

AUTOMPG DATASET REPORT

Analysis of dataset using r

**Sakshi Jaiswal**

Contents

[Introduction 2](#_Toc8561671)

[Load Data 3](#_Toc8561672)

[Explore Data 4](#_Toc8561673)

[Splitting the Data 7](#_Toc8561674)

[Multiple Regression 7](#_Toc8561675)

[Measuring Accuracy 30](#_Toc8561676)

[Conclusion 30](#_Toc8561677)

# 

# Introduction

In this project we will investigate the impact of a number of automobile engine factors on the vehicle’s mpg. The dataset auto-mpg.csv contains information for 398 different automobile models. Information regarding the number of cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name as well as mpg are contained in the file.

Using multiple linear regression to determine the relationship between mpg and different combinations of independent variables.

Consider all possible combinations of independent variables. Here we will report all the appropriate information regarding regression. We have to investigate -

1) Multiple R-squared

2) Adjusted R-squared

3) Complete Linear Regression equation

# Load Data

First using read.csv command we can upload the dataset in R studio.

#### ## 1) Load data ####### ##open file autompg<-read.csv(file.choose()) #view data #View(autompg) #structure of data #str(autompg) #five point summary of data summary(autompg)

## mpg cylinder displacement horsepower   
## Min. : 9.00 Min. :3.000 Min. : 68.0 150 : 22   
## 1st Qu.:17.50 1st Qu.:4.000 1st Qu.:104.2 90 : 20   
## Median :23.00 Median :4.000 Median :148.5 88 : 19   
## Mean :23.51 Mean :5.455 Mean :193.4 110 : 18   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:262.0 100 : 17   
## Max. :46.60 Max. :8.000 Max. :455.0 75 : 14   
## (Other):288   
## weight acceleration model.year origin   
## Min. :1613 Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:2224 1st Qu.:13.82 1st Qu.:73.00 1st Qu.:1.000   
## Median :2804 Median :15.50 Median :76.00 Median :1.000   
## Mean :2970 Mean :15.57 Mean :76.01 Mean :1.573   
## 3rd Qu.:3608 3rd Qu.:17.18 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :5140 Max. :24.80 Max. :82.00 Max. :3.000   
##   
## car.name   
## ford pinto : 6   
## amc matador : 5   
## ford maverick : 5   
## toyota corolla: 5   
## amc gremlin : 4   
## amc hornet : 4   
## (Other) :369

# Explore Data

Explore data is a phase where we can explore the data like view structure of your dataset, view correlation between variables and you can make scatterplot between the variables.

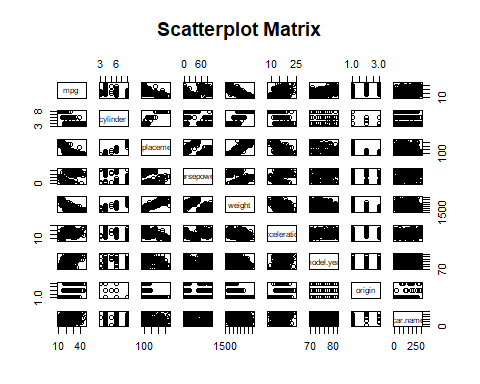
For exploring the data, I am using some commands:

Pairs: used for scatterplot of multiple variables.

Cor: used for correlation value matrix of data

#correlation Matrix of data  
  
  
#cor(correlations)

###### 2 ) Explore data #######  
  
#When we have more than two variables and we want to find the correlation between one variable versus the remaining ones we use scatterplot matrix  
  
pairs(~mpg + cylinder + displacement + horsepower + weight + acceleration + model.year+origin + car.name,data=autompg,main="Scatterplot Matrix")



##### Data Cleansing

Data Cleansing is a phase where we are cleaning our data like removing missing values, and conversions also takes place in this phase.

In this dataset horsepower column contains 6 missing values. So for removing this 6 missing values I will use this process after that I will re-sequence the dataset.

