Prediction Of Car Price

Aniruddha Mukherjee

CONTENTS

DESCRIPTION OF THE DATASET:

I have a data on price of 186 cars and some others features of the car namely wheel base, length ,width, height, curb weight, engine size, bore, stroke ,compression ratio, horsepower, peak rpm, city mpg, highway mpg.

Let us first rename the variables
Y= Price
x1=Wheelbase
x2=length
x3=width
x4=height
x5=curb weight

x6=engine size
x7=bore
x8=stroke
x9=compression ratio
x10=horse power
x11=peak rpm
x12=city mpg
x13=high way mpg

```
D=read.csv("dataset.csv")
View(D)
attach(D)
summary(D)

##    wheel_base          length           width            height
##   Min.    :86.60     Min.    :141.1    Min.    :60.30    Min.    :47.80
##   1stQu.:  94.50     1stQu.:166.3      1stQu.:64.00      1stQu.:52.00
##   Median:  96.90     Median:173.0      Median:65.40      Median:54.10
##   Mean    :  98.56   Mean    :173.6    Mean    :65.78    Mean    :53.81
##   3rdQu.:101.20      3rdQu.:181.7      3rdQu.:66.50      3rdQu.:55.50
##   Max.    :115.60    Max.    :202.6    Max.    :71.70    Max.    :59.80
##    curb_weight      engine_size         bore             stroke
##   Min.    :1488     Min.    :61.0     Min.    :2.680    Min.    :2.19
##   1stQu.:2128       1stQu.:98.0       1stQu.:3.150      1stQu.:3.10
##   Median:2405       Median:110.0      Median:3.310      Median:3.29
##   Mean    :2533     Mean    :124.8    Mean    :3.322    Mean    :3.25
##   3rdQu.:2921       3rdQu.:141.0      3rdQu.:3.580      3rdQu.:3.41
##   Max.    :4066     Max.    :326.0    Max.    :3.940    Max.    :4.17
##   compression_ratio   horsepower       peak_rpm        city_mpg
##   Min.    :7.00        Min.    :48.0    Min.    :4150    Min.    :13.00
##   1stQu.:  8.60        1stQu.:  70.0    1stQu.:4800      1stQu.:21.00
##   Median:  9.00        Median:  94.0    Median:5100      Median:25.00
##   Mean    :10.15       Mean    :101.4   Mean    :5106    Mean    :25.63
##   3rdQu.:9.40          3rdQu.:116.0     3rdQu.:5500      3rdQu.:30.00
##   Max.    :23.00       Max.    :262.0   Max.    :6600    Max.    :49.00
##    highway_mpg         price
##   Min.    :17.00     Min.    :5118
##   1stQu.:25.00       1stQu.:7609
##   Median:30.00       Median:9988
##   Mean    :31.09     Mean    :12524
##   3rdQu.:36.00       3rdQu.:15998
##   Max.    :54.00     Max.    :37028
```

Here we get the summary of the original dataset .I start working on the newly named data set.

```
D=read.csv("dataset.csv")
View(D)
attach(D)

##The following object is masked from D(pos=3): ##
##      price

summary(D)


##          x1                x2              x3              x4
##   Min.    : 86.60    Min.    :141.1    Min.    :60.30    Min.    :47.80
##   1stQu.:   94.50    1stQu.:166.3    1stQu.:64.00    1stQu.:52.00
##   Median:   96.90    Median:173.0    Median:65.40    Median:54.10
##   Mean    : 98.56    Mean    :173.6    Mean    :65.78    Mean    :53.81
##   3rdQu.:101.20    3rdQu.:181.7    3rdQu.:66.50    3rdQu.:55.50
##   Max.    :115.60    Max.    :202.6    Max.    :71.70    Max.    :59.80
##          x5              x6              x7              x8
##   Min.    :1488    Min.    :61.0     Min.    :2.680    Min.    :2.19
##   1stQu.:2128    1stQu.:98.0     1stQu.:3.150    1stQu.:3.10
##   Median:2405    Median:110.0    Median:3.310    Median:3.29
##   Mean    :2533    Mean    :124.8    Mean    :3.322    Mean    :3.25
##   3rdQu.:2921    3rdQu.:141.0    3rdQu.:3.580    3rdQu.:3.41
##   Max.    :4066    Max.    :326.0    Max.    :3.940    Max.    :4.17
##          x9              x10             x11             x12
##   Min.    :7.00     Min.    :48.0     Min.    :4150    Min.    :13.00
##   1stQu.:8.60     1stQu.:70.0     1stQu.:4800    1stQu.:21.00
##   Median:9.00     Median:94.0     Median:5100    Median:25.00
##   Mean    :10.15    Mean    :101.4    Mean    :5106    Mean    :25.63
##   3rdQu.:9.40     3rdQu.:116.0    3rdQu.:5500    3rdQu.:30.00
##   Max.    :23.00    Max.    :262.0    Max.    :6600    Max.    :49.00
##         x13             price
##   Min.    :17.00    Min.    :5118
##   1stQu.:25.00    1stQu.:7609
##   Median:30.00    Median:9988
##   Mean    :31.09    Mean    :12524
##   3rdQu.:36.00    3rdQu.:15998
##   Max.    :54.00    Max.    :37028
```

