Multiple Regression Craigs List Automobile Price Prediction:

Dataset: This information is on 45,425 automobiles offered for sale on Craig's List in the United States. The variables in the data set are:

1. REGION: The region (defined by Craig's List) where the car was for sale.
2. PRICE: The asking price given in the ad for the car.
3. YEAR: The model year of the auto for sale.
4. MAKE: The manufacturer of the car.
5. MODEL: The particular model of the car.
6. CONDITION: The seller-defined condition of the car.
7. CYLINDERS: The number of cylinders of the car's engine.
8. FUEL: The fuel type the car uses, gasoline or diesel.
9. ODOMETER: The odometer reading (miles) on the car.
10. PAINT COLOR: The color of the car.

Problem Statement: Create a primary data set for the analysis of n=250 randomly selected cars. Use the random number seed. The characteristics of this primary data set will be:

a. Only vehicles with MAKE of "cadillac".
b. Only cars from the 2006 through 2011 model years (inclusive).
c. Only cars with engines of 6 or 8 cylinders.

Build a Multiple Linear Regression Model in R and answer the following questions.

Analysis

1. Conduct a multiple linear regression on your random sample with PRICE as the dependent variable and ODOMETER, YEAR, and CYLINDERS as the independent variables.
2. Report the beta coefficients and associated p values and beta coefficient confidence intervals from the model.
3. Conduct appropriate analyses and give interpretations to determine if your model is a good fit to the data in the primary data set.
4. Assess the model's conformance with the LINE assumptions of regression.
5. Throckmorton P. Gildersleeve of Summerfield, Tennessee would like to sell his 2011 Cadillac DT. He says the vehicle is in "excellent" condition and has 175,757 miles on the odometer. Mr. Gildersleeve has not shared details of his Cadillac's engine because he thinks that all 2011 DTS cars had the same famous engine. Determine what price he should ask for the car. Do you believe your pricing advice to the Great Gildersleeve is accurate and usable? Give reasoning for the conclusions.

I. Preprocessing

```
#Author: Suryateja Chalapati

#Importing required libraries
rm(list=ls())
library(rio)
library(moments)
library(dplyr)

#Setting the working directory and importing the dataset
setwd("C:/Users/surya/Downloads")

df = import("Price Data.xlsx", sheet = "Sheet 1")
colnames(df)=tolower(make.names(colnames(df)))
df <- df %>% filter(make == "cadillac", year %in% 2006:2011, cylinders %in% 6:8)
attach(df)

#Setting seed and data sampling
set.seed(36991670)
```

```r
data_sample = data.frame(df[sample(1:nrow(df), 250, replace = FALSE),])
attach(data_sample)

#Checking for factor variable
data_sample$cylinders = as.factor(data_sample$cylinders)
is.factor(data_sample$cylinders)
```

```
## [1] TRUE
```

```r
data_sample$year <- as.numeric(data_sample$year)
```

II. Analysis

```r
#Analysis_[1,2]
#Multiple Regression
lin_reg=lm(price~odometer+year+cylinders, data=data_sample)
summary(data_sample$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    6995   10397   10963   13995   39900
```

```r
summary(lin_reg)
```

```
##
## Call:
## lm(formula = price ~ odometer + year + cylinders, data = data_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9883.6 -2187.8   520.4  2191.5 23189.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.681e+06  3.347e+05  -8.009 4.61e-14 ***
## odometer    -2.356e-02  6.813e-03  -3.458 0.000641 ***
## year         1.340e+03  1.665e+02   8.046 3.64e-14 ***
## cylinders8   5.660e+03  5.939e+02   9.531  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4244 on 246 degrees of freedom
## Multiple R-squared:  0.3392, Adjusted R-squared:  0.3311
## F-statistic: 42.09 on 3 and 246 DF,  p-value: < 2.2e-16
```

```r
confint(lin_reg)
```

```
##                     2.5 %        97.5 %
## (Intercept) -3.340114e+06 -2.021543e+06
## odometer    -3.697740e-02 -1.014041e-02
## year         1.011882e+03  1.667916e+03
## cylinders8   4.490329e+03  6.829852e+03
```

