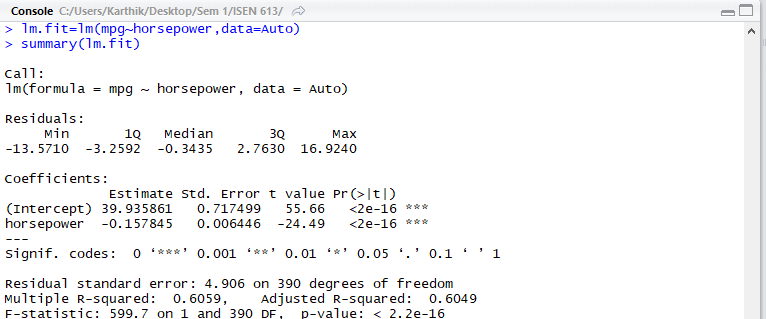
Problem

1)

a)



1. yes, there is a clear evidence of a relationship between the predictor and the response variable, as the p-value corresponding to the F- statistic almost close to 0 at significant levels.
2. R^2 value is 60.5% which suggests that 60.5% of the variability in mpg can be explained by horsepower. Moreover, with a really low p value and \*\*\*, the relationship is even stronger
3. As the coefficient of horsepower is negative, the relationship is negative. Higher the horsepower, mpg will be lesser. (-0.158)
4. The predictor (horsepower) coefficient estimate is negative. Hence there is a negative relationship with mpg. With an increase of 1 horsepower, the mpg goes gown by 0.158 units. Hence the fuel efficiency decreases with an increase in horsepower.

v) 24.46708

> predict(lm.fit,data.frame(horsepower=98),interval="confidence")

fit lwr upr

1 24.46708 23.97308 24.96108

> predict(lm.fit,data.frame(horsepower=98),interval="prediction")

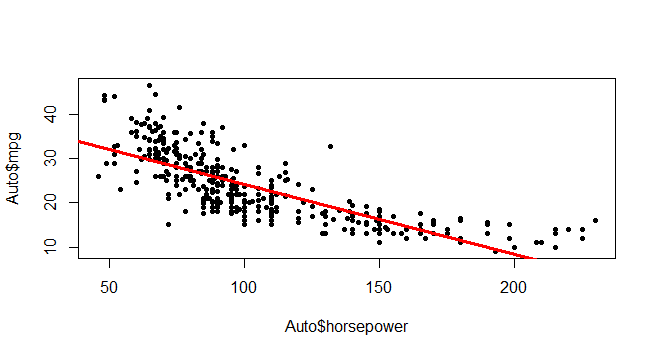
fit lwr upr

1 24.46708 14.8094 34.12476

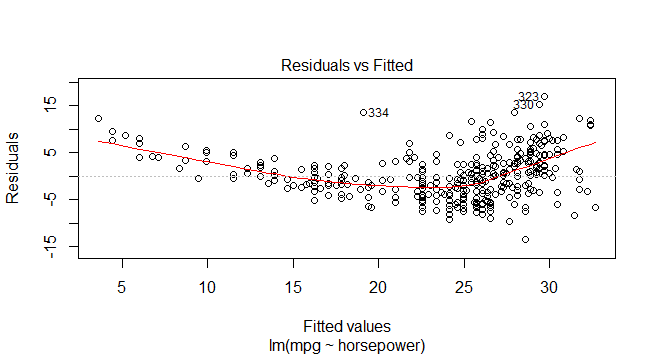
b)

plot(Auto$horsepower,Auto$mpg,pch=20,col="black")

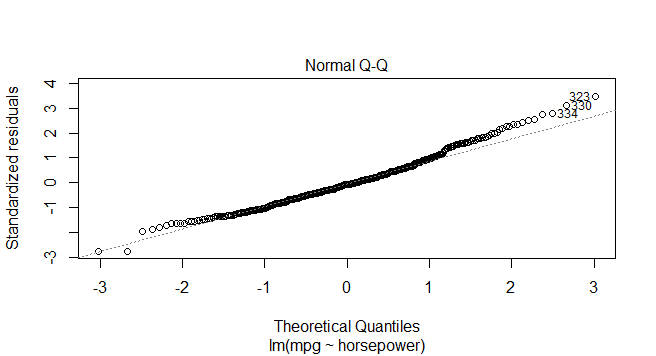
abline(lm.fit,lwd=3,col="red")



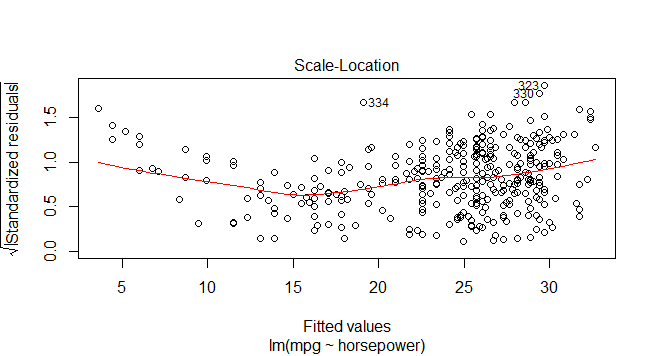
c)