Let's first fit a linear model to the data set and then examine the quality of the fit.

The linear model we want to fit is $y=\beta_0+\beta_1 x1+\beta_2 x2+\ldots+\beta_{13} x13+\epsilon$

Where the basic assumptions are

1. $\epsilon \sim N_{13}(0,\sigma^2 I_{13})$

and

2. $X=[x1,x2,\ldots,x13]$ is the data matrix having rank 14.

```
fit=lm(price~highway_mpg+city_mpg+peak_rpm+horsepower+compression_ratio+stroke          +  bo
summary(fit)

##
##Call:
##lm(formula=price~highway_mpg+city_mpg+peak_rpm+horsepower+
         compression_ratio+stroke+bore+engine_size+curb_weight+
         height+width+length+wheel_base)
##
##Residuals:
##        Min         1QMedian            3Q
         Max##-8841.3-1494.4-221.71433.910815.2
##
##Coefficients:
##                         EstimateStd.ErrortvaluePr(>|t|)
##(Intercept)            -4.972e+04   1.559e+04   -3.190    0.00169**
##highway_mpg             2.247e+02   1.572e+02    1.429    0.15476
##city_mpg               -2.698e+02   1.761e+02   -1.532    0.12732
##peak_rpm                1.690e+00   6.250e-01    2.703    0.00756**
##horsepower              3.994e+01   1.696e+01    2.356    0.01962*
##compression_ratio       2.619e+02   8.117e+01    3.226    0.00150**
##stroke                 -3.202e+03   7.847e+02   -4.0806.89e-05***
##bore                   -1.624e+03   1.233e+03   -1.317    0.18961
##engine_size             9.721e+01   1.698e+01    5.7254.54e-08***
##curb_weight             3.151e+00   1.684e+00    1.871    0.06305 .
##height                  2.438e+02   1.336e+02    1.824    0.06987 .
##width                   6.638e+02   2.417e+02    2.747    0.00667**
##length                 -8.223e+01   5.533e+01   -1.486    0.13907
##wheel_base              2.392e-01   1.020e+02    0.002    0.99813
##---
##Signif.codes:     0'***'0.001'**'0.01'*'0.05'.'0.1''1 ##
##Residualstandarderror:2894on171degreesoffreedom
##MultipleR-squared:      0.8401,AdjustedR-squared:      0.828
##F-statistic:69.12on13and171DF,             p-value:<2.2e-16
```

Comments about the fit

1. After fitting the model we see that 82 percent of the total variability is explained by the linear regression of y on x1, x2, x3...x13.

2. If all the features of the car is zero then the car has no price so we should not take a intercept model here intuitively and that is also very evident from the p value corresponding to the intercept term. Here the p value is greater than 0.05 so clearly the hypothesis of intercept is zero is not being rejected .So we should consider a non intercept model.

3. And now seeing the other p value we conclude that the variables Highway mpg, city_mpg , bore ,length and wheel base are not important to estimate the

