The Open University of Hong Kong

School of Science and Technology

BSc (Hons) in Data Science

STAT S460F Advanced Topics In Data Mining

2020-2021

Project title: Car Acceleration

by

Zheng Zequan

12307824

Supervisor: Dr. Tahir Mahmood

Date: January 2021

**Content**

**1 Introduction of the Project 3**

1.1 Research Background ······················································ 3

1.2 Problem Description ······················································· 3

1.3 Aims and Objectives ······················································· 3

**2 Data description 4**

**3 Exploratory and Descriptive Analysis of Data 5**

3.1 Acceleration ······························································· 5

3.2 MPG ········································································ 6

3.3 Displacement ······························································· 6

3.4 Horsepower ······························································· 7

3.5 Weight ··································································· 7

3.6 Cylinders ··································································· 8

**4 Correlation Plot 9**

**5 Q-Q Plot 10**

**6 Splitting the Data 10**

**7 Multiple Linear Regression Models 11**

**8 Ridge and Lasso Regression Model 12**

8.1 Ridge Regression Model ················································12

8.2 Lasso Regression Model ················································13

**9 Model Selection and Conclusion 15**

**10 Limitations 16**

**11 Recommendations 16**

**1. Introduction of The Project**

* 1. **Research Background**

Car have become the most popular conveyance for people and save more time if we drive the car instead of walking, because of the rapid car acceleration. Acceleration is the rate of change of velocity as a function of time. Every year the car industry moved forward. Therefore, the car acceleration got faster and faster. Especially, the sport car.

For me, acceleration is like the angry, fire-breathing dragon of motion variables. Therefore, in this final project, I would like to apply different statistical modelling methods for studying the Car Acceleration dataset. Hopefully, I can find out the research how the way to accelerate the car acceleration.

* 1. **Problem Description**

The problem is the stronger engine or motor, the heavier weight. Various factors will influence each other. How can we increase the car acceleration and make a balance between those various factors?

* 1. **Aims and Objectives**

Studying and understanding the relationship between acceleration and those variables. Then, finding which factors are more affected to car acceleration. So that the effective can be taken to reduce the expenditures for car industry. Finally, I choose some appropriate regression models for the dataset and analyze these several regression models to determine the best regression model.

**2. Data Description**

For Car Acceleration dataset, there have 394 observations. The dataset focuses on the car’s acceleration and five factors which may affect the acceleration, including MPG, Cylinders, Displacement, Horsepower and Weight, show below **TABLE 2.1**. “Acceleration” would be the dependent variable and the other five variables belong to independent variables. (The variables of “Model”,” Year” and “Origin” are just introducing the name and year of cars and the country of cars production, which are not effect on the study, so they are not use).

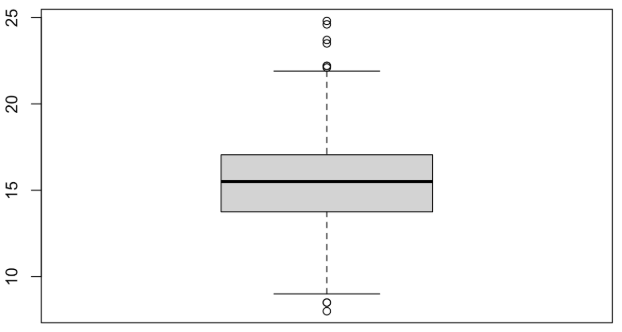
**TABLE 2.1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variable | Property | Description |
| Dependent | Acceleration | Numeric | The cars acceleration (m/s^2) |
| Independent | MPG | Numeric | How far car is able to travel for every gallon of fuel it uses. (miles per gallon) |
| Cylinders | Numeric | The power unit of car engine (cylinder) |
| Displacement | Numeric | The total volume of all the cylinders in car engine (cubic centimeters cc) |
| Horsepower | Numeric | The metric used to indicate the power produced by a car's engine (hp) |
|  | Weight | Numeric | The weight of cars (kg) |
| Disuse | Model | Category | The model of name of cars |
| Origin | Category | The country of cars production |
|  | Year | Numeric | The year of cars production |

