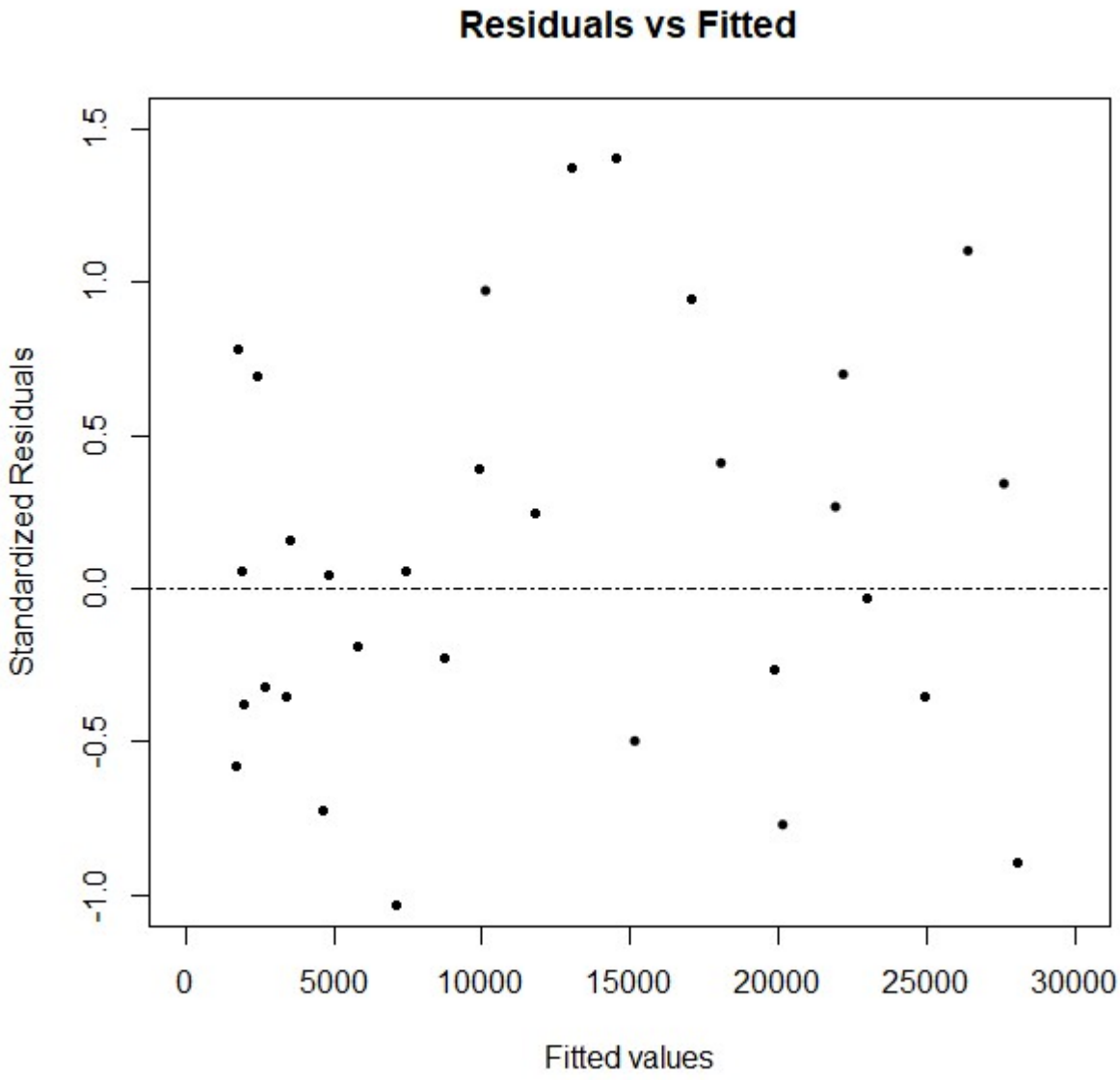
```
        Shapiro-Wilk normality test

data:  Z.stdres
W = 0.95871, p-value = 0.1417
```

Test statistic, W = 0.95871, p-value = 0.1417¿0.05(α)
So we fail to reject the null hypothesis at 5% level of significance and conclude on the basis of the given data that the distribution of errors is not significantly different from Normal Distribution. So, our assumption is true.

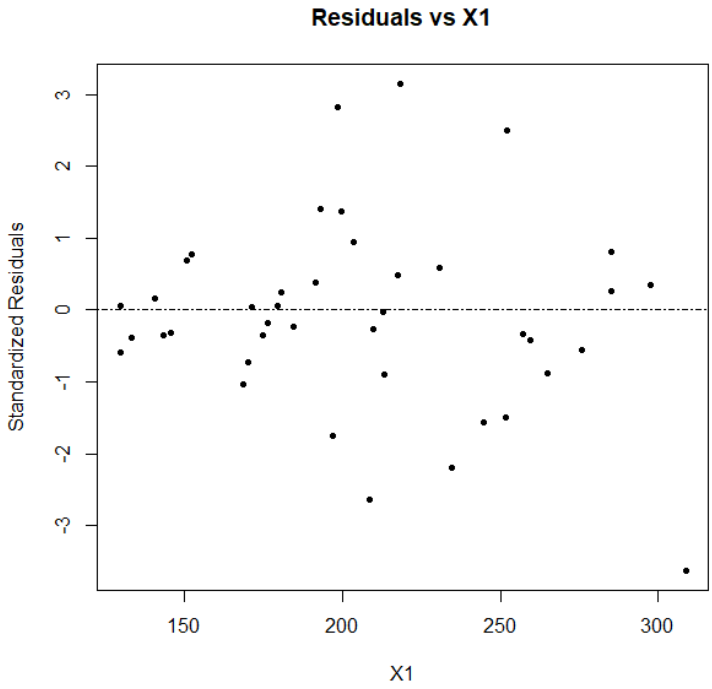
Inspection of Homoscedastic Assumptions of Errors

Residuals vs Fitted Plot

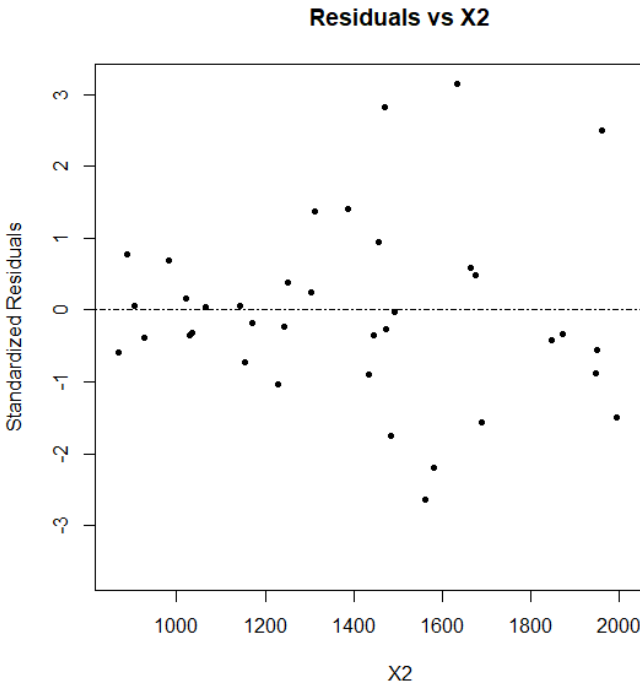
Here we plot the residuals against the fitted responses. If the errors are homoscedastic then we would expect a horizontal band and completely random pattern around $\hat{e_i} = 0$ line. If any pattern is detected this will indicate that the variances may be non constant.



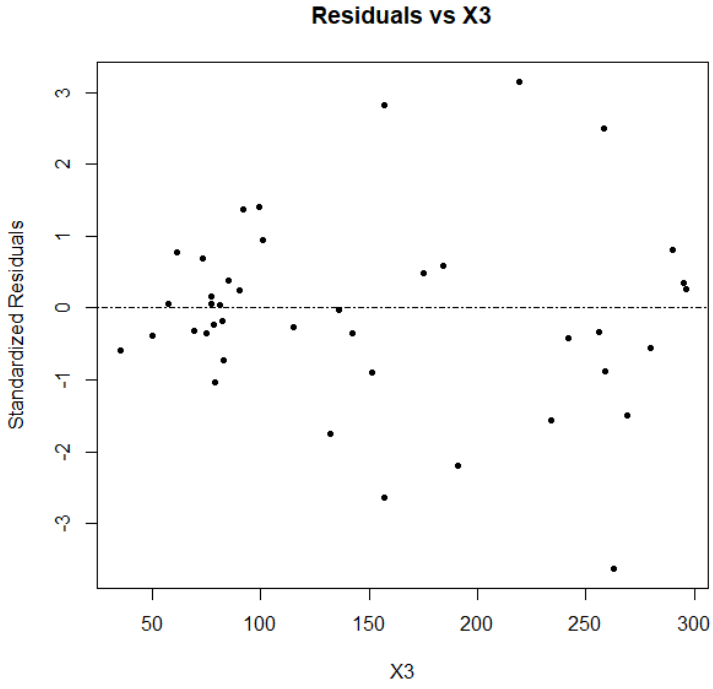
We have plotted the Residuals against the Fitted responses. We obtained a more or less random pattern among the residuals about the horizontal band. So we can conclude that, the assumption based on homoscedasticity is true in our Model.