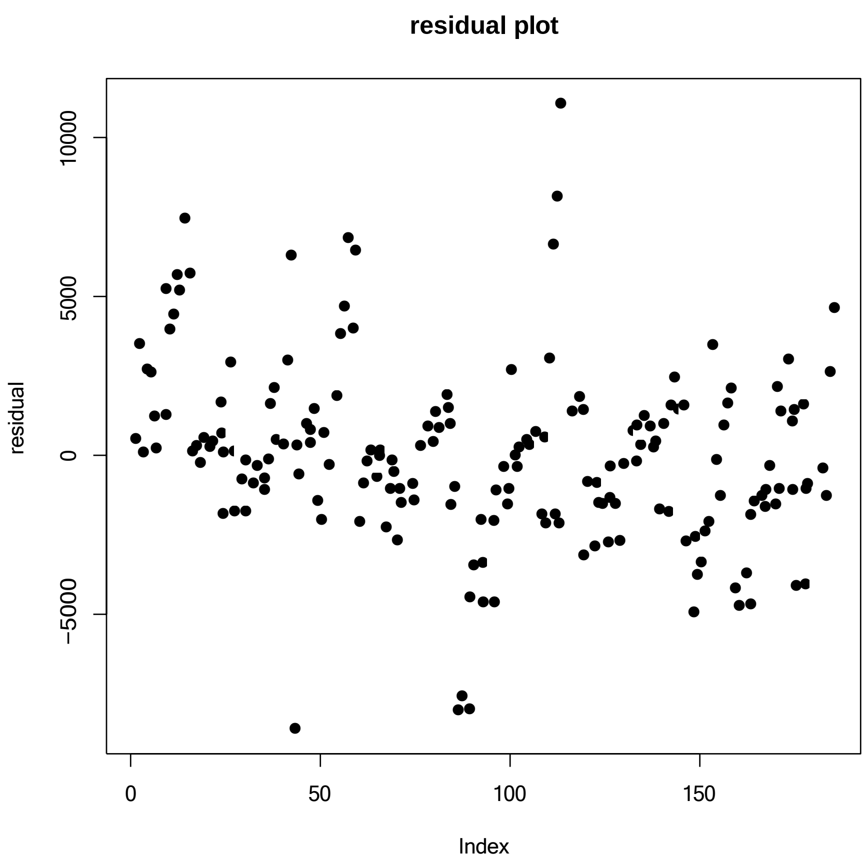
Price of the car.

So we eliminate these variables and try to fit the model.

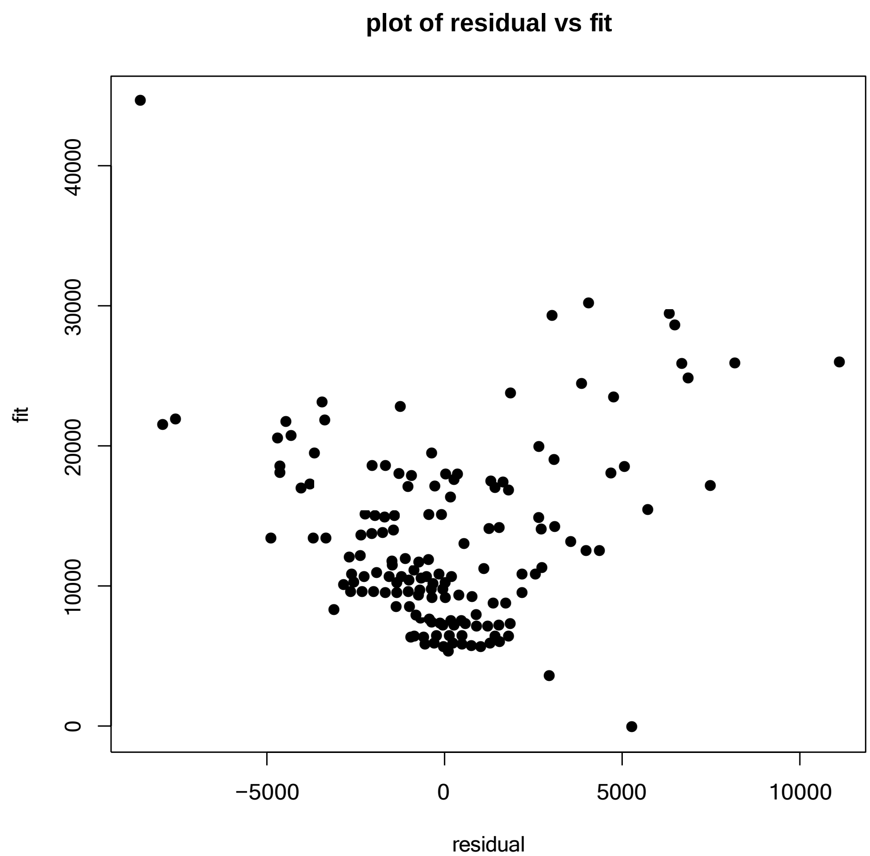
```
D=read.csv("dataset.csv")
attach(D)

##The following object is masked from D(pos=3): ##
##      price
##The following objects arem asked fromD(pos=4): ##
##      compression ratio, curb weight,enginesize,height, ##
        horsepower,peak rpm,price,stroke,width

fit=lm(price~peak_rpm+horsepower+compression_ratio+stroke+

engine_size+curb_weight)

summary(fit)

##
##Call:
##lm(formula=price~peak_rpm+horsepower+compression_ratio+##
         stroke+engine_size+curb_weight+height+width)
##
##Residuals:
##      Min      1QMedian        3Q      Max
##   -8603    -1542      -88      1377    11067
##
## Coefficients:
##                      Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)        -5.229e+04   1.328e+04   -3.939  0.000118 ***
## peak_rpm            1.955e+00   6.003e-01    3.256  0.001355 **
## horsepower          4.207e+01   1.489e+01    2.825  0.005279 **
## compression_ratio   2.554e+02   6.643e+01    3.845  0.000168 ***
## stroke             -2.837e+03   7.411e+02   -3.828  0.000179 ***
## engine_size         1.002e+02   1.674e+01    5.983   1.2e-08 ***
## curb_weight         1.536e+00   1.338e+00    1.148  0.252489
## height              1.528e+02   1.176e+02    1.299  0.195575
## width               4.953e+02   2.066e+02    2.397  0.017567 *
## ---
##Signif. codes:   0'***'0.001'**'0.01'*'0.05'.'0.1''1
##
##Residual standard error:2910on176degreesoffreedom
##Multiple R-squared:     0.8336,AdjustedR-squared:     0.826
##F-statistic:110.2on8and176DF,           p-value:<2.2e-16
```

