EFFECTS OF THE MAJOR ECONOMICAL VARIABLES ON THE CHANGE IN GROSS NATIONAL INCOME(GNI) (AT CURRENT PRICES) OF INDIA (1981-2021)

Submitted By:

Shiladitya Bose(211379)

Sourav Pal(211394)

Krishnendu Paul(211322)

COURSE PROJECT ON 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 timeperiod. 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

Sourav Pal

Krishnendu 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.

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

where : **PERSONAL CONSUMPTION (C), BUSINESS INVESTMENT (I), GOVERNMENT SPENDING (G), EXPORTS - IMPORTS (X)**

1.2 OBJECTIVE

- Collected data on GNI (at current price) and on 20 other economic variables for past 40 years and performed Data Cleansing task.
- 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.
- Also working on to solve multicollinearity problems using VIF and Variance Decomposition Method. then apply stepwise selection and then Ridge regression to introduce bias and remove Multicollinearity problem.
- Also applying LASSO technique to select regressors and compare the results obtained from both Ridge and LASSO technique. Finally we will come to a decision to which model will be preferred most to serve our purpose.

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 40 years i.e 1981-2021 from Handbook of Statistics on Indian Economy available at http://www.rbi.org.in. and World Bank National Accounts Data.

2.2 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:

- Missing value Treatment
- Duplicate data Treatment

2.2.1 LOADING DATA AND TREATMENT OF MISSING AND DUPLICATE VALUES

MISSING VALUE TREATMENT

```
[1] 0
```

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

DETECTION AND REMOVAL OF DUPLICATE VALUES

```
[1] 0
```

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

2.3 DATASET DESCRIPTION

Our data consists the information on the following variables:

```
[1] 40 22
```

Here, Our dataset contain 880 values.

```
'data.frame': 40 obs. of  22 variables:
 $ Year: chr  "1981-82" "1982-83" "1983-84" "1984-85" ...
 $ Y   : int  1728 1926 2241 2508 2831 3166 3592 4249 4875 5686 ...
 $ X1  : num  133 130 152 146 150 ...
 $ X2  : num  928 905 888 1035 982 ...
 $ X3  : int  50 57 61 69 73 75 77 83 81 82 ...
 $ X4  : int  22 20 21 19 17 22 24 25 28 32 ...
 $ X5  : num  1719 1723 1858 1984 2125 ...
 $ X6  : num  240 271 312 358 430 ...
 $ X7  : num  130 151 178 217 246 ...
 $ X8  : num  127 156 194 256 294 ...
 $ X9  : num  35.6 42 45.9 48.7 60.7 ...
 $ X10 : num  139 155 180 213 239 ...
 $ X11 : num  445 521 600 717 786 ...
 $ X12 : num  306 353 406 503 583 ...
 $ X13 : num  68.9 78.3 88.6 99.5 112.9 ...
 $ X14 : num  804 978 1111 1329 1592 ...
 $ X15 : num  78.1 88 97.7 117.4 109 ...
 $ X16 : num  136 143 158 171 197 ...
 $ X17 : num  0.78 0.66 0.1 0.23 1.36 1.51 2.72 1.25 4.06 4.2 ...
 $ X18 : num  40.2 47.8 59.7 72.4 78.2 ...
 $ X19 : int  964 1258 1338 1452 1449 2024 2893 2460 2595 3181 ...
 $ X20 : num  154 176 206 239 265 ...
```

RESPONSE VARIABLE (**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** (**X1**) In **Million Metric Tonnes** : 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** (**X2**) in **Million Metric tonnes**: 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** (**X3**) in **Million Metric tonnes**:

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** (**X4**) in **Million Metric tonnes**: [**Data source : Ministry of Petroleum and Natural Gas, Government of India, PPAC**]
- **GOLD PRICE IN MUMBAI(INDIA)** (**X5**) in **Rupees**: 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** (**X6**) in **Billion Rupees**: 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** (**X7**) in **Billion Rupees**: 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** (**X8**) in **Billion Rupees**: 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** (**X9**) in **Billion Rupees**: 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**]
- **GOVERNMENT'S DEVELOPMENTAL AND NON-DEVELOPMENTAL EXPENDITURE** (**X11**) in **Billion Rupees**: 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** (**X12**) in **Billion Rupees**: 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** (**X13**) in **Billion Rupees**: 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** (**X14**) in **Billion Rupees**: 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** (**X15**) in **Billion Rupees**: 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** (**X16**) in **Billion Rupees**: 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** (**X17**) in **Billion Rupees**: 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** (**X18**) in **Billion Rupees**: 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** (**X19**) in **Crore Rupees**: 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** (**X20**) in **Billion Rupees**: 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:

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

Where, ϵ be the error term in the model.

With the following assumptions;

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

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

and

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

3.2 NORMAL EQUATIONS

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

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

$$\text{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

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:

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

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

$$X^t X \beta = X^t Y$$

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

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

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

4 ORDINARY LEAST SQUARE FITTING

4.1 TEST FOR SIGNIFICANCE OF REGRESSORS

$$H_0 : \beta_1 = \beta_2 \ldots = \beta_{20} = 0 \qquad against$$

$$H_A : at\, least\, one\, \beta_j \neq 0$$

```
Call:
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
-1116.50  -215.01   -54.35   170.63  1167.94

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.434e+03  2.647e+03  -0.920 0.369307
X1          -4.542e+00  1.310e+01  -0.347 0.732578
X2           5.607e+00  3.471e+00   1.615 0.122719
X3          -3.057e+01  1.971e+01  -1.551 0.137425
X4           1.905e+01  3.097e+01   0.615 0.545699
X5          -2.304e-01  2.467e-01  -0.934 0.362032
X6           1.174e+00  1.090e+00   1.077 0.294843
X7          -1.354e+00  5.259e-01  -2.576 0.018520 *
X8          -1.580e+00  1.213e+00  -1.303 0.208193
X9           6.834e-01  9.775e-01   0.699 0.492975
X10         -3.411e+00  1.143e+00  -2.985 0.007609 **
X11          3.712e+00  1.308e+00   2.837 0.010534 *
X12         -7.759e-01  8.620e-01  -0.900 0.379332
X13          3.235e+00  9.034e-01   3.581 0.001994 **
X14          1.960e-01  2.364e-01   0.829 0.417282
X15          1.950e+00  8.482e-01   2.299 0.033050 *
X16         -3.679e-01  3.691e-01  -0.997 0.331467
X17          8.364e-02  1.252e+00   0.067 0.947442
X18         -1.451e+00  3.413e-01  -4.251 0.000432 ***
X19         -2.915e-02  3.206e-02  -0.909 0.374572
X20          1.275e+00  7.184e-01   1.775 0.091987 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 670.6 on 19 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 1.626e+04 on 20 and 19 DF,  p-value: < 2.2e-16
```

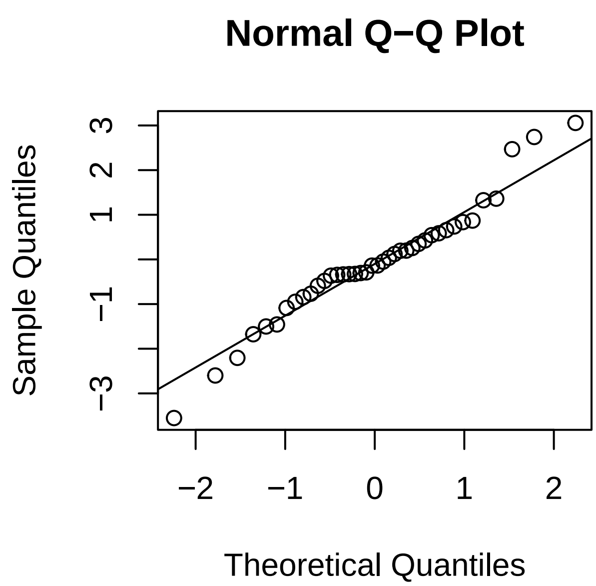
F statistics = 1.626e+04
p-value = 2.2e-16 < 0.05
So, we reject the null hypothesis at 5% level of significance and conclude on the basis of the given data that all the variables are not insignificant in explaining the GNI data.

5 STANDARDIZED THE RESIDUALS & INSPECTION OF THE NORMALITY ASSUMPTION OF ERRORS

5.1 INSPECTION OF THE NORMALITY ASSUMPTION OF ERRORS

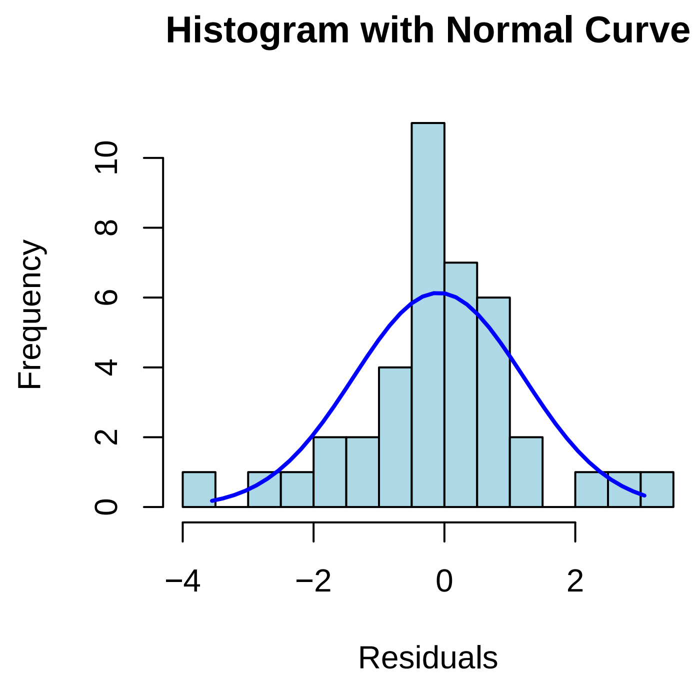
5.1.1 Q-Q PLOT

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.

5.1.2 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.

5.1.3 SHAPIRO-WILK TEST FOR NORMALITY

Here, the null hypothesis is,

H_0: ERRORS ARE NORMALLY DISTRIBUTED

against

H_A : 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.

$\hat{e_{(i)}}$ is the ith order statistic.

$\bar{e}$ is the sample mean.