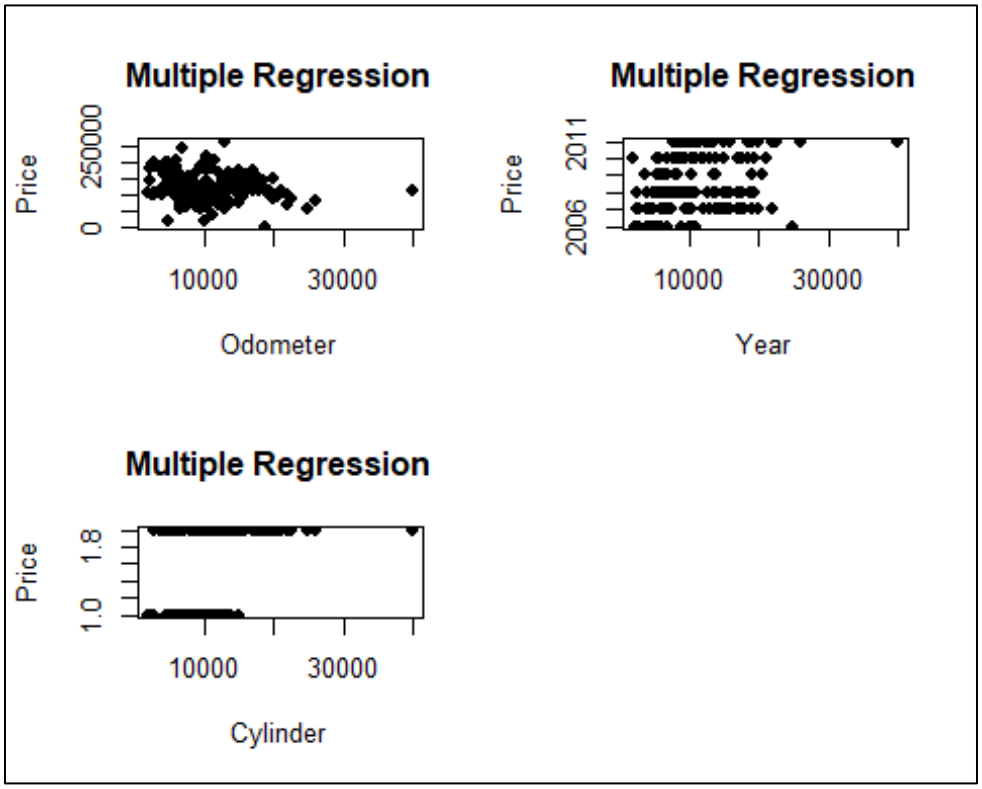
- From the data in the "Cylinders" column, we can say that it is a factor variable and R treats it as such.
- By default, R treats the "Year" column as character. So, I have converted the column into numeric (If not done, it would treat each year as a factor variable).

From the Multiple Regression analysis:

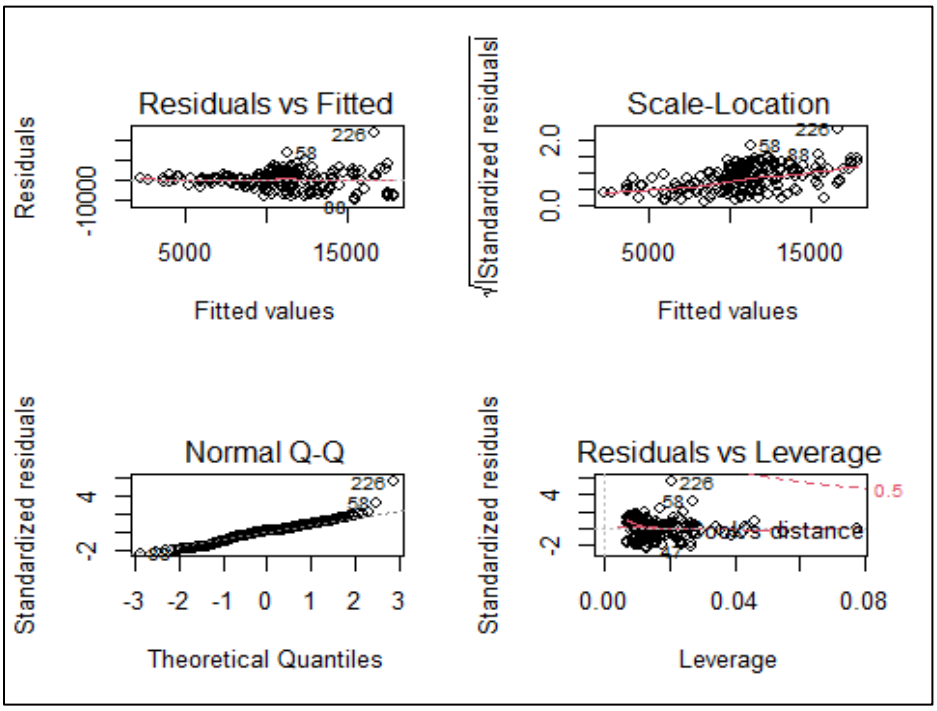
- The estimated B0 – Coefficients for the values are small.
- For all the variables, the p-value is < 0.05. This states we can reject the null hypothesis and accept the alternate, so we can say the B – Coefficients are not zero and have a value.