##### 3) Cleansing the data ############  
  
  
  
#Removing Null Values from horsepower Column   
  
autompg<-autompg[-c(33,127,331,337,355,375),]  
  
#view of dataset after removing column  
View(autompg)

And here in this dataset horsepower and car.name is given as factor so we will convert it as numeric using this process.

#Re-Sequence Horsepower column   
row. Names(autompg)<-seq(length=nrow(autompg))  
  
#Convert horsepower into numeric  
  
autompg$horsepower<- as.numeric(autompg$horsepower)  
  
#Convert car name into numeric  
  
autompg$car.name<- as.numeric(autompg$car.name)

# Splitting the Data

###### Here we will split the data into training and test and we used 80-20 ratio for this. Like here I have 398 rows in this dataset , so I will used 80% for training and 20% for test.

###### 4) splitting the dataset into training and test ###########  
  
training<-autompg[1:318,]  
  
test<-autompg[319:392,] ##### 6 was missing values in horsepower after removing that total 392 rows

# Multiple Regression

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is –

y = B0 + B1X + B2X2 +...Bn X n

Here In this dataset response variable is mpg and predictor are cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name.

Using this Predictor, we can make lots of combination for multiple regression model.

So, after the analysis, Here I made 34 multiple regression model.

Here is the list of my models whose r square value is higher. The **square of the correlation, *r2*,** is the fraction of the variation in values of *y* that is explained by the least-squares regression of *y* on *x*.

The maximum ***r2*** I found is82.32 % which is my first model. In this model I used all explanatory variables.

It means max 82.32 % variance explained by my model.

1. autoModel1 <- lm (mpg ~ cylinder + displacement + horsepower + weight + acceleration + model. Year + origin + car.name, data = autompg)

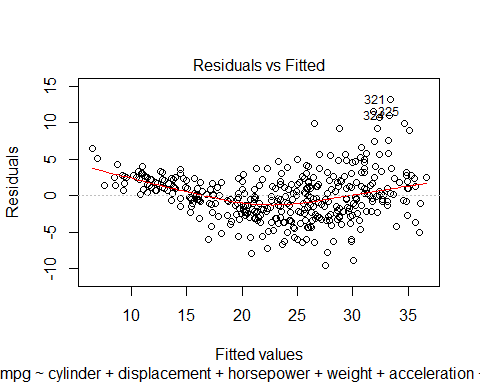
Multiple R-squared: 0.8232

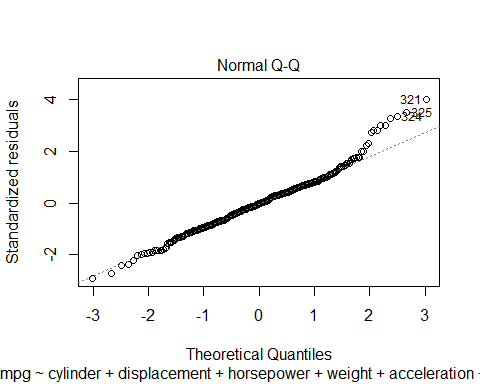
Here Multiple Regression equation for this model is:

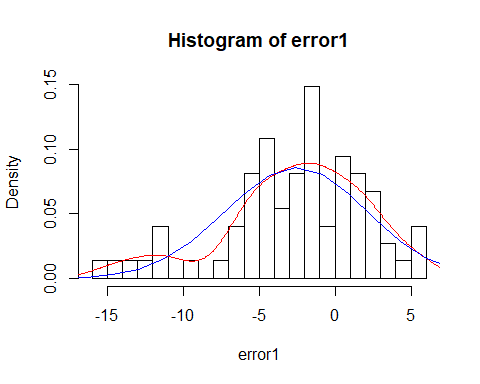
Y = B0 + B1 cylinder + B2 displacement +B3 horsepower +B4 weight +B5 acceleration + B6 model. Year +B7 origin + B8 car.name

