# Lasso & Ridge Regression (Module - 8)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Nukala Ayyappa Bharthwaj**

**Batch Id: DSWDMCOS 21012022**

**Topic: Lasso Ridge Regression**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Lasso and Ridge Regression Algorithm**
   3. **Train and Test the data and compare RMSE values tabulate R-Squared values, RMSE for different models in documentation and provide your explanation on it**
3. **Briefly explain the model output in the documentation.** **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

An Analytics Company has been tasked by a crucial job of finding out what factors does affect a startup company and will it be profitable to do so or not. For this, they have collected some historical data and would like to applying supervised predictive learning algorithm such as Lasso Ridge Regression on it and provide brief insights about their data. Predict Profit, given different attributes for various startup companies.

**Sol:**

**Business Objective:** To predict the profits of the company with other factors by using Lasso-Ridge regression model.

**Constraints:** Lack of analysis of the company data.

**Data Types:** All the given data is in numeric format except the states column and the complete datacan be used for the analysis.

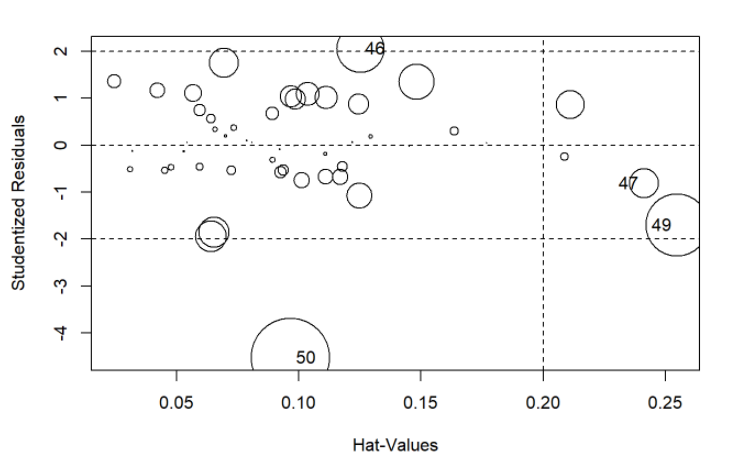
**Data Cleaning:** Since the states column is in Non-Numeric format the same is converted to numeric data for doing the analysis.

**Exploratory Data Analysis:** the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

**Multiple Linear Regression:** after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as profit of the company.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.9496.

Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 49,50 and that are found by using the following influential plot.

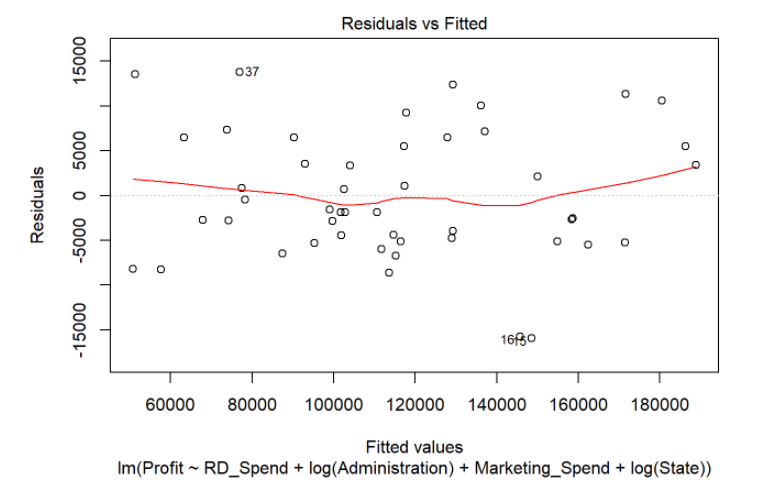


After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.