- The Regression Equation is **y [Price] = – 0.000002681 – 0.02356 + 0.00134*Year + 0.00566*Cylinders**.
- The R-sq is 0.3392, that means all the X variables explains about 33.92% of variation in Y. Based on this, we can say the model is not a good measure of price based on the variables we have taken for our analysis.
- If Year and Cylinders is zero then price decreases, which is something we cannot interpret.
- But we can say there is a direct proportionality between (year, price) and (cylinders, price).

```
#Analysis_3

par(mfrow=c(2,2))
plot(data_sample$price,data_sample$odometer,
     pch=19,main="Multiple Regression",xlab="Odometer",ylab="Price")
plot(data_sample$price,data_sample$year,
     pch=19,main="Multiple Regression",xlab="Year",ylab="Price")
plot(data_sample$price,data_sample$cylinders,
     pch=19,main="Multiple Regression",xlab="Cylinder",ylab="Price")
par(mfrow=c(1,1))
```



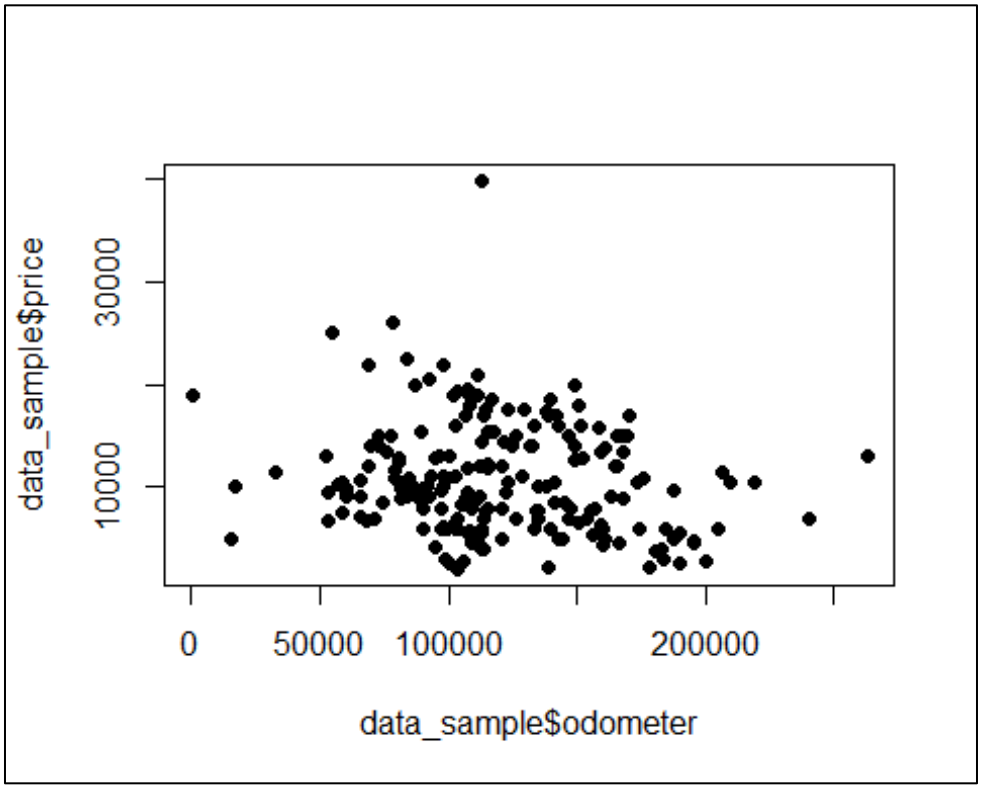

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(lin_reg)
```




```
par(mfrow=c(1,1))
```

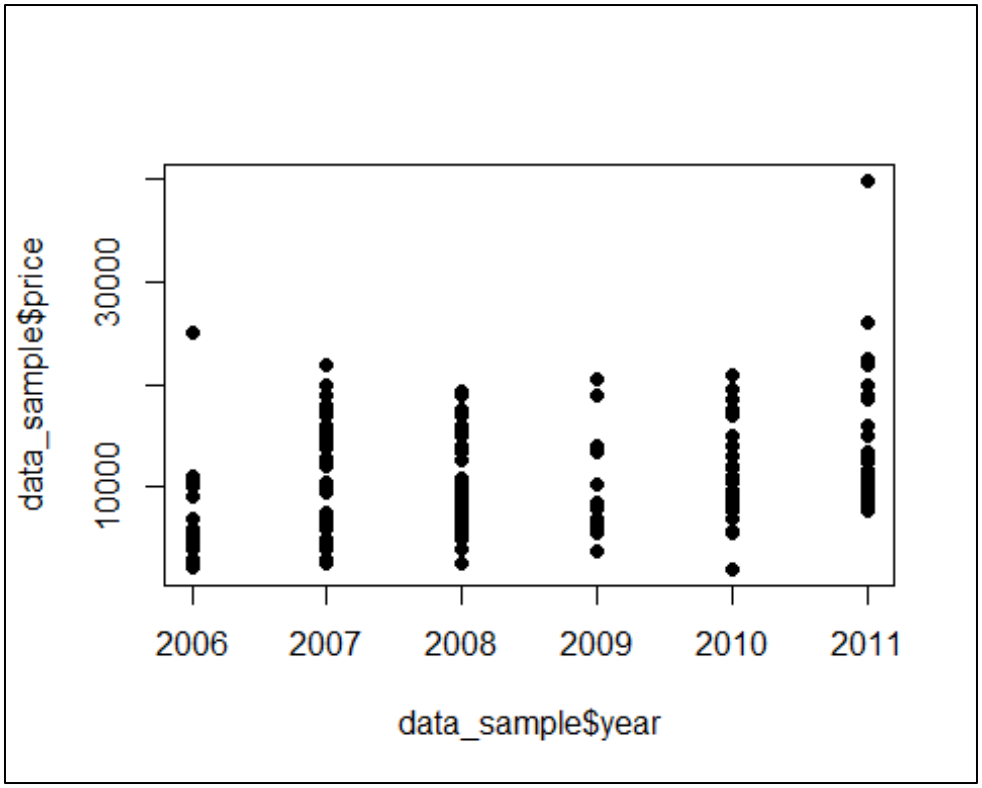
- Based on our analysis, we can say our model is not a good fit for our data.
- We also see heteroscedasticity from the residual plots.