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{C} \qquad and$$

$$C = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$$

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:  data1.stdres
W = 0.96202, p-value = 0.1962
```

Test statistic, W = 0.96202 and the p-value is 0.1962>0.05(α)

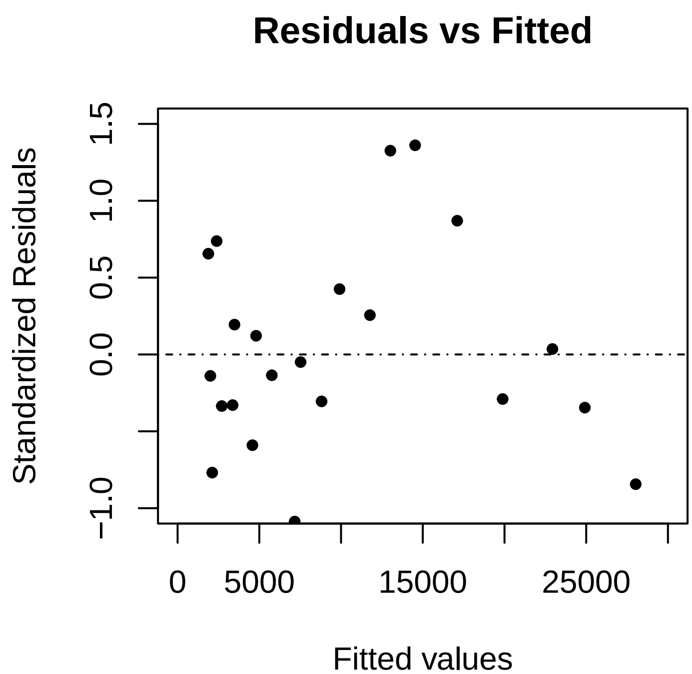
So, we fail to reject the null hypothesis at 5% level of significance and conclude on the basis of the given data that the distribution of errors is not significantly different from Normal Distribution.

So, our assumption is true.

5.2 INSPECTION OF HOMOSCEDASTIC ASSUMPTION OF ERRORS

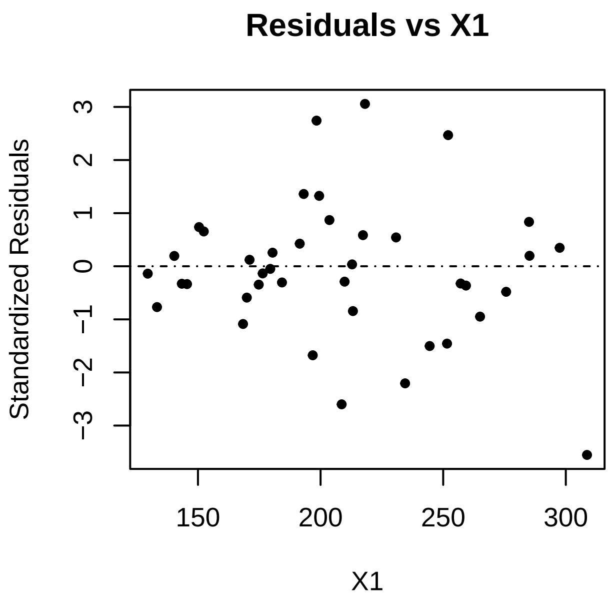
5.2.1 RESIDUAL VS FITTED PLOT

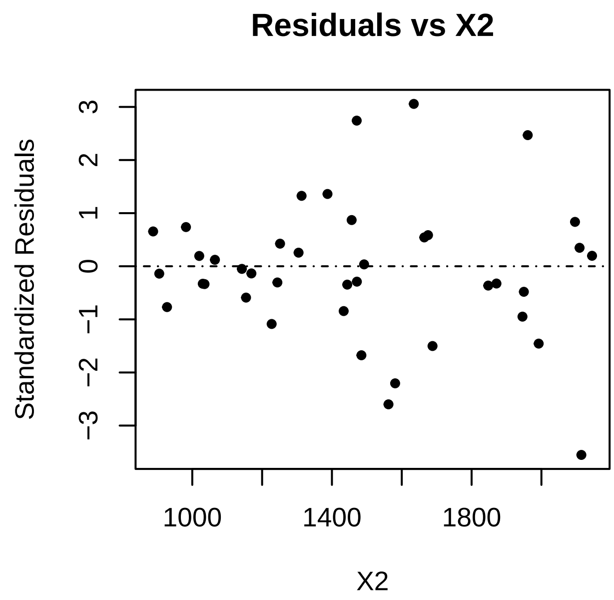
Here we plot the residuals against the fitted responses. If the errors are homoscedastic, then we would expect a horizontal band and completely random pattern around $\hat{e_i} = 0$ line. If any pattern is detected this will indicate that the variances may be non constant.

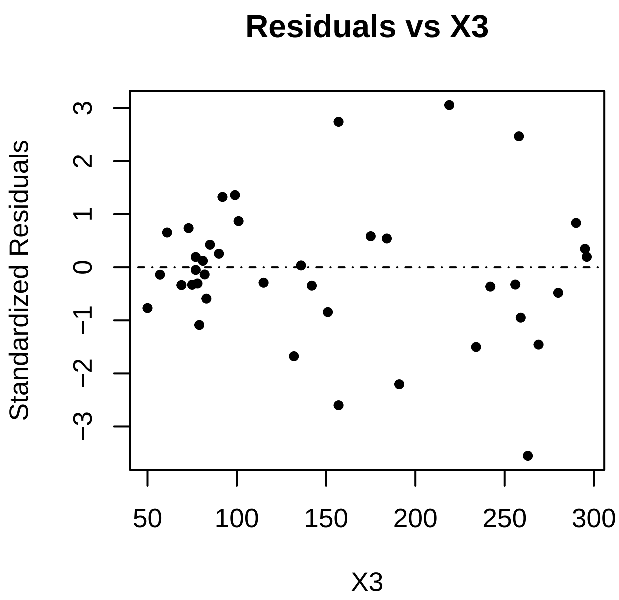


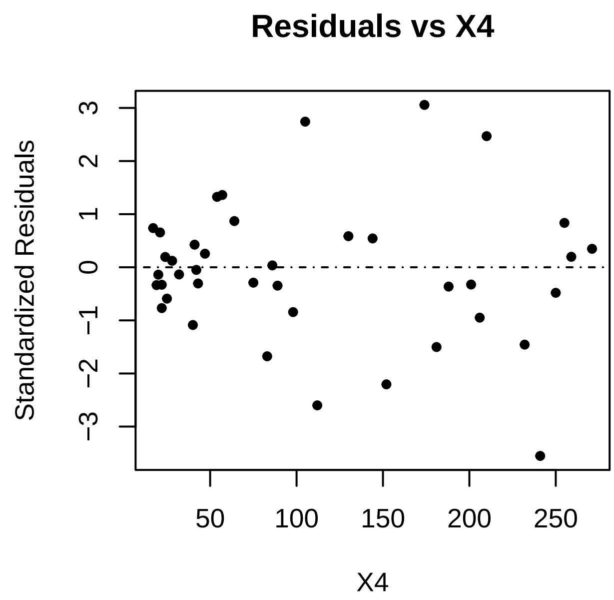
From the diagram, we can see a more or less random pattern among the residuals about the horizontal band. So we can conclude that, the assumption based on homoscedasticity is true in our Model.

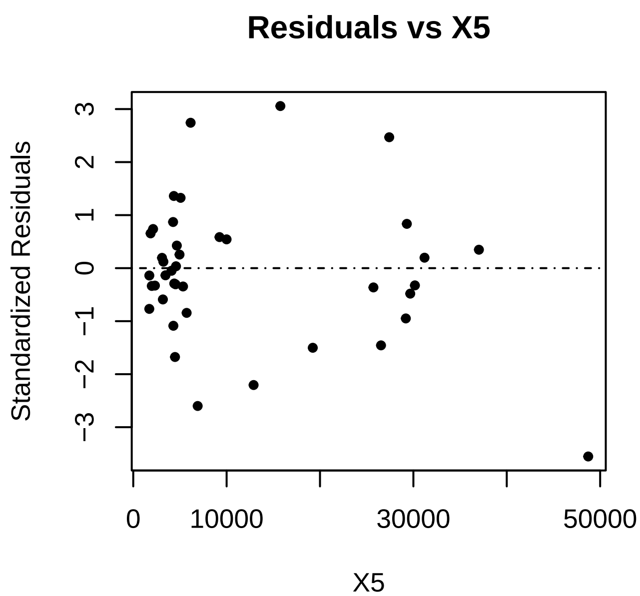
5.2.2 RESIDUAL VS EACH REGRESSORS

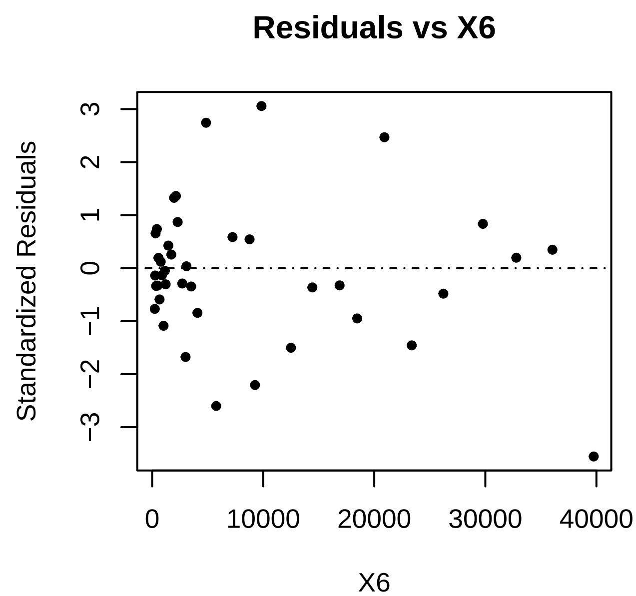


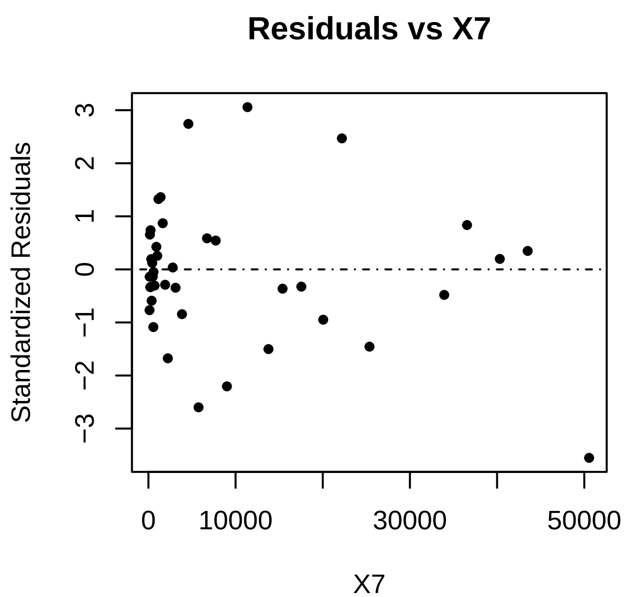


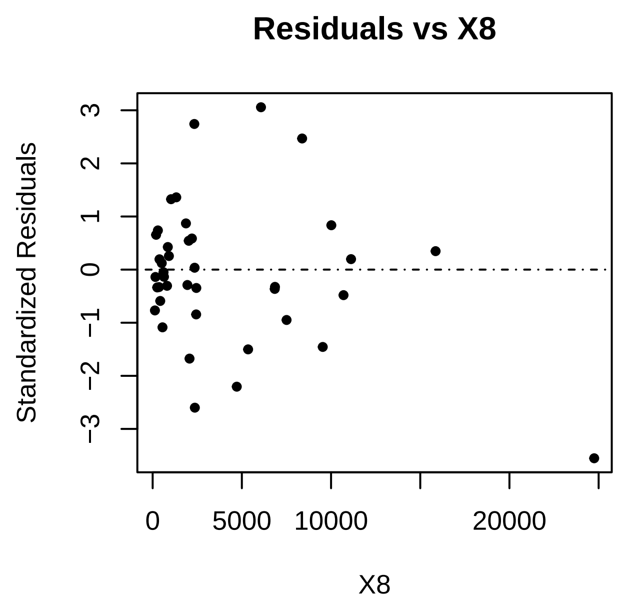


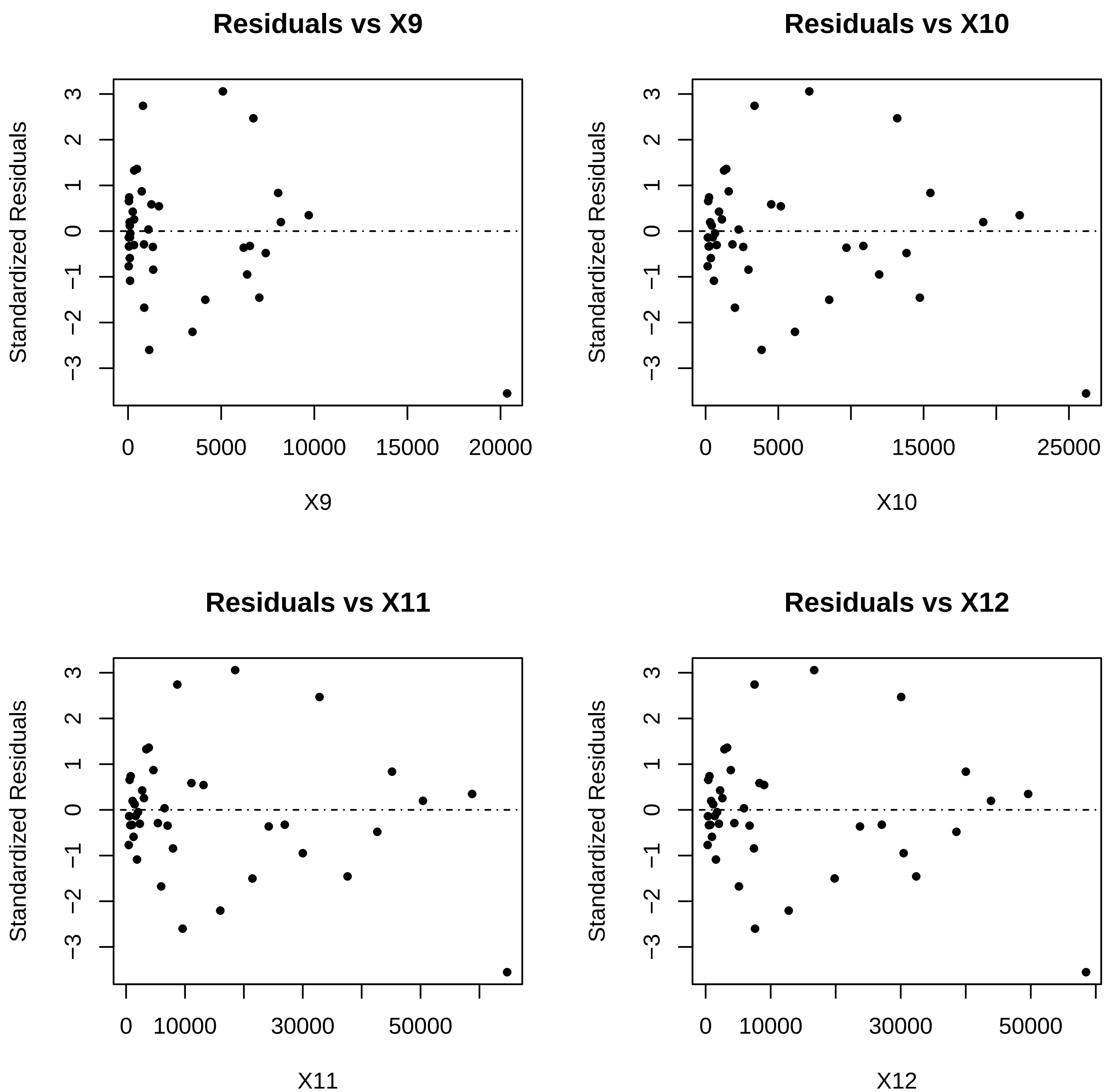


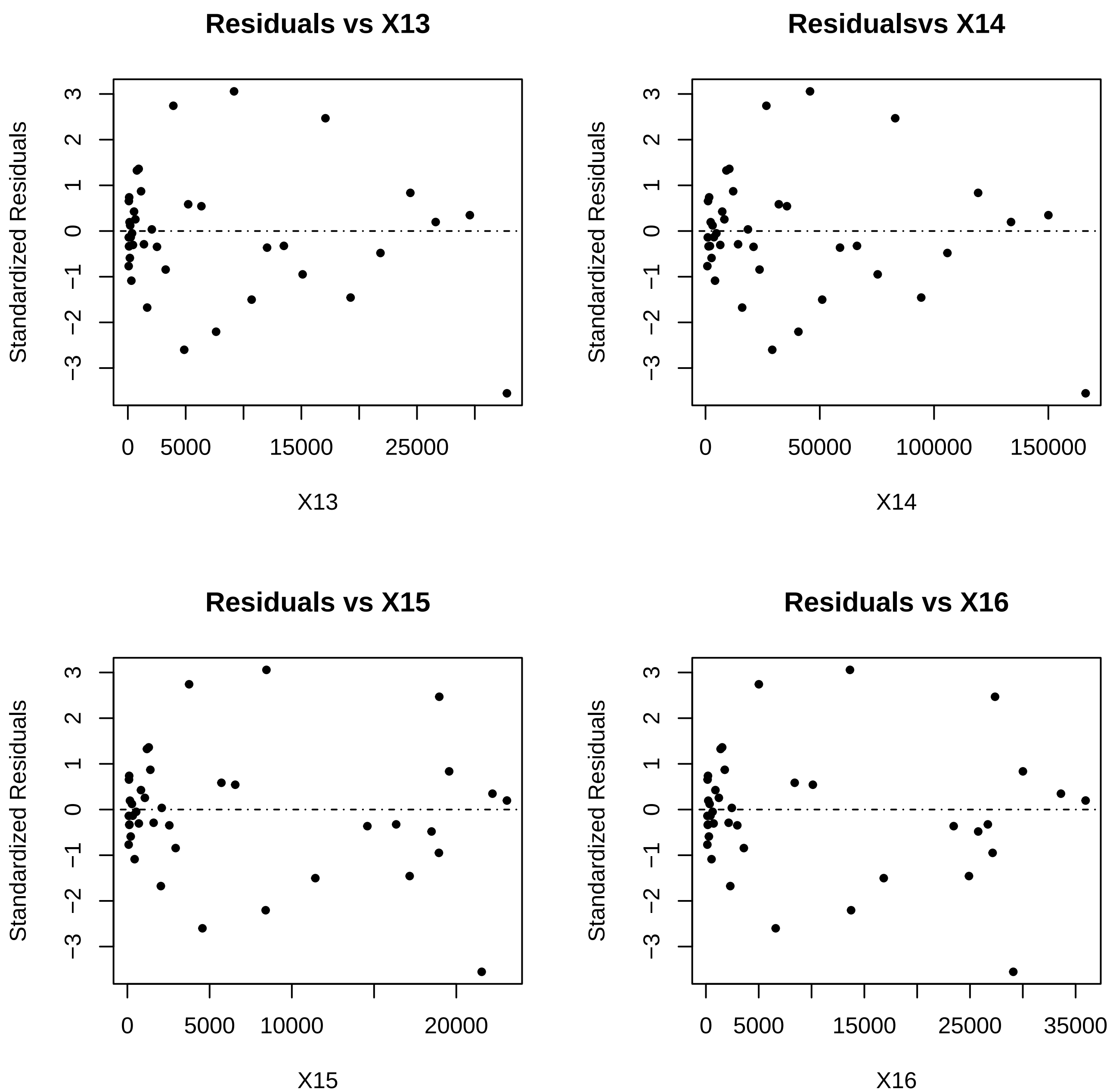

Residuals vs X5
Standardized Residuals
3
2
1
0
−1
−2
−3
0
10000
30000
50000
X5

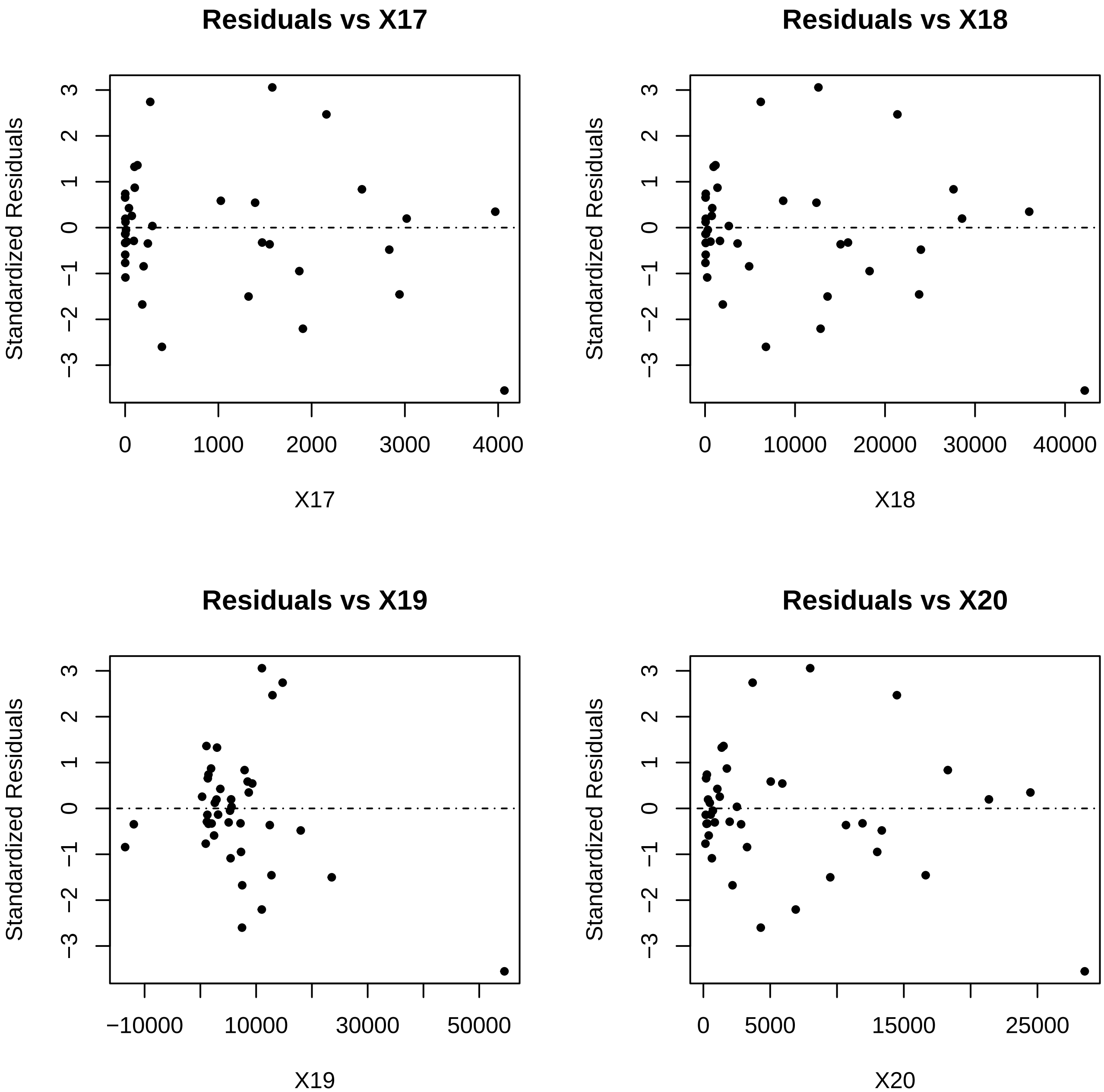

Residuals vs X6
Standardized Residuals
3
2
1
0
−1
−2
−3
0
10000
20000
30000
40000
X6


Residuals vs X7
Standardized Residuals
3
2
1
0
−1
−2
−3
0
10000
30000
50000
X7


Residuals vs X8
Standardized Residuals
3
2
1
0
−1
−2
−3
0
5000
10000
20000
X8


Residuals vs X9
Standardized Residuals
X9
Residuals vs X10
Standardized Residuals
X10
Residuals vs X11
Standardized Residuals
X11
Residuals vs X12
Standardized Residuals
X12


Residuals vs X13
Standardized Residuals
X13
Residualsvs X14
Standardized Residuals
X14
Residuals vs X15
Standardized Residuals
X15
Residuals vs X16
Standardized Residuals
X16



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.

5.3 BREUSCH-PAGAN TEST FOR HETEROSCEDASTICITY

In our assumptions for the MLR model, we assume, $\varepsilon_i \sim N(0,\sigma^2), \forall\ i=1(1)n$ i.e. the homoscedasticity of the random errors. We can check the validity of our assumption by looking at the residual plot for the model. A random pattern in the residual plot implies the homoscedasticity of the errors.

From the residual plot we can see that there is a random pattern in the residual plot and hence we can conclude that our errors are homoscedastic. Also, we can conduct the Breusch – Pagan test to check the homoscedasticity of the residuals. This test investigates whether the estimated variance of the residuals from the regression are dependent on the values of the independent variables. Here we test

H_0: RESIDUALS ARE HOMOSCEDASTIC AGAINST,

H_1: H_0 IS NOT TRUE.

The result of the test is