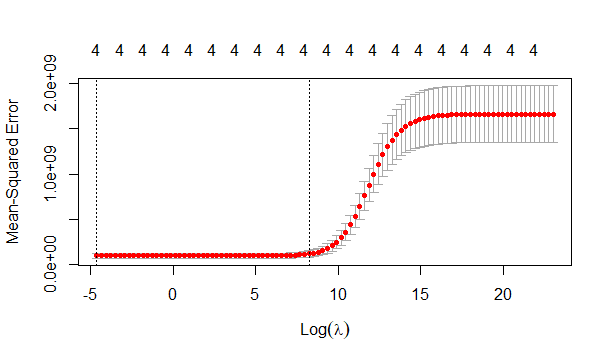
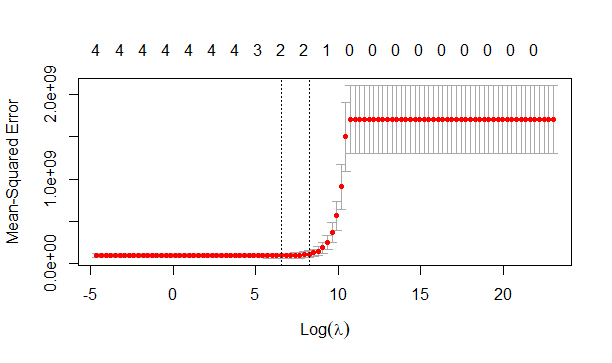
After applying Logarithmic transformations the R^2 value is 0.9451 and for the exponential transformations R^2 value is 0.9557 and for the quadratic model is 0.9567.

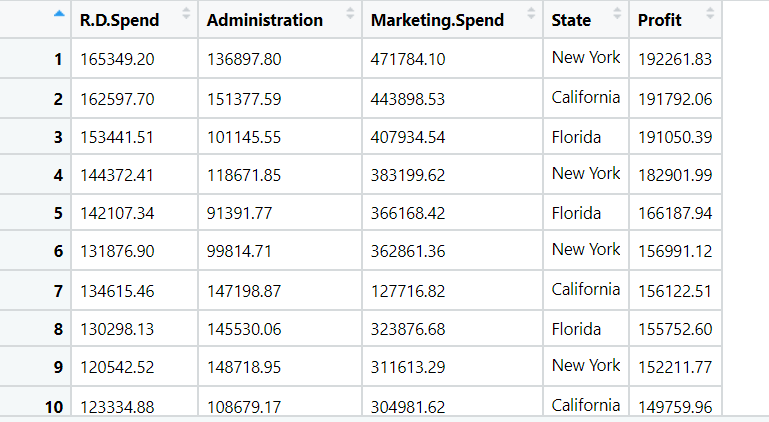
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.

The following residual plot is done to know the accuracy of the model.



**Lasso-Ridge Regression:** After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.9321 and for the Ridge model it is increased slightly and R^2 value is 0.9624. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





**Problem Statement: -**

Officeworks, is a leading retail store in Australia, with numerous outlets around the country, the manager would like to improve their customer experience by providing them online predictive prices about their gadgets/ Laptops if they wants to sell them. To improve this experience the manager would like us to build a model which is sustainable and accurate enough, to get the objective achieved. Apply Lasso Ridge Regression model on the dataset and predict Price, given other attributes and tabulate R squared ,RMSE and correlation values.

**Sol:**

**Business Objective:** To predict the Price of the car with other factors by using Lasso-Ridge regression model.

**Constraints:** Lack of analysis of the car sales data of the company

**Data Types:** The first column in the given data is I’d which is not useful for doing the analysis and the remaining data is used for doing the analysis.