```
#Linearity
plot(data_sample$odometer,data_sample$price,
     pch=19)
```

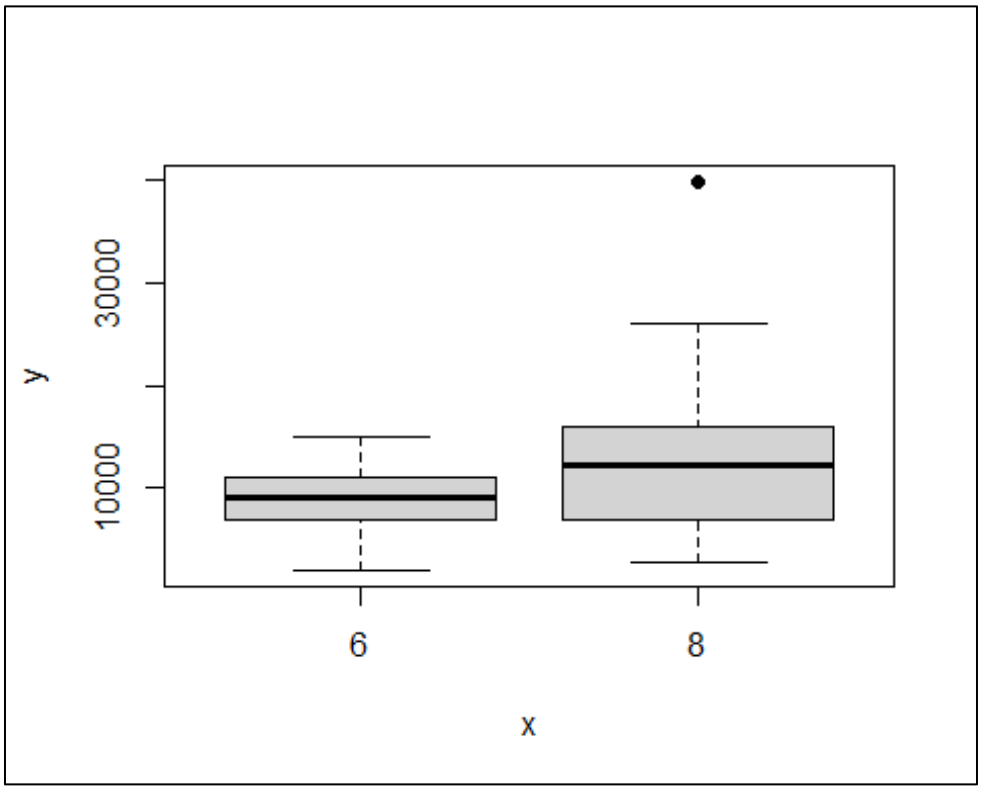


- Data follows linearity but there is an outlier at price = 40,000, pulling