**3. Exploratory and Descriptive Analysis of Data**

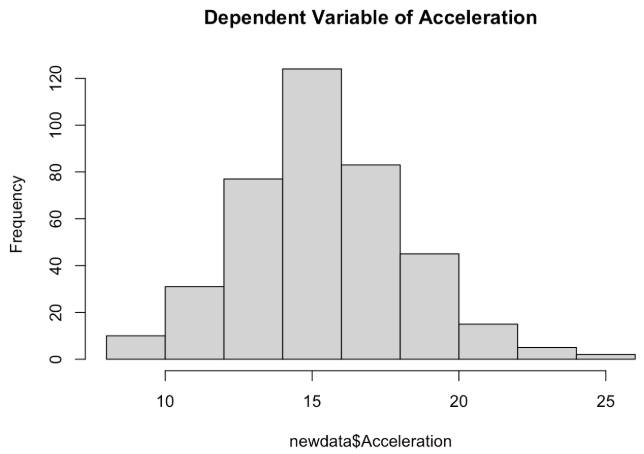
* 1. **Acceleration**

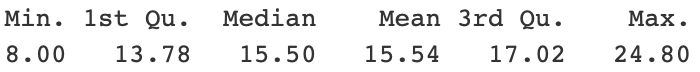
**FIGURE 3.1**



**FIGURE 3.1** The boxplot showing that they are some outliers which outliers in higher level more than outliers in lower level.

**FIGURE 3.2**

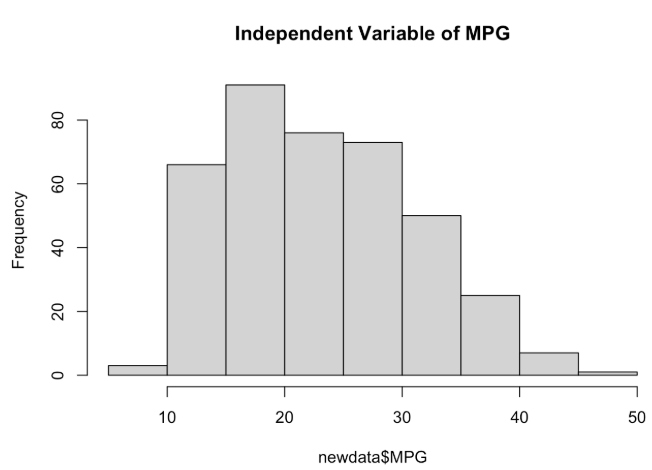




**FIGURE 3.2** The histogram is right skewed. The Median and Mean are very close.

* 1. **MPG**

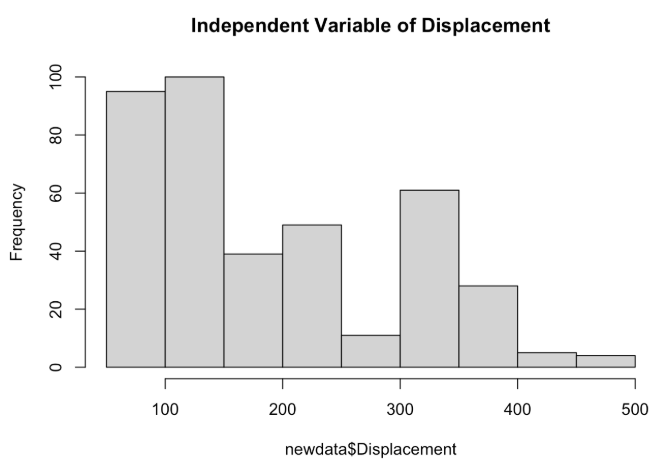
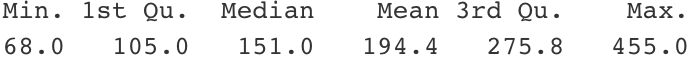
**FIGURE 3.3**



**FIGURE 3.3** The histogram is right skewed. The trend is significantly decrease after 20 mpg.

* 1. **Displacement**

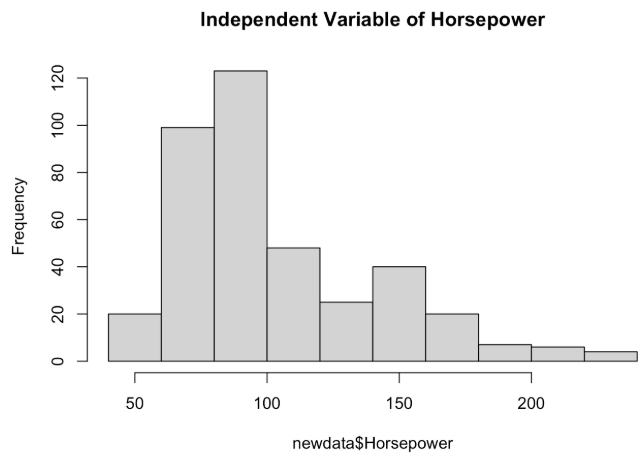
**FIGURE 3.4**

****

**FIGURE 3.4** The histogram is asymmetric. The maximum and minimum value of displacement have a big difference (68 and 455 respectively).