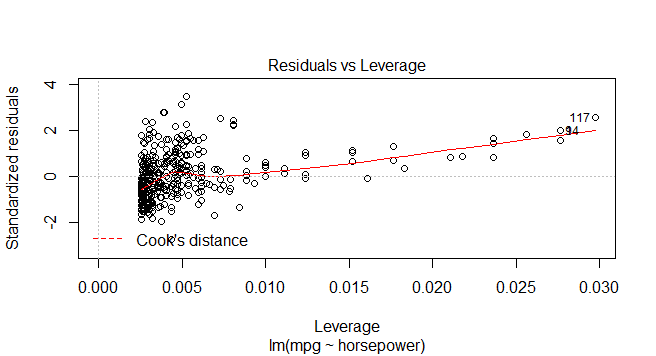
This graph shows a non-linear relationship between the predictor and the response variables.



The residuals of the errors are normally distributed , but there is a slight skew towards the right over theoretical quantiles of 2 to 3.



This graph suggests that the constant variance of error assumption becomes false as there is a funnel shape detected in the graph of increasing residuals towards the right (heteroscedasticity).

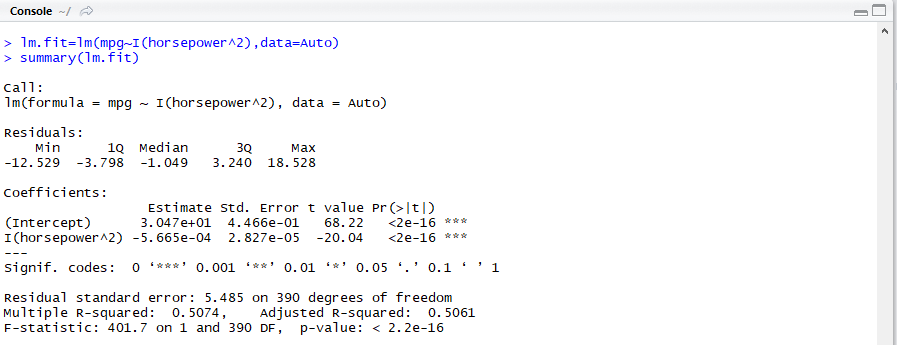


There are a few leverage points like 117 and 94 with h value h>>(p+1)/no. of observations = (1+1)/392=0.005. assuming “>>” as a factor of 3 => all values out of 0.015 in the leverage axis.

The outliers are such that the standardized value is out of the +/-3 range which are quite a few here.

**FOR X^2**

a)



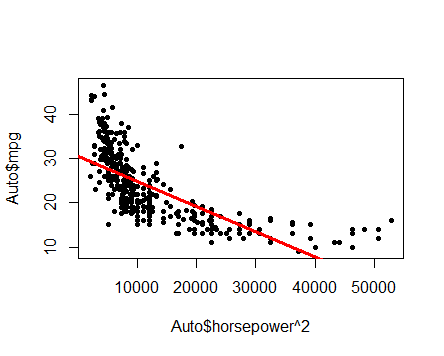
1. yes, there is a clear evidence of a relationship between the predictor and the response variable, as the p-value corresponding to the F- statistic almost close to 0 at significant levels.
2. R^2 value is 50.61% which suggests that 50.61% of the variability in mpg can be explained by the predictor variables.
3. As the coefficient of horsepower^2 is negative, there is a negative relationship of mpg with it.
4. With an increase of 1 horsepower^2, the mpg goes down 5.06e-04 units. Hence the fuel efficiency decreases with an increase in horsepower^2.

v) 25.02512

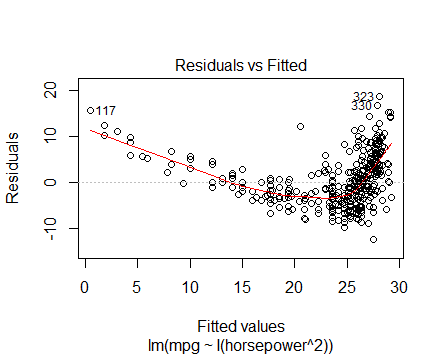
|  |
| --- |
| predict(lm.fit,data.frame(horsepower=98),interval="confidence")  fit lwr upr  1 25.02512 24.45883 25.5914  > predict(lm.fit,data.frame(horsepower=98),interval="prediction")  fit lwr upr  1 25.02512 14.22603 35.8242 |
|  |
| |  | | --- | |  | |

b) {plot(Auto$horsepower^2,Auto$mpg,pch=20,col="black")

abline(lm.fit,lwd=3,col='red')}

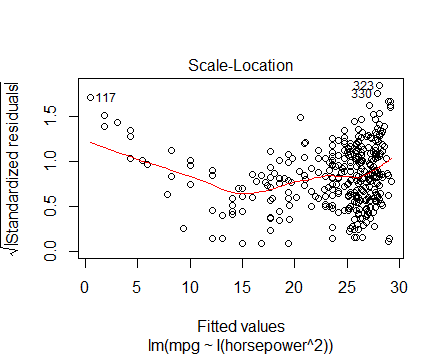


c)

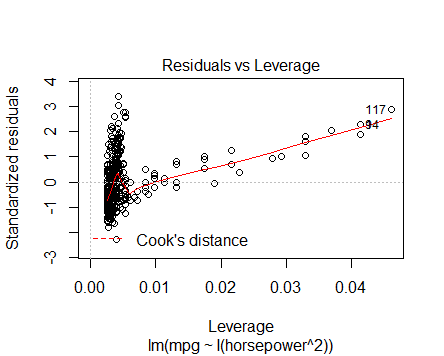


This graph shows a non linear relationship between the predictor and the response variables.