the regression line towards itself.

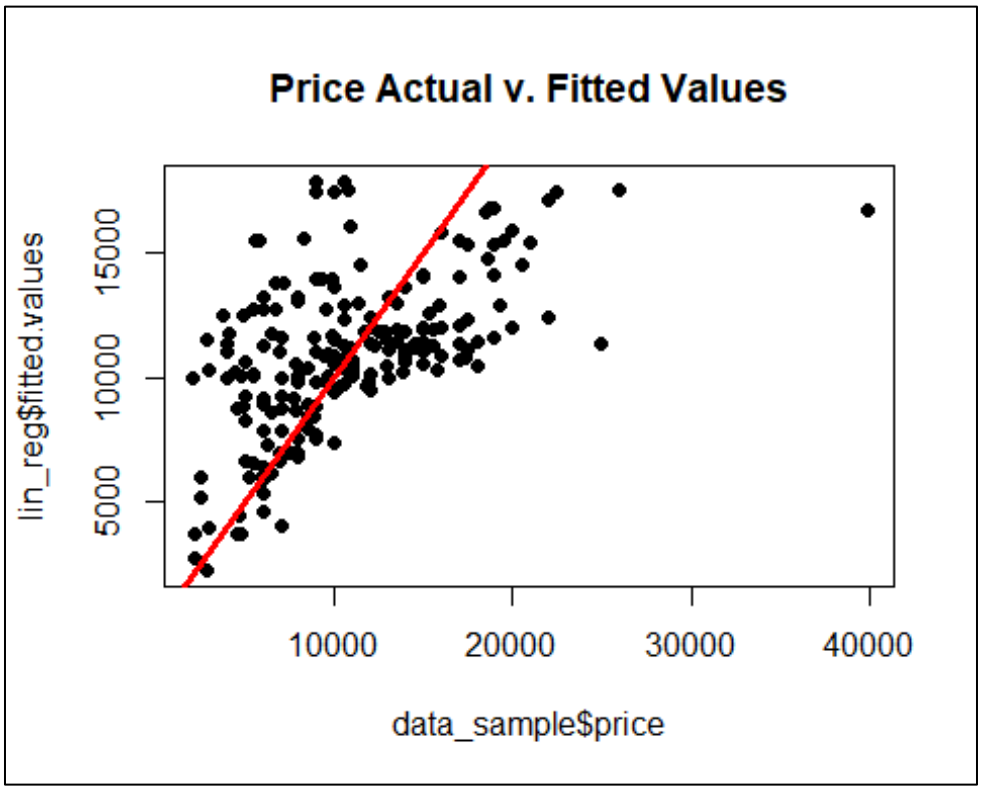
```
plot(data_sample$year,data_sample$price,
     pch=19)
```



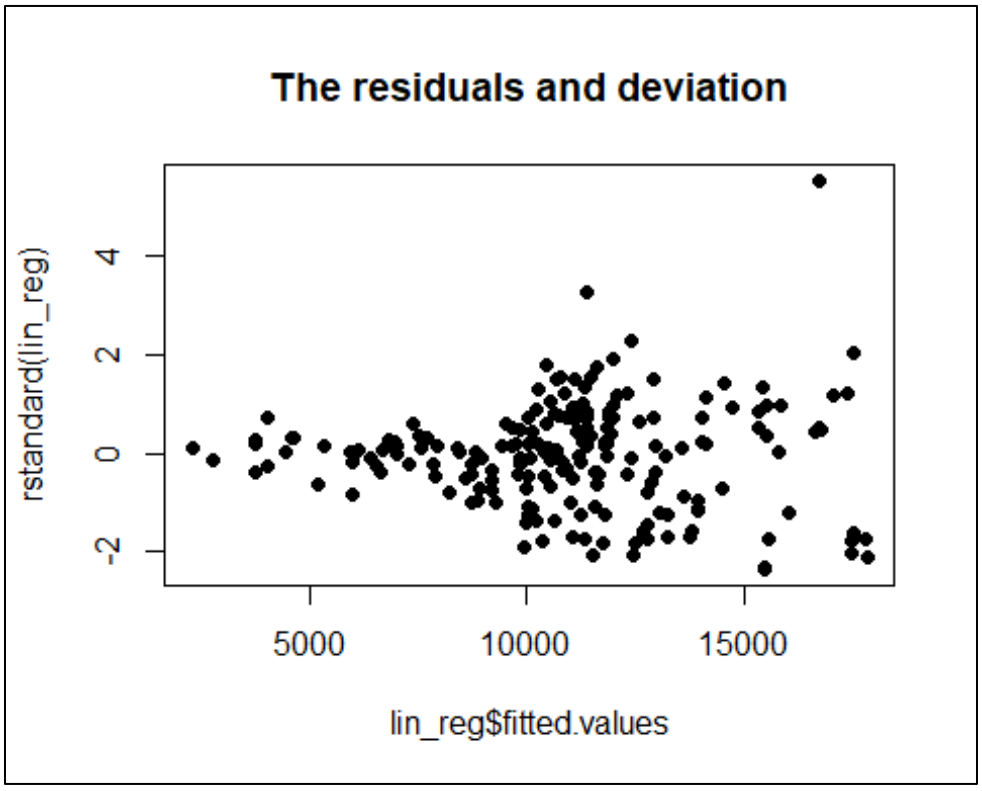