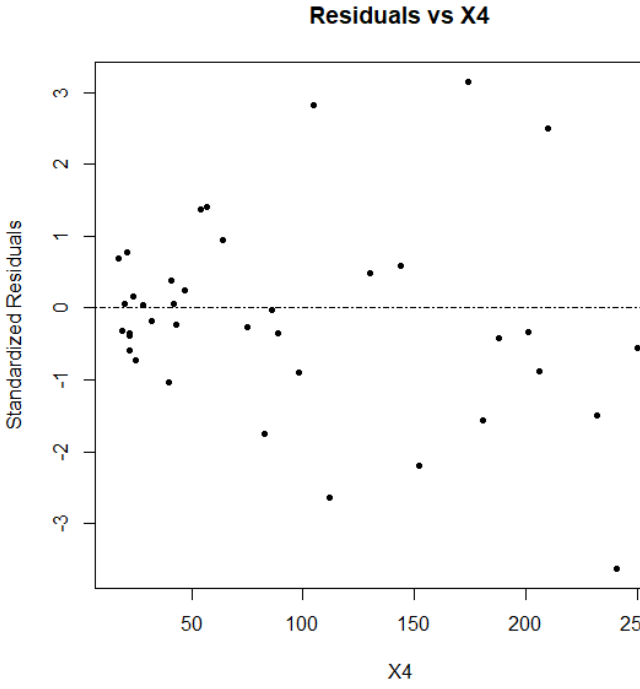
(a) X1



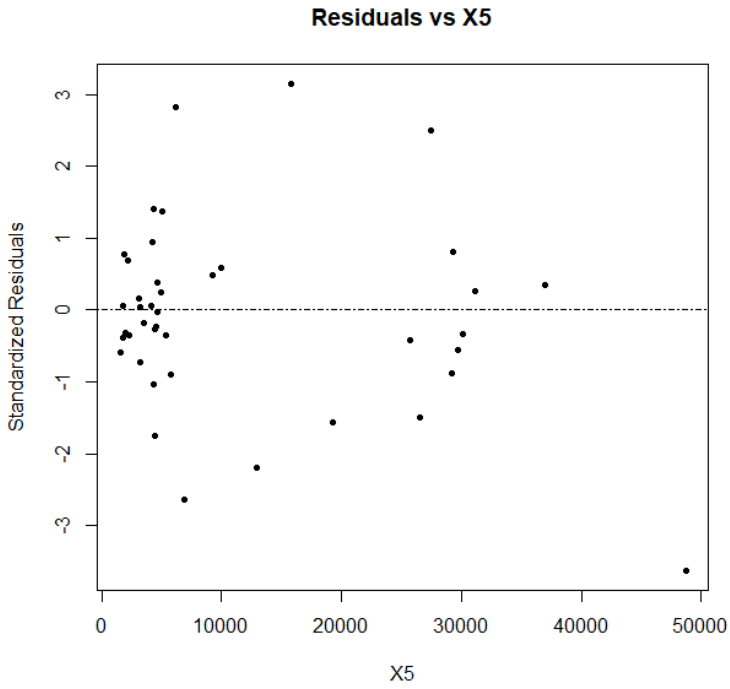
(b) X2



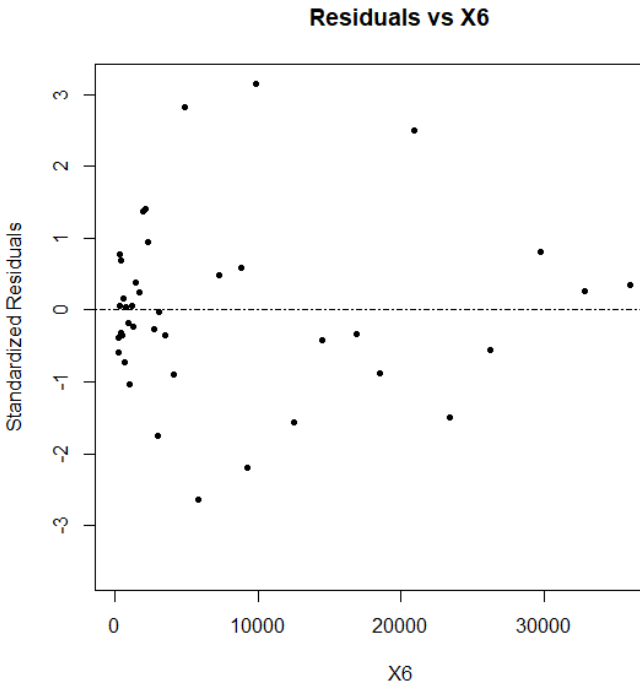
(a) X3



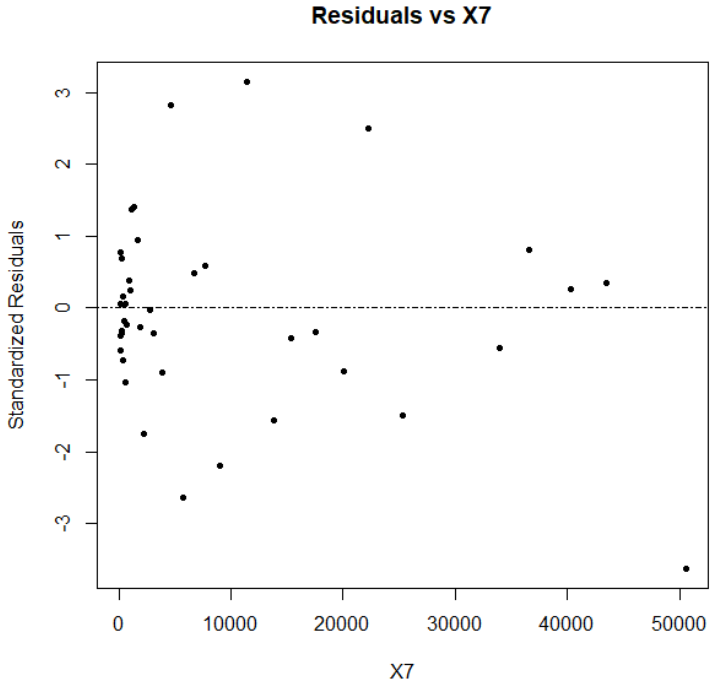
(b) X4



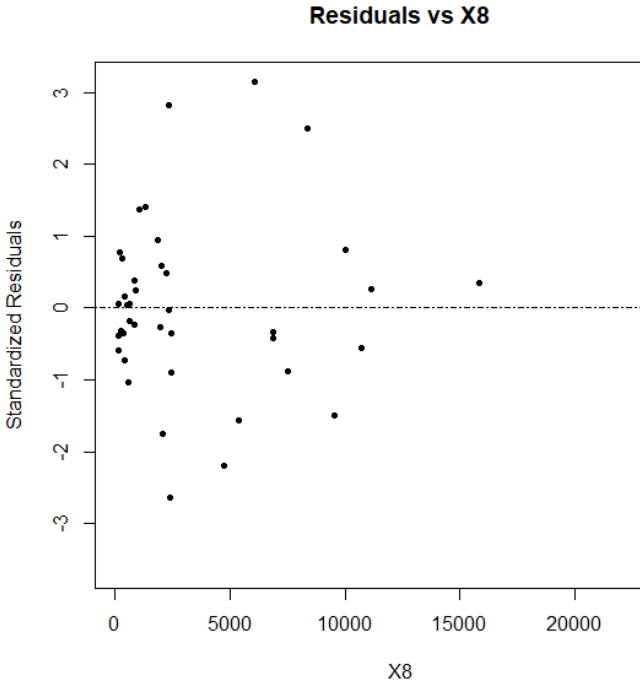
(a) X5



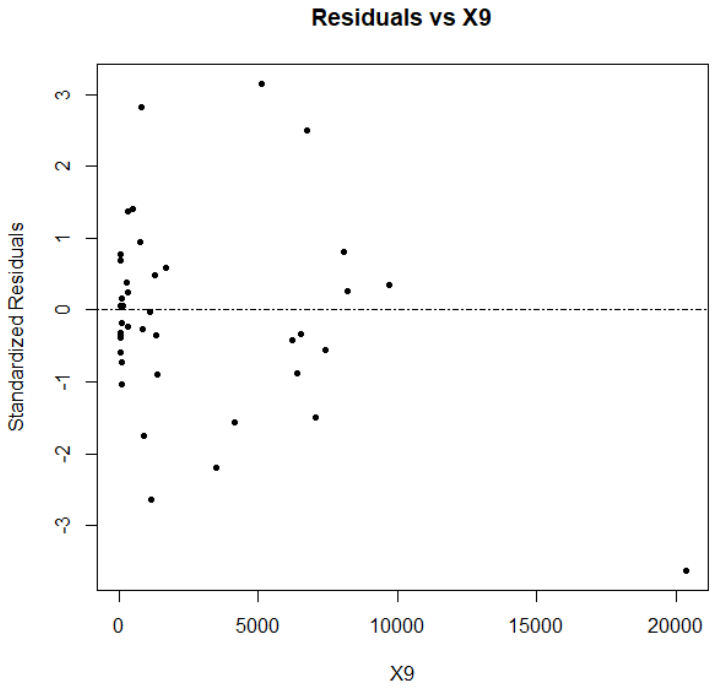
(b) X6



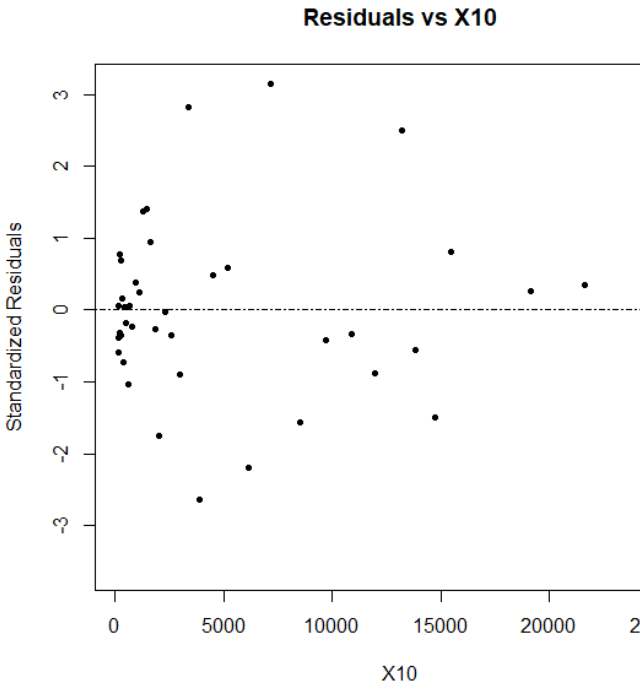
(a) X7



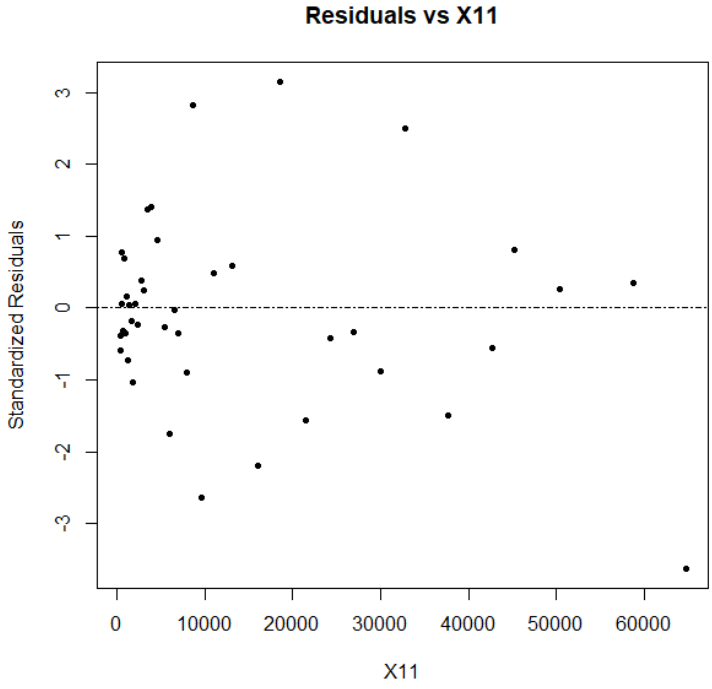
(b) X8



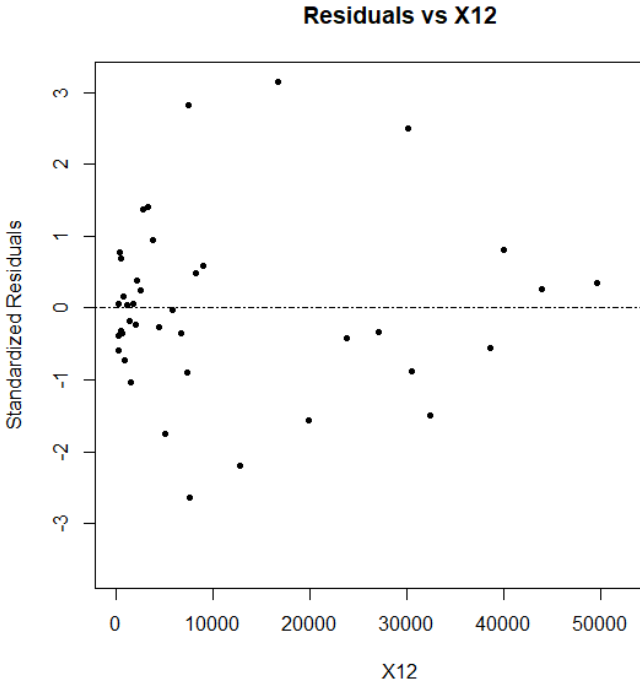
(a) X9



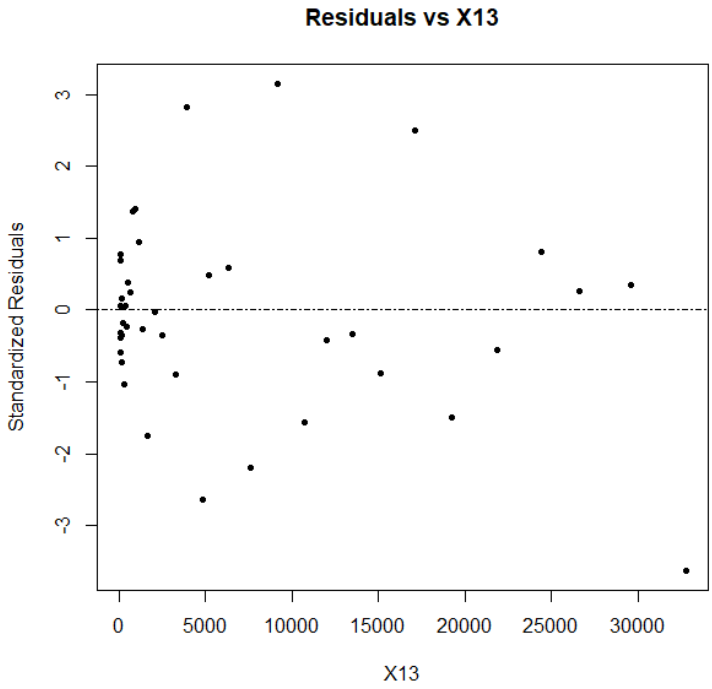
(b) X10



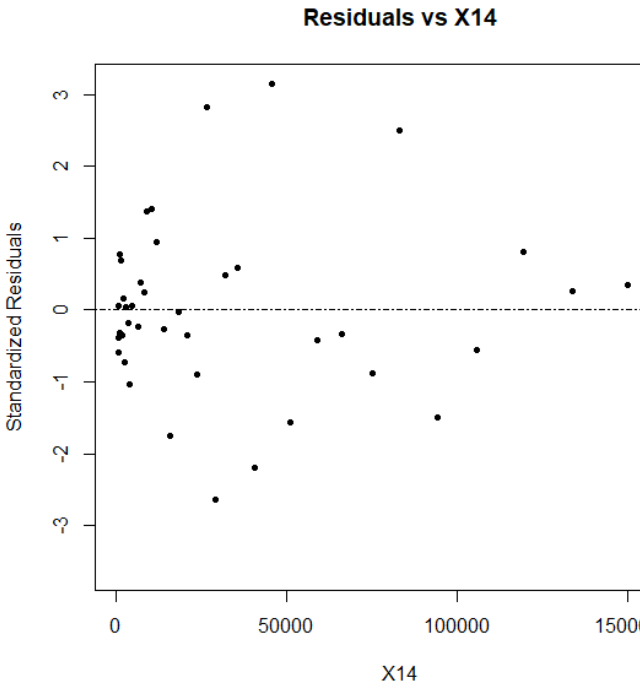
(a) X11



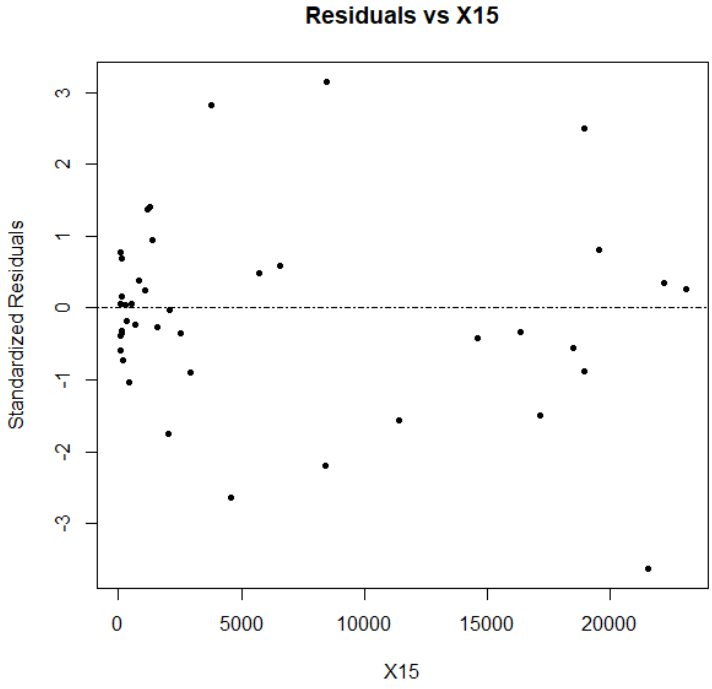
(b) X12



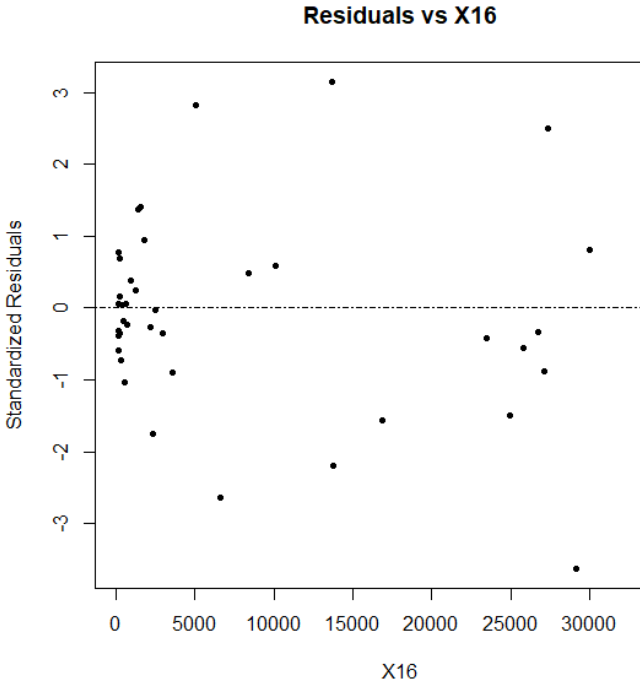
(a) X13



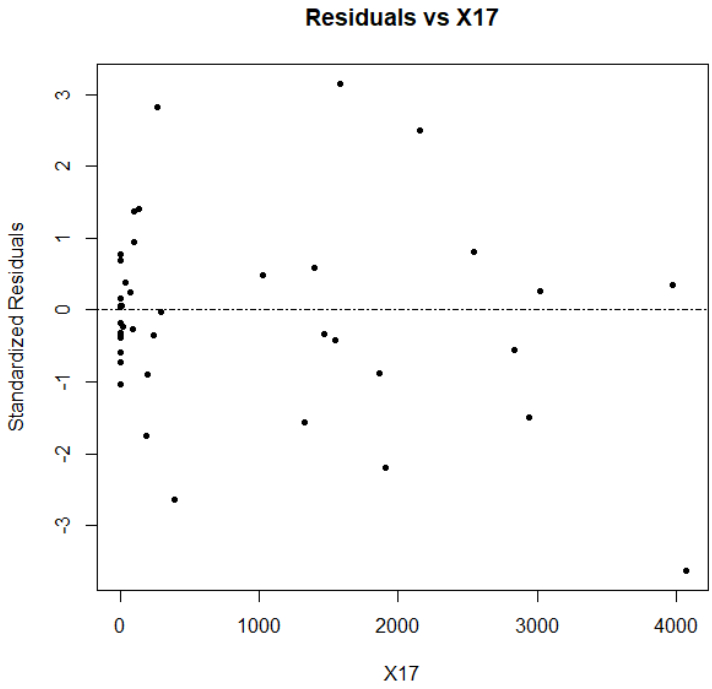
(b) X14



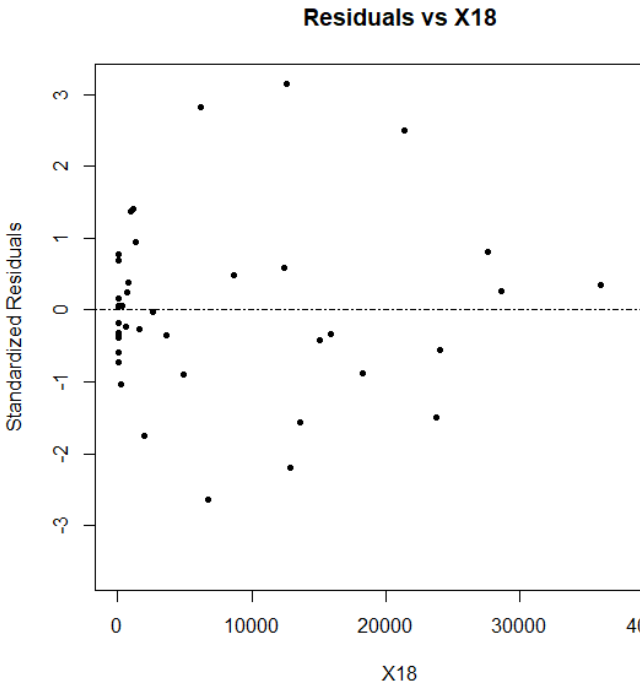
(a) X15



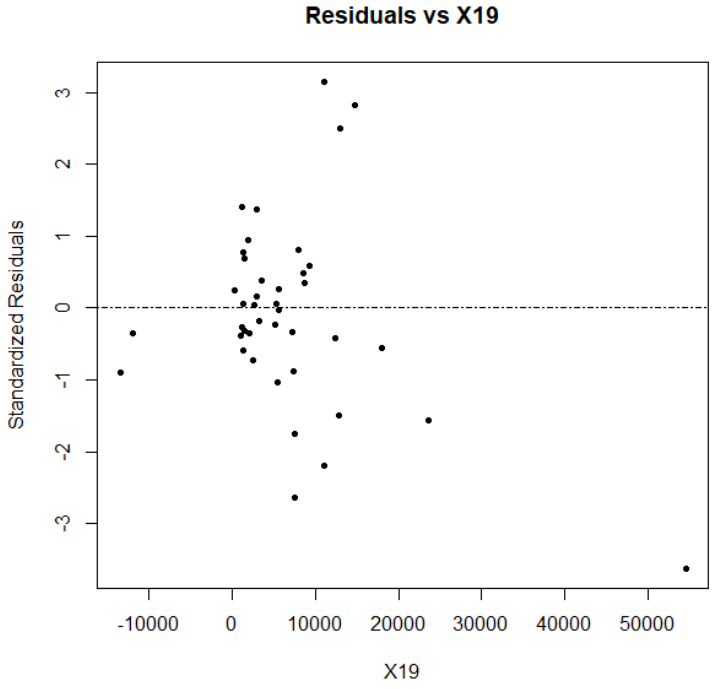
(b) X16



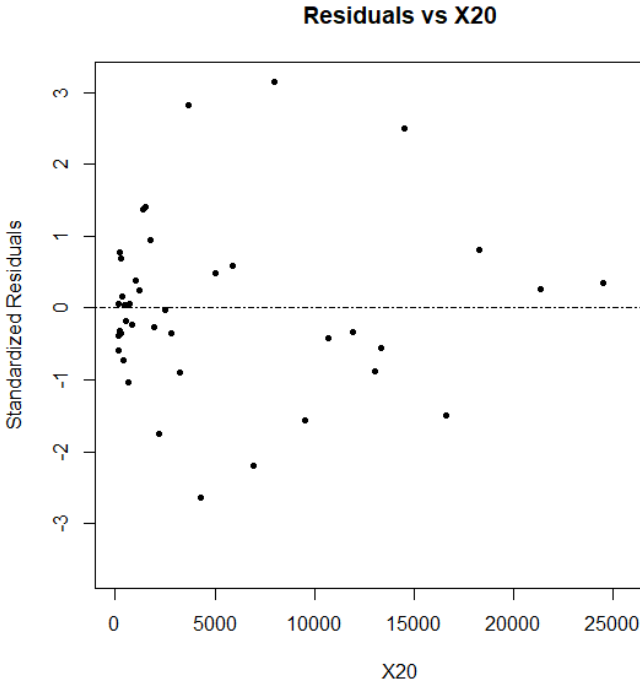
(a) X17



(b) X18



(a) X19



(b) X20

Hence we can observe that the residual v/s regressor plot for each regressor exhibits random behaviour which supports our previous conlusion about homoscedasticity of errors.

Breusch-Pagan Test for Heteroscedasticity

The Breusch-Pagan-Godfrey Test (sometimes shorted to the Breusch-Pagan test) is a test for heteroscedasticity of errors in regression. Heteroscedasticity means "differently scattered"; this is opposite to homoscedastic, which means "same scatter." Homoscedasticity in regression is an important assumption; if the assumption is violated, you won't be able to use regression analysis.
The test statistic for the Breusch-Pagan-Godfrey test is:
N * R^2 (with k degrees of freedom) Where:
n = sample size
R^2 = (Coefficient of Determination) of the regression of squared residuals from the original regression.
k = number of independent variables.
The test statistic approximately follows a chi-square distribution.

The null hypothesis for this test is that the error variances are all equal.
The alternate hypothesis is that the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease).
A small chi-square value (along with an associated small p-value) indicates the null hypothesis is true (i.e. that the variances are all equal).

Note that the Breush-Pagan test measures how errors increase across the explanatory variable, Y. The test assumes the error variances are due to a linear function of one or more explanatory variables in the model. That means heteroskedasticity could still be present in your regression model, but those errors (if present) are not correlated.

```
        studentized Breusch-Pagan test

data:  Z
BP = 24.972, df = 20, p-value = 0.2025
```

So, p-value is 0.2025
So, we fail to reject the null hypothesis at 5% level of significance and conclude on the basis of the given data the distribution of errors is not heteroscedastic. So, our assumption is true.

Inspection of Autocorrelation among the Errors

For our dataset, n=41, p=10, α=0.05
Hence we would like to test,
H_0: $\rho = 0$ against : $\rho > 0$ using Durbin-Watson test to check the existance of serial correlation in the data.

```
        Durbin-Watson test

data:  Z
DW = 2.4789, p-value = 0.5369
alternative hypothesis: true autocorrelation is greater than 0
```

In our model the value of Durbin-Watson Statistic is d=2.4789. It indicates that there may exist a small negative correlation among the errors as $2 < 2.4789 < 4$.
But the p-value of the test is $0.5369 > 0.05(\alpha)$, so we fail to reject the null hypothesis and conclude on the basis of the given data that the errors in our new model are independent.

Multicollinearity

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related. There can be more than one reason behind multicollinearity, such as:

- The data collection method employed
- Model specification using too many regressors
- An over-defined model etc.

The consequences of multicollinearity being present in the model can be severe. When one or more regressors are linearly related with each other, the design matrix becomes ill-conditioned producing regression coefficients with large standard errors which can potentially damage the prediction capability of the model. There can be other problems like significant variable becoming insignificant one or regression coefficients appearing with wrong signs from what is expected.

Detection

There are several methods for knowing the presence of multicollinearity in the model. One such method is to calculate the VIFs of the model.
VIF or Variance Inflation Factor for the j-th regressor is defined as:
$VIF_j = \frac{1}{1-R_j^2}$, j=1(1)p
Where R_j^2 is the multiple R_j^2 obtained from regressing X_j on other regressors. The VIF value of 5 or more is an indicator of multicollinearity. Large values of VIF indicate multicollinearity leading to poor estimates of associated regression coefficients.
We started our initial analysis with 20 regressors. So there is a high likelihood of multicollinearity being present the preliminary model.

```
vif(Z)
```

```
##          X1          X2          X3          X4          X5          X6
##   36.065715   159.090803   158.892103   516.412830   805.333713 13048.433985
##          X7          X8          X9         X10         X11         X12
##  4647.587622  3396.993754  1421.568137  5189.343300 47101.759326 16252.230524
##         X13         X14         X15         X16         X17         X18
##  6274.528562 10060.011682  3938.029479  1666.822734   198.491496  1266.896867
##         X19         X20
##    9.359149  2535.671589
```

Multicollinearity Diagnostics with Variance Decomposition