This graph suggests that the constant variance of error assumption is false as there is a funnel shape detected in the graph of increasing residuals towards the right.



As there is no linear line, it has a non constant variance.



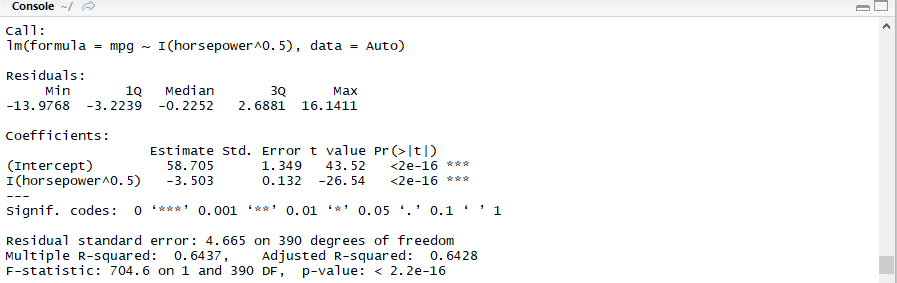
There are many leverage points with h value h>>(p+1)/no. of observations = (1+1)/392=0.005.

Assuming >> as a factor of 3, leverage above 0.015 are high leverage points like 117,94 and a few others.

There are outliers points such that the standardized residual is out of the +/3 range.

**FOR X^1/2**

a)



1. yes, there is a clear evidence of a relationship between the predictor and the response variable, as the p-value corresponding to the F- statistic almost close to 0 at significant levels.
2. R^2 value is 64.28% which suggests that 64.28% of the variability in mpg can be explained by the predictor variables.
3. As the coefficient of horsepower^0.5 is negative, the relationship of mpg is negative with it.
4. With an increase of 1 horsepower^0.5, the mpg goes down by 3.503 units. Hence the fuel efficiency decreases with an increase in horsepower.
5. 24.02206

> predict(lm.fit,data.frame(horsepower=98),interval="confidence")

fit lwr upr

24.02206 23.55687 24.48724

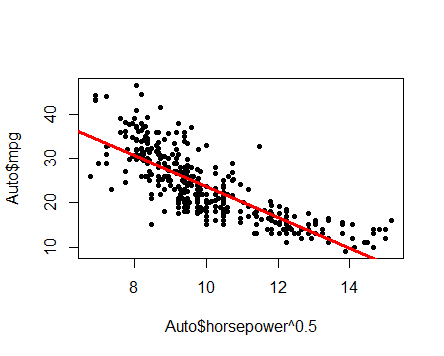
> predict(lm.fit,data.frame(horsepower=98),interval="prediction")

fit lwr upr

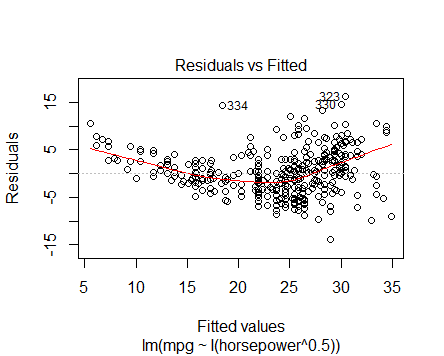
24.02206 14.83892 33.20519

b) {plot(Auto$horsepower^0.5,Auto$mpg,pch=20,col="black")

abline(lm.fit,lwd=3,col='red')}

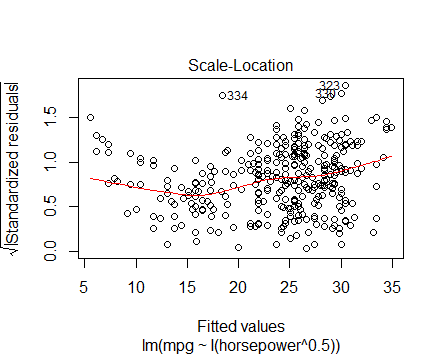


c)

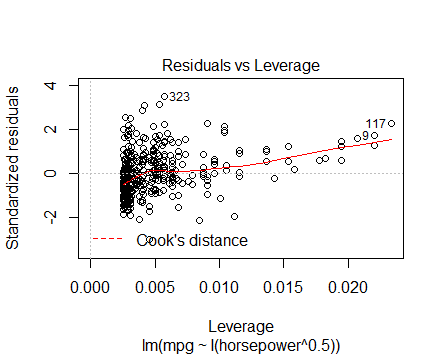


This graph shows a non linear relationship between the predictor and the response variables.