```
Loading required package:  zoo

Attaching package:  'zoo'
The following objects are masked from 'package:base':
      as.Date, as.Date.numeric


studentized Breusch-Pagan test

data:  data1
BP = 25.01, df = 20, p-value = 0.2011
```

So, p-value is 0.2011.

As the p–value of the test is $0.2011 > 0.05$

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.

5.4 INSPECTION OF AUTOCORRELATION AMONG THE ERRORS

For our dataset, **n=40, p=20,** $\boldsymbol{\alpha = 0.05}$

Hence we would like to test,

$H_0 : \rho = 0$

against

$H_A : \rho \neq 0$

using **DURBIN-WATSON** test to check the existance of serial correlation in the data.

```
Durbin-Watson test

data:  data1
DW = 2.5148, p-value = 0.8777
alternative hypothesis: true autocorrelation is not 0
```

In our model the value of **DURBIN-WATSON** Statistic is d=2.5148. .
& the **p-value** of the test is $0.8777 > 0.05(\alpha)$,
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.

6 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.

6.1 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.

6.1.1 VARIANCE INFLATION FACTOR

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.

```
Loading required package:  carData

         X1          X2          X3          X4          X5          X6 
  34.232569  149.338235  225.922469  589.885012  807.514276 12862.250643 
         X7          X8          X9         X10         X11         X12 
4592.137581 3349.536917 1406.502449 5159.092700 46511.252714 16250.121413 
        X13         X14         X15         X16         X17         X18 
6201.474363 9932.797271 3895.334785 1662.014538  197.201113  1275.958482 
        X19         X20 
   9.296434 2503.744623 
```

6.1.2 MULTICOLLINEARITY DIAGONISTIC 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 $X^T X$ 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.

```
Call:
eigprop(mod = data1)

   Eigenvalues        CI (Intercept)     X1     X2     X3     X4     X5     X6
1      18.8946    1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.4081    3.6631      0.0004 0.0002 0.0001 0.0001 0.0000 0.0000 0.0000
3       0.4629    6.3889      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
```