After knowing the presence of multicollinearity in our model, we would like to know the group(s) of variables responsible for it. For doing this we can use Variance Decomposition Method.
Variance Decomposition Method is a method to identify subsets that are involved in multi-collinearity. Variance decomposition proportions, defined as
$\pi_{kj} = \frac{\frac{v_{kj}^2}{l_k}}{\sum_{k=1}^{p} \frac{v_{kj}^2}{l_k}}$,$\forall$ k, j=1(1)p
where, l_1,l_2,...,l_p are eigen values of X^TX and v_1,v_2,...,v_p are corresponding orthonormal eigen vectors and $v_j = (v_{j1},v_{j2},...,v_{jp})^T$, j=1(1)p.
Now a variance decomposition table is formed with the π_{kj} values along with a column containing the corresponding condition indices arranged in ascending order. So, large proportion in a row corresponding to the maximum condition index indicates the presence of multicollinearity among the corresponding regressors.

Step 1:

```
Call:
eigprop(mod = Z)

   Eigenvalues         CI (Intercept)     X1     X2     X3     X4     X5     X6
1      18.8749     1.0000       0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.4246     3.6399       0.0004 0.0002 0.0001 0.0001 0.0000 0.0000 0.0000
3       0.4632     6.3837       0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4       0.1109    13.0452       0.0011 0.0001 0.0000 0.0008 0.0007 0.0001 0.0000
5       0.0523    18.9974       0.0000 0.0000 0.0000 0.0000 0.0002 0.0025 0.0000
6       0.0303    24.9641       0.0028 0.0000 0.0000 0.0027 0.0027 0.0001 0.0001
7       0.0192    31.3939       0.0150 0.0001 0.0003 0.0070 0.0058 0.0013 0.0000
8       0.0088    46.2121       0.0004 0.0000 0.0008 0.0008 0.0022 0.0022 0.0000
9       0.0051    60.6483       0.0341 0.0211 0.0020 0.0112 0.0015 0.0153 0.0000
10      0.0030    78.6949       0.0813 0.1019 0.0045 0.0316 0.0000 0.0138 0.0008
11      0.0025    87.5832       0.0184 0.0138 0.0126 0.0231 0.0000 0.0781 0.0000
12      0.0015   112.7880       0.0006 0.0013 0.0116 0.0652 0.0016 0.0183 0.0000
13      0.0013   121.6536       0.0017 0.2751 0.0692 0.0224 0.0030 0.0285 0.0021
14      0.0011   132.8060       0.0833 0.0540 0.0007 0.5565 0.3003 0.0071 0.0000
15      0.0006   173.6441       0.0003 0.3288 0.1103 0.0003 0.0180 0.0071 0.0001
16      0.0003   250.7635       0.0246 0.0059 0.0493 0.0051 0.0246 0.0201 0.0018
17      0.0001   369.7304       0.0273 0.0460 0.0006 0.0046 0.0097 0.0007 0.0343
18      0.0001   397.3837       0.2618 0.0007 0.2179 0.0673 0.0003 0.0430 0.1776
19      0.0001   525.7086       0.0179 0.0854 0.1138 0.0094 0.0024 0.2003 0.0000
20      0.0000   704.3006       0.2042 0.0257 0.3096 0.1908 0.4364 0.1573 0.0783
21      0.0000  1586.5912       0.2247 0.0401 0.0967 0.0009 0.1906 0.4042 0.7049
```

```
       X7     X8     X9    X10    X11    X12    X13    X14    X15    X16    X17
1  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
3  0.0000 0.0000 0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4  0.0001 0.0002 0.0003 0.0000 0.0000 0.0000 0.0000 0.0000 0.0002 0.0006 0.0002
5  0.0001 0.0001 0.0025 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0003 0.0109
6  0.0007 0.0005 0.0013 0.0000 0.0000 0.0000 0.0001 0.0000 0.0002 0.0004 0.0224
7  0.0001 0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0009 0.0616
8  0.0016 0.0002 0.0006 0.0019 0.0000 0.0002 0.0001 0.0000 0.0000 0.0004 0.0242
9  0.0009 0.0041 0.0235 0.0005 0.0001 0.0002 0.0007 0.0000 0.0001 0.0008 0.0155
10 0.0008 0.0085 0.0077 0.0013 0.0000 0.0006 0.0001 0.0000 0.0016 0.0000 0.0000
11 0.0006 0.0000 0.0170 0.0001 0.0000 0.0001 0.0002 0.0001 0.0019 0.0100 0.0490
12 0.0026 0.0028 0.0001 0.0000 0.0001 0.0000 0.0003 0.0005 0.0404 0.0137 0.0264
13 0.0000 0.0165 0.0049 0.0017 0.0000 0.0000 0.0000 0.0000 0.0006 0.0487 0.0498
14 0.0010 0.0002 0.0000 0.0000 0.0000 0.0000 0.0001 0.0000 0.0020 0.0001 0.0149
15 0.0012 0.0311 0.0127 0.0053 0.0000 0.0053 0.0010 0.0044 0.0199 0.1253 0.0186
16 0.0335 0.0041 0.0002 0.1906 0.0009 0.0098 0.0286 0.0112 0.0003 0.0001 0.0020
17 0.0274 0.0396 0.0003 0.0446 0.0035 0.0594 0.1888 0.0934 0.0013 0.0005 0.2053
18 0.2006 0.0027 0.0259 0.0023 0.0046 0.0024 0.1334 0.0179 0.0133 0.0241 0.0089
19 0.1922 0.0005 0.0009 0.0322 0.0258 0.0197 0.3204 0.3347 0.0075 0.0488 0.2021
20 0.5028 0.3579 0.1012 0.2816 0.0158 0.1736 0.2423 0.0110 0.3772 0.2813 0.0083
21 0.0339 0.5309 0.8006 0.4379 0.9492 0.7287 0.0836 0.5267 0.5332 0.4438 0.2797
```

```
      X18    X19    X20
1  0.0000 0.0001 0.0000
2  0.0000 0.0001 0.0000
3  0.0000 0.1246 0.0000
4  0.0000 0.0881 0.0000
5  0.0004 0.0409 0.0000
6  0.0007 0.0707 0.0000
7  0.0002 0.0072 0.0000
8  0.0026 0.0072 0.0125
9  0.0084 0.0048 0.0000
10 0.0142 0.0279 0.0003
11 0.0572 0.0096 0.0000
12 0.0617 0.0045 0.0196
13 0.0199 0.0008 0.0004
14 0.0023 0.0002 0.0017
15 0.0161 0.0083 0.0167
16 0.0037 0.0319 0.0961
17 0.0179 0.1559 0.0748
18 0.0101 0.1120 0.0369
19 0.4275 0.0254 0.0001
20 0.1362 0.1985 0.6720
21 0.2210 0.0812 0.0689

================================
Row 14==> X3, proportion 0.556520 >= 0.50
Row 21==> X6, proportion 0.704882 >= 0.50
Row 20==> X7, proportion 0.502775 >= 0.50
Row 21==> X8, proportion 0.530888 >= 0.50
Row 21==> X9, proportion 0.800570 >= 0.50
Row 21==> X11, proportion 0.949243 >= 0.50
Row 21==> X12, proportion 0.728736 >= 0.50
Row 21==> X14, proportion 0.526703 >= 0.50
Row 21==> X15, proportion 0.533200 >= 0.50
Row 20==> X20, proportion 0.672016 >= 0.50
```

- So the subsets (X3), (X7,X20) and (X6,X8,X9,X11,X12,X14,X15) are involved in Multicollinearity.