This graph suggests that the constant variance of error assumption is false as there is a funnel shape detected in the graph of increasing residuals towards the right.



As there is no linear line, it has a non constant variance.



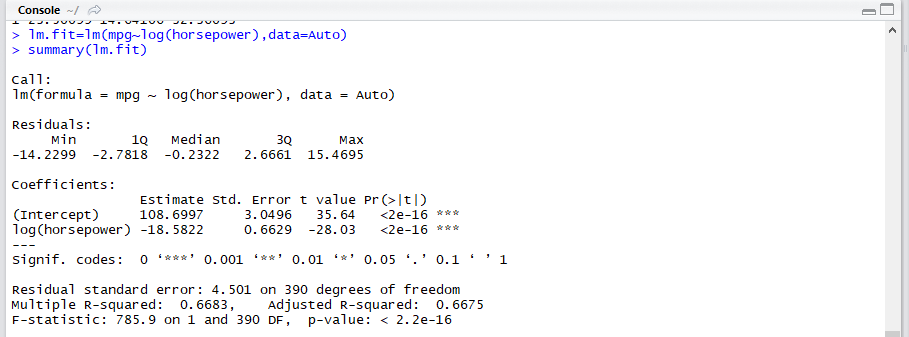
There are many leverage points with h value h>>(p+1)/no. of observations = (1+1)/392=0.005.

Assuming >> as a factor of 3, leverage above 0.015 are high leverage points like 117,94 and a few others.

There are outliers points such that the standardized residual is out of the +/3 range like 323

**FOR Log(x)**

a)



1. yes, there is a clear evidence of a relationship between the predictor and the response variable, as the p-value corresponding to the F- statistic almost close to 0 at significant levels.
2. R^2 value is 66.83% which suggests that 66.83% of the variability in mpg can be explained by the predictor variables.
3. As the coefficient of log(horsepower) is negative, the relationship of mpg is negative with it.
4. With an increase of 1 log(horsepower), the mpg goes gown by 18.582 units. Hence the fuel efficiency decreases with an increase in horsepower.

v) 23.05

> predict(lm.fit,data.frame(horsepower=98),interval="confidence")

fit lwr upr

1 23.50099 23.05405 23.94794

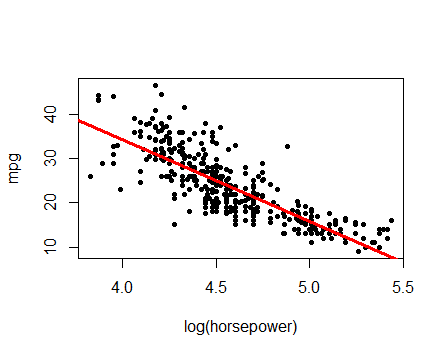
> predict(lm.fit,data.frame(horsepower=98),interval="prediction")

fit lwr upr

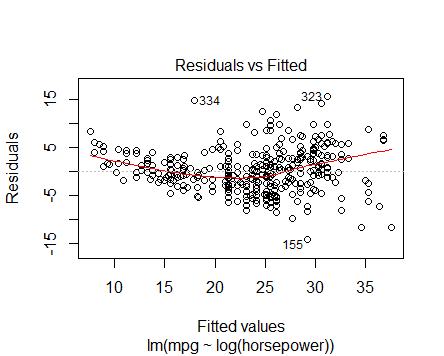
1 23.50099 14.64106 32.36093

b) {plot(log(horsepower),mpg,pch=20,col="black")

abline(lm.fit,lwd=3,col='red')}

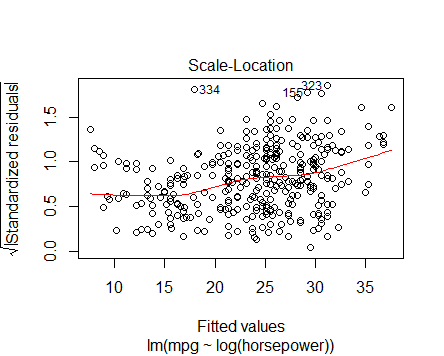


c)

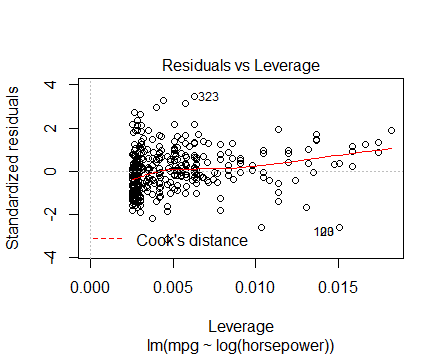


This graph shows a linear relationship between the predictor and the response variables.

This graph suggests that the constant variance of error assumption is false as there is a funnel shape detected in the graph of increasing residuals towards the right.



As there is no straight line, there is a non constant variance



There are outliers such that the standardized residual value >3 like 323. There are also some high leverage points whose value >> (p+1)/3. Here we take points greater than 0.015.