```
4       0.1100   13.1053      0.0009 0.0001 0.0000 0.0005 0.0006 0.0001 0.0000
5       0.0523   19.0076      0.0000 0.0000 0.0000 0.0000 0.0001 0.0024 0.0000
6       0.0299   25.1354      0.0020 0.0000 0.0000 0.0015 0.0022 0.0001 0.0001
7       0.0182   32.2196      0.0142 0.0000 0.0003 0.0048 0.0065 0.0016 0.0000
8       0.0088   46.2374      0.0003 0.0000 0.0008 0.0005 0.0018 0.0023 0.0000
9       0.0050   61.4161      0.0307 0.0200 0.0020 0.0100 0.0003 0.0137 0.0000
10      0.0030   78.7772      0.0754 0.1053 0.0049 0.0275 0.0002 0.0124 0.0008
11      0.0025   87.5885      0.0152 0.0117 0.0123 0.0126 0.0001 0.0782 0.0000
12      0.0015  113.6172      0.0000 0.0041 0.0165 0.0133 0.0119 0.0173 0.0001
13      0.0013  122.0139      0.0000 0.3002 0.0660 0.0004 0.0024 0.0370 0.0020
14      0.0007  167.0597      0.1121 0.2290 0.0291 0.4345 0.1786 0.0019 0.0000
15      0.0006  177.7019      0.0778 0.1242 0.0831 0.2651 0.2198 0.0059 0.0002
16      0.0003  251.4914      0.0342 0.0050 0.0532 0.0000 0.0365 0.0189 0.0018
17      0.0001  371.8260      0.0389 0.0487 0.0002 0.0150 0.0017 0.0009 0.0370
18      0.0001  397.4558      0.2448 0.0009 0.2202 0.0403 0.0005 0.0415 0.1755
19      0.0001  527.7724      0.0076 0.0920 0.1065 0.0162 0.0070 0.1982 0.0000
20      0.0000  705.6900      0.1647 0.0224 0.3081 0.1532 0.3452 0.1596 0.0799
21      0.0000 1591.5178      0.1807 0.0362 0.0967 0.0045 0.1845 0.4079 0.7026
       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 0.0108
6  0.0007 0.0005 0.0014 0.0000 0.0000 0.0000 0.0001 0.0000 0.0001 0.0004 0.0255
7  0.0001 0.0002 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0010 0.0604
8  0.0016 0.0002 0.0006 0.0019 0.0000 0.0002 0.0001 0.0000 0.0000 0.0004 0.0246
9  0.0010 0.0046 0.0242 0.0005 0.0001 0.0002 0.0008 0.0000 0.0001 0.0008 0.0124
10 0.0007 0.0080 0.0075 0.0013 0.0000 0.0005 0.0001 0.0000 0.0014 0.0000 0.0000
11 0.0007 0.0000 0.0169 0.0001 0.0000 0.0001 0.0002 0.0001 0.0020 0.0100 0.0499
12 0.0021 0.0018 0.0000 0.0000 0.0001 0.0000 0.0003 0.0005 0.0428 0.0179 0.0150
13 0.0000 0.0175 0.0045 0.0017 0.0000 0.0000 0.0000 0.0001 0.0000 0.0422 0.0424
14 0.0007 0.0152 0.0052 0.0030 0.0000 0.0014 0.0000 0.0013 0.0067 0.0563 0.0675
15 0.0041 0.0160 0.0078 0.0021 0.0000 0.0038 0.0022 0.0032 0.0142 0.0692 0.0007
16 0.0317 0.0039 0.0003 0.1900 0.0009 0.0099 0.0268 0.0109 0.0002 0.0002 0.0038
17 0.0279 0.0395 0.0003 0.0433 0.0037 0.0576 0.1837 0.0965 0.0008 0.0003 0.1949
18 0.1995 0.0029 0.0260 0.0025 0.0045 0.0025 0.1365 0.0193 0.0130 0.0236 0.0081
19 0.1963 0.0005 0.0009 0.0306 0.0256 0.0199 0.3350 0.3289 0.0066 0.0478 0.2075
20 0.4992 0.3623 0.1027 0.2813 0.0163 0.1738 0.2328 0.0105 0.3753 0.2815 0.0066
21 0.0336 0.5267 0.7989 0.4416 0.9488 0.7300 0.0813 0.5285 0.5363 0.4476 0.2697
```

```
     X18    X19    X20
1  0.0000 0.0001 0.0000
2  0.0000 0.0001 0.0000
3  0.0000 0.1248 0.0000
4  0.0000 0.0887 0.0000
5  0.0004 0.0406 0.0000
6  0.0007 0.0719 0.0000
7  0.0001 0.0054 0.0000
8  0.0026 0.0072 0.0124
9  0.0084 0.0055 0.0000
10 0.0130 0.0268 0.0003
11 0.0573 0.0101 0.0000
12 0.0557 0.0049 0.0202
13 0.0263 0.0005 0.0013
14 0.0086 0.0032 0.0032
15 0.0073 0.0052 0.0142
16 0.0036 0.0333 0.0958
17 0.0169 0.1520 0.0761
18 0.0099 0.1151 0.0353
19 0.4231 0.0245 0.0003
20 0.1401 0.2001 0.6700
21 0.2259 0.0799 0.0707

===============================
Row 21==> X6, proportion 0.702558 >= 0.50
Row 21==> X8, proportion 0.526671 >= 0.50
Row 21==> X9, proportion 0.798869 >= 0.50
Row 21==> X11, proportion 0.948756 >= 0.50
Row 21==> X12, proportion 0.730031 >= 0.50
Row 21==> X14, proportion 0.528517 >= 0.50
Row 21==> X15, proportion 0.536278 >= 0.50
Row 20==> X20, proportion 0.670038 >= 0.50
```

STEP 1:

- So, the subsets (X6,X8,X9,X11,X12,X14,X15) and (X20) are involved in Multicollinearity.
- In the first subset VIF of X11 is the highest and in the second subset the VIF of X20 is highest.
- We drop the variables X11 and X20 and again fit a model.

```
Call:
eigprop(mod = olsreg_1)

   Eigenvalues       CI (Intercept)     X1     X2     X3     X4     X5     X6
1      16.9364   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.3776   3.5063      0.0005 0.0002 0.0001 0.0001 0.0000 0.0000 0.0000
3       0.4601   6.0670      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
4       0.1077  12.5426      0.0012 0.0001 0.0000 0.0005 0.0009 0.0001 0.0000
5       0.0515  18.1365      0.0000 0.0000 0.0000 0.0000 0.0002 0.0043 0.0001
6       0.0296  23.9233      0.0029 0.0000 0.0000 0.0017 0.0038 0.0002 0.0003
7       0.0181  30.6227      0.0183 0.0000 0.0004 0.0050 0.0105 0.0030 0.0000
8       0.0049  59.0293      0.0471 0.0261 0.0034 0.0128 0.0007 0.0239 0.0000
9       0.0048  59.5130      0.0014 0.0002 0.0044 0.0034 0.0113 0.0092 0.0002
10      0.0030  75.3081      0.0901 0.1096 0.0037 0.0245 0.0013 0.0260 0.0019
11      0.0025  82.9896      0.0200 0.0128 0.0160 0.0142 0.0003 0.1396 0.0001
12      0.0013 114.8664      0.0000 0.2862 0.0967 0.0009 0.0002 0.0288 0.0054
13      0.0011 125.8783      0.0003 0.0816 0.0025 0.0174 0.0077 0.0634 0.0012
14      0.0007 159.3576      0.1933 0.1485 0.0136 0.6405 0.4382 0.0003 0.0003
15      0.0005 177.4788      0.0678 0.2106 0.1685 0.1042 0.2542 0.0025 0.0001
16      0.0002 305.2139      0.0620 0.0406 0.0115 0.0008 0.0301 0.0611 0.1200
17      0.0001 369.7613      0.2644 0.0121 0.1264 0.0043 0.0060 0.0570 0.1548
18      0.0001 442.4489      0.2075 0.0005 0.4111 0.1560 0.1927 0.0271 0.6162
19      0.0001 551.8366      0.0230 0.0708 0.1416 0.0139 0.0419 0.5536 0.0994
       X7     X8     X9    X10    X12    X13    X14    X15    X16    X17    X18
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 0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0001 0.0000
3  0.0000 0.0000 0.0003 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0000 0.0000
4  0.0001 0.0010 0.0024 0.0000 0.0000 0.0000 0.0000 0.0007 0.0016 0.0002 0.0000
5  0.0003 0.0003 0.0178 0.0000 0.0000 0.0001 0.0001 0.0006 0.0010 0.0191 0.0006
6  0.0014 0.0018 0.0108 0.0000 0.0001 0.0002 0.0000 0.0006 0.0010 0.0370 0.0008
7  0.0002 0.0006 0.0000 0.0000 0.0001 0.0000 0.0002 0.0005 0.0027 0.0981 0.0001
8  0.0015 0.0181 0.1864 0.0040 0.0007 0.0009 0.0001 0.0005 0.0023 0.0184 0.0088
9  0.0076 0.0006 0.0003 0.0357 0.0023 0.0001 0.0006 0.0008 0.0041 0.0749 0.0271
10 0.0023 0.0359 0.0715 0.0140 0.0017 0.0002 0.0000 0.0099 0.0001 0.0010 0.0106
11 0.0010 0.0000 0.1316 0.0006 0.0003 0.0003 0.0001 0.0088 0.0287 0.0717 0.0704
12 0.0004 0.0466 0.0344 0.0021 0.0003 0.0001 0.0000 0.0323 0.1828 0.0382 0.0058
13 0.0007 0.0102 0.0001 0.0832 0.0037 0.0002 0.0002 0.2649 0.0788 0.0138 0.0697
14 0.0025 0.0366 0.0217 0.0331 0.0008 0.0004 0.0006 0.0028 0.0603 0.0990 0.0098
15 0.0058 0.1205 0.0913 0.2704 0.0061 0.0020 0.0027 0.0116 0.1665 0.0002 0.0336
16 0.3823 0.0148 0.0008 0.0944 0.0206 0.2942 0.0551 0.0664 0.0146 0.0038 0.0117
```

```
17 0.2616 0.1246 0.1580 0.1577 0.2021 0.2564 0.1413 0.1202 0.0767 0.0933 0.0003
18 0.0058 0.5415 0.0483 0.0121 0.4733 0.0721 0.0135 0.2996 0.0842 0.2173 0.1264
19 0.3264 0.0470 0.2240 0.2928 0.2878 0.3728 0.7854 0.1798 0.2946 0.2138 0.6241
      X19
1  0.0002
2  0.0002
3  0.1467
4  0.1126
5  0.0409
6  0.0849
7  0.0074
8  0.0055
9  0.0010
10 0.0317
11 0.0116
12 0.0028
13 0.0060
14 0.0017
15 0.0133
16 0.0071
17 0.3709
18 0.1203
19 0.0352

===============================
Row 14==> X3, proportion 0.640472 >= 0.50
Row 19==> X5, proportion 0.553646 >= 0.50
Row 18==> X6, proportion 0.616229 >= 0.50
Row 18==> X8, proportion 0.541467 >= 0.50
Row 19==> X14, proportion 0.785429 >= 0.50
Row 19==> X18, proportion 0.624145 >= 0.50
```

- So, the subsets (X3), (X6,X8) and (X5,X14,X18) are involved in Multicollinearity.
- In the first subset VIF of X3 is the highest, in the second subset the VIF of X6 is highest and in the third subset the VIF of X14 is highest.
- We drop the variables X3 and X14 and again fit a model.

```
Call:
eigprop(mod = olsreg_2)

   Eigenvalues       CI (Intercept)     X1     X2     X4     X5     X7     X8
1      14.1022   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.2482   3.3613      0.0013 0.0003 0.0002 0.0000 0.0000 0.0000 0.0000
3       0.4467   5.6190      0.0000 0.0000 0.0000 0.0001 0.0000 0.0000 0.0001
4       0.0979  12.0037      0.0016 0.0001 0.0000 0.0023 0.0001 0.0003 0.0022
5       0.0480  17.1336      0.0000 0.0000 0.0000 0.0007 0.0075 0.0006 0.0001
6       0.0251  23.7262      0.0063 0.0000 0.0003 0.0110 0.0000 0.0064 0.0020
7       0.0151  30.5187      0.0485 0.0009 0.0029 0.0483 0.0064 0.0005 0.0015
8       0.0047  54.6159      0.0284 0.0080 0.0005 0.0011 0.0605 0.0010 0.0206
9       0.0043  57.2390      0.0355 0.0035 0.0076 0.0547 0.0014 0.0120 0.0205
10      0.0027  72.8626      0.1037 0.0706 0.0012 0.0049 0.2129 0.0022 0.0345
11      0.0022  80.6316      0.3857 0.1644 0.0353 0.0839 0.0668 0.0128 0.0107
12      0.0012 108.5758      0.0337 0.1761 0.2512 0.0190 0.0381 0.0017 0.1000
13      0.0010 116.6061      0.0051 0.1480 0.0309 0.0716 0.1254 0.0040 0.0286
14      0.0005 161.6779      0.0407 0.3187 0.4655 0.1132 0.0016 0.0131 0.2663
15      0.0002 305.5805      0.3059 0.0556 0.1208 0.3701 0.0323 0.9349 0.0696
16      0.0001 379.1282      0.0035 0.0539 0.0837 0.2190 0.4470 0.0105 0.4433
       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 0.0003
3  0.0005 0.0000 0.0000 0.0000 0.0001 0.0002 0.0001 0.0000 0.1894
4  0.0063 0.0001 0.0000 0.0000 0.0009 0.0021 0.0003 0.0000 0.1715
5  0.0214 0.0000 0.0000 0.0001 0.0008 0.0015 0.0525 0.0020 0.0203
6  0.0228 0.0000 0.0003 0.0007 0.0003 0.0004 0.0595 0.0009 0.1004
7  0.0006 0.0000 0.0003 0.0001 0.0004 0.0032 0.1525 0.0000 0.0056
8  0.1401 0.0136 0.0031 0.0004 0.0000 0.0076 0.1598 0.0740 0.0019
9  0.2540 0.0903 0.0000 0.0004 0.0019 0.0000 0.0290 0.0076 0.0113
10 0.0046 0.0101 0.0008 0.0000 0.0216 0.0097 0.0232 0.1195 0.0459
11 0.1649 0.0017 0.0022 0.0005 0.0000 0.0288 0.1223 0.0475 0.0000
12 0.0691 0.0037 0.0006 0.0000 0.0670 0.2604 0.0923 0.0018 0.0082
13 0.0000 0.1864 0.0034 0.0011 0.2552 0.0583 0.0163 0.1187 0.0036
14 0.1972 0.4804 0.0086 0.0002 0.0155 0.2591 0.0402 0.1061 0.0092
15 0.0676 0.2130 0.0951 0.5862 0.1393 0.0103 0.0907 0.0365 0.0749
16 0.0508 0.0005 0.8856 0.4102 0.4971 0.3585 0.1612 0.4853 0.3570
```

```
================================
Row 15==> X7, proportion 0.934889 >= 0.50
Row 16==> X12, proportion 0.885615 >= 0.50
Row 15==> X13, proportion 0.586226 >= 0.50
```

STEP 2:

- So the subsets (X7,X13) and (X12) are involved in Multicollinearity.
- In the first subset VIF of X7 is the highest and in the second subset the VIF of X12 is highest.
- We drop the variables X7 and X12, and again fit a model.