- In the first subset VIF of X3 is the highest, in the second subset the VIF of X20 is highest and in the third subset the VIF of X11 is highest.

- We drop the variables X3,X20 and X11, and again fit a model.

Step 2:

```
> eigprop(olsreg_1)

Call:
eigprop(mod = olsreg_1)

   Eigenvalues       CI (Intercept)     X1     X2     X4     X5     X6     X7
1      16.0339   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.2953   3.5183      0.0007 0.0003 0.0001 0.0000 0.0000 0.0000 0.0000
3       0.4547   5.9381      0.0000 0.0000 0.0000 0.0001 0.0000 0.0000 0.0000
4       0.1017  12.5578      0.0010 0.0000 0.0000 0.0016 0.0001 0.0000 0.0001
5       0.0515  17.6460      0.0000 0.0000 0.0000 0.0003 0.0044 0.0001 0.0003
6       0.0282  23.8318      0.0021 0.0000 0.0001 0.0061 0.0000 0.0004 0.0018
7       0.0167  30.9865      0.0293 0.0010 0.0017 0.0288 0.0025 0.0000 0.0000
8       0.0049  57.4068      0.0196 0.0100 0.0001 0.0002 0.0363 0.0000 0.0010
9       0.0046  59.2326      0.0234 0.0054 0.0047 0.0508 0.0022 0.0002 0.0081
10      0.0028  75.0939      0.1027 0.0889 0.0006 0.0090 0.0783 0.0025 0.0015
11      0.0023  83.0985      0.1008 0.0508 0.0218 0.0353 0.0868 0.0011 0.0023
12      0.0013 111.9366      0.0001 0.2901 0.1148 0.0001 0.0295 0.0063 0.0004
13      0.0011 123.4498      0.0134 0.1118 0.0026 0.0658 0.0643 0.0008 0.0003
14      0.0006 170.0100      0.0060 0.3258 0.2127 0.0745 0.0021 0.0000 0.0024
15      0.0002 295.6720      0.0687 0.0398 0.0154 0.0348 0.0624 0.1184 0.3681
16      0.0001 353.1828      0.4437 0.0080 0.2498 0.0221 0.0373 0.3482 0.2598
17      0.0001 392.4441      0.1581 0.0028 0.2487 0.6296 0.0568 0.3542 0.0247
18      0.0001 533.9000      0.0305 0.0652 0.1271 0.0410 0.5368 0.1679 0.3293
```

```
       X8     X9    X10    X12    X13    X14    X15    X16    X17    X18    X19
1  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0002
2  0.0000 0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0000 0.0001
3  0.0000 0.0004 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0000 0.0000 0.1505
4  0.0012 0.0030 0.0000 0.0000 0.0000 0.0000 0.0010 0.0020 0.0002 0.0000 0.1139
5  0.0003 0.0179 0.0000 0.0000 0.0001 0.0001 0.0006 0.0010 0.0195 0.0006 0.0413
6  0.0015 0.0096 0.0000 0.0001 0.0002 0.0001 0.0003 0.0005 0.0539 0.0010 0.0845
7  0.0009 0.0001 0.0000 0.0001 0.0000 0.0003 0.0003 0.0024 0.0819 0.0000 0.0035
8  0.0139 0.1019 0.0058 0.0029 0.0002 0.0001 0.0000 0.0066 0.0921 0.0351 0.0012
9  0.0073 0.1235 0.0405 0.0003 0.0008 0.0005 0.0011 0.0001 0.0085 0.0033 0.0075
10 0.0363 0.0156 0.0090 0.0019 0.0001 0.0000 0.0184 0.0033 0.0000 0.0287 0.0368
11 0.0026 0.1505 0.0005 0.0014 0.0007 0.0001 0.0019 0.0267 0.1052 0.0654 0.0043
12 0.0509 0.0360 0.0026 0.0003 0.0001 0.0000 0.0321 0.1898 0.0433 0.0071 0.0027
13 0.0071 0.0009 0.0791 0.0037 0.0004 0.0002 0.2749 0.0920 0.0036 0.0661 0.0064
14 0.1645 0.1087 0.3134 0.0078 0.0010 0.0031 0.0120 0.2237 0.0175 0.0471 0.0143
15 0.0143 0.0007 0.0845 0.0223 0.3236 0.0584 0.0661 0.0138 0.0051 0.0114 0.0053
16 0.0287 0.1095 0.1856 0.0848 0.2645 0.1114 0.0446 0.0446 0.0259 0.0106 0.2552
17 0.5915 0.0817 0.0003 0.6221 0.0022 0.0302 0.3923 0.1197 0.3137 0.1145 0.2282
18 0.0787 0.2400 0.2786 0.2522 0.4060 0.7955 0.1544 0.2737 0.2294 0.6088 0.0442

================================
Row 17==> X4, proportion 0.629561 >= 0.50
Row 18==> X5, proportion 0.536817 >= 0.50
Row 17==> X8, proportion 0.591523 >= 0.50
Row 17==> X12, proportion 0.622107 >= 0.50
Row 18==> X14, proportion 0.795526 >= 0.50
Row 18==> X18, proportion 0.608850 >= 0.50
```

- So the subsets (X4,X8,X12) and (X5,X14,X18) are involved in Multicollinearity.

- In the first subset VIF of X4 is the highest and in the second subset the VIF of X14 is high-

est.

- We drop the variables X4 and X14, and again fit a model.

Step 3:

```
> eigprop(olsreg_2)

Call:
eigprop(mod = olsreg_2)

   Eigenvalues       CI (Intercept)     X1     X2     X5     X6     X7     X8
1      14.1141   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.2586   3.3487      0.0019 0.0003 0.0003 0.0000 0.0000 0.0000 0.0000
3       0.4403   5.6615      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001
4       0.0890  12.5947      0.0014 0.0000 0.0000 0.0005 0.0000 0.0002 0.0027
5       0.0491  16.9520      0.0003 0.0000 0.0000 0.0075 0.0002 0.0005 0.0006
6       0.0250  23.7630      0.0009 0.0000 0.0002 0.0002 0.0007 0.0039 0.0010
7       0.0090  39.6784      0.2647 0.0154 0.0407 0.0089 0.0000 0.0016 0.0051
8       0.0048  54.0053      0.0423 0.0094 0.0004 0.0703 0.0001 0.0008 0.0224
9       0.0032  65.9304      0.1776 0.0502 0.0117 0.0008 0.0002 0.0125 0.0438
10      0.0027  72.6477      0.0080 0.0117 0.0170 0.3229 0.0017 0.0003 0.0198
11      0.0015  95.9313      0.0574 0.0203 0.0001 0.0057 0.0130 0.0018 0.0195
12      0.0013 105.0834      0.0000 0.3049 0.3129 0.0643 0.0109 0.0005 0.0878
13      0.0007 141.4358      0.0004 0.4413 0.4326 0.0416 0.0039 0.0006 0.0669
14      0.0003 214.9895      0.0208 0.0441 0.0078 0.0012 0.1075 0.0853 0.5230
15      0.0002 295.4778      0.0559 0.0291 0.1485 0.0646 0.6756 0.8896 0.0566
16      0.0001 370.0791      0.3683 0.0733 0.0278 0.4116 0.1860 0.0022 0.1509
```

```
       X9    X10    X12    X13    X15    X16    X17    X18    X19
1  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0004
2  0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0000 0.0001
3  0.0005 0.0000 0.0000 0.0000 0.0001 0.0002 0.0001 0.0000 0.2164
4  0.0065 0.0001 0.0000 0.0000 0.0017 0.0038 0.0004 0.0000 0.1727
5  0.0285 0.0000 0.0000 0.0001 0.0005 0.0010 0.0496 0.0018 0.0528
6  0.0109 0.0000 0.0004 0.0004 0.0000 0.0000 0.1797 0.0027 0.0928
7  0.0161 0.0041 0.0001 0.0000 0.0001 0.0012 0.0749 0.0059 0.0025
8  0.1631 0.0074 0.0043 0.0004 0.0000 0.0077 0.1878 0.0802 0.0025
9  0.3083 0.0915 0.0001 0.0000 0.0030 0.0002 0.0640 0.0005 0.0224
10 0.0612 0.0000 0.0016 0.0000 0.0274 0.0192 0.0725 0.1465 0.0376
11 0.0303 0.1700 0.0187 0.0018 0.0821 0.0578 0.0753 0.1878 0.0012
12 0.0570 0.0078 0.0002 0.0002 0.0277 0.2226 0.0954 0.0256 0.0029
13 0.1127 0.0825 0.0000 0.0065 0.2977 0.3595 0.0205 0.0079 0.0278
14 0.1465 0.1913 0.2301 0.1058 0.2336 0.0123 0.1438 0.0380 0.0243
15 0.0315 0.4449 0.0017 0.1042 0.0147 0.0085 0.0063 0.1532 0.0608
16 0.0268 0.0004 0.7427 0.7805 0.3113 0.3061 0.0296 0.3499 0.2829

================================
Row 15==> X6, proportion 0.675631 >= 0.50
Row 15==> X7, proportion 0.889610 >= 0.50
Row 14==> X8, proportion 0.523039 >= 0.50
Row 16==> X12, proportion 0.742650 >= 0.50
Row 16==> X13, proportion 0.780471 >= 0.50
```

- So the subsets (X8), (X6,X7) and (X12,X13) are involved in Multicollinearity.

- In the first subset VIF of X8 is the highest, in the second subset the VIF of X7 is highest and in the third subset the VIF of X13 is highest

- We drop the variables X8, X7 and X13, and again fit a model.

Step 4:

```
> eigprop(olsreg_3)

Call:
eigprop(mod = olsreg_3)

   Eigenvalues       CI (Intercept)     X1     X2     X5     X6     X9    X10
1      11.2643   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0000
2       1.1724   3.0996      0.0037 0.0003 0.0004 0.0000 0.0000 0.0003 0.0000
3       0.4236   5.1567      0.0000 0.0000 0.0000 0.0000 0.0001 0.0008 0.0000
4       0.0612  13.5667      0.0035 0.0000 0.0000 0.0004 0.0008 0.0450 0.0013
5       0.0446  15.8950      0.0000 0.0000 0.0000 0.0138 0.0011 0.0500 0.0001
6       0.0148  27.6296      0.0031 0.0000 0.0001 0.0109 0.0331 0.0834 0.0068
7       0.0077  38.2308      0.6750 0.0290 0.0749 0.0087 0.0076 0.0139 0.0014
8       0.0042  51.8754      0.1739 0.0249 0.0001 0.2393 0.0036 0.1345 0.0195
9       0.0025  67.0968      0.0013 0.0019 0.0249 0.4952 0.0052 0.3272 0.0166
10      0.0019  77.0072      0.0051 0.0626 0.0375 0.0175 0.0429 0.1178 0.7517
11      0.0014  88.9269      0.1029 0.2072 0.0927 0.0079 0.0937 0.0412 0.0571
12      0.0009 111.6674      0.0079 0.5063 0.6391 0.1965 0.3709 0.0235 0.1431
13      0.0005 147.1269      0.0235 0.1677 0.1302 0.0097 0.4412 0.1623 0.0023
```

```
      X12    X15    X16    X17    X18    X19
1  0.0000 0.0000 0.0000 0.0001 0.0000 0.0009
2  0.0000 0.0000 0.0000 0.0003 0.0001 0.0007
3  0.0000 0.0001 0.0003 0.0002 0.0000 0.3331
4  0.0004 0.0040 0.0089 0.0006 0.0011 0.2817
5  0.0002 0.0006 0.0010 0.1229 0.0061 0.0336
6  0.0107 0.0000 0.0017 0.2914 0.0002 0.1479
7  0.0016 0.0000 0.0014 0.0180 0.0140 0.0149
8  0.0380 0.0048 0.0450 0.2623 0.1208 0.0003
9  0.0402 0.0266 0.0050 0.2401 0.3843 0.0443
10 0.0095 0.0002 0.0125 0.0412 0.2061 0.0053
11 0.0748 0.2018 0.3193 0.0102 0.1612 0.0048
12 0.1264 0.0605 0.0178 0.0103 0.0221 0.0001
13 0.6981 0.7014 0.5870 0.0024 0.0841 0.1325

================================
Row 12==> X1, proportion 0.506315 >= 0.50
Row 12==> X2, proportion 0.639124 >= 0.50
Row 10==> X10, proportion 0.751747 >= 0.50
Row 13==> X12, proportion 0.698132 >= 0.50
Row 13==> X15, proportion 0.701392 >= 0.50
Row 13==> X16, proportion 0.586961 >= 0.50
```

- So the subsets (X10), (X1,X2) and (X12,X15,X16) are involved in Multicollinearity.
- In the first subset VIF of X10 is the highest, in the second subset the VIF of X2 is highest and in the third subset the VIF of X15 is highest
- We drop the variables X2, X10 and X15, and again fit a model.

Step 5:

```
> eigprop(olsreg_4)

Call:
eigprop(mod = olsreg_4)

   Eigenvalues      CI (Intercept)     X1     X5     X6     X9    X12    X16    X17
1       8.5979  1.0000      0.0001 0.0001 0.0001 0.0000 0.0002 0.0000 0.0001 0.0002
2       0.8885  3.1107      0.0068 0.0025 0.0000 0.0001 0.0006 0.0000 0.0001 0.0005
3       0.3977  4.6498      0.0000 0.0000 0.0000 0.0001 0.0007 0.0001 0.0023 0.0005
4       0.0490 13.2416      0.0019 0.0001 0.0011 0.0000 0.1396 0.0007 0.0427 0.0305
5       0.0405 14.5733      0.0035 0.0000 0.0196 0.0021 0.0004 0.0000 0.0874 0.1000
6       0.0141 24.7331      0.0013 0.0003 0.0106 0.0533 0.0897 0.0194 0.0120 0.2732
7       0.0052 40.7993      0.7281 0.7328 0.0032 0.0107 0.0037 0.0020 0.0104 0.0439
8       0.0040 46.4515      0.1180 0.0477 0.2330 0.0004 0.1915 0.0431 0.4341 0.3354
9       0.0023 61.5423      0.0373 0.1339 0.6012 0.0068 0.4397 0.0302 0.4107 0.2062
10      0.0009 99.5474      0.1030 0.0826 0.1312 0.9265 0.1340 0.9045 0.0001 0.0097
```

```
      X18    X19
1  0.0000 0.0017
2  0.0001 0.0009
3  0.0001 0.3893
4  0.0003 0.3114
5  0.0087 0.0269
6  0.0010 0.1491
7  0.0556 0.0282
8  0.2220 0.0002
9  0.5131 0.0558
10 0.1991 0.0365

==============================
Row 7==> X1, proportion 0.732820 >= 0.50
Row 9==> X5, proportion 0.601233 >= 0.50
Row 10==> X6, proportion 0.926498 >= 0.50
Row 10==> X12, proportion 0.904520 >= 0.50
Row 9==> X18, proportion 0.513094 >= 0.50
```

- So the subsets (X1), (X5,X18) and (X6,X12) are involved in Multicollinearity.
- In the first subset VIF of X1 is the highest, in the second subset the VIF of X5 is highest and in the third subset the VIF of X6 is highest
- We drop the variables X1, X5 and X6, and again fit a model.