Findings:

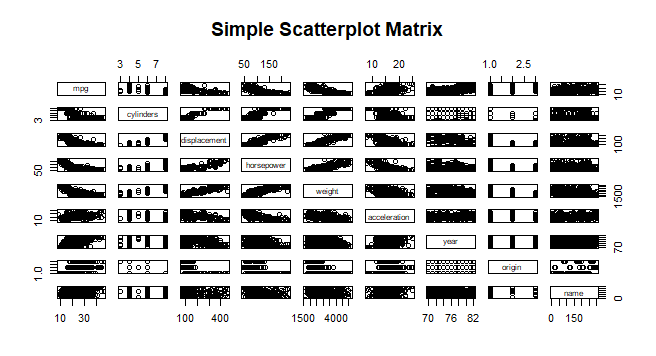
The Log transformation helped reduce the non linearity and give a more linear fit. Also, it can be found that the R^2 value increased in the Log(X) transformation model compared to the model without any transformation.

The other transformations did not make the non linearity any better compared to the model where there was no transformation.

For this model, log transformation seems like a good transformation for avoiding non linearity between predictors and response variables.

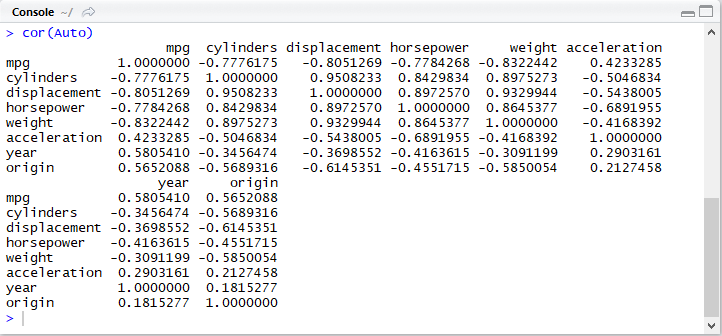
2)

a)

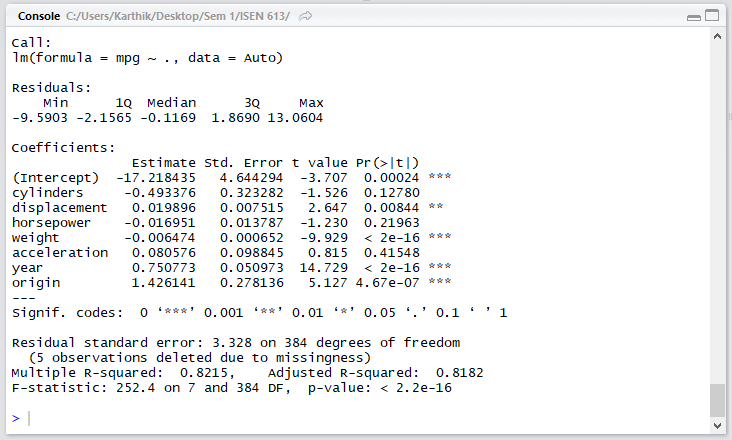


Assuming, mpg as the response, there seems to be a negative relationship from predictors like cylinders, displacement, horsepower, weight and a positive relationship from predictors like acceleration and year. Origin and name do not seem to have an effect from the very look of the scatter plot above. Also the relationships are weaker relative to other predictors on mpg from acceleration and year .

b)



c)



1. As the F-statistic (252.4) is way greater than 1 and the overall p value is close to 0 almost, there is a relationship of at least one predictor to the response.
2. There is a relation between the predictor and the response if the p value is almost near 0 or less than the threshold p value significantly.

Hence:

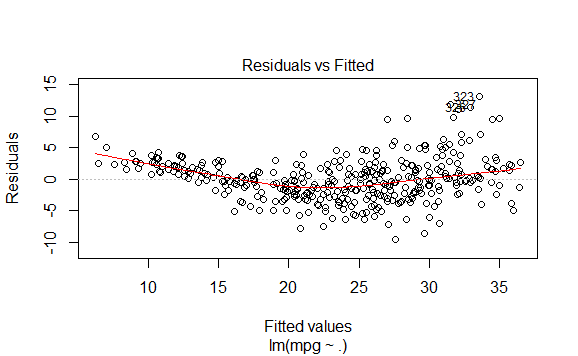
Response mpg is related to : Displacement at 0.001 significance

Weight, year, origin at near 0 significance

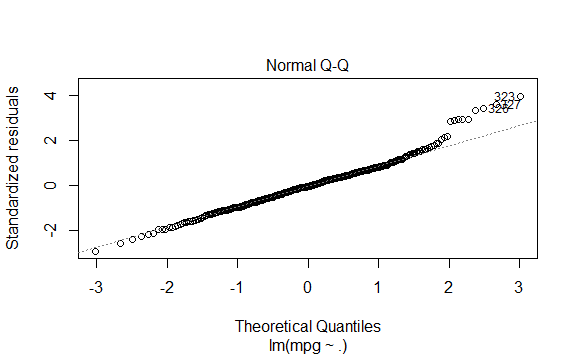
1. The coefficient of year variable is positive and corresponding p value is significant at near 0 (\*\*\*), hence there is a positive relationship between year and the mpg. With an increase of 1 year or from one year to the immediate next year keeping all other variables constant, there is an increase of 0.75 units in mpg.