```
plot(data_sample$cylinders,data_sample$price,
     pch=19)
```




```
plot(data_sample$price,lin_reg$fitted.values,pch=19,main="Price Actual v. Fitted Values")
abline(0,1,col="red",lwd=3)
```

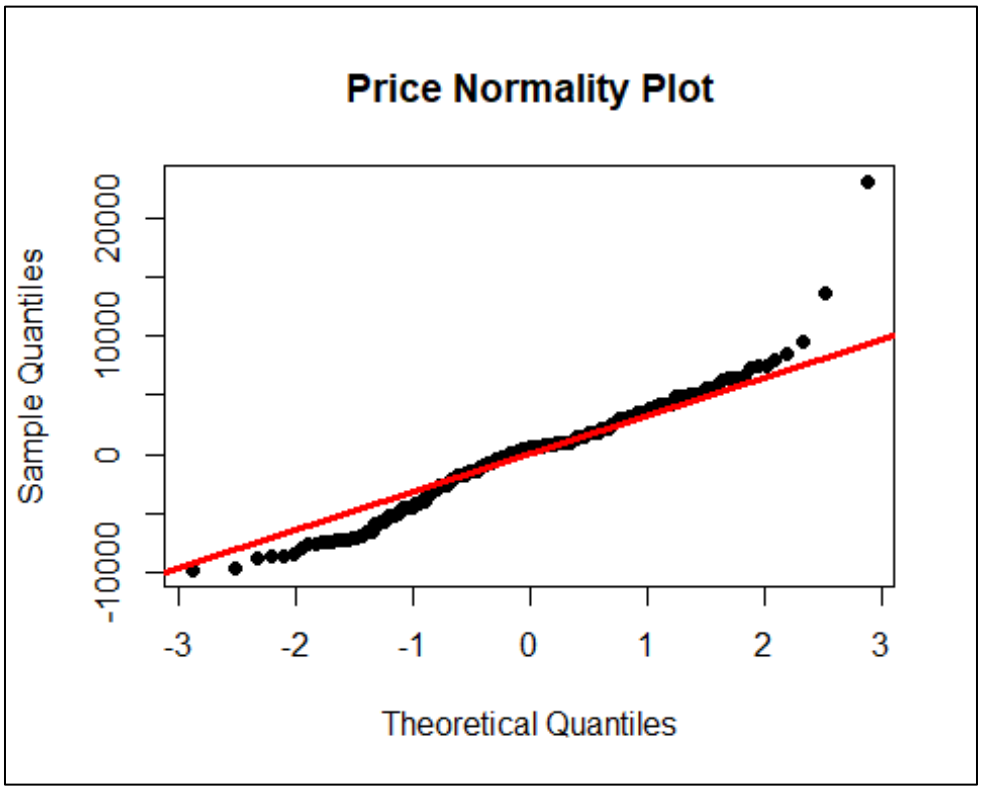