Step 6:

```
> eigprop(olsreg_5)

Call:
eigprop(mod = olsreg_5)

  Eigenvalues      CI (Intercept)     X9    X12    X16    X17    X18    X19
1      5.9496  1.0000      0.0065 0.0007 0.0002 0.0008 0.0004 0.0001 0.0042
2      0.5794  3.2045      0.7618 0.0017 0.0001 0.0003 0.0005 0.0002 0.0000
3      0.3803  3.9552      0.0058 0.0009 0.0006 0.0097 0.0010 0.0003 0.4378
4      0.0485 11.0760      0.0805 0.3259 0.0034 0.1534 0.0170 0.0000 0.3775
5      0.0313 13.7945      0.0320 0.0282 0.0053 0.3884 0.2016 0.0152 0.0365
6      0.0075 28.1108      0.0998 0.6220 0.4144 0.3558 0.2615 0.0952 0.1036
7      0.0034 41.8669      0.0135 0.0206 0.5759 0.0916 0.5181 0.8890 0.0404

================================
Row 6==> X9, proportion 0.622007 >= 0.50
Row 7==> X12, proportion 0.575942 >= 0.50
Row 7==> X17, proportion 0.518085 >= 0.50
Row 7==> X18, proportion 0.889041 >= 0.50
```

- So the subsets (X9) and (X12,X17,X18) are involved in Multicollinearity.
- In the first subset VIF of X9 is the highest, and in the second subset the VIF of X18 is highest
- We drop the variables X9 and X18, and again fit a model.

Step 7:

```
> eigprop(olsreg_6)

Call:
eigprop(mod = olsreg_6)

  Eigenvalues      CI (Intercept)    X12    X16    X17    X19
1      4.0541  1.0000      0.0186 0.0013 0.0021 0.0020 0.0144
2      0.5331  2.7578      0.9099 0.0016 0.0023 0.0042 0.0049
3      0.3688  3.3155      0.0078 0.0022 0.0128 0.0028 0.7366
4      0.0288 11.8654      0.0602 0.0000 0.5872 0.6496 0.1774
5      0.0153 16.3016      0.0035 0.9949 0.3956 0.3414 0.0667

================================
Row 5==> X12, proportion 0.994857 >= 0.50
Row 4==> X16, proportion 0.587226 >= 0.50
Row 4==> X17, proportion 0.649594 >= 0.50
Row 3==> X19, proportion 0.736577 >= 0.50
```

In this way to remove the variables contributing Multicollinearity, we almost end up all important variables for our model building exercise. Still the proportion of variability of X12,X16,X17 and X19 are much higher than 0.5. So, we decided to go for stepwise selection method to get better subset model.

7.VARIABLE SELECTION

When we fit a MLR model, we use the p-value in the ANOVA table to determine whether the model, as a whole, is significant. A natural question arises which regressors, among a larger set of all potential regressors, are important. We could use the individual p-values of the regressors and refit the model with only significant terms. But the p-values of the regressors are adjusted for the other terms in the model. So, picking out the subset of significant regressors can be somewhat challenging. This procedure of identifying the best subset of regressors to include in the model, among all possible subsets of regressors, is referred to as variable selection.

One approach is to start with a model containing only the intercept. Then using some chosen model fit criterion we slowly add terms to the model, one at a time, whose inclusion gives the most statistically significant improvement of the the model, and repeat this process until none improves the model to a statistically significant extent. This procedure is referred to as forward selection.

Another alternative is backward elimination. Here we start with the full model, then based on some model fit criterion we slowly remove variables one at a time, whose deletion gives the most statistically insignificant deterioration of the model fit, and repeat this process until no further variables can be deleted without a statistically insignificant loss of fit.

A third classical approach is stepwise selection. This is a combination of forward selection (FS) and backward elimination (BE). We start with FS, but at each step we recheck all regressors already entered, for possible deletion by BE method, this is because of the fact that regressor added at an earlier step may now be unnecessary in presence of new regressor.

Here we use stepwise selection method based on partial F-test & AIC criterions to determine the best subset model.

On the Basis of Partial F-Test

On the basis of the Step-wise Selection method

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	X6	addition	0.993	0.993	2236.4230	819.2618	5033.6
2	X4	addition	0.997	0.997	944.7400	786.7498	3347.6
3	X18	addition	0.999	0.999	401.8710	755.4481	2260.2
4	X19	addition	0.999	0.999	260.6150	741.0638	1876.3
5	X13	addition	0.999	0.999	206.2040	734.2496	1709.0
6	X9	addition	0.999	0.999	144.6290	723.3298	1481.2
7	X7	addition	1.000	1.000	79.0710	704.8192	1170.7
8	X11	addition	1.000	1.000	45.7800	689.8387	966.5

As we can see from the above stepwise selection summary we are losing most of our important variables, hence we go for stepwise selection based on Information Theoretic Criterion to obtain a better model.

On the Basis of Information Theoretic Criterion

Our MLR model is

$$Y = X\beta + \epsilon$$

Where we assume that $\epsilon \sim$ N(0,σ^2) and Y$\sim N_n(X\beta,\sigma^2 I_n)$
The likelihood function given by,

$$L(\beta,\sigma^2|y) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}e^{-\frac{(y-X\beta)^T(y-X\beta)}{2}}$$

So, the general form of the penalized likelihood function is given by,

$$\begin{aligned} &-2ln\hat{L} + \text{penalty term} \\ &= \text{nln(SSRes)} + \text{penalty term} \end{aligned}$$

Where,

$$\hat{L} = max_{\beta,\sigma^2} L(\beta,\sigma^2|y) = L(\hat{\beta_{mle}},\hat{\sigma^2}_{mle})$$

1.1 Akaike Information Criterion

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given dataset. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.
AIC is founded on information theory: it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model.
AIC does not provide a test of a model in the sense of testing a null hypothesis, so it can tell nothing about the absolute quality of the model.If all the candidate models fit poorly, AIC will not give any warning of that.

Definition

Suppose that we have a statistical model of some n data. Then the AIC value of the model is the given by,
$AIC = -2ln(\hat{L}) + 2k$ Where,
K = The number of estimated parameters in the model
$\hat{L}$ = The maximized value of the likelihood function for the model
At first we consider all the subset models excluding one regressor at a time, and calculate the AIC value for each of those subset models. Then we discard the variable for which the subset model has the minimum AIC value.
Firstly, the method considered the Full 20 parameter model in the first step.

```
> AIC<-stepAIC(Z,direction="both")
Start:  AIC=544.84
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
    X12 + X13 + X14 + X15 + X16 + X17 + X18 + X19 + X20

       Df Sum of Sq      RSS    AIC
- X17   1      6020  8699102 542.87
- X1    1     36720  8729801 543.02
- X4    1     82490  8775571 543.23
- X9    1    230392  8923473 543.92
- X5    1    333543  9026624 544.39
- X14   1    335936  9029017 544.40
- X19   1    379175  9072257 544.59
- X12   1    430604  9123685 544.83
<none>               8693081 544.84
- X6    1    512675  9205757 545.19
- X16   1    513196  9206277 545.19
- X8    1    760617  9453698 546.28
- X3    1   1003237  9696318 547.32
- X2    1   1171465  9864546 548.03
- X20   1   1468141 10161222 549.24
- X15   1   2478836 11171918 553.13
- X7    1   2979617 11672699 554.93
- X11   1   3713281 12406362 557.43
- X10   1   4199725 12892807 559.00
- X13   1   5852955 14546036 563.95
- X18   1   8638518 17331599 571.13
```

The AIC corresponding to the Full Model is 544.84.
In this step this method compares the AICs by discarding each variable from the full model with the AIC of the full model.
From the table it can be observed that, the AIC corresponding to the model with 19 regressors after discarding the X17 variable is lower than the full model and also it is minimum among all 19 regressor model.

```
Step:  AIC=542.87
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
    X12 + X13 + X14 + X15 + X16 + X18 + X19 + X20

       Df Sum of Sq      RSS    AIC
- X1    1     44297  8743399 541.08
- X4    1     81104  8780206 541.25
- X9    1    306689  9005791 542.29
- X5    1    327556  9026657 542.39
- X14   1    336760  9035862 542.43
- X19   1    390334  9089435 542.67
<none>               8699102 542.87
- X6    1    528992  9228094 543.29
- X16   1    571585  9270686 543.48
- X12   1    694645  9393747 544.02
- X8    1    824715  9523817 544.59
+ X17   1      6020  8693081 544.84
- X3    1   1010282  9709383 545.38
- X2    1   1171730  9870832 546.05
- X20   1   1477322 10176424 547.30
- X7    1   2973599 11672701 552.93
- X15   1   3068061 11767162 553.26
- X10   1   4497249 13196350 557.96
- X11   1   6164351 14863452 562.83
- X13   1   6291458 14990559 563.18
- X18   1   8661879 17360981 569.20
```

In the next step considers the subset model by discarding X17 from the full model. The AIC corresponding to that model is 542.87.

AIC will be calculated after discarding each of the variable from the current subset model. The AIC corresponding to the model after discarding X1 from the current subset model is minimum, and in the next step this variable will be deleted from the model.

```
Step:  AIC=541.08
Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 +
    X13 + X14 + X15 + X16 + X18 + X19 + X20

       Df Sum of Sq      RSS    AIC
- X4    1     77261  8820660 539.44
- X9    1    306570  9049969 540.49
- X14   1    325193  9068591 540.58
- X5    1    375131  9118530 540.80
- X19   1    435098  9178497 541.07
<none>               8743399 541.08
- X6    1    555832  9299231 541.61
- X16   1    565527  9308926 541.65
- X12   1    672113  9415512 542.12
- X8    1    823440  9566839 542.77
+ X1    1     44297  8699102 542.87
+ X17   1     13597  8729801 543.02
- X3    1   1028205  9771604 543.64
- X2    1   1139343  9882742 544.10
- X20   1   1434753 10178152 545.31
- X15   1   3029670 11773068 551.28
- X7    1   3279506 12022905 552.14
- X10   1   4525851 13269249 556.18
- X11   1   6128883 14872282 560.86
- X13   1   7735165 16478563 565.06
- X18   1   8753084 17496483 567.52
```

In the next step considers the subset model by discarding X1 and X17 from the full model. The AIC corresponding to that model is 541.08.
AIC will be calculated after discarding each of the variable from the current subset model. The AIC corresponding to the model after discarding X4 from the current subset model is minimum, and in the next step this variable will be deleted from the model.

```
Step:  AIC=539.44
Y ~ X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X19 + X20

       Df Sum of Sq      RSS    AIC
- X19   1    362380  9183040 539.09
- X5    1    376558  9197219 539.15
<none>               8820660 539.44
- X6    1    517350  9338010 539.78
- X16   1    533421  9354081 539.85
- X14   1    592362  9413022 540.11
- X12   1    710147  9530807 540.62
+ X4    1     77261  8743399 541.08
- X9    1    857161  9677822 541.24
+ X1    1     40454  8780206 541.25
+ X17   1     11167  8809493 541.39
- X3    1    981223  9801883 541.77
- X20   1   1400196 10220857 543.48
- X2    1   1533213 10353873 544.01
- X8    1   2415218 11235879 547.36
- X15   1   2970365 11791025 549.34
- X10   1   4587916 13408576 554.61
- X7    1   5702429 14523089 557.88
- X13   1   8538838 17359498 565.20
- X18   1   8679586 17500246 565.53
- X11   1  11058706 19879366 570.76
```

In the next step considers the subset model by discarding X1,X4 and X17 from the full model. The AIC corresponding to that model is 539.44.
AIC will be calculated after discarding each of the variable from the current subset model.The AIC corresponding to the model after discarding X19 from the current subset model is minimum, and in the next step this variable will be deleted from the model.

```
Step:  AIC=539.09
Y ~ X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
- X5    1    425413  9608453 538.95
<none>               9183040 539.09
- X16   1    460582  9643622 539.10
- X6    1    483348  9666388 539.19
+ X19   1    362380  8820660 539.44
- X12   1    656689  9839730 539.92
+ X1    1     92600  9090440 540.68
+ X17   1     38257  9144783 540.92
+ X4    1      4543  9178497 541.07
- X14   1   1114420 10297460 541.79
- X9    1   1249191 10432231 542.32
- X2    1   1255421 10438461 542.34
- X3    1   1298824 10481864 542.51
- X20   1   1368221 10551261 542.79
- X15   1   2845821 12028861 548.16
- X8    1   3506159 12689199 550.35
- X10   1   5110035 14293075 555.23
- X13   1   8196237 17379277 563.25
- X7    1   8662984 17846024 564.33
- X18   1   9475623 18658663 566.16
- X11   1  12491470 21674510 572.30
```

In the next step considers the subset model by discarding X1,X4,X17 and X19 from the full model.
The AIC corresponding to that model is 539.09.
AIC will be calculated after discarding each of the variable from the current subset model.The AIC corresponding to the model after discarding X5 from the current subset model is minimum, and in the next step this variable will be deleted from the model.

```
Step:  AIC=538.95
Y ~ X2 + X3 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13 + X14 +
    X15 + X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
- X6    1    140050  9748503 537.54
<none>               9608453 538.95
+ X5    1    425413  9183040 539.09
+ X19   1    411235  9197219 539.15
+ X1    1    165614  9442839 540.24
- X2    1    830407 10438861 540.35
+ X17   1     16126  9592327 540.88
+ X4    1      7339  9601114 540.92
- X3    1   1068096 10676549 541.27
- X9    1   2817014 12425467 547.49
- X20   1   3130013 12738466 548.51
- X16   1   3833656 13442109 550.71
- X8    1   5058231 14666684 554.29
- X12   1   7458918 17067371 560.50
- X13   1   8209365 17817818 562.27
- X7    1   8517744 18126197 562.97
- X14   1   8620534 18228987 563.20
- X15   1   9778536 19386989 565.73
- X10   1  12413429 22021882 570.95
- X11   1  26945574 36554027 591.73
- X18   1  31473422 41081876 596.52
```

In the next step considers the subset model by discarding X1,X4,X17,X5 and X19 from the full model. The AIC corresponding to that model is 538.95.
AIC will be calculated after discarding each of the variable from the current subset model.The AIC corresponding to the model after discarding X6 from the current subset model is minimum, and in the next step this variable will be deleted from the model.

```
Step:  AIC=537.54
Y ~ X2 + X3 + X7 + X8 + X9 + X10 + X11 + X12 + X13 + X14 + X15 +
    X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
<none>               9748503 537.54
+ X19   1    361182  9387321 537.99
+ X1    1    168597  9579906 538.83
+ X6    1    140050  9608453 538.95
+ X4    1     85041  9663462 539.18
+ X5    1     82115  9666388 539.19
+ X17   1        41  9748463 539.54
- X2    1   1302027 11050531 540.68
- X3    1   2483930 12232433 544.85
- X20   1   3449395 13197899 547.96
- X9    1   4229535 13978039 550.32
- X16   1   4418481 14166984 550.87
- X12   1   7888403 17636906 559.85
- X13   1   8104151 17852654 560.35
- X7    1   8812516 18561019 561.94
- X14   1   9323859 19072362 563.06
- X8    1   9797513 19546017 564.06
- X15   1  10942376 20690879 566.40
- X10   1  12472494 22220998 569.32
- X18   1  31727896 41476399 594.91
- X11   1  43570116 53318619 605.21
```