This also means that the fuel efficiency of the car increases each year.

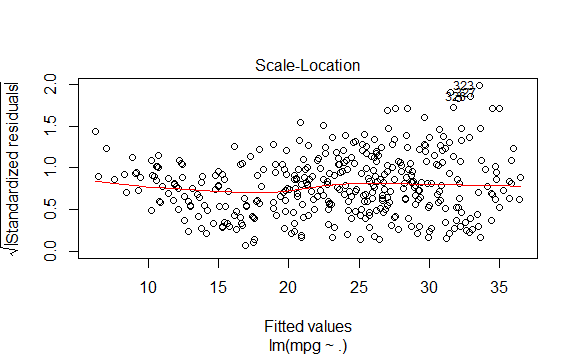
d)



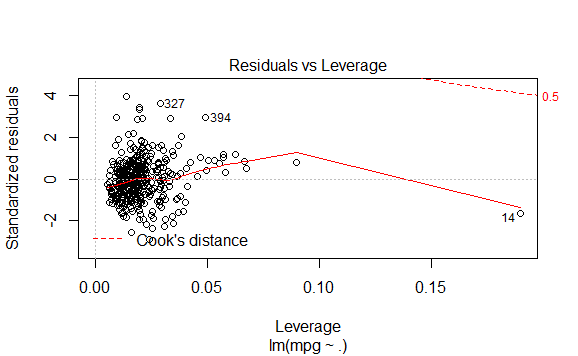
It can be found in this graph that there is a non linear relationship between the predictor and the response variables.



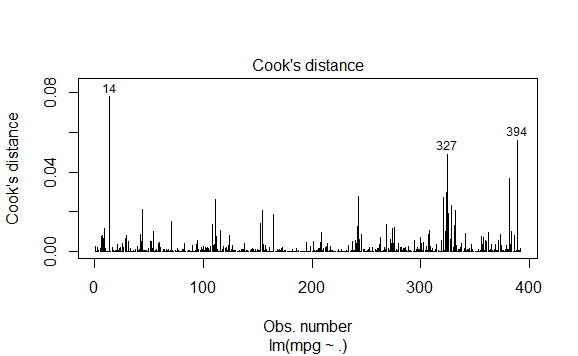
This graph shows that the residuals are normally distributed as is assumed in the linear regression models, but seems to be skewed on the right.



This plot suggests that the model does not have a constant variance of error as the standardized error/ residuals are following a funnel shape. Hence there is heteroskedacity in the model.



This graph shows that there is hardly one leverage point (14th observation) such that h>>(p+1)/n = (7+1)/392 =0.02. Assuming “>>” as a factor of 3, h>0.06 points are high leverage points. This also suggests a few outliers above and below +/- 3.



According to cook’s distance, point 14 seem to be the outlier along with border line outliers like observation 327 and 394.

e)

> install.packages("VIF")

> library(car)

> vif(lm.fit3)

cylinders displacement horsepower weight acceleration year

10.737535 21.836792 9.943693 10.831260 2.625806 1.244952

origin

1.772386

If VIF=1, there is absolutely no collinearity. It is difficult to obtain VIF=1 in real world as there is some collinearity in data.

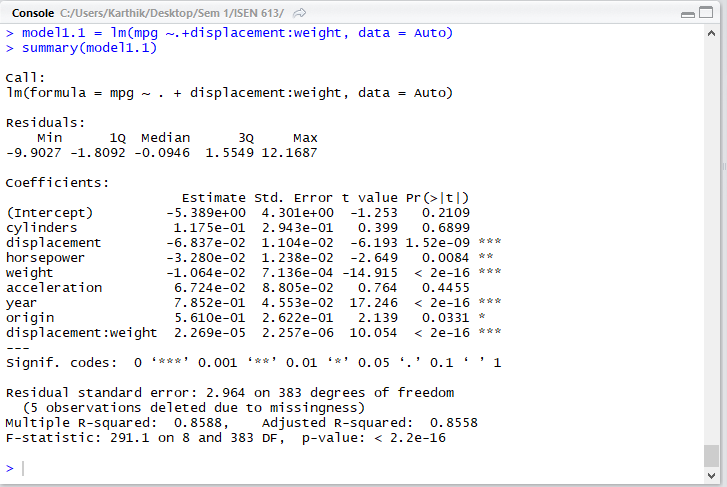
Hence, as a rule of thumb, a VIF that is more than 5 or 10 is considered problematic.

Here, cylinders, displacement, horsepower, weight have high collinearity as their VIF is almost 10 and above.