PLOT OF FRESIDUAL AND RESIDUAL VS FIT

```
y=fitted(fit)
e=residuals(fit)
plot(e,main="residualplot",ylab="residual")
```




```
plot(e,y,xlab="residual",ylab="fit",main="plot of residual vs fit")
```

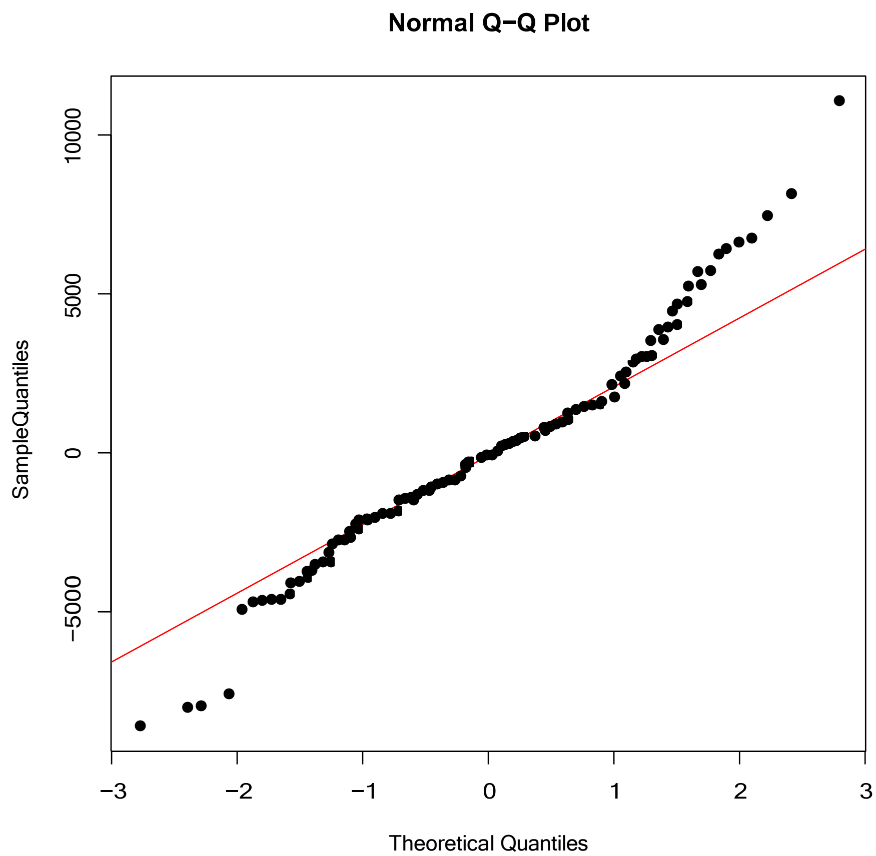


Residual plot is not uniformly scattered around zero and also they are not in certain bandwidth and residual vs fit plot is not also showing a random pattern so we can suspect that all the assumption of the regression is not being satisfied. That is probably the error mean is not zero and the the homoscedasticity assumption is violated.