Finally considers the subset model by discarding X1,X4,X5,X6,X17 and X19 from the full model.
The AIC corresponding to that model is 537.54. AIC will be calculated after discarding each of the variable from the current subset model.
If any one of the variables is discarded from the current subset model the AIC is higher than the current model. So, no variable will be discarded any more, the current model is our final model.

```
> AIC

Call:
lm(formula = Y ~ X2 + X3 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X20, data = regression_data)

Coefficients:
(Intercept)           X2           X3           X7           X8           X9
 -1694.3526       3.7615     -28.8492      -1.5838      -2.7344       1.6184
        X10          X11          X12          X13          X14          X15
    -4.2760       4.9207      -1.4857       3.2643       0.4834       2.5439
        X16          X18          X20
    -0.6579      -1.7458       1.5042
```

So, our best subset model chosen by AIC is given by,

```
> summary(AIC)

Call:
lm(formula = Y ~ X2 + X3 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X20, data = regression_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1275.90  -247.09   -19.06   128.86  1112.49

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.694e+03  1.572e+03  -1.078  0.29088
X2           3.761e+00  2.019e+00   1.863  0.07373 .
X3          -2.885e+01  1.121e+01  -2.574  0.01611 *
X7          -1.584e+00  3.267e-01  -4.848 5.02e-05 ***
X8          -2.734e+00  5.349e-01  -5.112 2.50e-05 ***
X9           1.618e+00  4.819e-01   3.359  0.00242 **
X10         -4.276e+00  7.414e-01  -5.768 4.49e-06 ***
X11          4.921e+00  4.565e-01  10.780 4.34e-11 ***
X12         -1.486e+00  3.239e-01  -4.587  0.00010 ***
X13          3.264e+00  7.021e-01   4.649 8.49e-05 ***
X14          4.834e-01  9.694e-02   4.987 3.48e-05 ***
X15          2.544e+00  4.709e-01   5.402 1.17e-05 ***
X16         -6.579e-01  1.916e-01  -3.433  0.00201 **
X18         -1.746e+00  1.898e-01  -9.199 1.17e-09 ***
X20          1.504e+00  4.959e-01   3.033  0.00543 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 612.3 on 26 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 2.832e+04 on 14 and 26 DF,  p-value: < 2.2e-16
```

```
> summary(AIC)

Call:
lm(formula = Y ~ X2 + X3 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X20, data = regression_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1275.90  -247.09   -19.06   128.86  1112.49

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.694e+03  1.572e+03  -1.078  0.29088
X2           3.761e+00  2.019e+00   1.863  0.07373 .
X3          -2.885e+01  1.121e+01  -2.574  0.01611 *
X7          -1.584e+00  3.267e-01  -4.848 5.02e-05 ***
X8          -2.734e+00  5.349e-01  -5.112 2.50e-05 ***
X9           1.618e+00  4.819e-01   3.359  0.00242 **
X10         -4.276e+00  7.414e-01  -5.768 4.49e-06 ***
X11          4.921e+00  4.565e-01  10.780 4.34e-11 ***
X12         -1.486e+00  3.239e-01  -4.587  0.00010 ***
X13          3.264e+00  7.021e-01   4.649 8.49e-05 ***
X14          4.834e-01  9.694e-02   4.987 3.48e-05 ***
X15          2.544e+00  4.709e-01   5.402 1.17e-05 ***
X16         -6.579e-01  1.916e-01  -3.433  0.00201 **
X18         -1.746e+00  1.898e-01  -9.199 1.17e-09 ***
X20          1.504e+00  4.959e-01   3.033  0.00543 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 612.3 on 26 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 2.832e+04 on 14 and 26 DF,  p-value: < 2.2e-16
```

As we can observe from the above summary, the adjusted R-squared of the model is 0.999. Now, we want to keep all the regressors in our selected model but their may be presence of multicollinearity.
So, we check for the presence of Multicollinearity by looking at the corresponding VIFs.

Multicollinearity Detection after AIC

Variable	VIF
X2	64.53264
X3	92.16279
X7	2151.69969
X8	792.78979
X9	414.69627
X10	2646.14101
X11	6898.62211
X12	2795.07020
X13	4554.24417
X14	2036.14923
X15	1464.04974
X16	546.00801
X18	480.16478
X20	1453.39455

As all the VIFs are higher than 5, we can say that the selected Subset model is also suffering from Multicollinearity.

Now as we obtained the best subset by AIC, we will keep all the variables in our model. For removal of multicollinearity we will use Ridge Regression.

RIDGE REGRESSION

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. coefficients are shrunk by the same factor (so none are eliminated).
A tuning parameter (λ) controls the strength of the penalty term. When $\lambda = 0$, ridge regression equals least squares regression. If $\lambda =$ inf, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and inf.
Ridge estimators theoretically produce new estimators that are shrunk closer to the "true" population parameters.
The ridge function fitting the ridge regression is given by,

$$R(\beta) = min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda(\beta)^T(\beta)$$

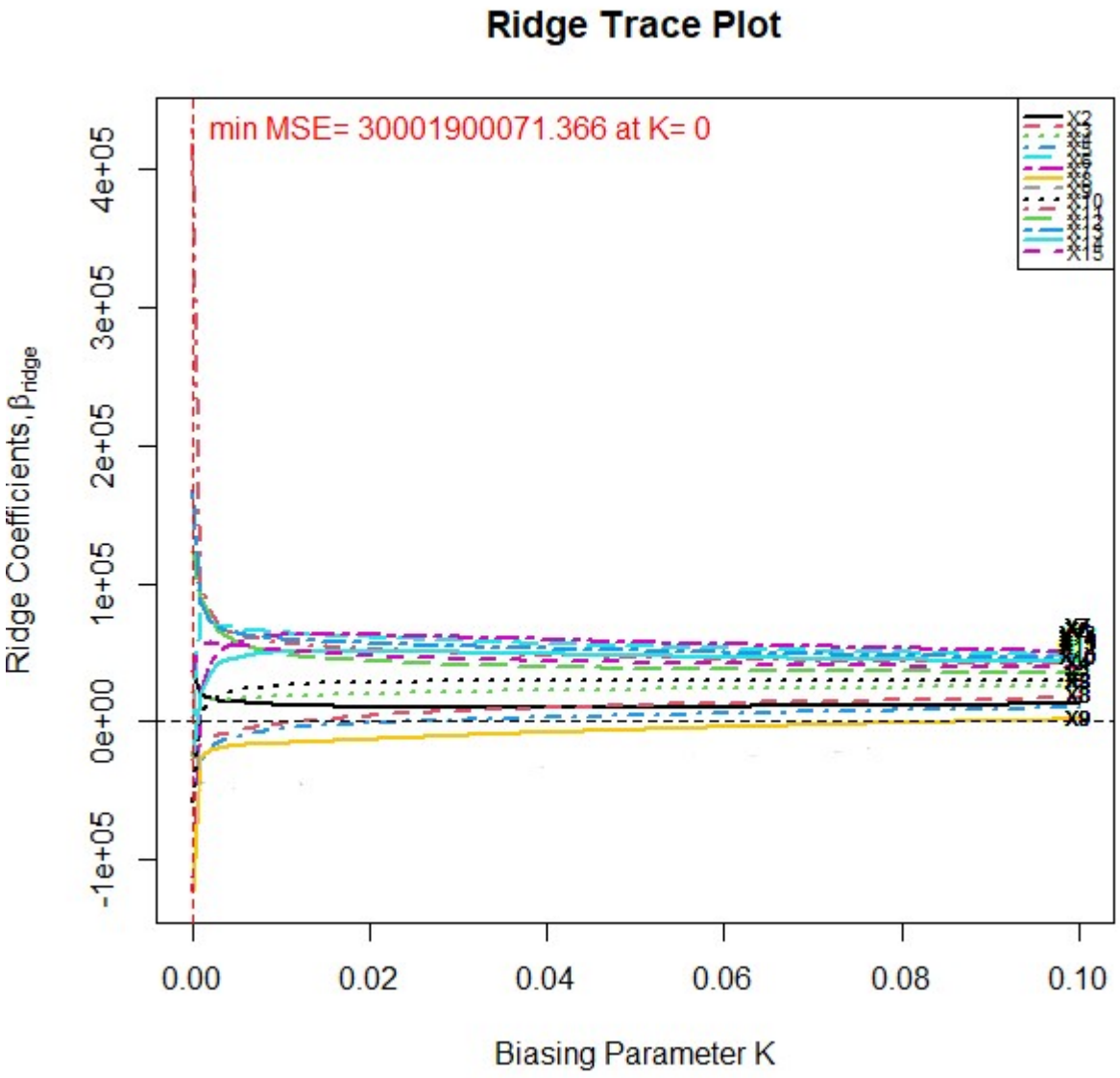
OLS regression uses the following formula to estimate coefficients:

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

Ridge regression adds a product of ridge parameter the identity matrix to the cross product matrix $(X^T X)$, forming a new matrix $(X^T X + \lambda I)$. The new formula is used to find the coefficients:

$$\tilde{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

To choose the value of λ, we have used a graphical method called ridge trace plot, a plot of estimated coefficients against a shrinkage parameter, to determine a favorable trade-off of bias against precision(inverse variance) of the estimates.



From the above plot it seems that the estimates of coefficients stabilizes for some value of between 0.020 and 0.025.

From the ridge trace plot we choose that value of λ for which VIFs all get stabilized (i.e. < 5). The estimate of obtained by this method is 0.021. Hence we fit a new model with this value of and inspect its adjusted R-squared value.

```
Call:
lmridge.default(formula = regression_data$Y ~ ., data = df, K = 0.021)


Coefficients: for Ridge parameter K= 0.021
                          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept              -6.1672e+03   -5.3052e+09  1.5866e+08     -33.4372   <2e-16 ***
X2                      6.7821e+00    1.6527e+04  5.4038e+03       3.0585   0.0043 **
X3                      9.5604e+00    5.0141e+03  6.1405e+03       0.8165   0.4197
X7                      8.5470e-01    7.4317e+04  4.6521e+03      15.9750   <2e-16 ***
X8                     -2.8890e-01   -9.3108e+03  4.7310e+03      -1.9681   0.0570 .
X9                     -1.8296e+00   -4.7346e+04  5.4839e+03      -8.6336   <2e-16 ***
X10                     5.0830e-01    2.1597e+04  3.6162e+03       5.9724   <2e-16 ***
X11                     5.0980e-01    5.6796e+04  2.4595e+03      23.0928   <2e-16 ***
X12                     4.5280e-01    4.5252e+04  3.8838e+03      11.6517   <2e-16 ***
X13                     1.0256e+00    6.0359e+04  2.5983e+03      23.2306   <2e-16 ***
X14                     1.9930e-01    5.6799e+04  2.9834e+03      19.0384   <2e-16 ***
X15                     6.6080e-01    3.2876e+04  4.5667e+03       7.1991   <2e-16 ***
regression_data...18.   4.3500e-01    3.2475e+04  4.7203e+03       6.8799   <2e-16 ***
regression_data...20.   2.0560e-01    1.4538e+04  6.2162e+03       2.3387   0.0252 *
regression_data...22.   5.4250e-01    2.5535e+04  5.4419e+03       4.6922   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge Summary
        R2     adj-R2   DF ridge           F        AIC        BIC
   0.99220    0.98840    5.06884 2219.46583  633.79586  794.73814
Ridge minimum MSE= 440703084618 at K= 0.021
P-value for F-test ( 5.06884 , 34.60437 ) = 2.108297e-42
------------------------------------------------------------------------
```

```
          R2 adj-R2 DF ridge         F      AIC      BIC
[1,] 0.9922 0.9884  5.06884 2219.466 633.7959 794.7381
```

VIFs for the new fitted model are:

```
> vif(model1)
             X2     X3      X7      X8      X9     X10     X11     X12     X13     X14
k=0.021 6.10392 7.8819 4.52383 4.67867 6.28627 2.73346 1.26444 3.15298 1.41119 1.86054
            X15 regression_data...18. regression_data...20. regression_data...22.
k=0.021 4.35938               4.65747               8.07731               6.19046
```

Observation:

- We observe that after fitting ridge regression model the VIFs have decreased significantly.

- The adjusted R-square is 98.84

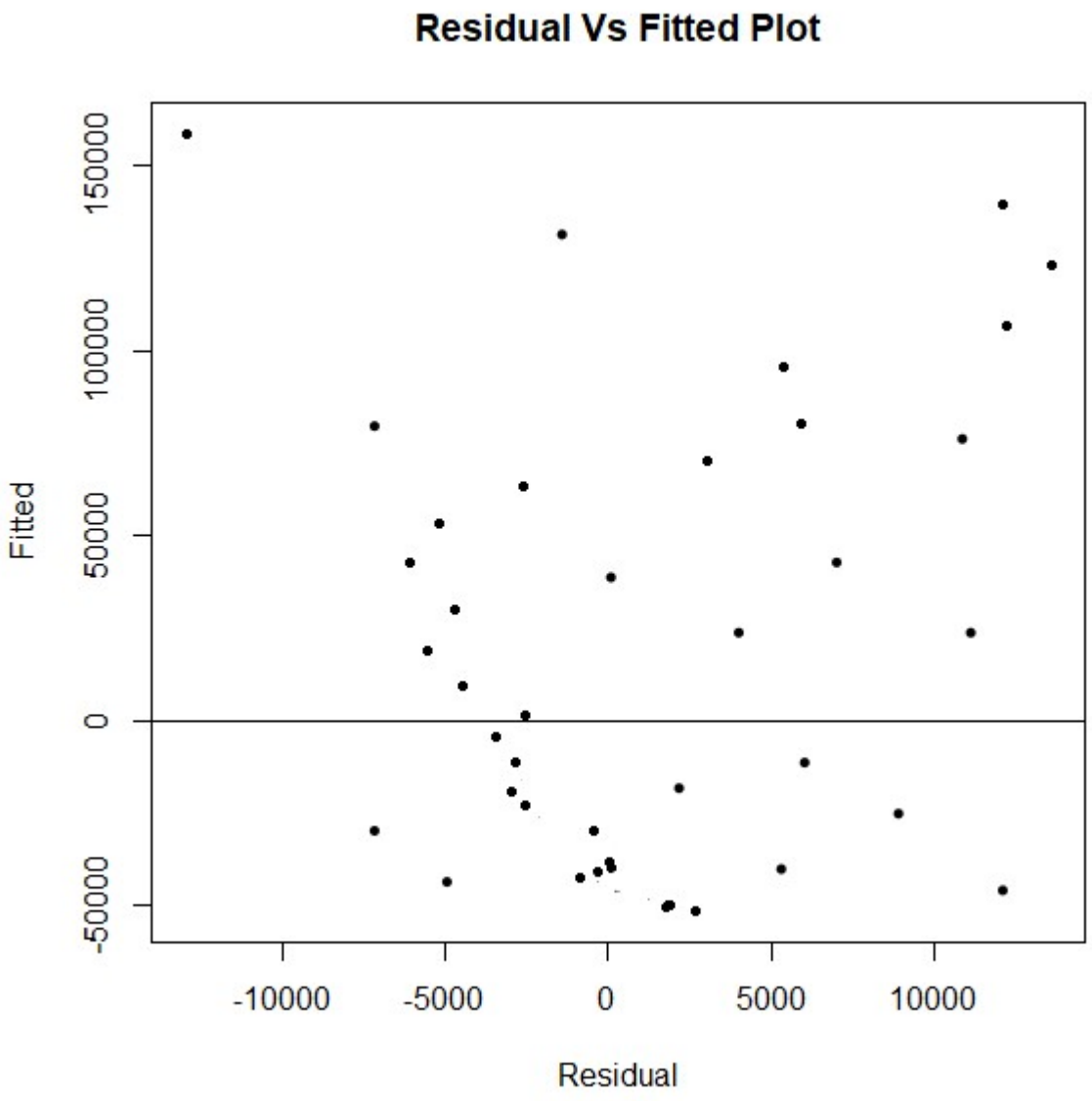
Now we perform residual analysis on our newly fitted model.