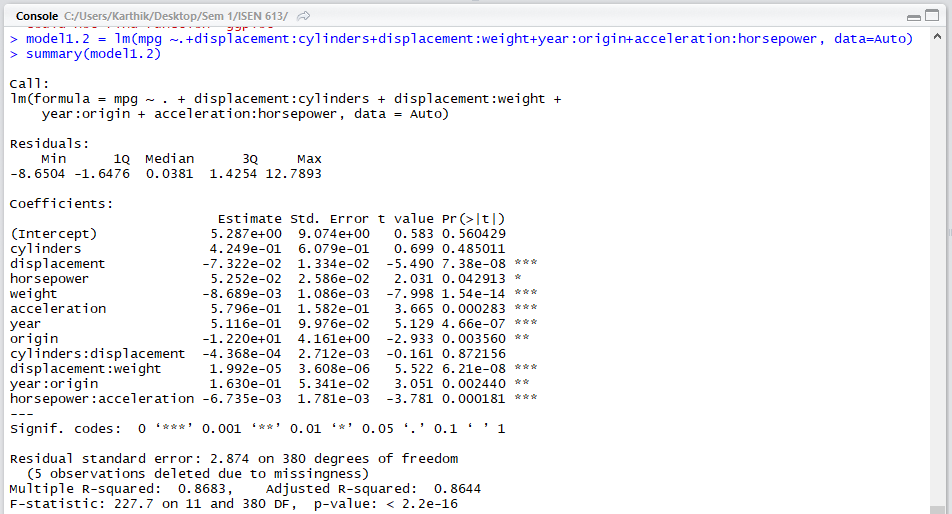
Whereas, acceleration, year, origin do not have high collinearity that affects the mpg response.

f)

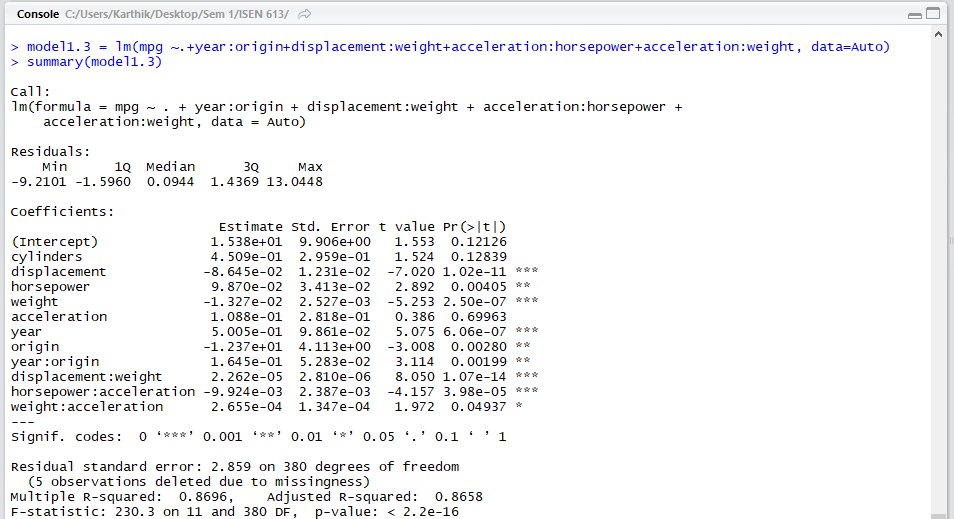
Model 1.1:



Model 1.2:



Model 1.3:



Fitting models with interaction terms one by one progressively to check if each addition of interaction has any effect on the improvement of the R^2 values along with being significant wrt p values and the model.

Finally using the domain knowledge of cars and the limited experimentation of the combination of design of experiments of the different interactions (here at max only 2 variable interaction is considered).

Hence, There is interaction between:

year:origin

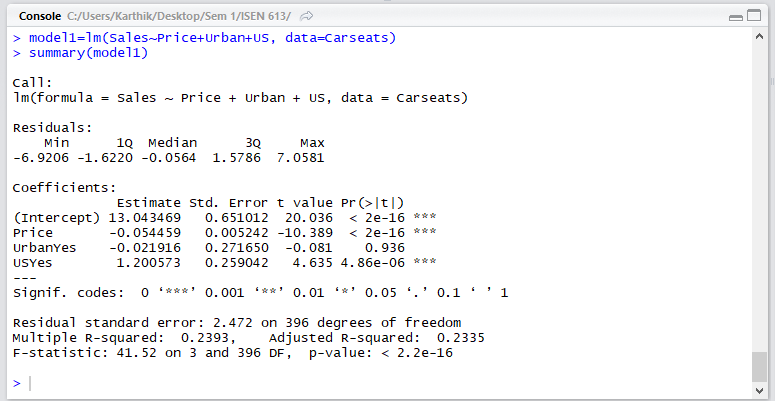
displacement:weight

acceleration:horsepower

acceleration:weight

Also, the predictors, acceleration and cylinders seem to not have significane on the response variable mpg.

3)



1. Coefficient of Price suggests that with 1 dollar increase in the price, the Sales ( in thousands) goes down by 0.054\*1000=54 .when all other predictors remain a constant.

- Coefficient of UrbanYes / Urban predictor suggests that , the unit sales in urban location is 21.9 units “less” on an “Average” than in a rural location (UrbanNo) when all other predictors remain a constant.