Y = -21.796585594 + (-0.268534045) cylinder + 0.016192212 displacement + 0.010228569 horsepower + (-0.006832202) weight + 0.149554354 acceleration + 0.771085956 model. Year + 1.230845717 origin + 0.003497711 car.name

Here is the residual graph and histogram for this model:







2. autoModel2 = lm (mpg ~ cylinder + displacement + horsepower + weight + acceleration + model. Year + origin, data = autompg)  
summary(autoModel2)

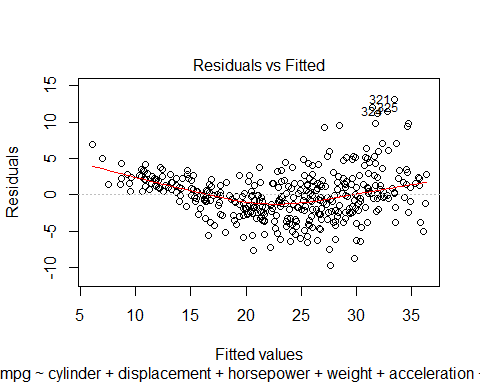
Multiple R-squared: 0.8218

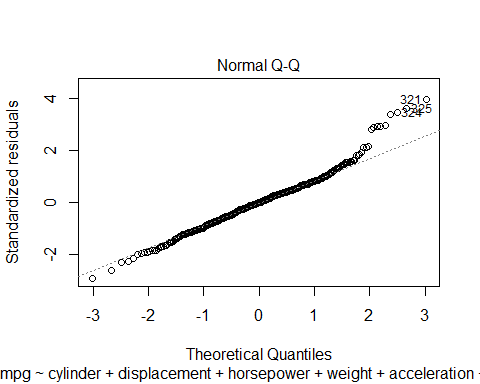
Here Multiple regression equation is:

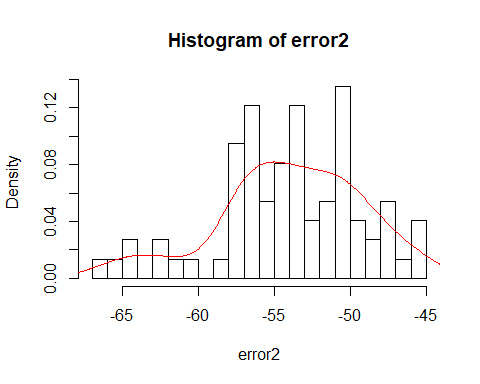
Y = B0 + B1 cylinder + B2 displacement +B3 horsepower +B4 weight +B5 acceleration + B6 model. Year +B7 origin

Y= -21.336705997 + (-0.273860004) cylinder + 0.015389602 displacement + 0.010718491 horsepower +( -0.006755643) weight + 0.148888529 acceleration + 0.768801925 model. Year + 1.343758663 origin

Here is the residual graph and histogram for this model:







In this histogram it’s more towards negative.

3. autoModel4 = lm(mpg ~ horsepower + weight + model.year, data = autompg)

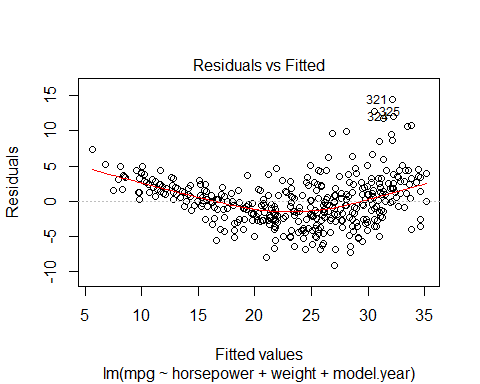
Multiple R-squared: 0.8097

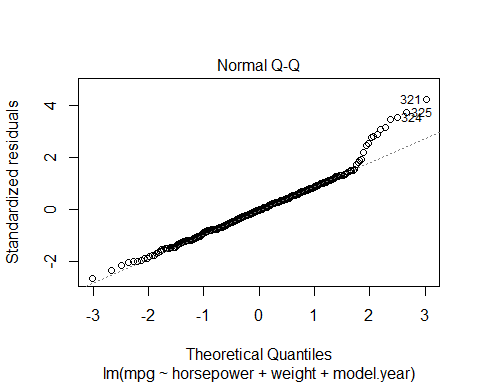
Here Multiple regression equation is:

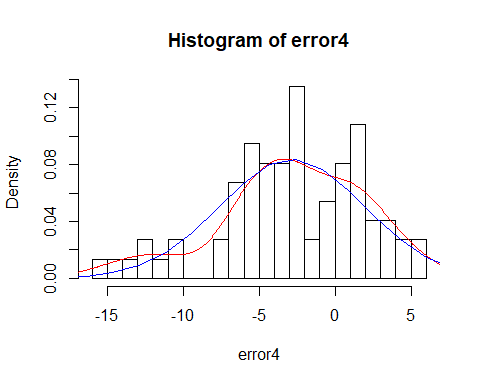
Y = B0 + B1 horsepower +B2 weight +B3 model. Year

Y= 40.444949303 +(-0.010265205) displacement + 0.006438145 horsepower + (-0.006106136) weight + 0.182890644 acceleration

Here is the residual graph and histogram for this model:







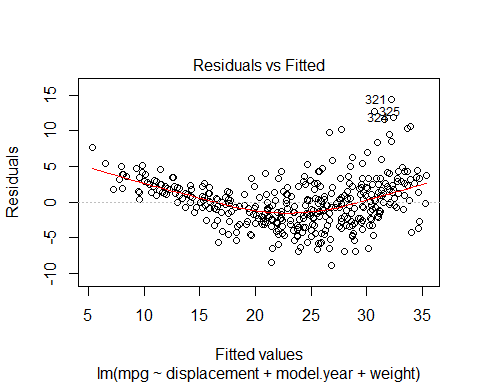
4. autoModel5 = lm (mpg ~ displacement + model.year + weight , data = autompg)

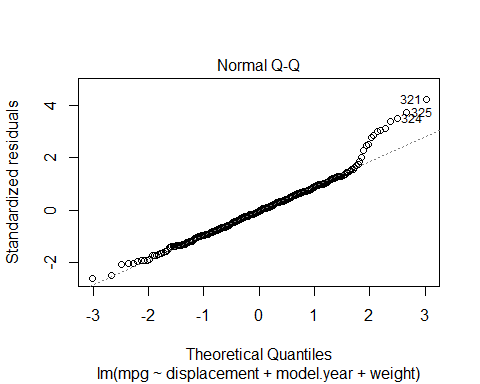
Multiple R-squared: 0.8082

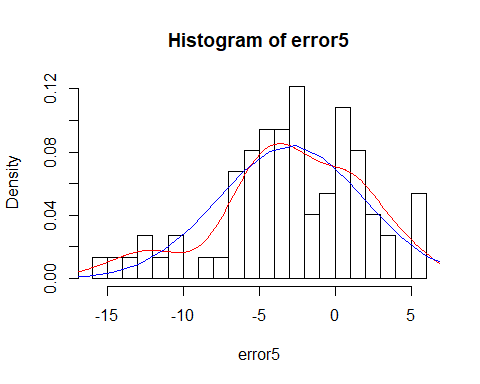
Y = B0 + B1displacement +B2 model. year +B3 weight

Y= -14.363338390+ 0.000283494 displacement + 0.758043499 model. year +(-0.006663689) weight

Here is the residual graph and histogram for this model:







5. autoModel6 = lm(mpg ~ acceleration + model.year + weight , data = autompg)