```


Call:
eigprop(mod = olsreg_3)

   Eigenvalues       CI (Intercept)     X1     X2     X4     X5     X8     X9
1      12.1950   1.0000      0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
2       1.1857   3.2070      0.0017 0.0003 0.0002 0.0000 0.0000 0.0001 0.0002
3       0.4392   5.2695      0.0001 0.0000 0.0000 0.0003 0.0000 0.0001 0.0004
4       0.0920  11.5117      0.0017 0.0000 0.0000 0.0042 0.0000 0.0062 0.0111
5       0.0455  16.3707      0.0000 0.0000 0.0000 0.0022 0.0150 0.0000 0.0188
6       0.0166  27.1179      0.0653 0.0007 0.0031 0.1298 0.0023 0.0044 0.0178
7       0.0119  32.0357      0.0057 0.0002 0.0008 0.0008 0.0185 0.0017 0.0431
8       0.0043  53.2437      0.1188 0.0228 0.0019 0.0209 0.1332 0.0990 0.3943
9       0.0028  66.0981      0.1429 0.0250 0.0330 0.1945 0.0050 0.0468 0.0030
10      0.0026  68.6347      0.0089 0.0242 0.0188 0.0080 0.4893 0.0500 0.0210
11      0.0017  83.9091      0.4367 0.4005 0.0000 0.1195 0.0878 0.0374 0.2053
12      0.0011 103.8041      0.1112 0.0921 0.2617 0.0586 0.0674 0.2102 0.1218
13      0.0010 110.5883      0.0005 0.1003 0.0402 0.0823 0.1812 0.1797 0.0196
14      0.0005 154.3052      0.1064 0.3338 0.6402 0.3788 0.0002 0.3644 0.1436
      X10    X13    X15    X16    X17    X18    X19
1  0.0000 0.0000 0.0000 0.0000 0.0001 0.0000 0.0008
2  0.0000 0.0001 0.0000 0.0000 0.0003 0.0001 0.0016
3  0.0000 0.0001 0.0002 0.0003 0.0002 0.0000 0.3224
4  0.0003 0.0004 0.0017 0.0025 0.0001 0.0001 0.3635
5  0.0001 0.0021 0.0021 0.0027 0.0926 0.0049 0.0104
6  0.0009 0.0168 0.0014 0.0033 0.0161 0.0000 0.0374
```

```
7  0.0095 0.0851 0.0002 0.0006 0.3930 0.0037 0.1098
8  0.0071 0.0103 0.0007 0.0058 0.0544 0.0621 0.0081
9  0.0820 0.2815 0.0005 0.0039 0.0155 0.1742 0.0071
10 0.0412 0.0394 0.0469 0.0312 0.0590 0.2092 0.0627
11 0.0375 0.2538 0.0107 0.0320 0.1222 0.0777 0.0002
12 0.0018 0.1789 0.1787 0.4021 0.1666 0.0332 0.0170
13 0.2188 0.0957 0.6254 0.0659 0.0630 0.2483 0.0116
14 0.6008 0.0358 0.1315 0.4496 0.0170 0.1864 0.0474

===============================
Row 14==> X2, proportion 0.640242 >= 0.50
Row 14==> X10, proportion 0.600792 >= 0.50
Row 13==> X15, proportion 0.625365 >= 0.50
```

STEP 3:

- So, the subsets (X2,X10) and (X15) are involved in Multicollinearity.
- In the first subset VIF of X2 is the highest, in the second subset the VIF of X15 is highest and in the third subset the VIF of X13 is highest.
- We drop the variables X2 and X15, and again fit a model.

```
Call:
eigprop(mod = olsreg_4)

   Eigenvalues      CI (Intercept)     X1     X4     X5     X8     X9    X10
1      10.4974  1.0000      0.0000 0.0000 0.0001 0.0000 0.0000 0.0001 0.0000
2       0.9273  3.3645      0.0040 0.0011 0.0002 0.0000 0.0001 0.0003 0.0000
3       0.4131  5.0411      0.0001 0.0000 0.0007 0.0000 0.0001 0.0003 0.0000
4       0.0821 11.3102      0.0034 0.0001 0.0141 0.0000 0.0073 0.0152 0.0002
5       0.0414 15.9285      0.0009 0.0000 0.0001 0.0236 0.0003 0.0163 0.0001
6       0.0153 26.1688      0.0575 0.0051 0.2830 0.0050 0.0066 0.0210 0.0022
7       0.0117 29.9165      0.0065 0.0017 0.0118 0.0198 0.0021 0.0405 0.0116
8       0.0043 49.6042      0.1402 0.0530 0.0174 0.1743 0.1050 0.4214 0.0071
9       0.0026 63.4225      0.1432 0.1472 0.1931 0.0242 0.0540 0.0021 0.0503
10      0.0023 67.9478      0.0000 0.0107 0.0278 0.5211 0.0340 0.0067 0.1826
11      0.0017 78.2233      0.6356 0.7801 0.2613 0.0880 0.0436 0.2169 0.0757
```

```
12      0.0008 114.3264      0.0086 0.0010 0.1903 0.1440 0.7469 0.2593 0.6701
      X13    X16    X17    X18    X19
1  0.0000 0.0000 0.0001 0.0000 0.0012
2  0.0001 0.0000 0.0003 0.0001 0.0018
3  0.0002 0.0007 0.0005 0.0001 0.3663
4  0.0001 0.0072 0.0027 0.0000 0.3480
5  0.0022 0.0141 0.0954 0.0049 0.0018
6  0.0271 0.0146 0.0077 0.0000 0.0553
7  0.0929 0.0012 0.4323 0.0041 0.1027
8  0.0069 0.0103 0.0447 0.0636 0.0074
9  0.2935 0.0083 0.0724 0.3728 0.0401
10 0.2130 0.2548 0.0104 0.0973 0.0576
11 0.2857 0.0147 0.1174 0.0480 0.0007
12 0.0781 0.6739 0.2161 0.4092 0.0171

===============================
Row 11==> X1, proportion 0.780052 >= 0.50
Row 10==> X5, proportion 0.521083 >= 0.50
Row 12==> X8, proportion 0.746928 >= 0.50
Row 12==> X10, proportion 0.670145 >= 0.50
Row 12==> X16, proportion 0.673901 >= 0.50
```

STEP 4:

- So, the subsets (X1), (X5) and (X8,X10,X16) 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 X8 is highest.
- We drop the variables X1, X5 and X8, and again fit a model.

```
Call:
eigprop(mod = olsreg_5)

  Eigenvalues      CI (Intercept)     X4     X9    X10    X13    X16    X17
1      7.8551  1.0000      0.0010 0.0002 0.0004 0.0000 0.0001 0.0003 0.0002
2      0.6184  3.5639      0.1601 0.0015 0.0017 0.0001 0.0002 0.0001 0.0005
3      0.4079  4.3884      0.0010 0.0014 0.0014 0.0001 0.0003 0.0042 0.0005
```

```
4      0.0592 11.5208      0.1286 0.0504 0.1687 0.0018 0.0011 0.0597 0.0022
5      0.0311 15.8864      0.0162 0.0073 0.1294 0.0000 0.0015 0.1978 0.1927
6      0.0139 23.7660      0.2894 0.4903 0.2774 0.0082 0.0647 0.1696 0.0249
7      0.0101 27.8377      0.2035 0.2265 0.1139 0.0143 0.0840 0.3623 0.4391
8      0.0023 57.8231      0.0002 0.1094 0.2351 0.2952 0.8399 0.0401 0.1429
9      0.0019 64.3856      0.2002 0.1129 0.0720 0.6804 0.0082 0.1658 0.1969
     X18    X19
1 0.0000 0.0023
2 0.0001 0.0014
3 0.0001 0.4188
4 0.0002 0.3646
5 0.0062 0.0001
6 0.0000 0.1656
7 0.0150 0.0214
8 0.3018 0.0044
9 0.6766 0.0214

===============================
Row 9==> X10, proportion 0.680355 >= 0.50
Row 8==> X13, proportion 0.839895 >= 0.50
Row 9==> X18, proportion 0.676625 >= 0.50
```

STEP 5:

- So the subsets (X10,X18) and (X13) are involved in Multicollinearity.
- In the first subset VIF of X10 is the highest and in the second subset the VIF of X13 is highest.
- We drop the variables X10 and X13, and again fit a model.

```
Call:
eigprop(mod = olsreg_6)

  Eigenvalues      CI (Intercept)     X4     X9    X16    X17    X18    X19
1      5.9181  1.0000      0.0022 0.0005 0.0009 0.0007 0.0004 0.0002 0.0046
2      0.5903  3.1662      0.1884 0.0014 0.0043 0.0003 0.0008 0.0003 0.0139
3      0.3885  3.9029      0.0086 0.0021 0.0006 0.0091 0.0014 0.0004 0.4376
4      0.0553 10.3412      0.1406 0.0464 0.3508 0.0616 0.0002 0.0019 0.4690
```

```
5      0.0308 13.8560      0.0169 0.0063 0.1491 0.2704 0.2195 0.0170 0.0027
6      0.0122 22.0625      0.6150 0.8512 0.0765 0.6578 0.0626 0.0020 0.0491
7      0.0047 35.5386      0.0283 0.0921 0.4178 0.0000 0.7151 0.9781 0.0231

===============================
Row 6==> X4, proportion 0.851172 >= 0.50
Row 6==> X16, proportion 0.657845 >= 0.50
Row 7==> X17, proportion 0.715109 >= 0.50
Row 7==> X18, proportion 0.978100 >= 0.50
```

STEP 6:

- So, the subsets (X4,X16) and (X17,X18) are involved in Multicollinearity.
- In the first subset VIF of X4 is the highest, and in the second subset the VIF of X18 is highest.
- We drop the variables X4 and X18, and again fit a model.

```
Call:
eigprop(mod = olsreg_7)

  Eigenvalues       CI (Intercept)     X9    X16    X17    X19
1      4.0296  1.0000      0.0176 0.0037 0.0032 0.0025 0.0108
2      0.5478  2.7123      0.8451 0.0083 0.0010 0.0029 0.0114
3      0.3477  3.4042      0.0016 0.0000 0.0363 0.0109 0.4710
4      0.0471  9.2535      0.1073 0.9108 0.2440 0.0276 0.4787
5      0.0279 12.0287      0.0284 0.0772 0.7155 0.9561 0.0281

===============================
Row 4==> X9, proportion 0.910782 >= 0.50
Row 5==> X16, proportion 0.715471 >= 0.50
Row 5==> X17, proportion 0.956079 >= 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 X19,X16 and X17 are much higher than 0.5. So, we decided to go for stepwise selection method to get better subset model.

6.1.3 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 ELEMINATION (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.