NORMALITY OF ERRORS

Now we plot the qqplot of the errors and perform Shapiro Wilks normality test to check the normality of the errors

```
e=residuals(fit)
qqnorm(e)
par(new=T)
qqline(e,col="red")
```




```
shapiro.test(e)

##
##   Shapiro-Wilk normality test
##
##data:     e
##W=0.96143,p-value=5.778e-05
```

qq plot suggests that all the points are not really near the qq line so the data in not coming from a normal distribution.

And Shapiro Wilks test p value is sufficiently small so we reject the hypothesis of normality of the data set... the data set is not really coming from a normal distribution.

MULTICOLLINEARITY:

Now we check whether in the data mear multicollinearity is present or not.

```
library(car)

##Warning:   package 'car' was built under Rversion 3.4.4
##Loadingrequiredpackage:        carData
##Warning:   package' car Data' was built under Rversion 3.4.4

vif(fit)

##            peak_rpm          horsepower   compression_ratio              stroke
##            1.716666            6.368360            1.494792            1.132973
##         engine_size          curb_weight              height               width
##            8.172037           10.130561            1.756628            3.900127
```

We see that horsepower , engine size and curb weight has variance influence factor 6, 8, 10 respectively so we conclude that these variables are responsible for multicollinearity. That is these variables are dependent on the another variables. So that rank of the design matrix is not full column rank..so one of the assumption of the Gauss Markoff setup is not being satisfied.

INDEPENDENCEOFERRORS:

Now we check whether the errors are really independent or not..we do it by performing Durbin Watson test.

```
library(car)
durbinWatsonTest(fit,alternative="two.sided")

##   lagAutocorrelationD-WStatisticp-value
##     1          0.5682531        0.8488679          0
##   Alternativehypothesis:rho!=0
```

Here we have tested H_0: ρ =0 ag H_1: $\rho \neq$0 Where ρ is the auto correlation. P value for the test is zero so we reject the null hypothesis. So autocorrelation is present in the data set. SO really the errors are not independently distributed Violating another assumption.

HETEROSCADASTICITY:

Now we check whether the data is really homoscedastic or not by Breusch-Pegan test.

```
library(lmtest)

##Warning:    package 'lmtest' was built under Rversion 3.4.4
##Loading required package:       zoo
##Warning:    package 'zoo' was built under Rversion 3.4.4
##
##Attachingpackage:      'zoo'
##The following objects are masked from 'package:base': ##
##      as.Date, as.Date.numeric

bptest(fit)

##
##   studentized Breusch-Pagan test
##
##data:   fit
##BP= 106.52, df = 8, p-value< 2.2e-16
```

This test have a p-value less that a significance level of 0.05, therefore we can reject the null hypothesis that the variance of the residuals is constant and infer that heteroscedasticity is indeed present, thereby confirming our graphical inference.

OUTLIARS AND INFLUENTIAL OBSERVATION:

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the dataset. Lever- age Points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation. An Influential Observation is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation. In particular, in regression analysis an influential point is one whose deletion has a large effect on the parameter estimates. We employ the following method to detect the presence of the above observations.

Covariance Ratio: Covariance ratio is a measure for detecting high leverage points and outliers. First, we detect the influential points by the hat diagonals and covariance ratios, and obtain the following detected points:

```
p=8
n=186
h<-hatvalues(fit)
outlier<-which(h>2*p/n)
COVARIANCE.RATIO<-covratio(fit)
cov.outlier<-which(abs(COVARIANCE.RATIO-1)>3*p/n)
influential.points<-sort(unique(c(outlier,cov.outlier)))
influential.points

##  [1]   1   7   8   9  14  32  39  41  42  43  57  58  59  60  75  86  87
##[18]  88  96  98 111 112 113 117 136 154 168 179 180
```

1, 7, 8, 9, 14, 32, 39, 41, 42, 43, 57, 58, 59, 60, 75, 86, 87, 88, 96, 98, 111, 112, 113, 117, 136, 154, 168, 179, 180 th data points are unusual observation they are coming from different distribution.

DEALING WITH THE UNUSUAL OBSERVATION:

Now we remove all the unusual observation and again fit the model and then check whether the model is better than the previous model or not.

```
D=read.csv("C:\\Users\\HP\\Desktop\\Kcsirproject\\md.csv") View(D)
D=as.matrix(D)
v=D[c(1,7,8,9,14,32,39,41,42,43,57,58,59,60,75,86,87,88,96,
111,112,113,117,136,154,168,179,180]
View(v)
```

v is the new data set after removing the unusual observations.