Inspection of Properties of Fitted Model After Ridge Regression

Check for Homoscedasticity Assumption of Errors

```
> cor(res,yhat)
             K=0.021
K=0.021 0.1009673
```

The correlation between fitted values and residuals is 0.1009673.

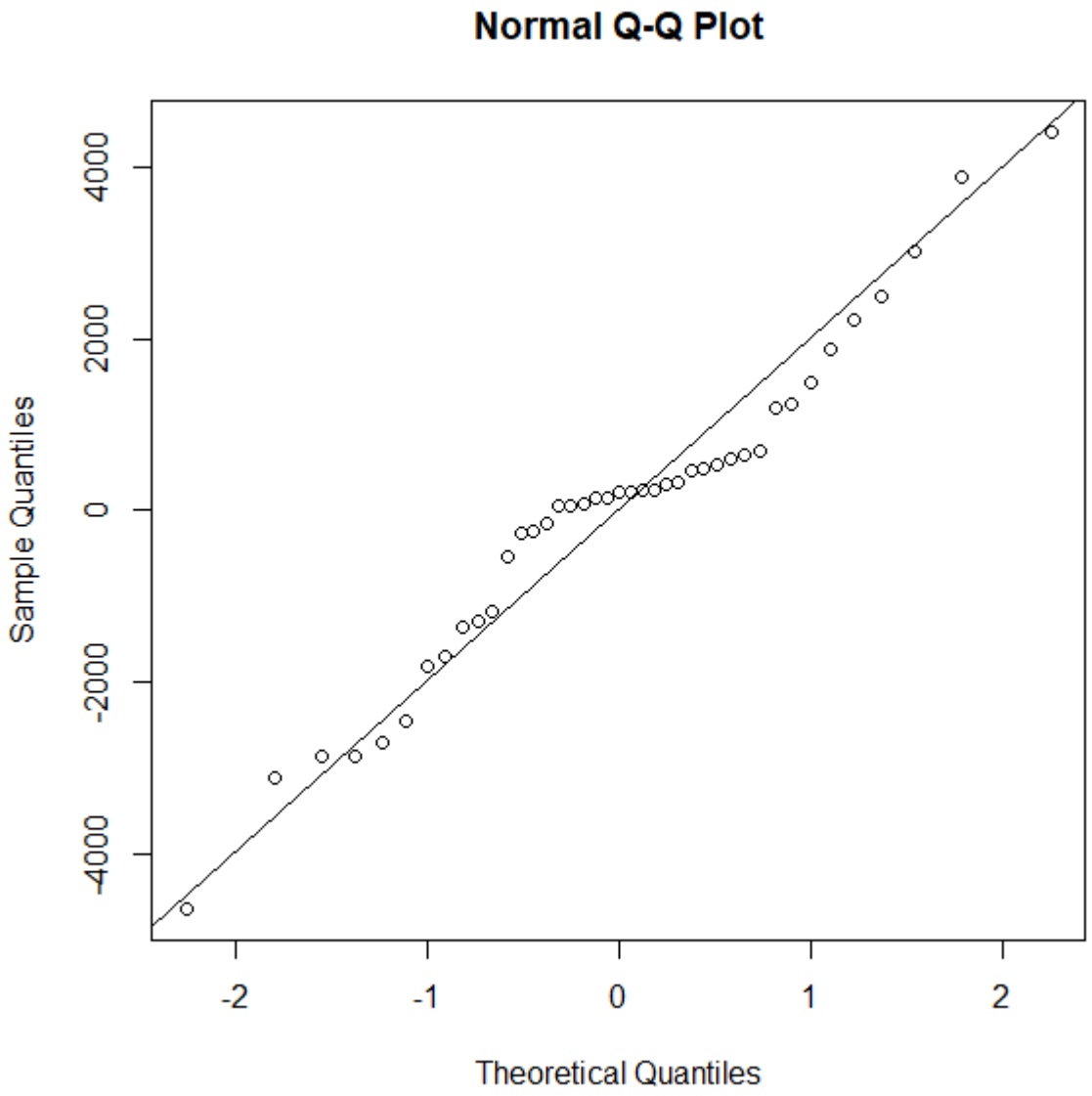


From the plot we cannot find any systematic behavior and the correlation between fitted values and residuals is nearly 0. Hence our assumption of homoscedasticity holds true. For more concrete evidence we perform Breusch-Pagan Test for heteroscedasticity.

```
        studentized Breusch-Pagan test

data:  model1
BP = 23.665, df = 14, p-value = 0.05028
```

As p-value=0.05028>0.05.
Hence, we conclude that there is no violation of homoscedasticity assumption in our model

Test for Normality Assumption of Errors



As we can see majority of points lies on the straight line. Hence no evidence of violation of normality assumption is found. To strengthen our judgement we further perform Shapiro-Wilk Test for normality.

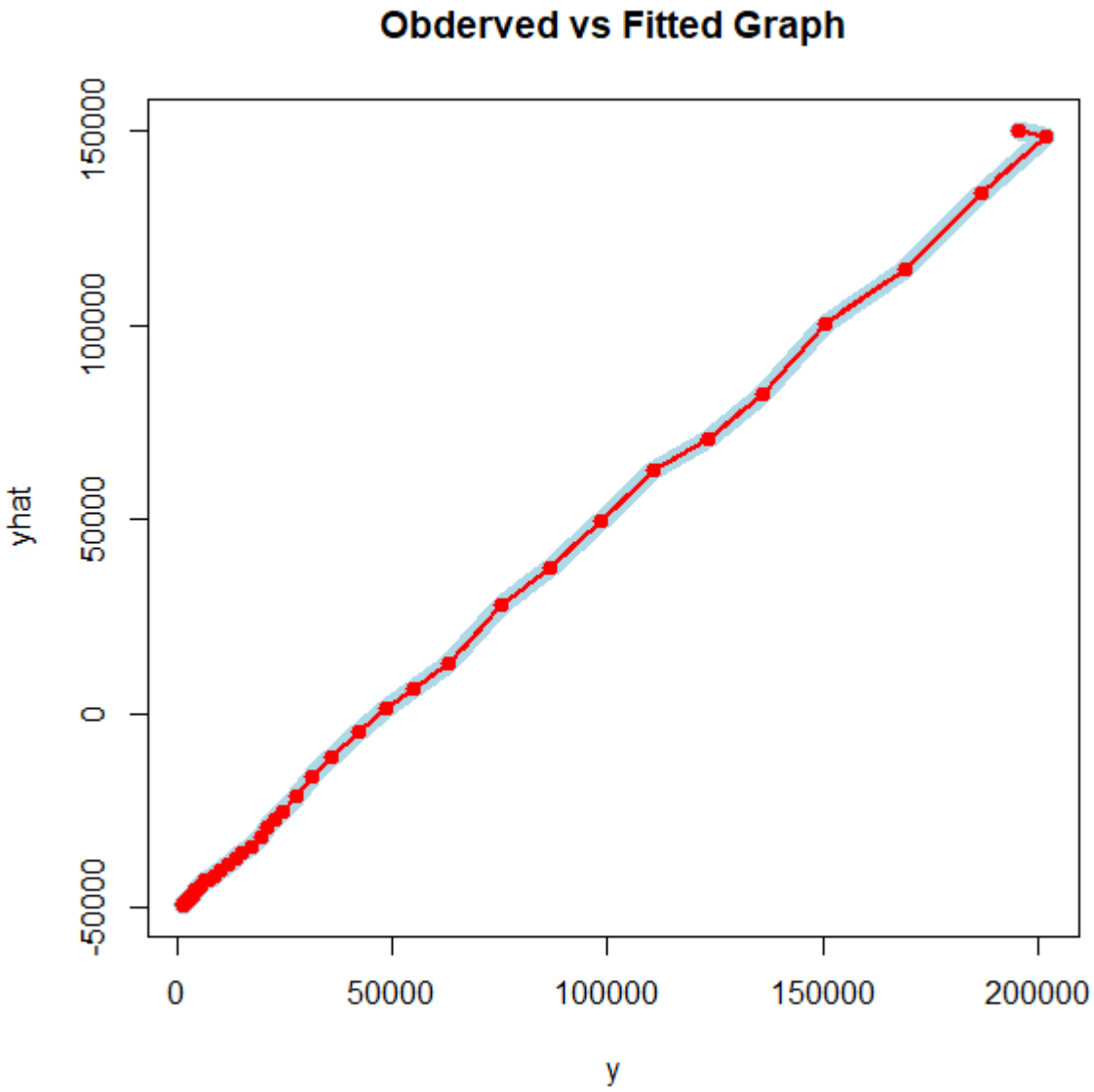
```
        Shapiro-Wilk normality test

data:  res
W = 0.9583, p-value = 0.1371
```

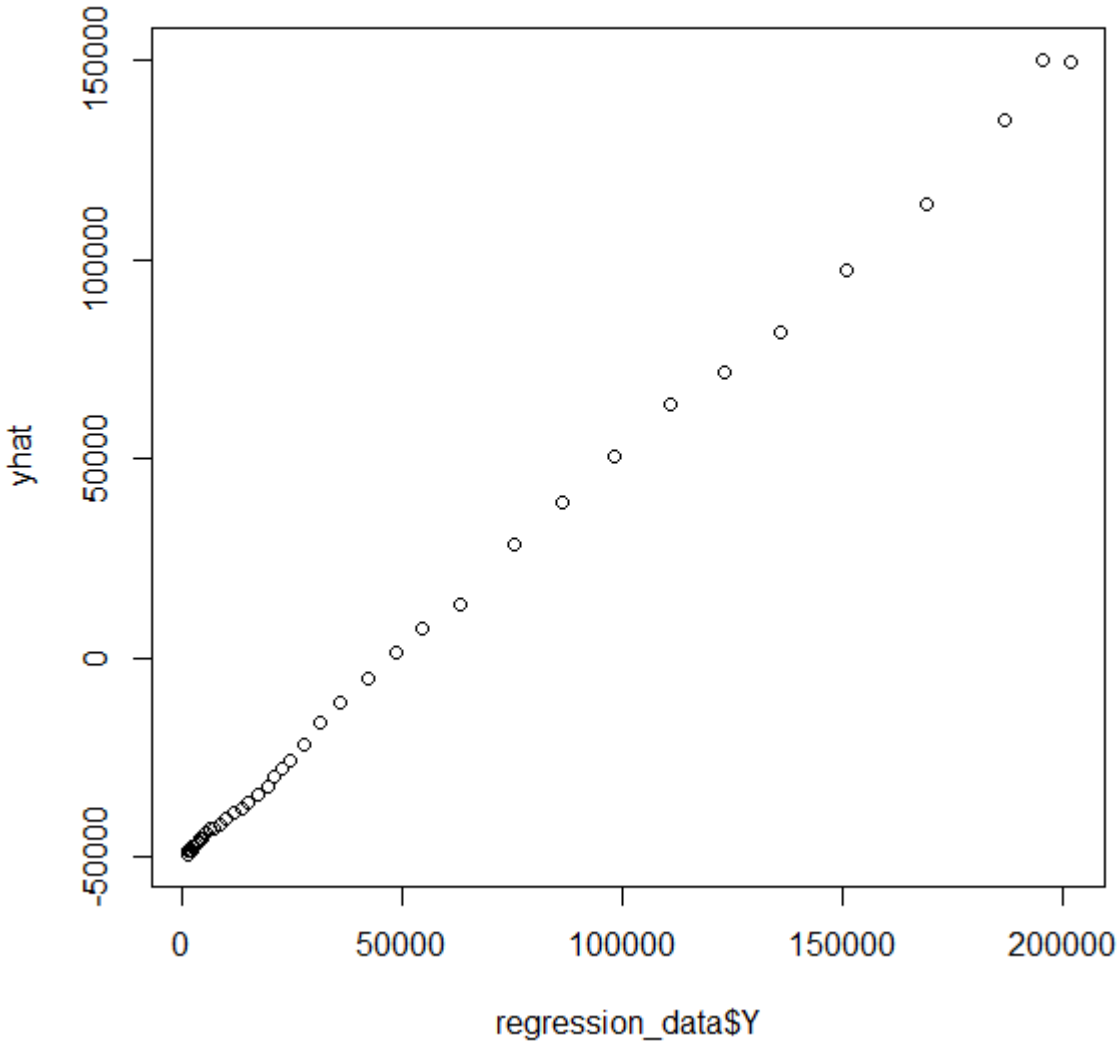
As we can see p-value= 0.1371 $>$ 0.05, hence Normality Assumption of error holds. Now we compare between observed and fitted responses.

```
> cor(regression_data$Y,yhat)
        K=0.021
[1,] 0.9994487
```

The correlation between fitted and observed response is 0.9994487, which indicates a good fit of the observed responses.

GRAPH BETWEEN OBSERVED AND FITTED RESPONSE





From the above graph, we conclude that our fitted values are approximately equal to observed values of response variable (GNI at Current Prices).

Final Fitted Model

Our final model after Ridge regression is given by,

$$\hat{Y} = (-6.1672e{+}03) + 6.7821e{+}00)X2 + (9.5604e{+}00)X3 + (8.5470e{-}01)X7 + (-2.8890e{-}01)X8 + (-1.8296e{+}00)X9 + (5.0830e{-}01)X10 + (5.0980e{-}01)X11 + (4.5280e{-}01)X12 + (1.0256e{+}00)X13 + (1.9930e{-}01)X14 + (6.6080e{-}01)X15 + (4.3500e{-}01)X16 + (2.0560e{-}01)X18 + (5.4250e{-}01)X20$$