* 1. **Horsepower**

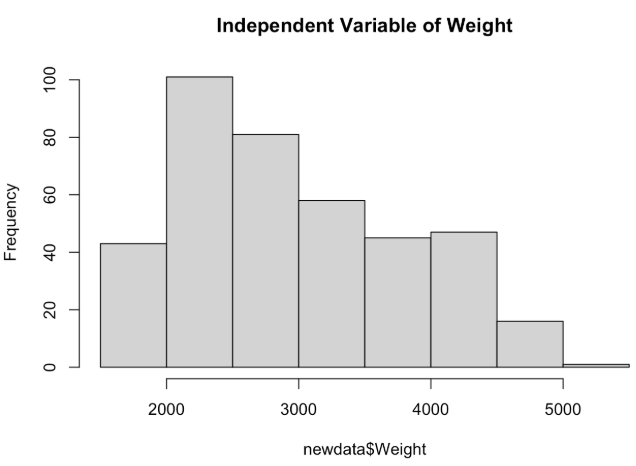
**FIGURE 3.5**

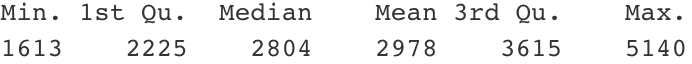
 

**FIGURE 3.5** The histogram is asymmetric. The Mean is greater than Median.

* 1. **Weight**

**FIGURE 3.6**

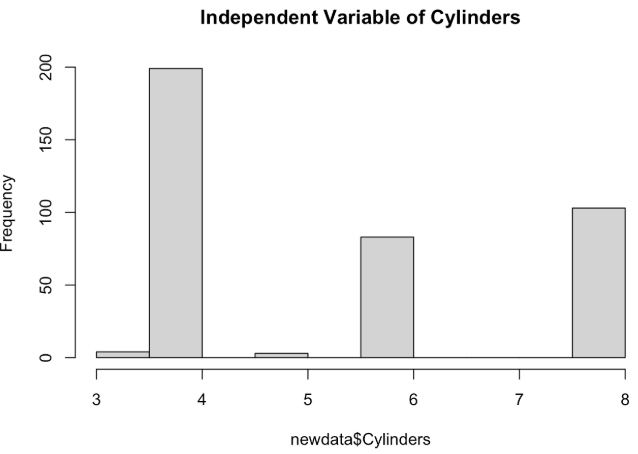




**FIGURE 3.6** The histogram is right skewed. The maximum and minimum value of Weight have a big difference (1613 and 5140 respectively).