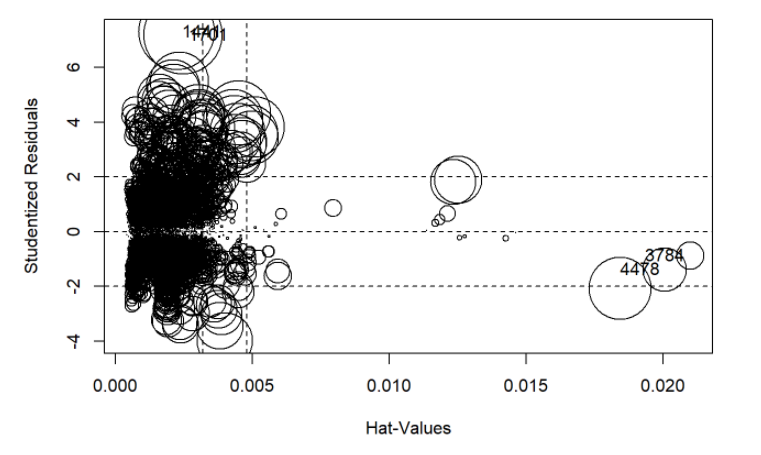
**Data Cleaning:** Since some of the columns in the given data is no-numeric the same is converted into numeric data so that they can be used for doing the analysis.

**Exploratory Data Analysis:** the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

**Multiple Linear Regression:** after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer part.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.7752.

Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.

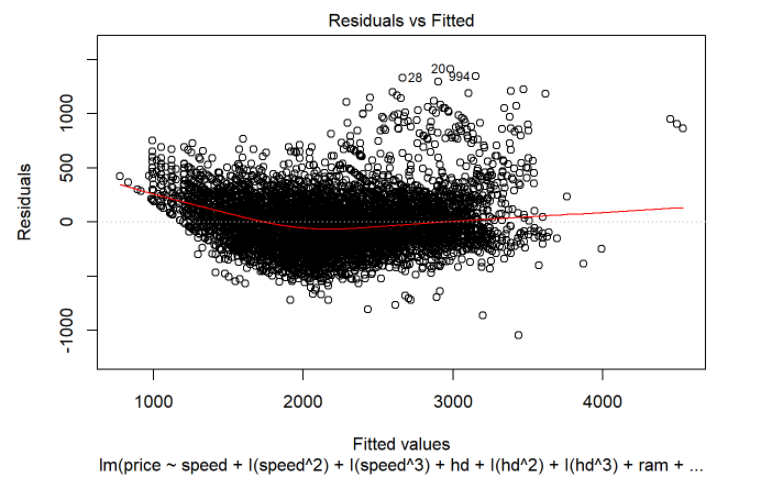


After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.

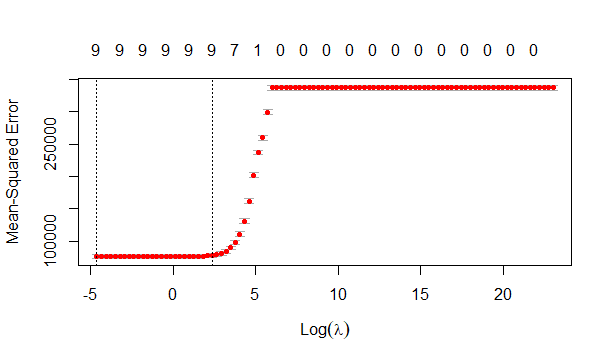
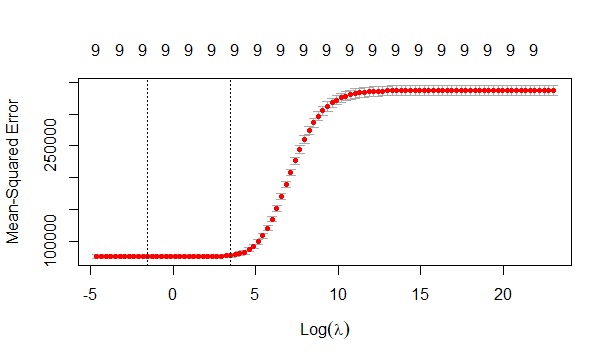
After applying Logarithmic transformations the R^2 value is 0.7441and for the exponential transformations R^2 value is 0.7833 and for the quadratic model is 0.8049.