6.1.4 ON THE BASIS OF THE PAIRED F-TEST

On the basis of the Step-wise Selection method:

```
Attaching package:  'olsrr'
The following object is masked from 'package:datasets':
      rivers


                                Stepwise Selection Summary
-------------------------------------------------------------------------------------------
                     Added/                  Adj.
Step    Variable    Removed     R-Square    R-Square     C(p)         AIC          RMSE
-------------------------------------------------------------------------------------------
   1        X6      addition       0.993       0.993    2142.7290    800.0794    5078.1387
   2        X4      addition       0.997       0.997     912.3880    768.7256    3391.7859
   3       X18      addition       0.999       0.999     385.5790    737.9984    2284.0885
   4       X19      addition       0.999       0.999     250.9710    724.1496    1900.1671
   5       X13      addition       0.999       0.999     199.1730    717.6479    1733.5387
   6        X9      addition       0.999       0.999     139.3370    706.9390    1501.1474
   7        X7      addition       1.000       1.000      75.7490    688.7256    1184.0581
   8       X11      addition       1.000       1.000      44.4090    674.4526     981.5820
```

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 the **INFORMATION THEORETIC CRITERION(ITC)** , 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,

$$-2ln\hat{L} + penalty\,term = nln(SSRes) + penalty\,term$$

Where,

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

6.1.5 AKAIKE INFORMATION CRITERION(AIC)}

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.

```
Attaching package:  'MASS'
The following object is masked from 'package:olsrr':
      cement

Start:  AIC=532.88
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      2007  8547618 530.89
- X1    1     54083  8599694 531.13
- X4    1    170244  8715855 531.67
- X9    1    219805  8765416 531.90
- X14   1    309257  8854868 532.30
- X12   1    364397  8910008 532.55
- X19   1    371898  8917509 532.59
- X5    1    392343  8937954 532.68
<none>               8545611 532.88
- X16   1    446745  8992356 532.92
- X6    1    521974  9067584 533.25
- X8    1    763455  9309066 534.30
- X3    1   1081816  9627427 535.65
- X2    1   1173608  9719219 536.03
- X20   1   1416429  9962040 537.02
- X15   1   2376289 10921900 540.70
- X7    1   2983834 11529445 542.86
- X11   1   3620431 12166042 545.01
- X10   1   4007749 12553359 546.26
- X13   1   5766657 14312268 551.51
- X18   1   8127906 16673517 557.62

Step:  AIC=530.89
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     60747  8608364 529.17
- X4    1    170315  8717933 529.68
- X9    1    277401  8825019 530.17
- X14   1    309453  8857071 530.31
```

```
- X19   1    379230  8926848 530.63
- X5    1    391711  8939329 530.68
<none>               8547618 530.89
- X16   1    484583  9032201 531.10
- X12   1    546402  9094020 531.37
- X6    1    553774  9101392 531.40
- X8    1    814925  9362543 532.53
+ X17   1      2007  8545611 532.88
- X3    1   1095907  9643524 533.72
- X2    1   1184088  9731706 534.08
- X20   1   1421513  9969130 535.04
- X15   1   2869345 11416963 540.47
- X7    1   2982766 11530383 540.86
- X10   1   4240386 12788004 545.01
- X11   1   5803376 14350993 549.62
- X13   1   6256749 14804367 550.86
- X18   1   8178411 16726028 555.74

Step:  AIC=529.17
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    157660  8766024 527.90
- X9    1    278920  8887284 528.45
- X14   1    297741  8906105 528.53
- X19   1    430399  9038763 529.13
<none>               8608364 529.17
- X5    1    443934  9052299 529.19
- X16   1    482339  9090703 529.36
- X12   1    528068  9136432 529.56
- X6    1    583320  9191684 529.80
- X8    1    814017  9422382 530.79
+ X1    1     60747  8547618 530.89
+ X17   1      8670  8599694 531.13
- X3    1   1084698  9693062 531.92
- X2    1   1128718  9737082 532.10
- X20   1   1360830  9969194 533.05
- X15   1   2827542 11435906 538.54
- X7    1   3314028 11922392 540.20
```

```
- X10   1   4215370 12823734 543.12
- X11   1   5742824 14351188 547.62
- X13   1   7772158 16380522 552.91
- X18   1   8307632 16915996 554.20

Step:  AIC=527.9
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    283235  9049259 527.17
- X5    1    413610  9179634 527.74
- X6    1    427144  9193168 527.80
<none>               8766024 527.90
- X16   1    467165  9233189 527.98
- X14   1    627280  9393304 528.67
- X12   1    628581  9394605 528.67
+ X4    1    157660  8608364 529.17
+ X1    1     48091  8717933 529.68
- X9    1    911103  9677128 529.86
+ X17   1      7837  8758187 529.86
- X3    1    935496  9701521 529.96
- X20   1   1203292  9969316 531.05
- X2    1   1581966 10347991 532.54
- X8    1   2375814 11141839 535.49
- X15   1   2791538 11557562 536.96
- X10   1   4412178 13178202 542.21
- X7    1   5557039 14323063 545.54
- X18   1   8273441 17039465 552.49
- X13   1   8586823 17352847 553.22
- X11   1  11085618 19851642 558.60

Step:  AIC=527.17
Y ~ X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13 +
    X14 + X15 + X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
- X6    1    371296  9420555 526.78
- X16   1    387417  9436676 526.85
<none>               9049259 527.17
```

```
- X5    1    486687  9535945 527.27
- X12   1    558527  9607786 527.57
+ X19   1    283235  8766024 527.90
+ X1    1     99806  8949452 528.73
+ X17   1     24832  9024427 529.06
+ X4    1     10496  9038763 529.13
- X20   1   1123561 10172820 529.85
- X14   1   1141501 10190760 529.92
- X9    1   1350513 10399772 530.74
- X2    1   1366301 10415560 530.80
- X3    1   1376534 10425792 530.84
- X15   1   2645336 11694595 535.43
- X8    1   3611288 12660547 538.60
- X10   1   4757244 13806503 542.07
- X13   1   8328057 17377316 551.27
- X7    1   8741570 17790828 552.21
- X18   1   8748217 17797476 552.23
- X11   1  12611524 21660783 560.09

Step:  AIC=526.78
Y ~ X2 + X3 + X5 + X7 + X8 + X9 + X10 + X11 + X12 + X13 + X14 +
    X15 + X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
- X5    1    183064  9603619 525.55
<none>               9420555 526.78
+ X6    1    371296  9049259 527.17
+ X19   1    227387  9193168 527.80
+ X1    1    146416  9274139 528.15
+ X4    1     43672  9376883 528.60
+ X17   1      3648  9416907 528.77
- X2    1   1261200 10681754 529.81
- X16   1   1324815 10745369 530.04
- X12   1   1620615 11041170 531.13
- X20   1   1910331 11330886 532.17
- X3    1   2369212 11789767 533.75
- X14   1   3095278 12515833 536.15
- X9    1   3892483 13313038 538.61
- X15   1   5288147 14708702 542.60
- X10   1   6685177 16105732 546.23
```

```
- X13   1   8260389 17680944 549.96
- X7    1   8383100 17803654 550.24
- X8    1  10105018 19525572 553.93
- X18   1  11218808 20639362 556.15
- X11   1  35357444 44777999 587.13

Step:  AIC=525.55
Y ~ X2 + X3 + X7 + X8 + X9 + X10 + X11 + X12 + X13 + X14 + X15 +
    X16 + X18 + X20

       Df Sum of Sq      RSS    AIC
<none>               9603619 525.55
+ X19   1    296072  9307547 526.30
+ X1    1    184126  9419493 526.78
+ X5    1    183064  9420555 526.78
+ X6    1     67673  9535945 527.27
+ X4    1     11559  9592059 527.50
+ X17   1         9  9603610 527.55
- X2    1   1292025 10895644 528.60
- X3    1   2590411 12194029 533.10
- X20   1   2987007 12590626 534.38
- X16   1   4172050 13775669 537.98
- X9    1   4354822 13958441 538.51
- X12   1   7619962 17223581 546.92
- X13   1   8244892 17848511 548.34
- X7    1   8659633 18263252 549.26
- X14   1   9403332 19006950 550.86
- X8    1   9924532 19528151 551.94
- X15   1  10506882 20110501 553.11
- X10   1  12195135 21798754 556.34
- X18   1  31410439 41014058 581.62
- X11   1  43593031 53196650 592.03

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
 -1464.1073       3.7472     -31.4570      -1.6490      -2.7892       1.6533
```

```
      X10         X11         X12         X13         X14         X15
  -4.2406      4.9220     -1.4667      3.3327      0.4961      2.5097
      X16         X18         X20
  -0.6438     -1.7395      1.4348
```

STEP 1 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 2 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 3 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 4 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 5 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 6 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.

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.

```
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
-1287.44  -249.30   -22.34   136.33  1113.32

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1464.1073  1634.3346  -0.896 0.378882
X2              3.7472     2.0433   1.834 0.078592 .
X3            -31.4570    12.1138  -2.597 0.015539 *
X7             -1.6490     0.3473  -4.748 7.16e-05 ***
X8             -2.7892     0.5488  -5.083 3.01e-05 ***
X9              1.6533     0.4910   3.367 0.002461 **
X10            -4.2406     0.7526  -5.634 7.29e-06 ***
X11             4.9220     0.4620  10.653 8.82e-11 ***
X12            -1.4667     0.3293  -4.454 0.000154 ***
X13             3.3327     0.7194   4.633 9.65e-05 ***
X14             0.4961     0.1003   4.948 4.27e-05 ***
X15             2.5097     0.4799   5.230 2.06e-05 ***
X16            -0.6438     0.1953  -3.296 0.002937 **
X18            -1.7395     0.1924  -9.043 2.35e-09 ***
X20             1.4348     0.5145   2.789 0.009975 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 619.8 on 25 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 2.719e+04 on 14 and 25 DF,  p-value: < 2.2e-16
```

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

```
        X2         X3         X7         X8         X9        X10        X11
  60.58406   99.90614 2345.47408  802.95236  415.53603 2620.89919 6791.36547
       X12        X13        X14        X15        X16        X18        X20
2776.45404 4603.74923 2092.06327 1459.81854  545.04009  474.52729 1503.78823
```

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.

6.2 MULTICOLLINEARITY DETECTION AFTER AIC

VARIABLE	VIF
X2	60.58406
X3	99.90614
X7	2345.47408
X8	802.95236
X9	415.53603
X10	2620.89919
X11	6791.36547
X12	2776.45404
X13	4603.74923
X14	2092.06327
X15	1459.81854
X16	545.04009
X18	474.52729
X20	1503.78823

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.

6.2.1 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 $\lambda = \infty$, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

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:

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

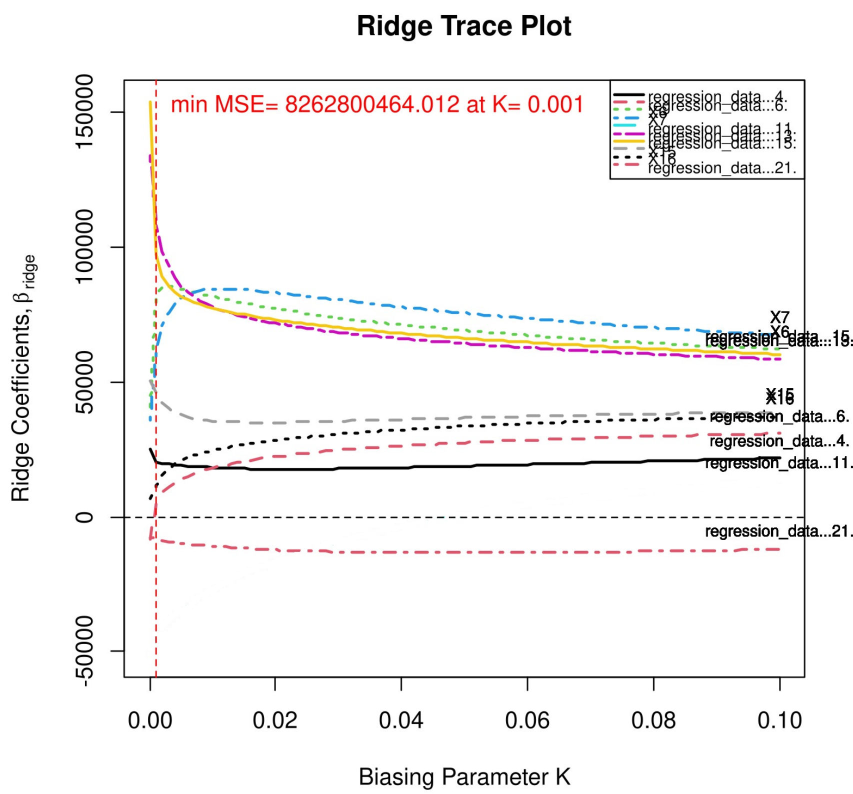
Ridge regression adds a product of ridge parameter & the identity matrix to the cross product matrix $(X^T X)$, forming a new matrix $(X^T X \; + \; \lambda I)$. The new formula is used to find the coefficients:

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

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.

```
- Attaching packages -------------------- tidyverse 1.3.1 -
v ggplot2 3.3.6     v purrr   0.3.4
v tibble  3.1.7     v dplyr   1.0.9
v tidyr   1.2.0     v stringr 1.4.0
v readr   2.1.2     v forcats 0.5.1
- Conflicts --------------------- tidyverse_conflicts() -
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x dplyr::recode() masks car::recode()
x dplyr::select() masks MASS::select()
x purrr::some()   masks car::some()

Attaching package:  'lmridge'
The following object is masked from 'package:car':
      vif
```



From the above plot it seems that the estimates of coefficients stabilizes for some value of λ between 0.025 and 0.035.