* 1. **Cylinders**

**FIGURE 3.7**

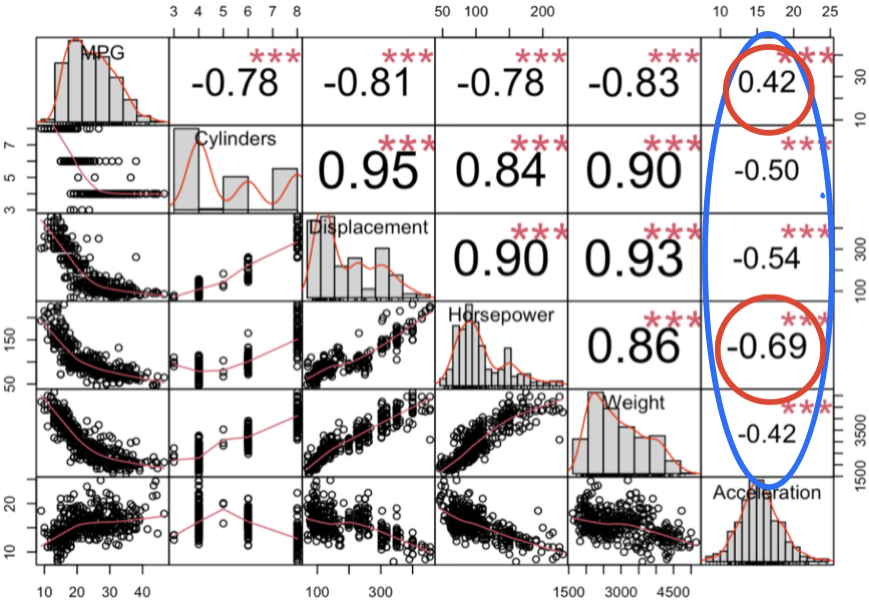


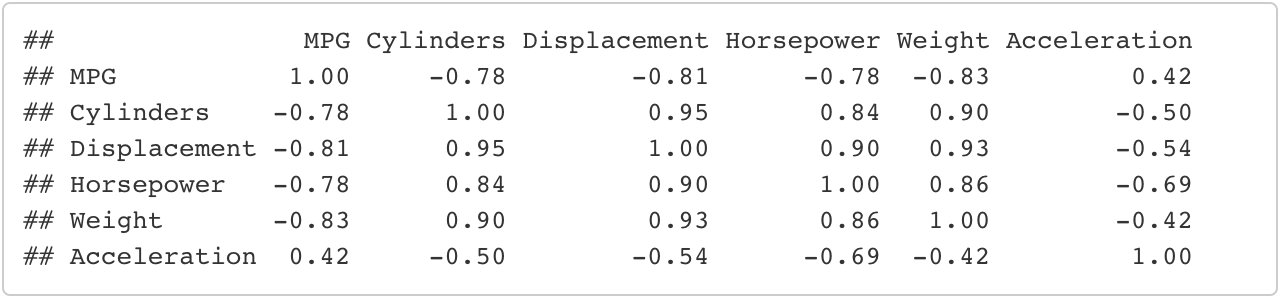


**FIGURE 3.7** Especially, independent variable of cylinders is difference of others. Because, cylinders have only 5 values (3, 4, 5, 6 and 8 respectively). Therefore, the histogram has interval unsurprisingly.

**4. Correlation Plot**

**FIGURE 4.1**

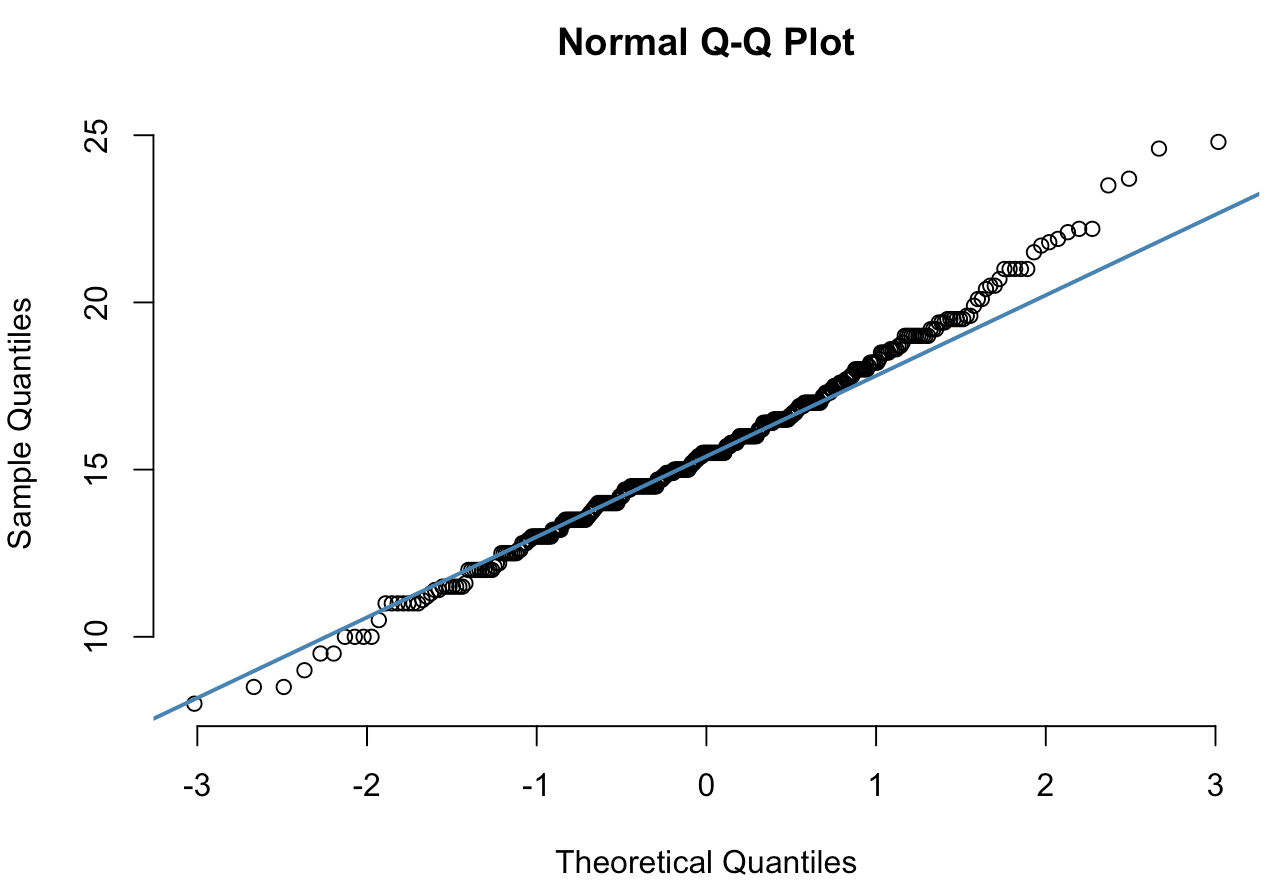




**FIGURE 4.1 Interpretation:** All the variables are asymmetric except the “Acceleration”. Only one positive correlation coefficient which is “MPG”, and it is highest correlation coefficient, but it is a weak positive relationship. Then, “Horsepower” is the lowest correlation coefficient.