CONCLUSION ABOUT THE RIDGE MODEL

R^2 and Adjusted R^2 are used to explain the overall adequacy of the model, where,

$$R^2 = 1 - \frac{SSRes}{SST}$$

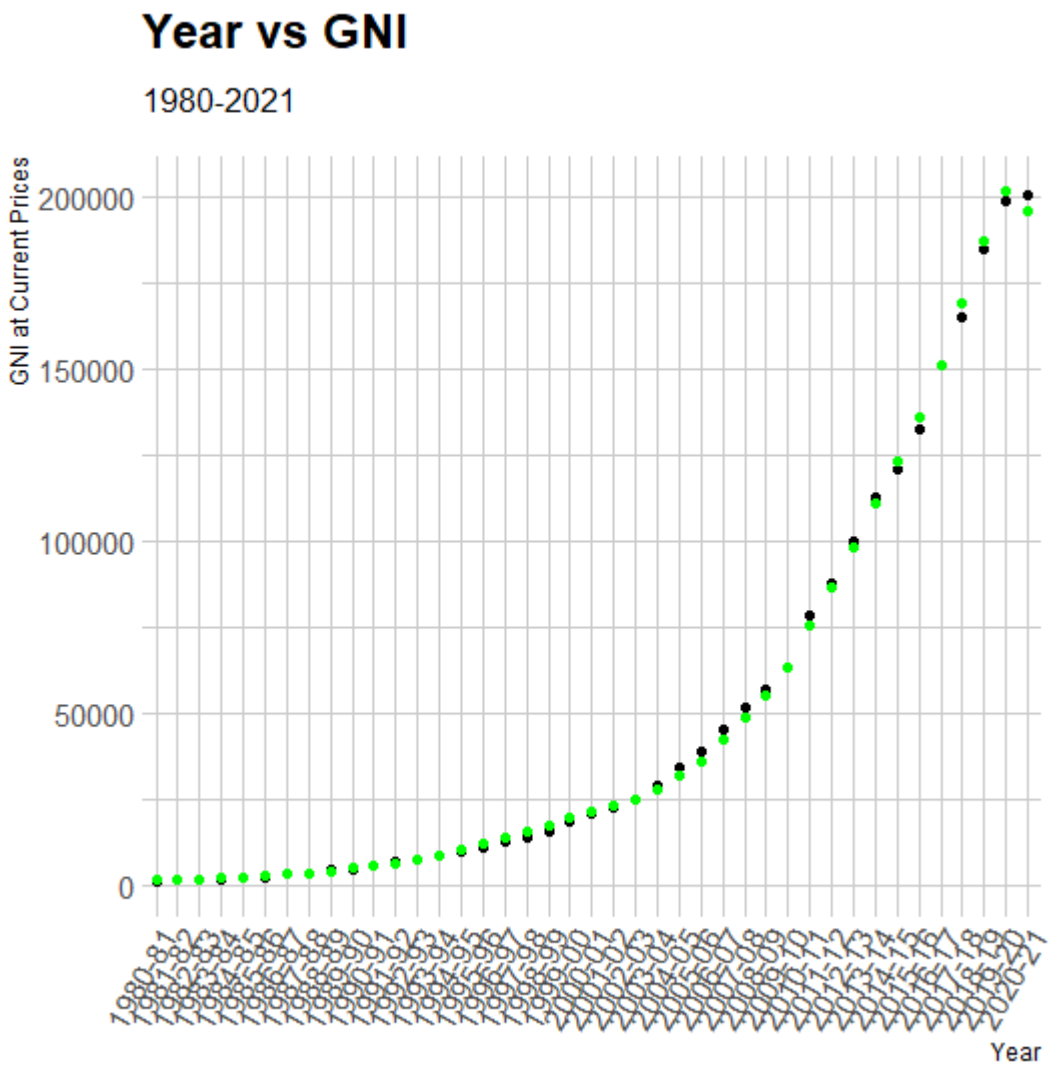
$$R^2_{Adj} = 1 - \frac{n-1}{n-p-1}\frac{SSRes}{SST}$$

As adjusted R-squared value is 0.9865, we can conclude that 98.88% variability of our response variable (GNI at current prices) can be explained by the regressors we included in the model.

Finally, from our analysis we come to conclude that agricultural production of commercial products , production of crude oil and petroleum , total savings deposit in commercial banks, gross fiscal deficit, combined net borrowing of both state and central government, Currency with public, total developmental and non-developmental expenditures of government, net bank credited to government, invested by LIC, combined liabilities of Central and State government, Export of principle commodities, Import of principle commodities, Foreign Exchange reserve in gold, foriegn currency assests etc and Currency in circulation these economical variables have effects on the change of Indian GNI at current prices. By optimizing these variables we can optimize the Indian GNI at current prices. We also see that gross fiscal deficit combined net borrowing of both state and central government have negative impacts on GNI.

Now, we visualize our fitted and observed responses for the time period 1980-2021.

GRAPHICAL OVERVIEW OF THE MODEL



Here **black** dots represent fitted and **green** dots represent observed values of Y.
We can see from the figure that our model is satisfactorily efficient in explaining the change in Indian GNI at current prices.
We are satisfied with our model, but we also furthur want to use LASSO technique if we get a better model or not than the previous one.

LASSO REGRESSION

LASSO Meaning

The word "LASSO" stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularisation of data models and feature selection.

REGULARIZATION

Regularization is an important concept that is used to avoid overfitting of the data, especially when the trained and test data are much varying.
Regularization is implemented by adding a "penalty" term to the best fit derived from the trained data, to achieve a lesser variance with the tested data and also restricts the influence of predictor variables over the output variable by compressing their coefficients.
In regularization, what we do is normally we keep the same number of features but reduce the magnitude of the coefficients. We can reduce the magnitude of the coefficients by using different types of regression techniques which uses regularization to overcome this problem.
Lasso regression is a type of Regularization that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.
In other words, Lasso regression performs L1 regularization technique, which adds a penalty equal to the absolute value of the magnitude of coefficients. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) doesn't result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

What is L1 Regularization

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) doesn't result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

Performing the Regression

Lasso solutions are quadratic programming problems, which are best solved with software . The goal of the algorithm is to minimize:
$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$
Which is the same as minimizing the sum of squares with constraint $\sum|\beta_j| \leq s$ Some of the β s are shrunk to exactly zero, resulting in a regression model that's easier to interpret.
A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

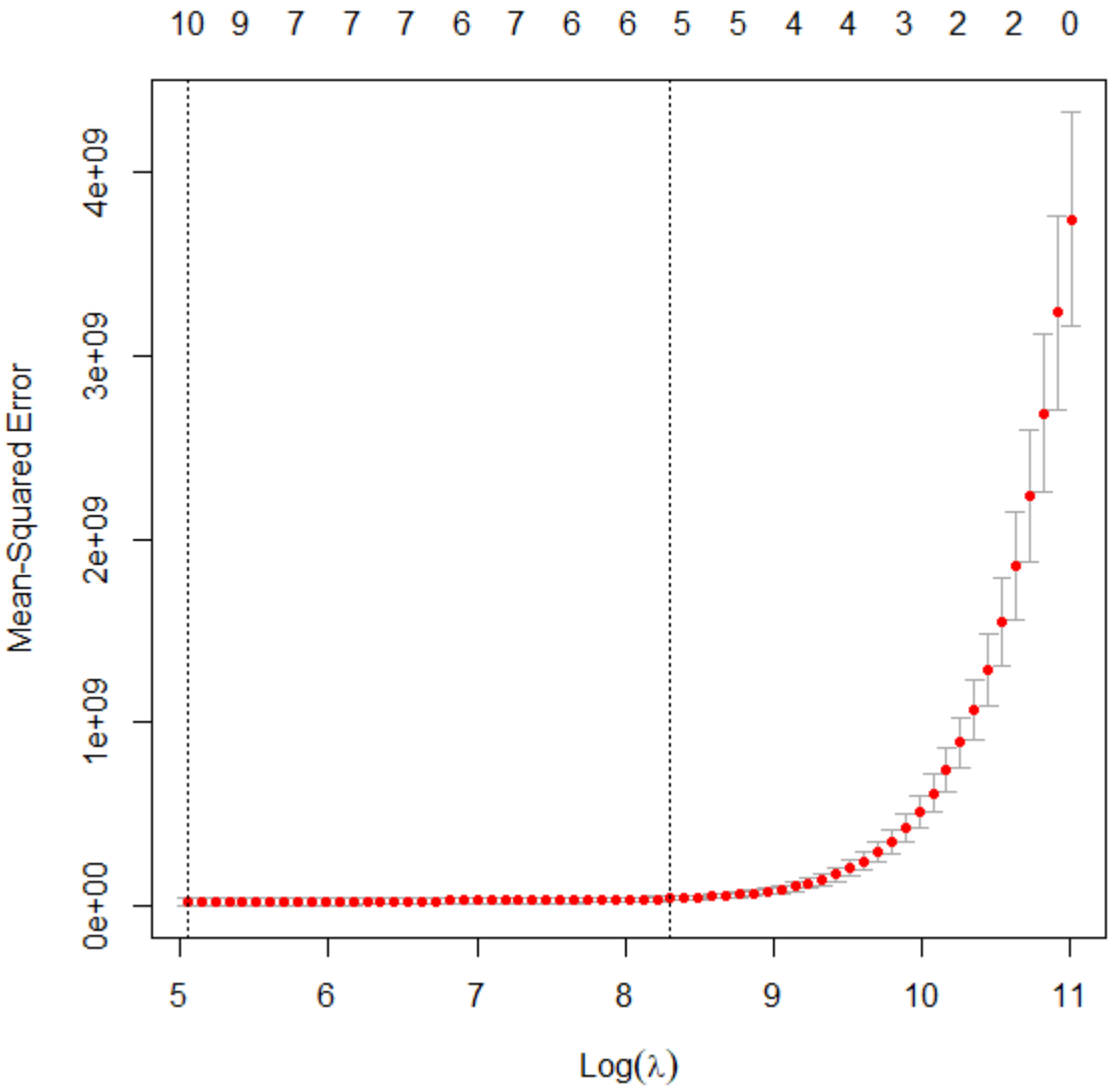
- When $\lambda = 0$, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when $\lambda = \infty$, all coefficients are eliminated).
- As λ increases, bias increases.
  λ As λ decreases, variance increases.

if an intercept is included in the model, it is usually left unchanged.

Analyze Final Model in LASSO

we analyze the final model produced by the optimal lambda value.
To determine what value to use for lambda, we'll perform k-fold cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE).
Generally for coding, automatically performs k-fold cross validation using k = 10 folds.
The lambda value that minimizes the test MSE turns out to be 155.7381

```
>   best_lambda
[1] 155.7381
```



No coefficient is shown for the predictor X1,X5,X8,X12,X13,X14,X17,X18 and X20, because the lasso regression shrunk the coefficient all the way to zero. This means it was completely dropped from the model because it wasn't influential enough.

```
>   coef(best_model)
21 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) -1.302346e+03
X1           .
X2           7.596611e-01
X3           7.729990e+00
X4           7.684219e+01
X5           .
X6           3.690882e+00
X7           6.081934e-01
X8           .
X9          -3.099209e-01
X10          .
X11          1.968277e-02
X12          .
X13          .
X14          .
X15          7.831350e-01
X16          4.947532e-02
X17          .
X18          .
X19         -2.091705e-01
X20          .
```

Lastly, we can calculate the R-squared of the model on the dataset

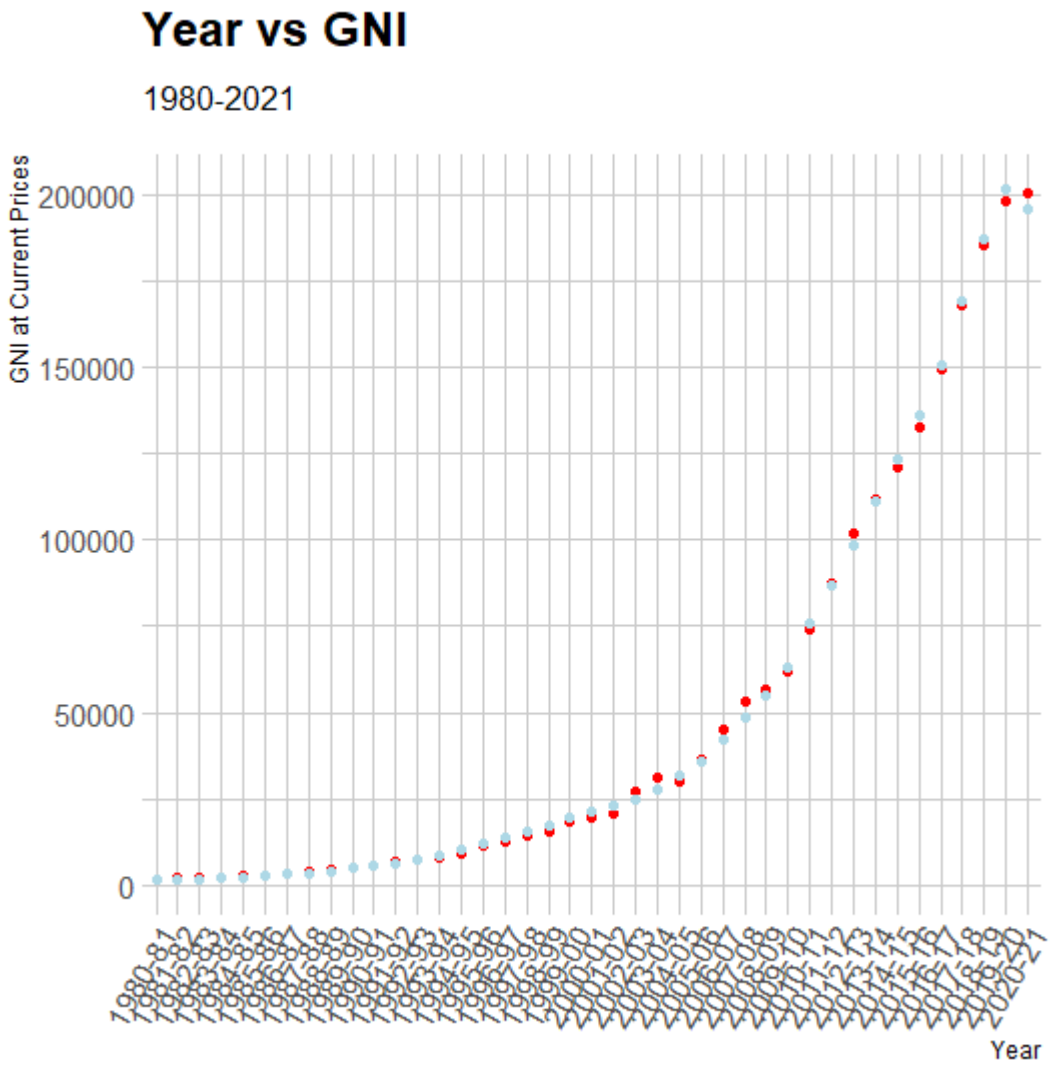
```
> SST
[1] 148657659532
> SSRes
[1] 136633245
> Adj_rsq
[1] 0.9987745
```

Here **red** dots represent fitted and **lightblue** dots represent observed values of Y.
The R-squared turns out to be 0.9987745. That is, the best model was able to explain 99.87 % of the variation in the response values of our dataset.

Final Fitted Model

Our final model after LASSO regression is given by,

$$\hat{Y} = (-1.302346e{+}03) + (7.596611e{-}01)X2 + (7.729990e{+}00)X3 + (7.684219e{+}01)X4 + (3.690882e00)X6 + (6.081934e{-}01)X7 + (-3.099209e{-}01)X9 + (1.968277e{-}02)X11 + (7.631350e{-}01)X15 + (4.947532e{-}02)X16 + (-2.091705e{-}01)X19$$



Note that this is a key difference between ridge regression and lasso regression. Ridge regression shrinks all coefficients towards zero, but lasso regression has the potential to remove predictors from the model by shrinking the coefficients completely to zero.

FINAL CONCLUSION ON LASSO REGRESSION

Finally, from our analysis we come to conclude that agricultural production of commercial products , production of crude oil and petroleum , Import of crude oil and petroleum, direct and indirect tax revenue, total savings deposit in commercial banks, combined net borrowing of both state and central government, total developmental and non-developmental expenditures of government, Export of principle commodities, Import of principle commodities Net inflow of aid these economical variables have effects on the change of Indian GNI at current prices. By optimizing these variables we can optimize the Indian GNI at current prices. We also see that combined net borrowing of both state and central government Net inflow of aid have negative impacts on GNI.

1.2 FINAL CONCLUSION

we see that the adjusted R-square for the which we get previously through ridge regression is 0.9884 which is less than the adjusted R-square obtained from the LASSO technique. This difference is very small. So both of our model fitting exercises are quite satisfactory. Now, we visualize our fitted and observed responses for the time period 1980-2021.

BIBLIOGRAPHY

- Lectures and lecture notes of MTH 416A Regression Analysis class of Dr. Sharmistha Mitra, Associate professor, Department of Mathematics and Statistics, IIT Kanpur

- Introduction to Linear Regression Analysis: D.C. Montgomery, Peck , Vinning

- different materials from internet for our project