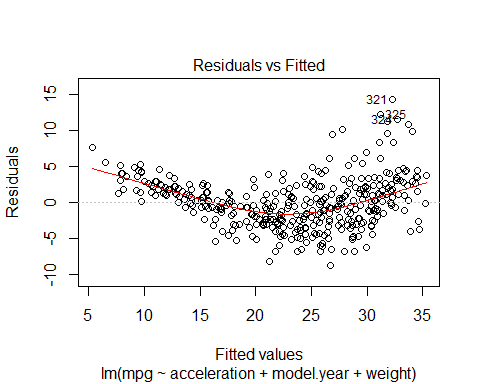
Multiple R-squared: 0.8086

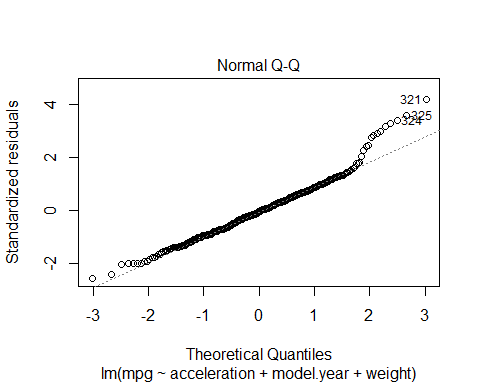
Y = B0 + B1 acceleration +B2 model. year +B3 weight

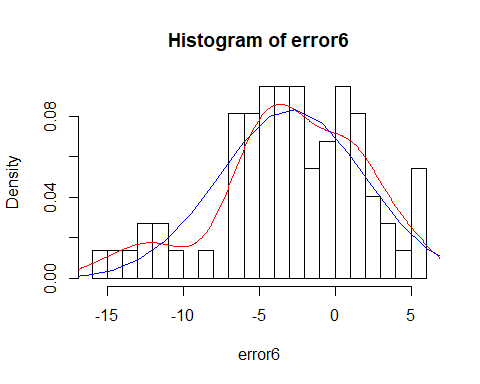
Y=-14.936555253 + 0.066359316 acceleration + 0.748446024 model.year +

(-0.006554126) weight

Here is the residual graph and histogram of model:







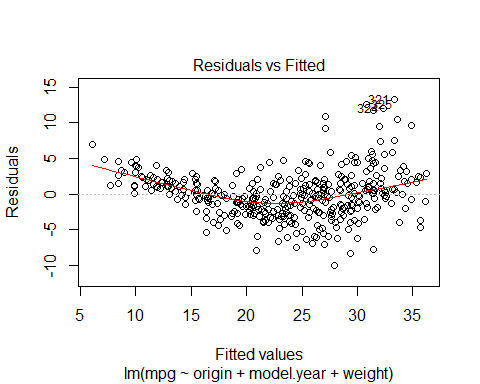
6. autoModel7 = lm(mpg ~ origin + model.year + weight , data = autompg)

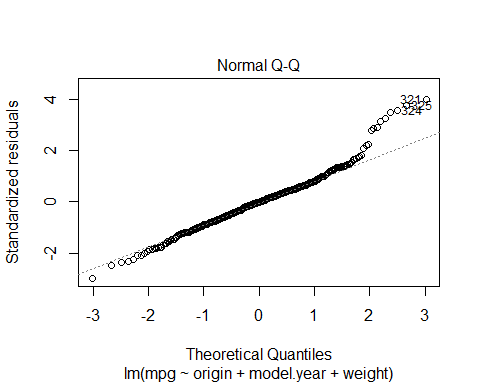
Multiple R-squared: 0.8175

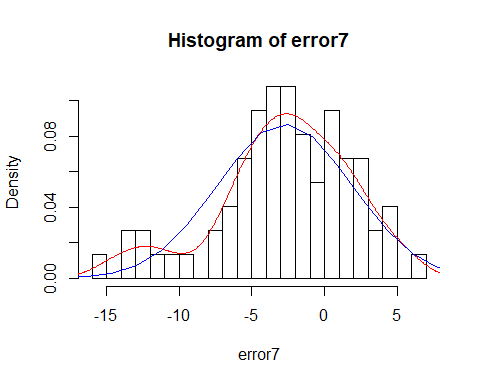
Y = B0 + B1origin +B2 model. year +B3 weight