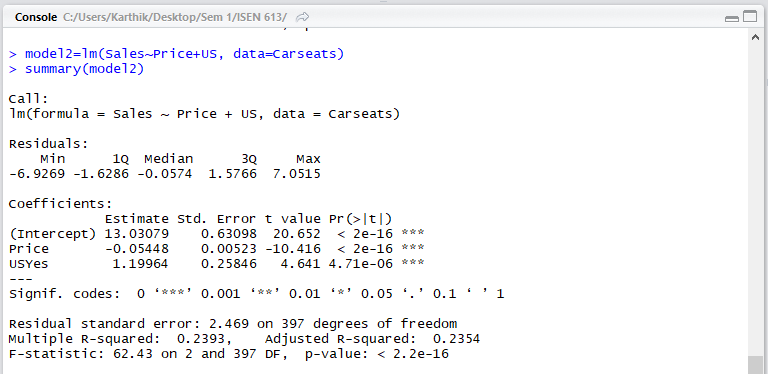
-Coefficient of US predicter suggests that the sales in US is on an Average 1200.57 units “more” than the sales in non-US stores when all other predictors remain a constant.

1. Sales (in thousands) = 13.0434 – 0.054(Price) – 0.0219(Urban) + 1.2(US)

Urban= 1 if Yes, else, 0 for Rural

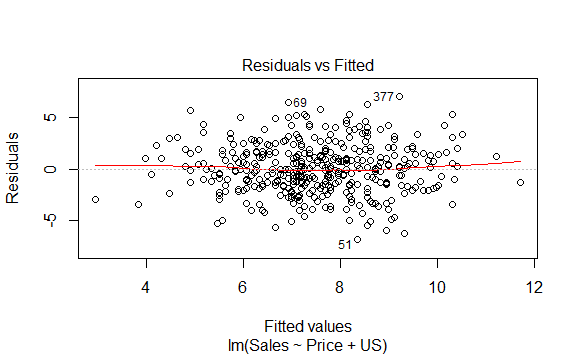
US=1 if Yes, else no for Non-US

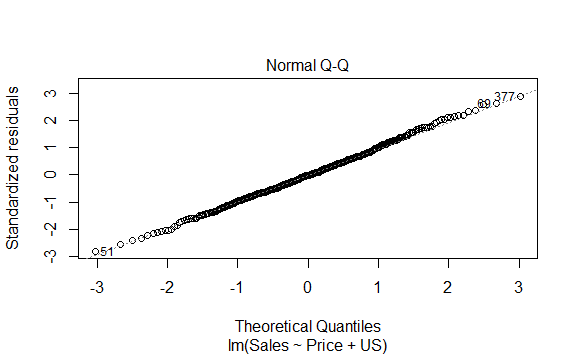
1. We can reject the null hypothesis 𝐻0: 𝛽𝑗=0 for predictors Price and US as they are significant at \*\*\* level.

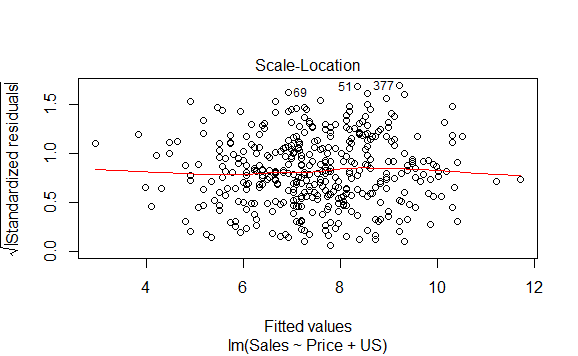


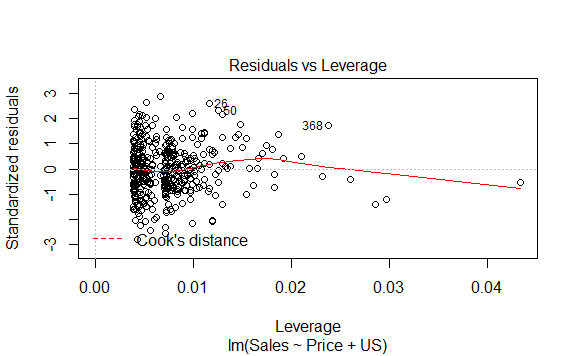
1. The model with lesser number of variables, with only Price and US predictors has a slightly better Adjusted R^2 value than the model which includes Urban predictor variable also. The adjusted R^2 is higher because an insignificant predictor is removed from the bigger model.

But both the models pretty much fit the model similarly with 23.93%variability being explained.

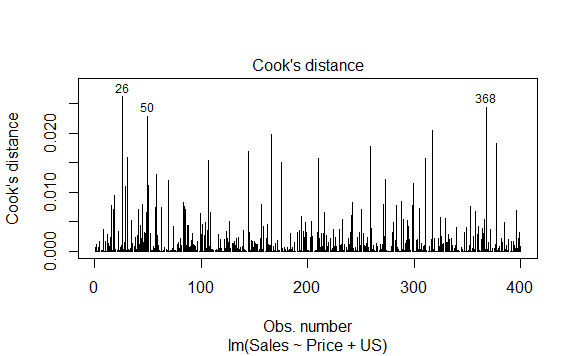








There are a few high leverage point such that h(Leverage) >> (2+1)/400=0.0075 “>>” as a factor of 3 => points outside 0.0225 leverage are leverage points (around 4-5 points)



From the above plot, it can be seen that there is an outlier at observations number 26,50, 368 and a few more.