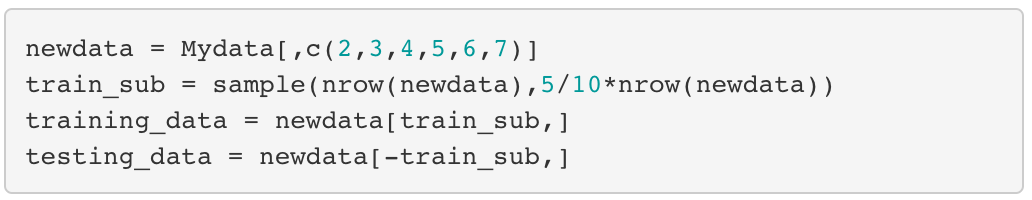
**5. Correlation Plot**

**FIGURE 5.1**



**FIGURE 5.1 Interpretation:** In the dataset for correlation plot, as all the points fall approximately this reference line, we can assume normality.

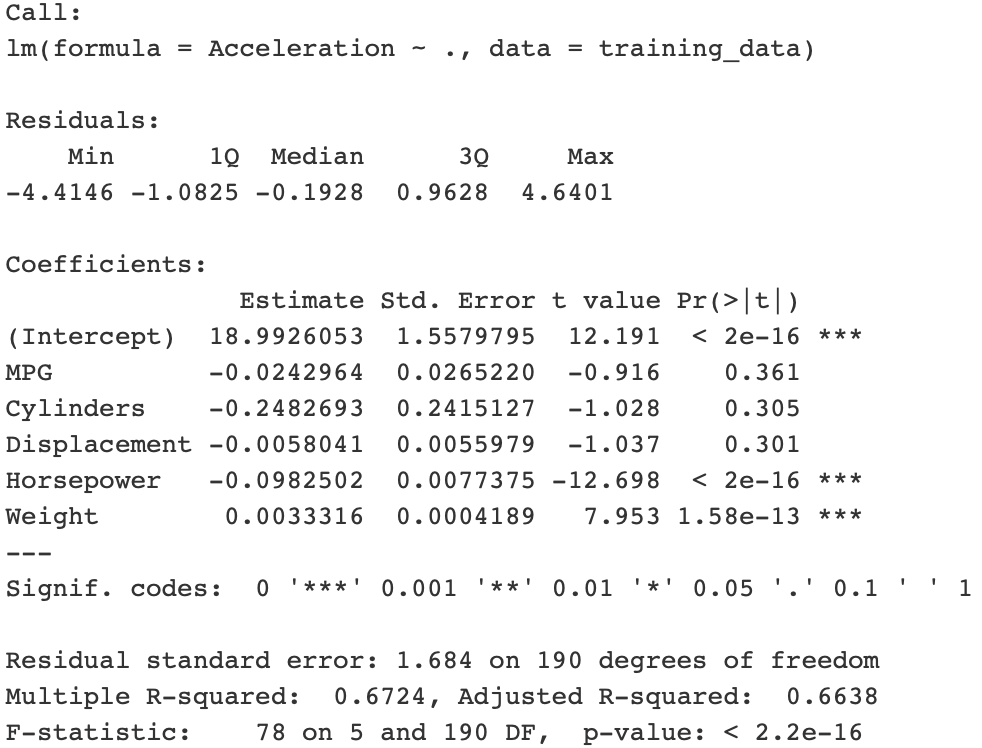
**6. Splitting the Data**



For the dataset, I’m splitting the data into training and testing data in 50-50 randomly.

**7. Multiple Linear Regression Model**

**FIGURE 7.1**



**FIGURE 7.1** We can see that the significance code, horsepower and weight are important predictors (Signif. Codes = \*\*\*).