```
#Independence
plot(lin_reg$fitted.values,rstandard(lin_reg),pch=19,main="The residuals and deviation")
```



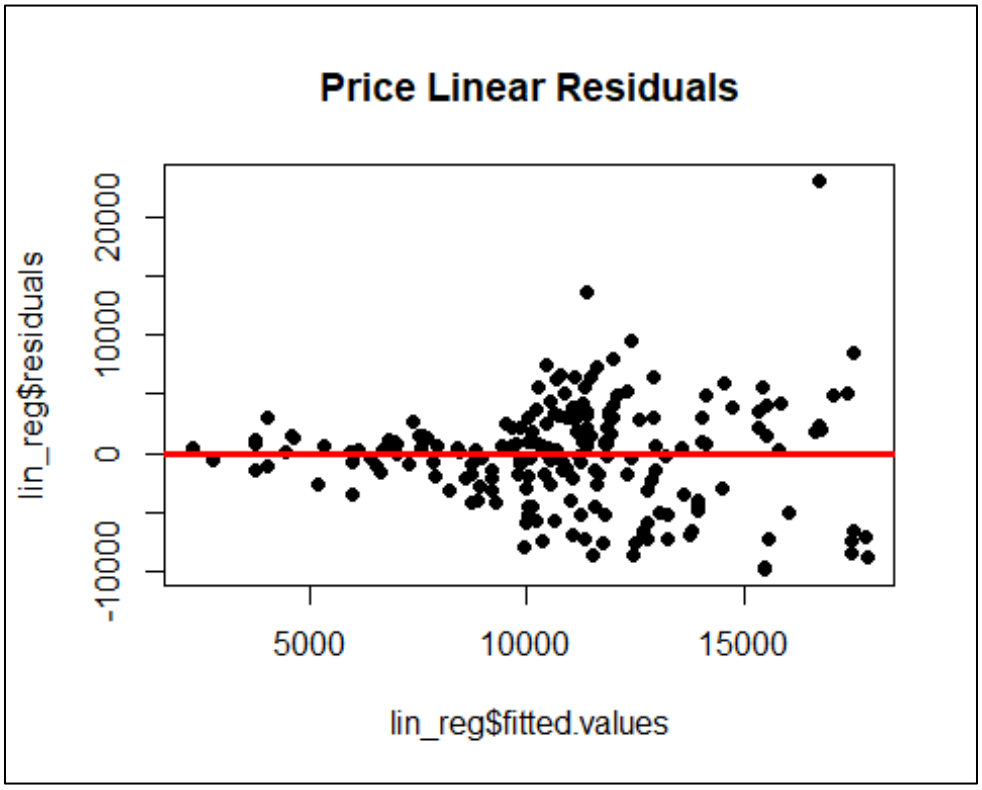
- The fitted vales with their residuals show no pattern and is random. So, it qualifies the Independence test.

```
#Normality
qqnorm(lin_reg$residuals,pch=19,main="Price Normality Plot")
qqline(lin_reg$residuals,col="red",lwd=3)
```



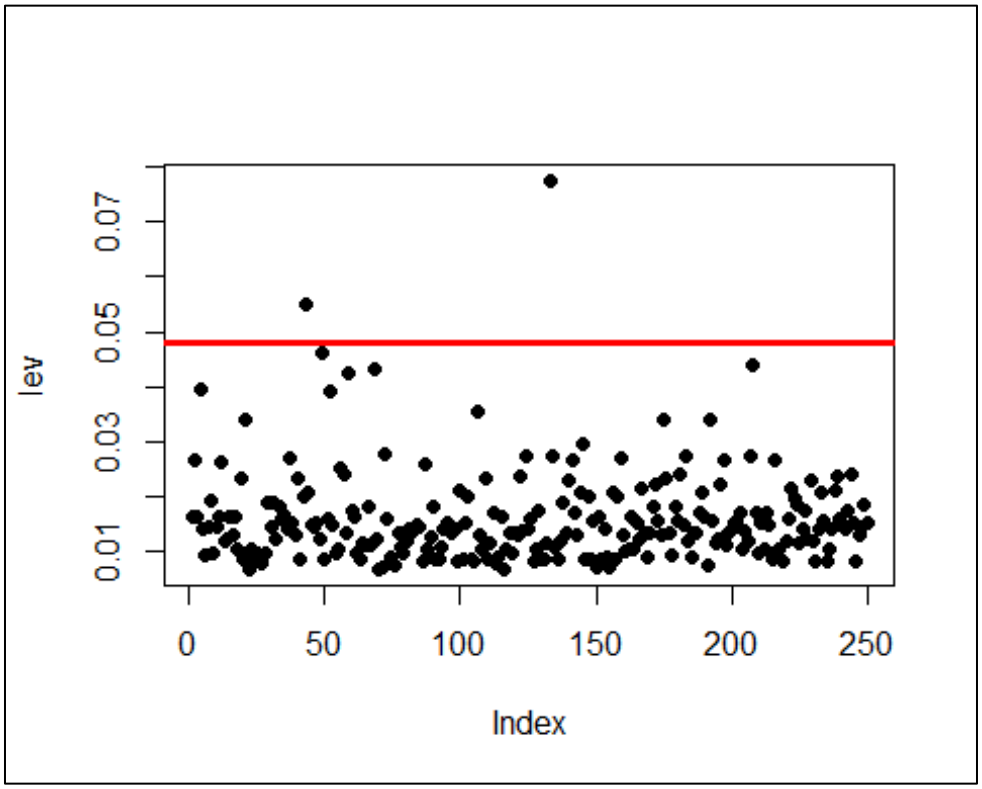
- From the plot the data almost follows a normal distribution but there are few values out in the tails.