Y= -18.045850149 + 1.150390789 origin + 0.757126111 model.year + (-0.005994118) weight

Here is the residual graph and histogram of model:







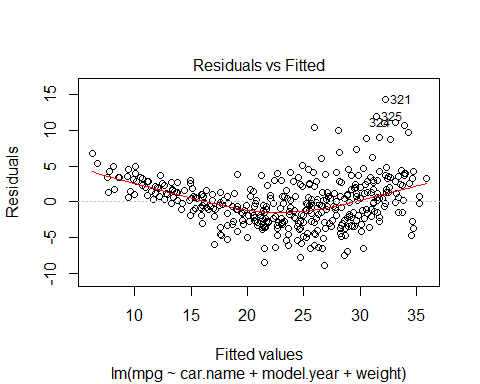
7. autoModel8 = lm (mpg ~ car.name + model.year + weight , data = autompg)

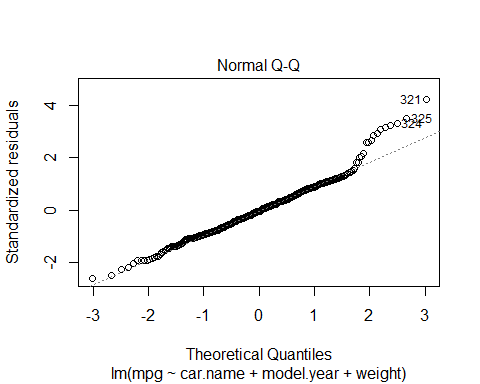
Multiple R-squared: 0.8118

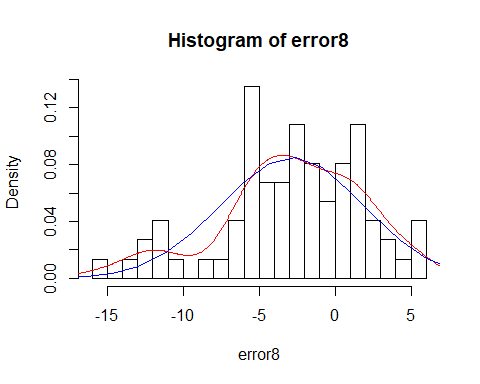
Y = B0 + B1 car.name +B2 model. year +B3 weight

Y= -15.615348708 + 0.005449309 car.name + 0.757626282 model.year +( -0.006487357 ) weight

Here is the residual graph and histogram of model:







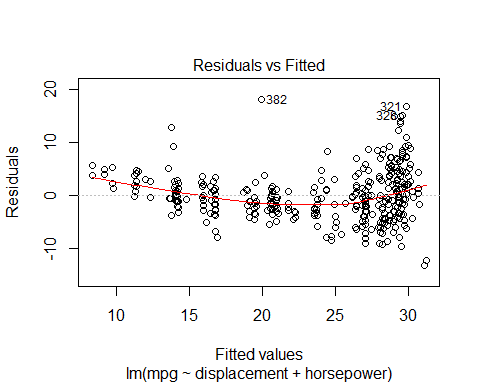
8. autoModel14= lm(mpg ~ displacement + horsepower , data = autompg)