**Interpretation:** A multiple linear regression model is derived between “Acceleration”, “MPG”, “Cylinders”, “Displacement”, “Horsepower” and “Weight”. The following regression model is obtained

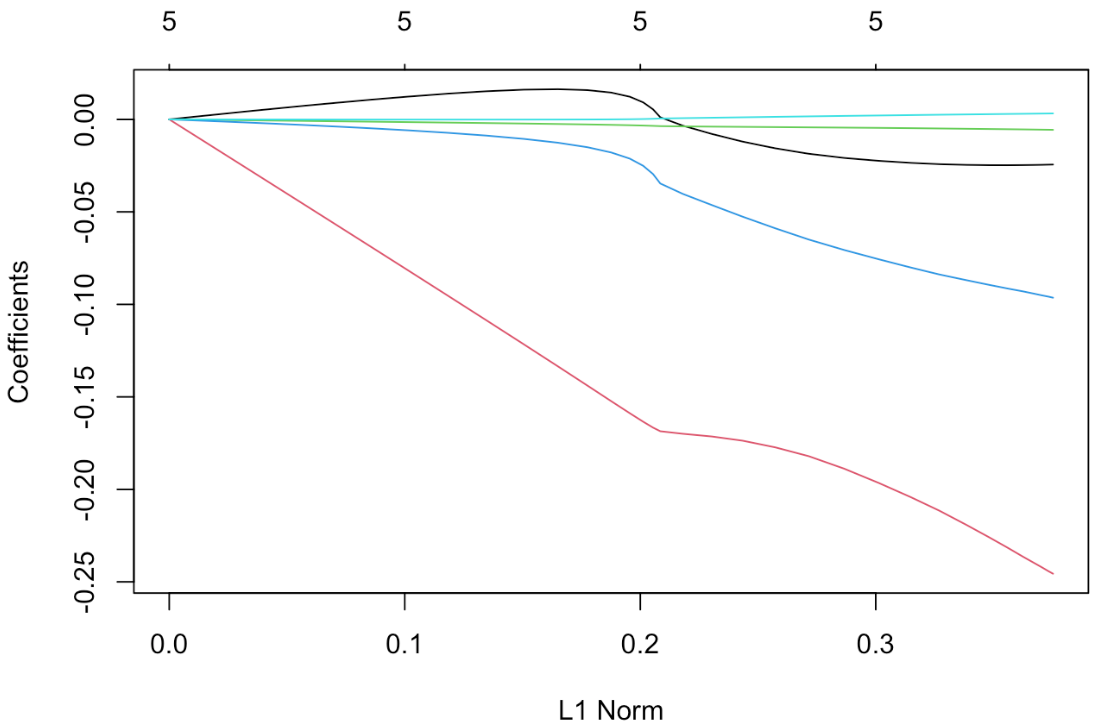
Y = 18.9926053 – 0.0242964(MPG) – 0.2482693(Cylinders) – 0.0058041(Displacement) – 0.0982502(Horsepower) + 0.0033316(Weight)

R-squared and Adjusted R-squared is equals to 0.6724 and 0.6638, respectively. Which means that the explanatory variables explained around 67% variation of the dependent variable. And the mean square error is 2.747682.