```
#Equality of Variances
plot(lin_reg$fitted.values,lin_reg$residuals,pch=19,main="Price Linear Residuals")
abline(0,0,col="red",lwd=3)
```



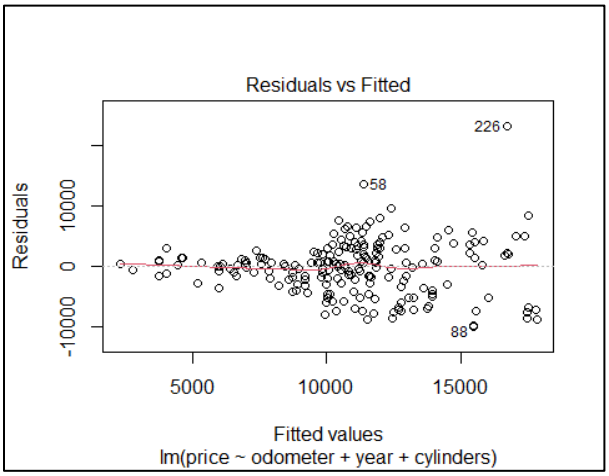
- There is a bit of a scattering across the plot. But, we do not have any pattern. So, it satisfies the equality of variance.

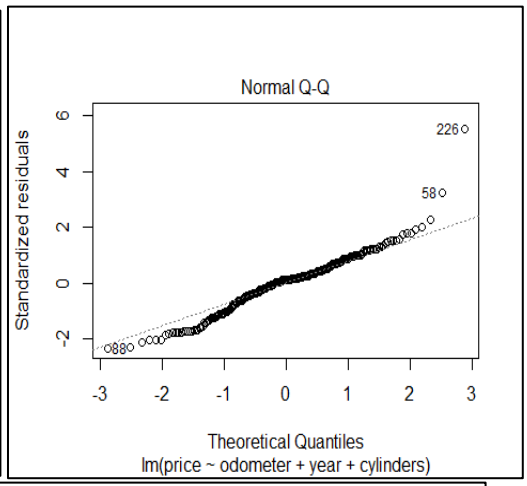
```r
#Identifying high leverage points.
lev=hat(model.matrix(lin_reg))
plot(lev,pch=19)
abline(3*mean(lev),0,col="red",lwd=3)
```

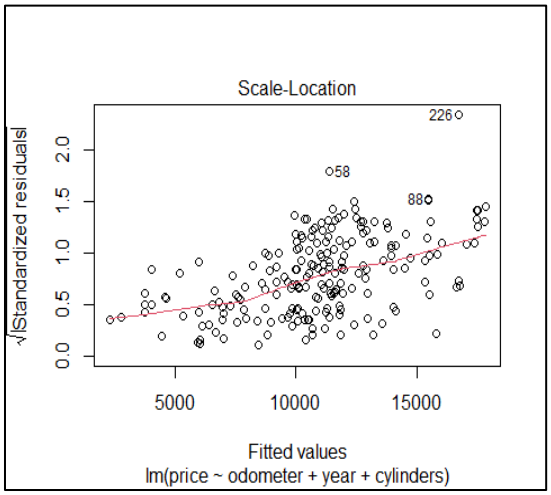


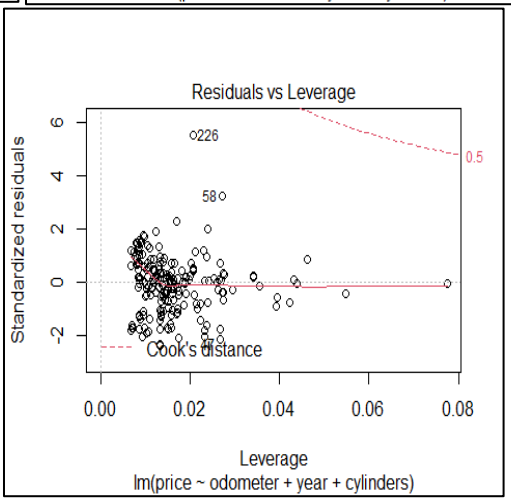

```r
plot(lin_reg)

#Analysis_5

df_predict <- data.frame('odometer'=175757,'year'=2011 ,'cylinders'=8)

df_predict$cylinders <- as.factor(df_predict$cylinders)

predict(lin_reg, df_predict)

## 15227.22
```

- The price prediction is 15227.22, but we cannot say that our prediction is accurate. Even though there seems to be a linear relationship and the data follows heteroscedasticity, due the low R-Sq and a wide range on the CI.


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ odometer + year + cylinders)
226
58
88


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(price ~ odometer + year + cylinders)
226
58
88


Scale-Location
√|Standardized residuals|
Fitted values
lm(price ~ odometer + year + cylinders)
226
58
88


Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
Leverage
lm(price ~ odometer + year + cylinders)
226
58