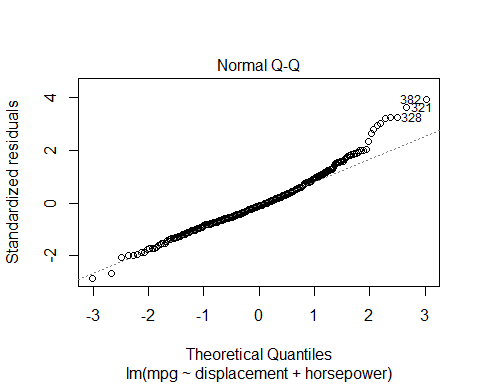
Multiple R-squared: 0.8097

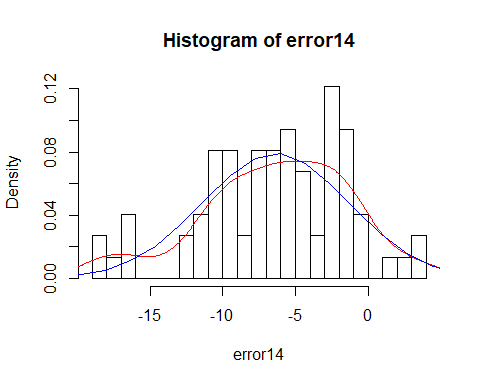
Y = B0 + B1 displacement +B2 horsepower

Y= 33.93597537 + (-0.05793190) displacement + 0.01481191 horsepower

Here is the residual graph and histogram of this model :







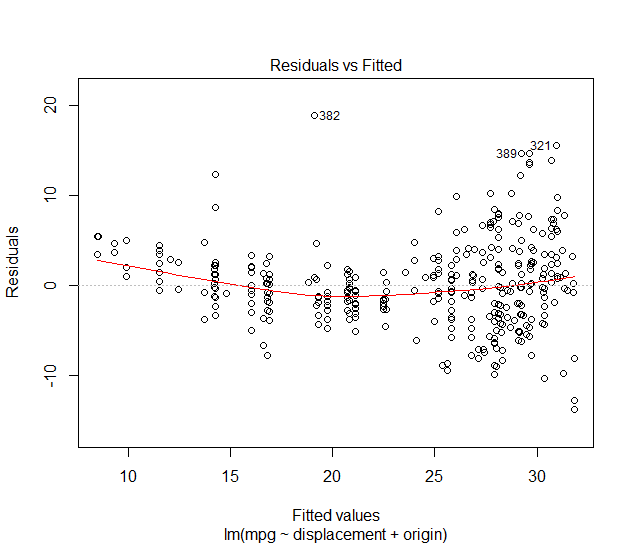
9. autoModel18= lm(mpg ~ displacement + origin, data = autompg)