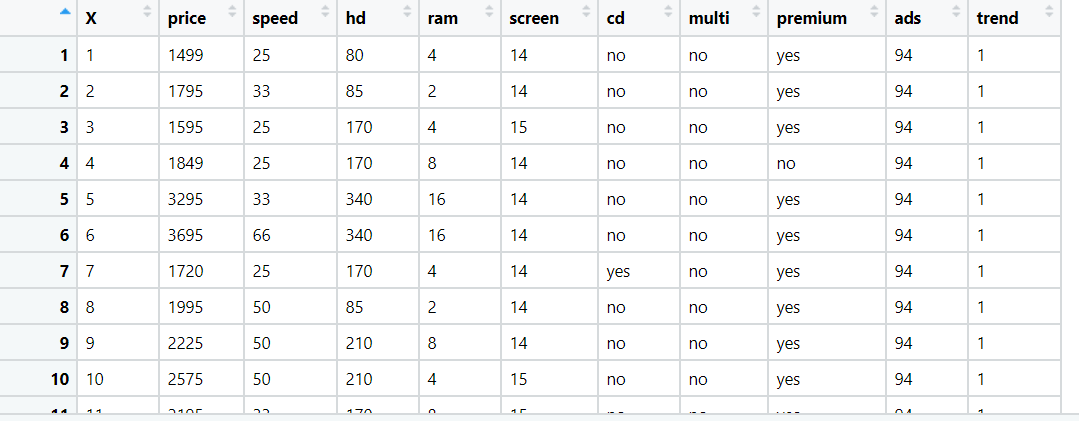
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.

The following residual plot is done to know the accuracy of the model.



**Lasso-Ridge Regression:** After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.7666 and for the Ridge model it is increased slightly and R^2 value is 0.8051. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:







**Problem Statement: -**

An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like to have an automated model which can predict the price of the car if user inputs the required factors. Help the business achieve the objective by applying Lasso and Ridge regression model on it.

Please use the below columns for the analysis purpose.

Price, Age\_08\_04, KM, HP, cc, Doors , Gears, Quarterly\_Tax, Weight

**Sol:**

**Business Objective:** To predict the Price of the computer parts with other factors by using Lasso-Ridge regression model.

**Constraints:** Lack of analysis of the sales data of the company

**Data Types:** The first column in the given data is I’d which is not useful for doing the analysis and the remaining data is used for doing the analysis.

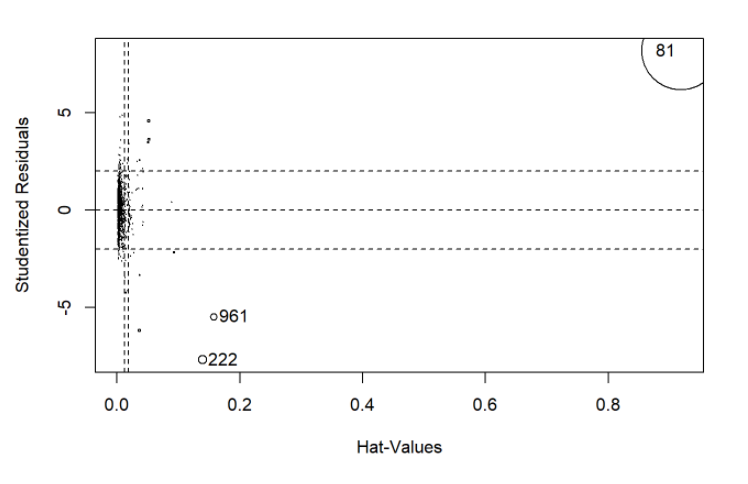
**Data Cleaning:** Since some of the columns in the given data is non-numeric the same is converted into numeric data so that they can be used for doing the analysis.