From the ridge trace plot we choose that value of λ for which VIFs all get stabilized (i.e. $<$ 5). The estimate of λ obtained by this method is 0.031. Hence we fit a new model with this value of λ and inspect its adjusted R-squared value.

```
Call:
lmridge.default(formula = regression_data$Y ~ ., data = df, K = 0.031)


Coefficients: for Ridge parameter K= 0.031
             Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept -6.6368e+03   -5.2186e+09  1.5061e+08     -34.6498   <2e-16 ***
X2         6.7110e+00    1.5845e+04  5.5791e+03       2.8400   0.0075 **
X3         1.5757e+01    8.0580e+03  5.8787e+03       1.3707   0.1793
X7         8.0400e-01    6.9482e+04  4.3591e+03      15.9394   <2e-16 ***
```

```
X8        -2.2440e-01   -7.1812e+03  4.3317e+03      -1.6578   0.1064
X9        -1.6076e+00   -4.1364e+04  5.4204e+03      -7.6311   <2e-16 ***
X10        5.4510e-01    2.2982e+04  3.1602e+03       7.2723   <2e-16 ***
X11        4.7750e-01    5.2789e+04  2.2682e+03      23.2736   <2e-16 ***
X12        4.2570e-01    4.2216e+04  3.4546e+03      12.2202   <2e-16 ***
X13        9.6460e-01    5.6388e+04  2.4292e+03      23.2126   <2e-16 ***
X14        1.9040e-01    5.3831e+04  2.7834e+03      19.3402   <2e-16 ***
X15        6.6090e-01    3.2611e+04  4.3037e+03       7.5774   <2e-16 ***
X16        4.4430e-01    3.2909e+04  4.5926e+03       7.1656   <2e-16 ***
X18        2.5200e-01    1.7688e+04  5.5798e+03       3.1701   0.0032 **
X20        5.5850e-01    2.6088e+04  4.8577e+03       5.3704   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge Summary
        R2     adj-R2   DF ridge          F        AIC        BIC
   0.98950    0.98420    4.56815 1609.91132  630.23090  785.50114
Ridge minimum MSE= 439612436334 at K= 0.031
P-value for F-test ( 4.56815 , 34.1518 ) = 4.771097e-39
-------------------------------------------------------------------
         R2 adj-R2 DF ridge        F      AIC      BIC
[1,] 0.9895 0.9842  4.56815 1609.911 630.2309 785.5011
```

VIFs for the new fitted model are:

```
             X2      X3      X7      X8      X9     X10     X11     X12     X13
k=0.031 4.79771 5.32677 2.92892 2.89217 4.52863 1.53932 0.79298 1.83949 0.90955
            X14     X15     X16    X18     X20
k=0.031 1.19411 2.85491 3.25102 4.7989 3.63723
```

6.2.2 OBSERVATION:

- We observe that after fitting ridge regression model the VIFs have decreased significantly.
- The adjusted R-square is 98.42%.

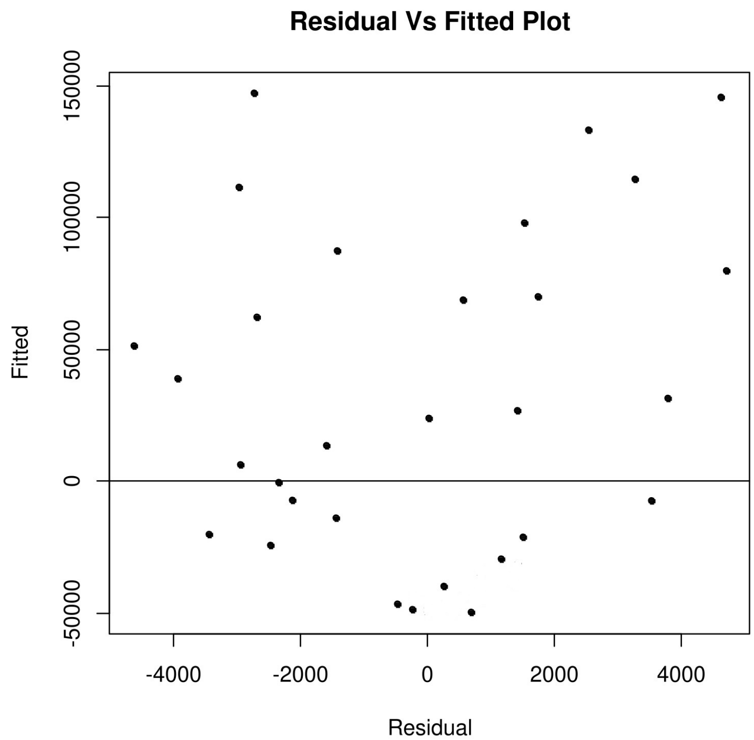
Now we perform residual analysis on our newly fitted model.

6.3 INSPECTION OF PROPERTIES OF FITTED MODEL AFTER RIDGE REGRESSION

6.3.1 CHECK FOR HOMOSCEDASTICITY ASSUMPTION OF ERRORS

```
          K=0.031
K=0.031 0.1161798
```

The correlation between fitted values and residuals is 0.1161798.



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 = 23.816, df = 14, p-value = 0.005028
```

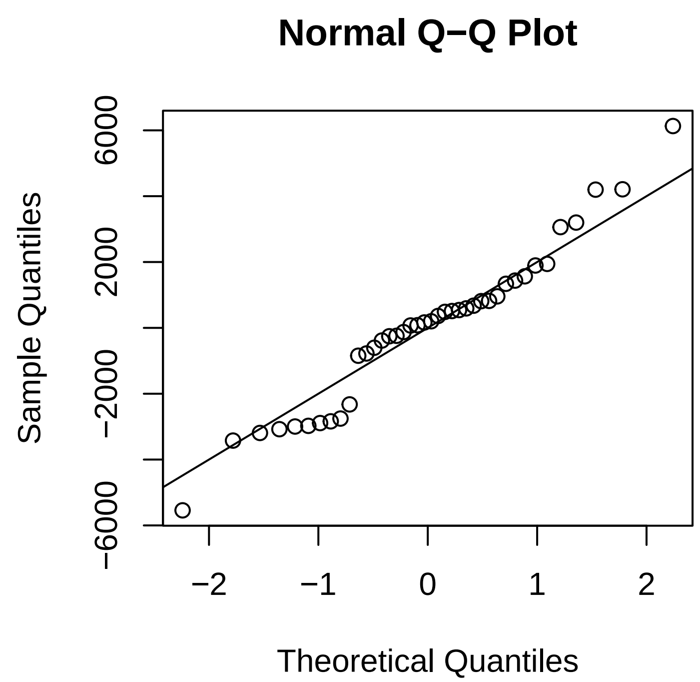
As p-value=0.05028>0.05.

Hence, we conclude that there is no violation of homoscedasticity assumption in our model.

6.3.2 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 = 0.96371, p-value = 0.2238
```



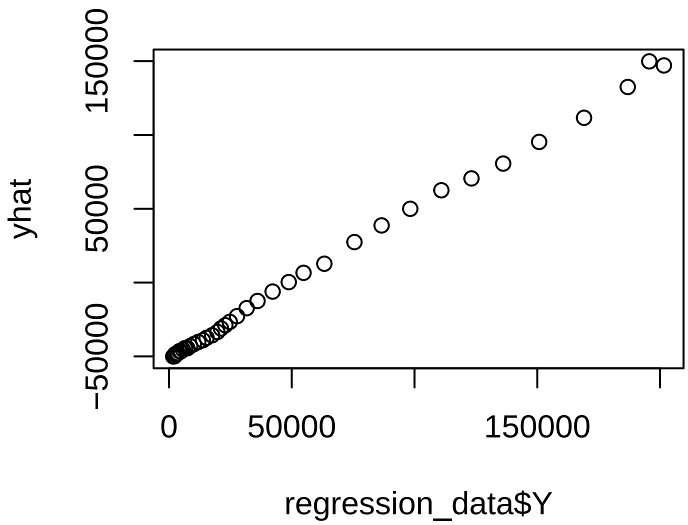
As we can see p-value= 0.2238 > 0.05, hence Normality Assumption of error holds.

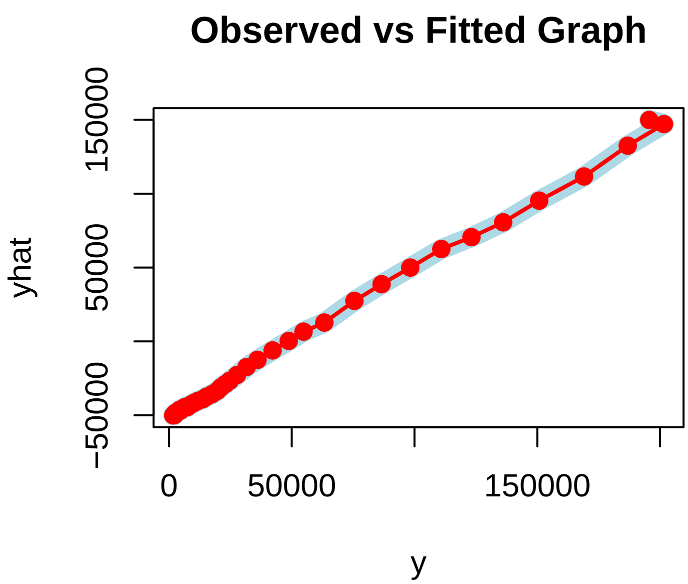
Now we compare between observed and fitted responses.

```
      K=0.031
[1,] 0.9992524
```

The correlation between fitted and observed response is 0.9992524, which indicates a good fit of the observed responses.

6.3.3 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).

6.3.4 FINAL FITTED MODEL USING RIDGE REGRESSION

Our final model after Ridge regression is given by,

$\hat{Y}$ **= -6636.8 + 6.711(X2) + 15.757(X3) + 0.804(X7) - 0.224(X8) - 1.608(X9) + 0.545(X10) + 0.478(X11) + 0.426(X12) + 0.965(X13) + 0.190(X14) + 0.661(X15) + 0.444(X16) + 0.252(X18) + 0.559(X20)**

6.3.5 CONCLUSION ABOUT THE RIDGE MODEL

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

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

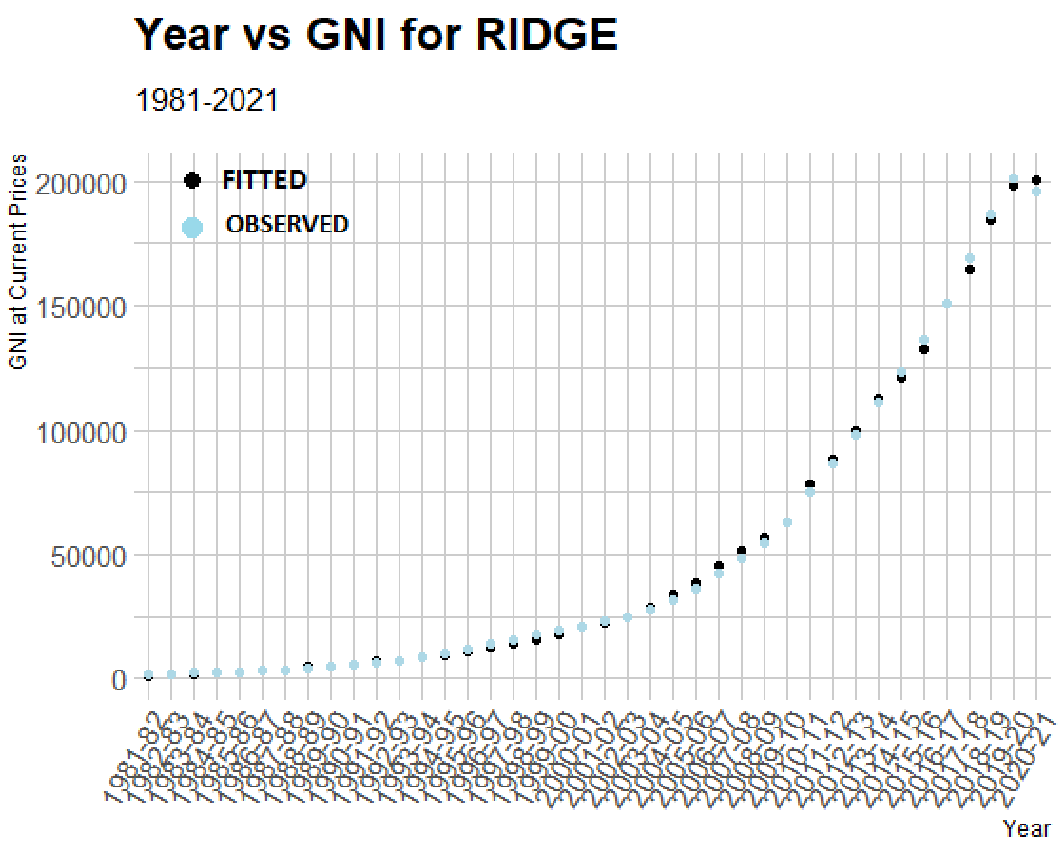
$$R^2_{Adj} = 1 - \frac{n-1}{n-p-1}\frac{SSRes}{SST}$$

As **adjusted R-squared** value is **0.9842**, we can conclude that **98.42%** 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 (X2)Agricultural production of commercial products , (X3)Production of crude oil and petroleum , (X7)Total savings deposit in commercial banks, (X8) Gross fiscal deficit, (X9)Combined net borrowing of both state and central government, (X10)Currency with public, (X11)Total developmental and non-developmental expenditures of government, (X12)Net bank credited to government, (X13)Invested by LIC, (X14)Combined liabilities of Central and State government,(X15) Export of principle commodities, (X16)Import of principle commodities, (X18)Foreign Exchange reserve in gold, foriegn currency assests etc and (X20)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 1981-2021.

6.3.6 GRAPHICAL OVERVIEW OF THE MODEL



We can see from the figure that our model is satisfactorily efficient in explaining the change in Indian GNI at current prices.

Here BLACK dots represent fitted and LIGHTBLUE DOTS represent observed values of Y.

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.

7 LASSO REGRESSION

7.1 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.

7.2 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.

7.3 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.

7.4 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.