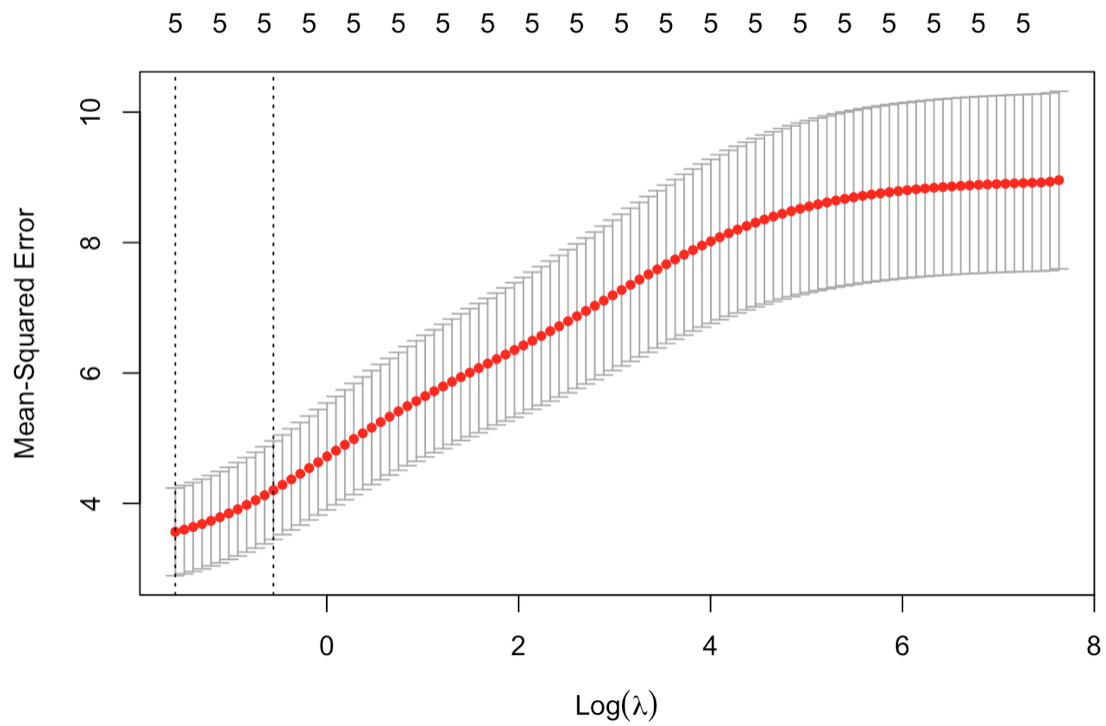
**8. Ridge and Lasso Regression Model**

* 1. **Ridge Regression Model**

**FIGURE 8.1**



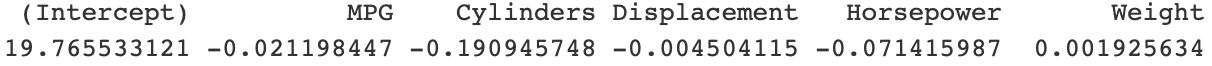
**FIGURE 8.2**



**FIGURE 8.1 and 8.2**

**Interpretation:** By using the cross validation, it is revealed that best lambda choice is 0.206244.

**FIGURE 8.3**



**FIGURE 8.3** None of coefficients are exactly zero. Therefore, ridge regression does not perform variable selection.

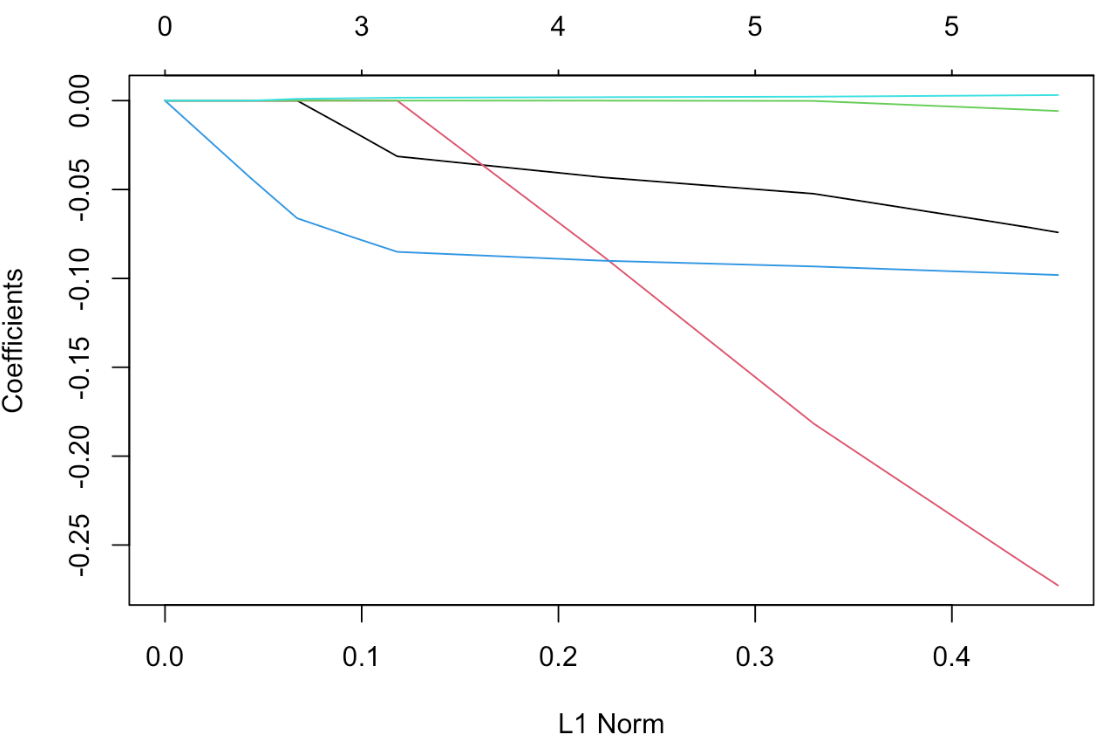
Hence the best ridge regression model is,

Y = 19.765533121 – 0.021198447(MPG) – 0.190945748(Cylinders) – 0.004504115(Displacement) – 0.071415987(Horsepower) + 0.001925634(Weight)