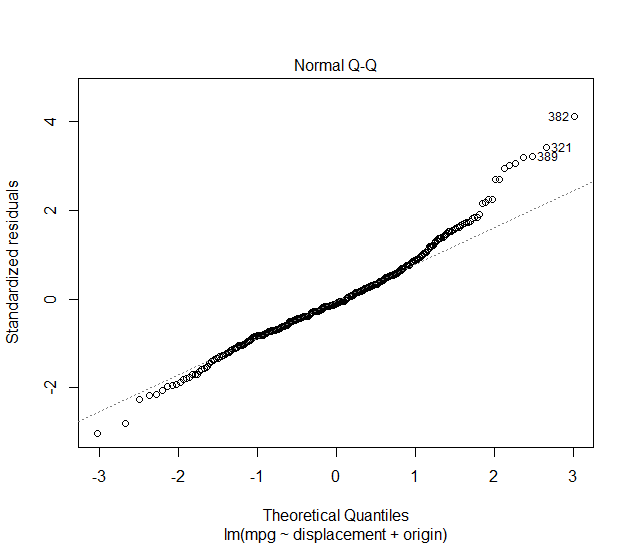
Multiple R-squared: 0.8118

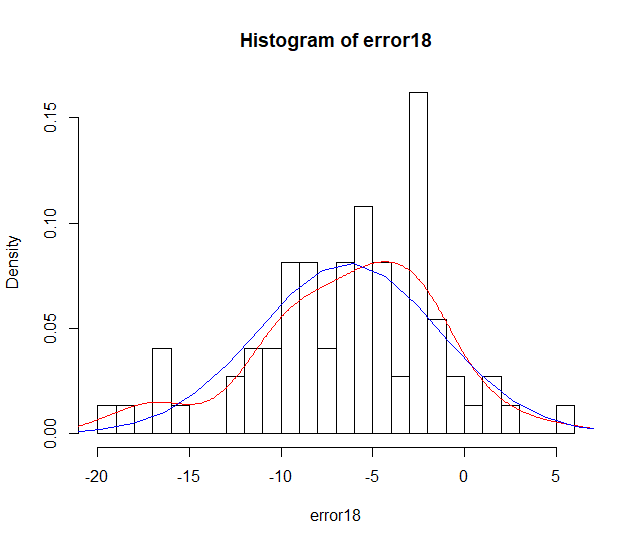
Y = B0 + B1 displacement +B2 origin

Y= 32.38346312 + (-0.05486425) displacement + 1.09653677 origin

Here is the residual graph and histogram of this model:





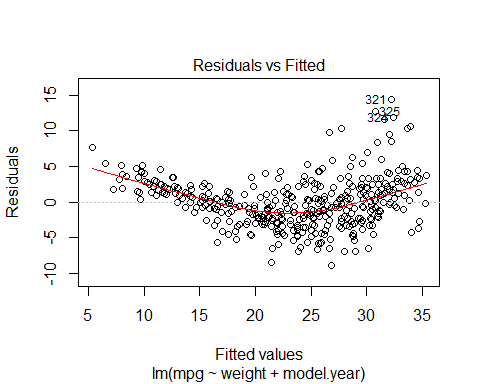


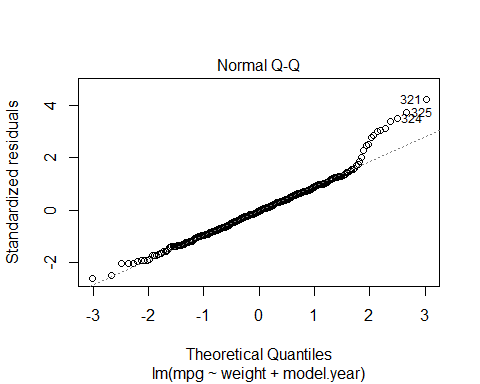
10. autoModel26= lm(mpg ~ weight + model.year , data = autompg)

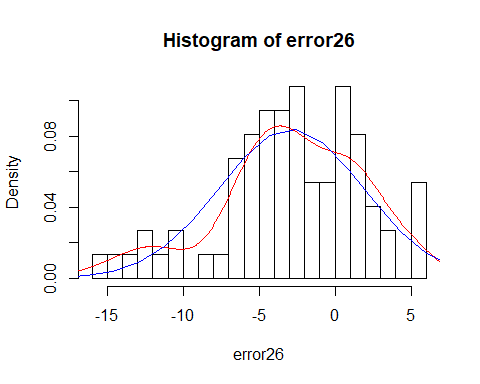
Multiple R-squared: 0.8082

Y = B0 + B1 weight +B2 model.year

Y=-14.347253018 + (-0.006632075) weight + 0.757318281 model.year







# Measuring Accuracy

In this phase we are calculating error and predicted values for every model. Using model on test dataset after that finding the values of errors.

After that creating normal curve of error and of model.

# Conclusion

The best Model is Model 1 here I used all 8 explanatory variable.

autoModel1 <- lm (mpg ~ cylinder + displacement + horsepower + weight + acceleration + model. Year + origin + car.name, data = autompg)

Multiple R-squared: 0.8232

Here Multiple Regression equation for this model is:

Y = B0 + B1 cylinder + B2 displacement +B3 horsepower +B4 weight +B5 acceleration + B6 model. Year +B7 origin + B8 car.name

Y = -21.796585594 + (-0.268534045) cylinder + 0.016192212 displacement + 0.010228569 horsepower + (-0.006832202) weight + 0.149554354 acceleration + 0.771085956 model. Year + 1.230845717 origin + 0.003497711 car.name