7.5 ANALYZE FINAL MODEL IN LASSO

we analyze the final model produced by the optimal lambda value.

```
Loading required package:  lattice

Attaching package:  'caret'
The following object is masked from 'package:purrr':
      lift
Loading required package:  Matrix

Attaching package:  'Matrix'
The following objects are masked from 'package:tidyr':
      expand, pack, unpack
Loaded glmnet 4.1-4

glmnet

40 samples
20 predictors

No pre-processing
Resampling: Bootstrapped (25 reps)
```

```
Summary of sample sizes: 40, 40, 40, 40, 40, 40, ...
Resampling results across tuning parameters:

  alpha  lambda      RMSE       Rsquared   MAE
  0.10     120.5222   6276.755  0.9923315   3158.342
  0.10    1205.2218   6405.228  0.9922066   3241.064
  0.10   12052.2178   7671.324  0.9908435   4154.871
  0.55     120.5222   5654.612  0.9931036   2981.508
  0.55    1205.2218   6082.458  0.9931084   3157.800
  0.55   12052.2178   9382.685  0.9911454   7035.373
  1.00     120.5222   5077.930  0.9943951   2742.547
  1.00    1205.2218   5755.994  0.9940649   3097.893
  1.00   12052.2178  13290.978  0.9935478  10755.995

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 120.5222.
```

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 lambda = 120.5222(for alpha=1).

```
                       s1
(Intercept) -657.52473142
X1             0.00000000
X2             0.31754409
X3             0.00000000
X4            88.69223779
X5             0.00000000
X6             3.93919843
X7             0.44428713
X8             0.00000000
X9            -0.23438763
X10            0.00000000
X11            0.01110223
X12            0.00000000
X13            0.01284893
X14            0.00000000
X15            0.32975752
X16            0.26955886
X17            0.00000000
```

```
X18            0.00000000
X19           -0.21154937
X20            0.00000000
```

No coefficient is shown for the predictor X1,X3,X5,X8,X10,X12,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. 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.

Now we check he correlation between fitted values and residuals.

```
[1] 0.1166227
```

The correlation coefficient between the fitted value and the residuals is , which can be neglacted and we can say there is no correlation between the fitted values and the residuals.

Now, we check the correlation between the fitted and observed response. which indicates a good fit of the observed responses.

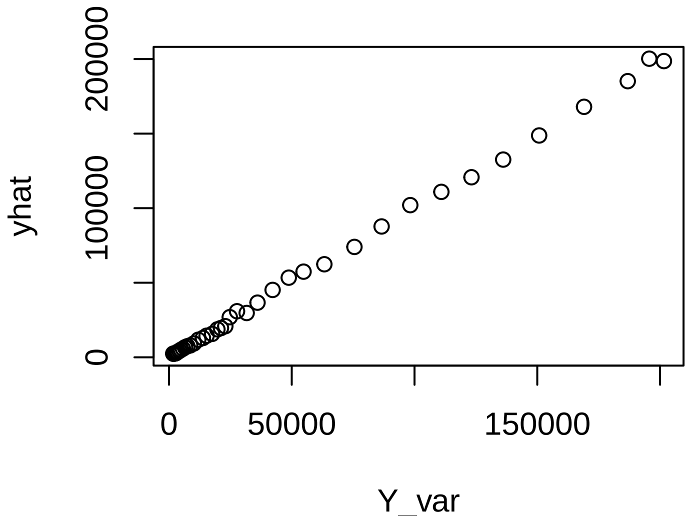
```
[1] 0.9995048
```

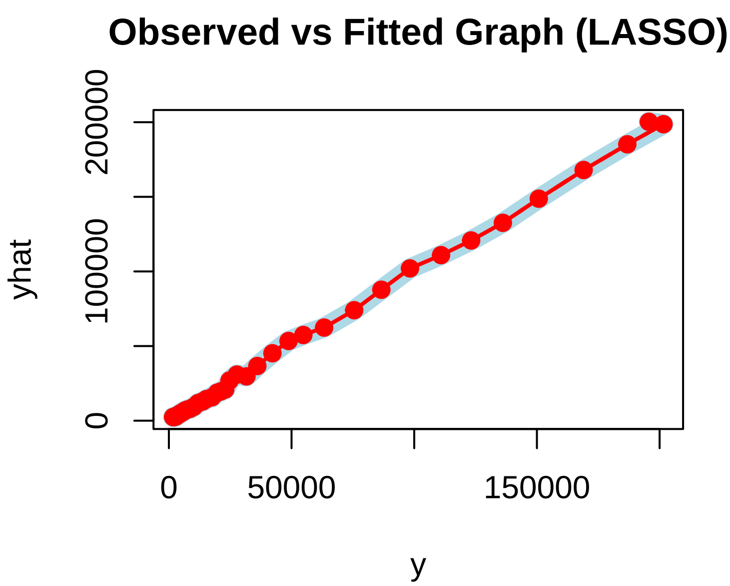
The correlation between the fitted and observed response is 0.9995048, which indicates a good fit of the observed responses.

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

```
[1] 0.9979677
```

TheAdjusted R-squared turns out to be 0.9990099. That is, the best model was able to explain 99.90 % of the variation in the response values of our dataset.


yhat
0
100000
200000
0
50000
150000
Y_var

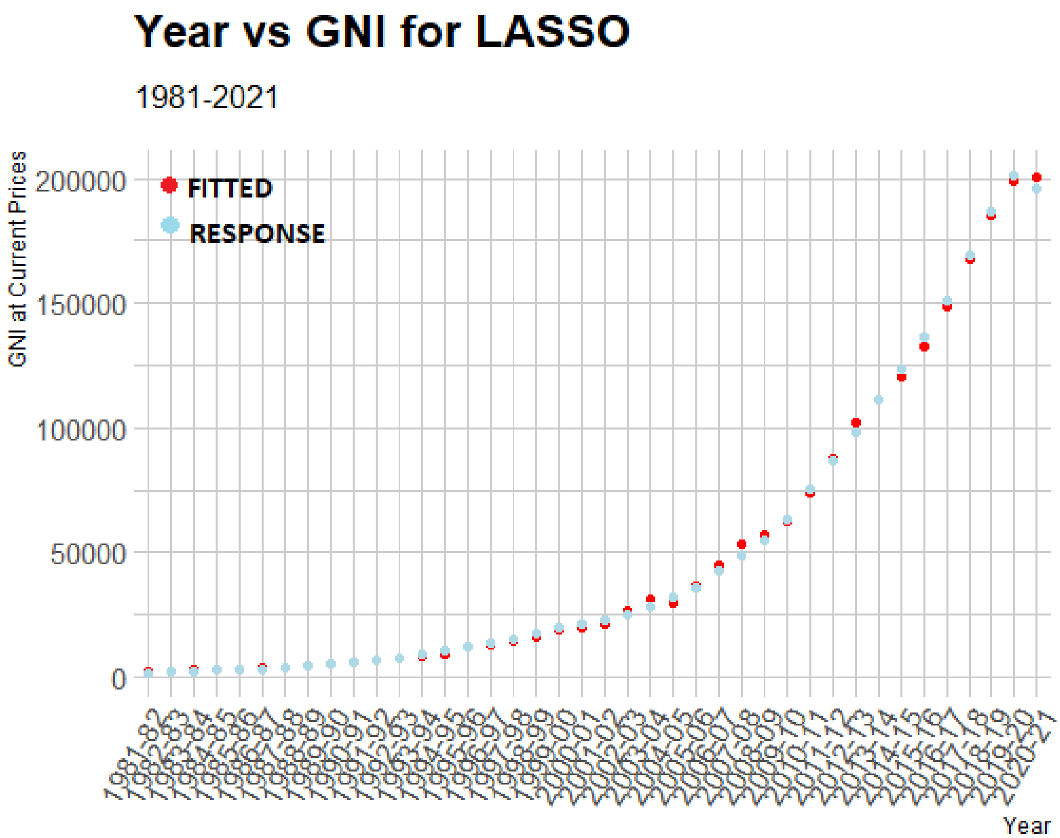

Observed vs Fitted Graph (LASSO)
yhat
0
100000
200000
0
50000
150000
y

7.6 FINAL FITTED LASSO MODEL

Our final model after LASSO regression is given by,

$$\mathbf{Y = -657.525 + (0.318)X2 + (88.692)X4 + (3.939)X6 + (0.444)X7 - (0.234)X9 + (0.011)X11 + (0.013)X13 + (0.330)X15 + (0.270)X16 - (0.212)X19}$$

7.6.1 GRAPHICAL OVERVIEW OF THE MODEL



Here RED DOTS represent fitted and LIGHTBLUE DOTS represent observed values of Y.

7.7 FINAL CONCLUSION ON LASSO REGRESSION

Finally, from our analysis we come to conclude that (X2)Agricultural production of commercial products , (X4)Import of crude oil and petroleum, (X6)Direct and indirect tax revenue, (X7)Total savings deposit in commercial banks, (X9)Combined net borrowing of both state and central government, (X11)Total developmental and non-developmental expenditures of government, (X13)Investment by lic, (X15)Export of principle commodities, (X16)Import of principle commodities, (X19)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.

8 FINAL CONCLUSION

	RIDGE REGRESSION	LASSO REGRESSION
X1:FOOD GRAINS	✖	✖
X2:COMMERCIAL GRAINS	✔	✔
X3:PRODUCTION OF OIL	✔	✖
X4:IMPORT OIL	✖	✔
X5:SPREAD OF GOLD PRICE	✖	✖
X6:TAX REVENUE	✖	✔
X7:SAVING DEPOSITS IN BANKS	✔	✔
X8:GROSS FISCAL DEFICIT	✔	✖
X9:NET BORROWING OF GOVT.	✔	✔
X10:CURRENCY WITH PUBLIC	✔	✖
X11:GOVT.'S EXPENDITURE	✔	✔
X12:NET BANK CREDITED TO GOVT.	✔	✖
X13:LIC INVESTMENT	✔	✔
X14:COMBINED LIABILITIES OF GOVT.	✔	✖
X15:EXPORT OF COMMODITIES	✔	✔
X16:IMPORT OF COMMODITIES	✔	✔
X17:F.D.I INFLOWS	✖	✖
X18:FOREIGN EXCHANGE RESERVES	✔	✖
X19:NET INFLOWOF AID	✖	✔
X20:CURRENCY IN CIRCULATION	✔	✖
Correlation b/w observed & Fitted	0.9992	0.9995
Adjusted R^2	0.984	0.998

As we see from the above table, the following points are arising

1. the adjusted R-square for the which we get previously through ridge regression is 0.984, that means model which is obtained through Ridge regression is expained 98.4% variability in the dataset. And the adjusted R-square for the which we get previously through **LASSO** regression is **0.9995**, that means

model which is obtained through Ridge regression is expained 99.95% variability in the dataset, which is greater than the model obtained trhough Ridge, but the difference is not very much significant. This difference is very small. So in this point of view both of our model fitting exercises are quite satisfactory.

2. The Correlation coefficient b/w the observed & Fitted response obtained through Ridge regression is 0.9992. And the Correlation coefficient b/w the observed & Fitted response obtained through LASSO regression is 0.9995, which is obviously greater than Ridge technique, but this this difference also not very much significant. So in this point of view both of our model fitting exercises are quite satisfactory.

3. But after the Ridge regression, we get the model which contains **14** Regressors, these are **(X2)**Agricultural Production of Commercial Products **, (X3)**Production of Crude Oil and Petroleum **, (X7)**Total Savings Deposit in Commercial Banks**, (X8)** Gross Fiscal Deficit**, (X9)**Combined Net Borrowing of Both State and Central Government**, (X10)**Currency with Public**, (X11)**Total Developmental and Non-Developmental Expenditures of Government**, (X12)**Net Bank Credited to Government**, (X13)**Investment by LIC**, (X14)**Combined Liabilities of Central and State Goverment**,(X15)** Export of Principle Commodities**, (X16)**Import of Principle Commodities**, (X18)**Foriegn Exchange Reserve in Gold, Foriegn Currency Assests etc. and **(X20)**Currency in Circulation, these are good Regressors. But on the other hand, the **LASSO** regression, select **10** Regressors, these are **(X2)**Agricultural Production of Commercial Products **, (X4)**Import of Crude Oil and Petroleum**, (X6)**Direct and Indirect Tax Revenue**, (X7)**Total Savings Deposit in Commercial Banks**, (X9)**Combined Net Borrowing of Both State and Central Government**, (X11)**Total Developmental and Non-Developmental Expenditures of Government**, (X13)**Investment by LIC**, (X15)** Export of Principle Commodities**, (X16**Import of Principle Commodities**, (X19)**Net Inflow of Aid, these are also good regressors. But, **RIDGE PERFORMS BETTER** in terms of more as well good regressors than **LASSO** technique.

So, finally we select the model obtained after **RIDGE REGRESSION**, and the Final Model is:

$\hat{Y}$ **= -6636.8 + 6.711(X2) + 15.757(X3) + 0.804(X7) - 0.224(X8) - 1.608(X9) + 0.545(X10) + 0.478(X11) + 0.426(X12) + 0.965(X13) + 0.190(X14) + 0.661(X15) + 0.444(X16) + 0.252(X18) + 0.559(X20)**

9 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
- An Introduction to the Statistical Learning with R (ISLR), Hastie, Tibshirani
- different materials from internet for our project