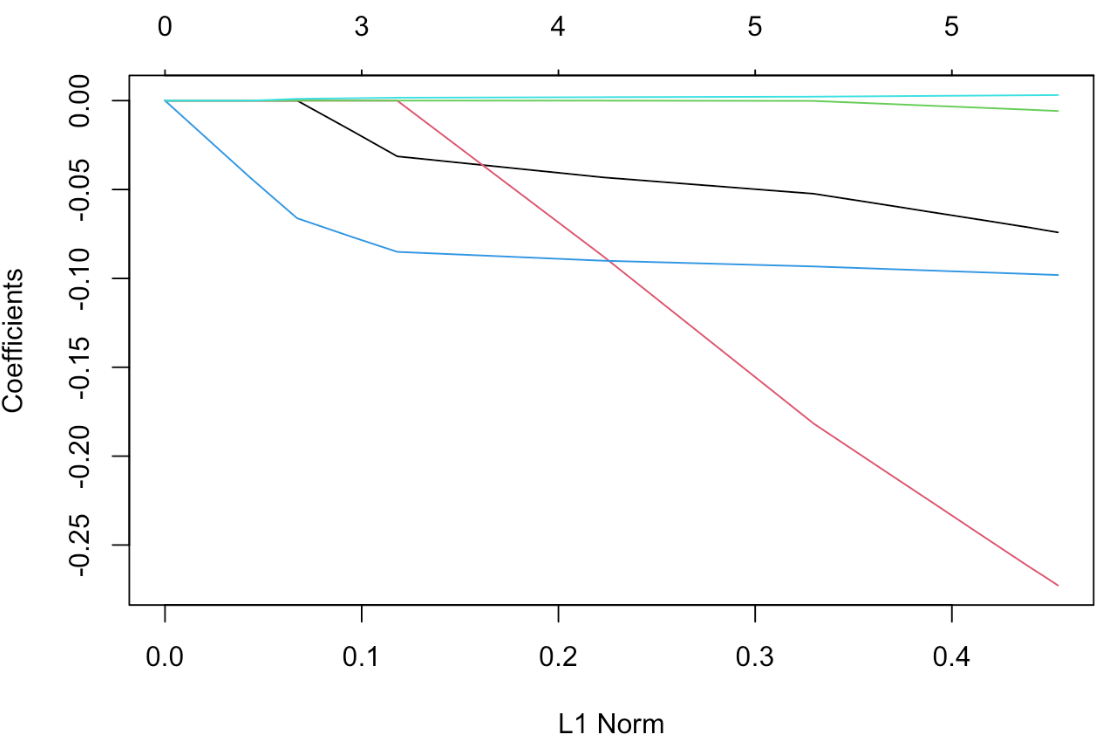
The mean square error is 2.942557.

* 1. **Lasso Regression Model**

**FIGURE 8.4**



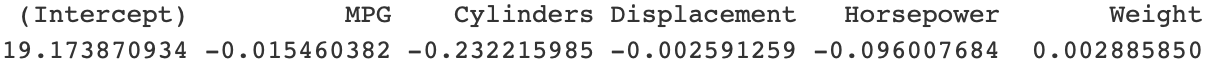
**FIGURE 8.5**



**FIGURE 8.4 and 8.5**

**Interpretation:** By using the cross validation, it is revealed that best lambda choice is 0.02856245.

**FIGURE 8.6**



**FIGURE 8.6** Lasso regression is selecting only the predictors with non-zero coefficients and we can see that they are both non-zero coefficients. So, all coefficients are selecting.

Hence the best Lasso regression model is,

Y = 19.173870934 – 0.015460382 (MPG) – 0.232215985(Cylinders) – 0.002591259 (Displacement) – 0.096007684(Horsepower) + 0.002885850(Weight)

The mean square error is 2.845647.

**9. Model Selection and Conclusion**

**TABLE 9.1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Multiple Linear Regression | Ridge Regression | Lasso Regression |
| MSE | 2.747682 | 2.942557 | 2.845647 |

**TABLE 9.1** The table is a comparison of the fitting results among the models by the mean square error. If the mean square error value is the smaller, then the model fitted should be better.

The result shows that the Multiple Linear Regression has the smallest mean square error. Therefore, Multiple Linear Regression is more suitable than other models for this dataset.

**In conclusion,** the multiple linear regression model is the best model for this dataset. Five selecting independent variables will not be removed. The horsepower and weight are more effect to the car acceleration.

**10. Limitations**

**There are four limitations in this dataset:**

* Only 394 observations and 5 independent variables
* Some outliers may affect the accuracy
* Some technical data cannot be informed
  + Shape of car
* Required the knowledge of automobile manufacturing
  + Aerodynamics

**11. Recommendations**

To improve the accuracy for this dataset. Firstly, it should increase the sample size of the dataset. Secondly, it should increase more professional variables.