**Exploratory Data Analysis:** the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

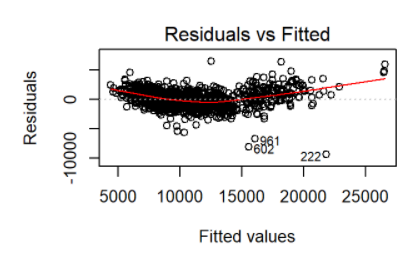
**Multiple Linear Regression:** after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.863.

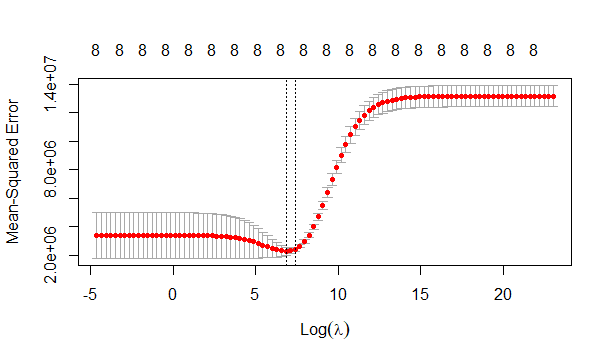
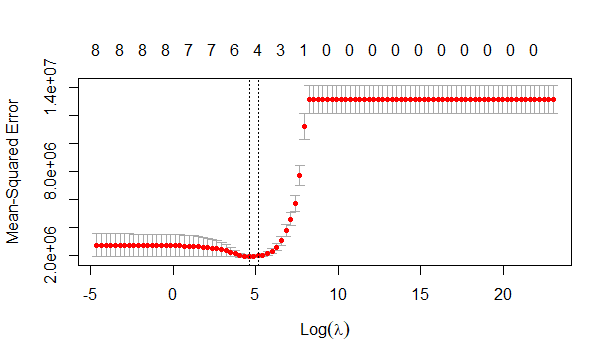
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.

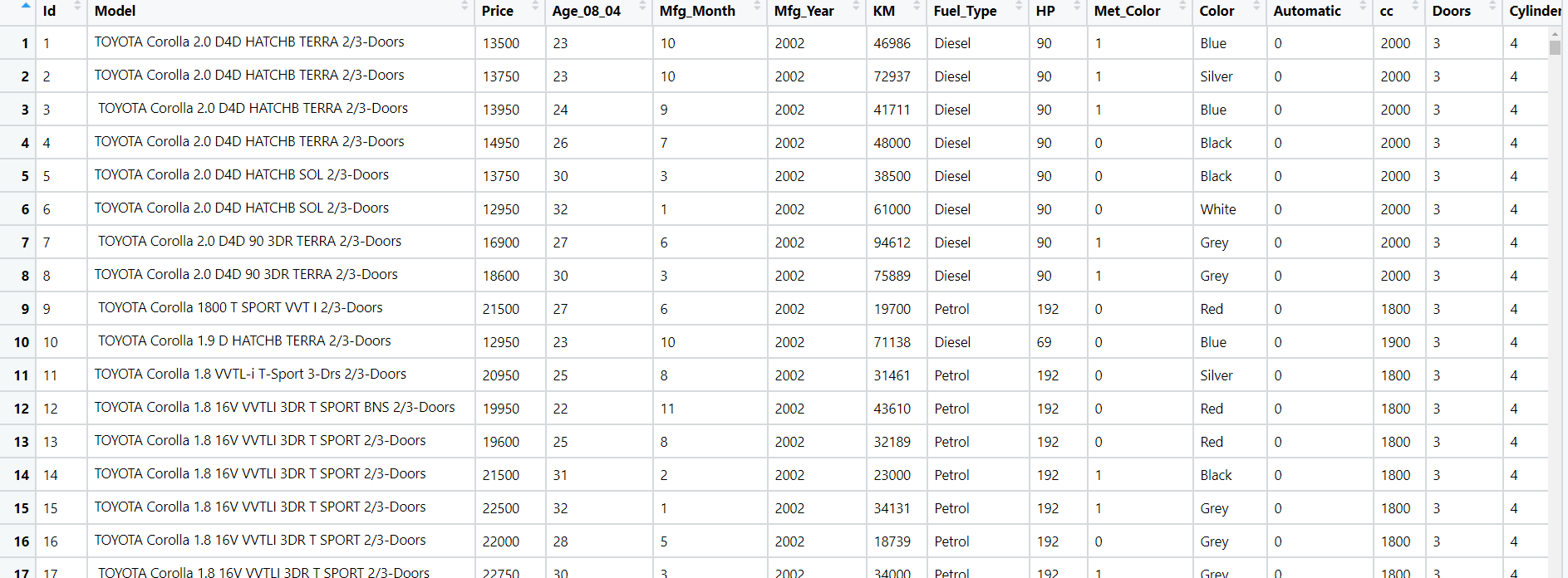


After observing the influential plot it is noted that 81st observation is effecting the complete analysis so the same observation is removed and then the basic regression model is done then the R^2 values is observed as 0.9291 so the same model is used for the future prediction in the data . The residual plot for the model is as follows.



**Lasso-Ridge Regression:** After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.8012 and for the Ridge model it is increased slightly and R^2 value is 0.9326. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





**Problem Statement: -**

Data of various countries and the factors affecting their Life expectancy has been recorded over past few decades. An analytics firm would like to know how it varies country wise and what other factors are influential in model building. Use your skills to analyze the data and build a Lasso and Ridge Regression model and also summarize the output of the model.

Snapshot the dataset is given below: -

**Business Objective:** To predict the Life Expectancy of the country with other factors by using Lasso-Ridge regression model.

**Constraints:** Lack of analysis of the life Expectancy data of different countries.

**Data Types:** All the data given is used for doing the analysis on the data.

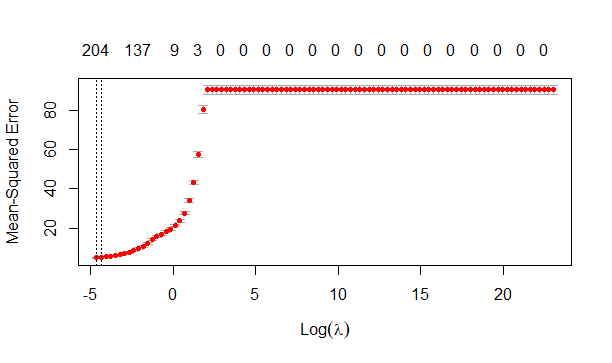
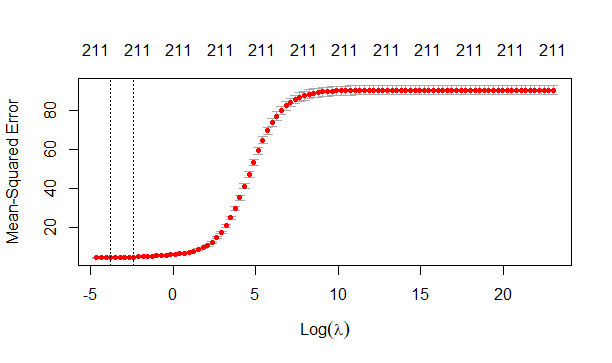
**Data Cleaning:** Since some of the columns in the given data is non-numeric the same is converted into numeric data so that they can be used for doing the analysis.

**Exploratory Data Analysis:** the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

**Multiple Linear Regression:** after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as Life Expectancy of the country.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.98.

**Lasso-Ridge Regression:** After applying the linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.9608 and for the Ridge model it is increased slightly and R^2 value is 0.9540. So finally Multi linear regression can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:



A screenshot of a cell phone

Description automatically generated