FITTING THE MODEL :

```
v=as.data.frame(v)
attach(v)

##The following objects are masked from D(pos=7): ##
##      compression_ratio, curb_weight, engine_size, height,
##      horsepower, peak_rpm,price, stroke, width
##The following object is masked from D(pos=8): ##
##      price
##The following objects are masked from D(pos=9): ##
##      compression_ratio,curb_weight,engine_size,height,
##      horsepower,peak_rpm,price,stroke,width

fit1=lm(price~peak_rpm+horse_power+compression_ratio+stroke+

engine_size+curb_weight+height+width)

summary(fit1)

##
##Call:
##lm(formula=price~peak_rpm+horsepower+compression_ratio+
          stroke+engine_size+curb_weight+height+width)
```

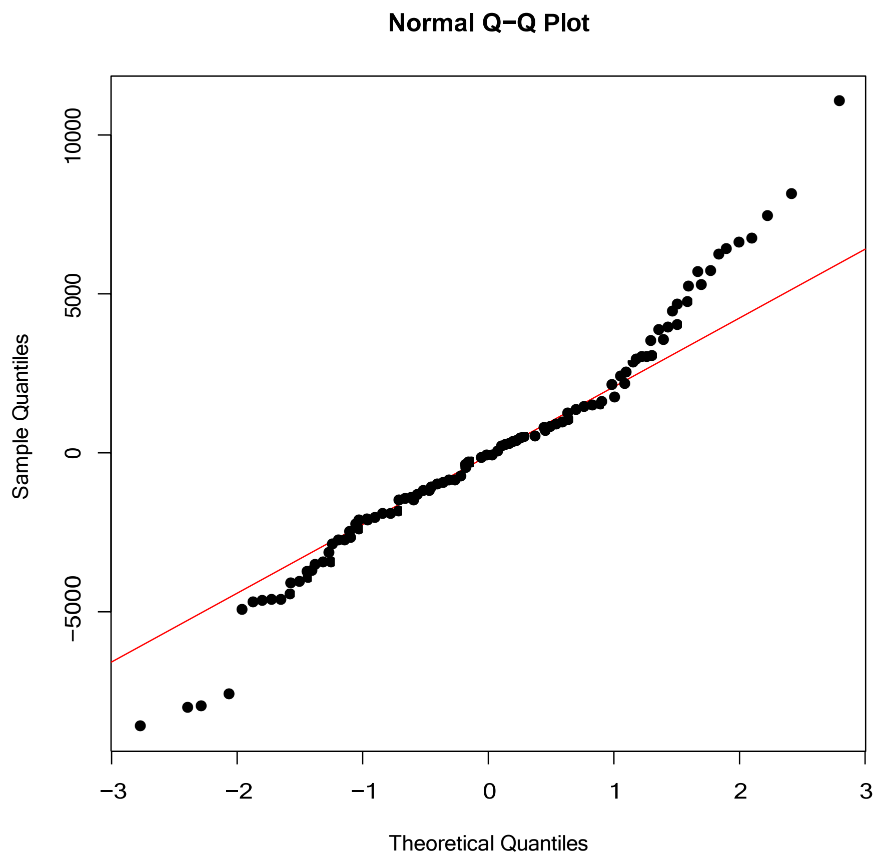
+ engine_size+curb_weigh

```
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -4687.6-1358.8   -109.7    911.7   6713.5
##
## Coefficients:
##                     EstimateStd.Error  tvalue  Pr(>|t|)
## (Intercept)        -3.798e+04  1.391e+04  -2.731 0.007092 **
## peak_rpm            1.477e+00  5.224e-01   2.828 0.005344 **
## horsepower          3.164e+01  1.781e+01   1.777 0.077650 .
## compression_ratio   2.202e+02  6.468e+01   3.405 0.000854 ***
## stroke             -2.293e+03  7.078e+02  -3.240 0.001478 **
## engine_size         6.541e+01  2.189e+01   2.988 0.003292 **
## curb_weight         3.486e+00  1.458e+00   2.392 0.018038 *
## height              9.719e+01  1.025e+02   0.948 0.344670
## width               3.397e+02  2.272e+02   1.495 0.136973
## ---
## Signif.codes:   0'***'0.001'**'0.01'*'0.05'.'0.1       ''1
##
## Residualstandarderror:2148on147degreesoffreedom
## MultipleR-squared:    0.8166,AdjustedR-squared:    0.8067
## F-statistic:81.84on8and147DF,          p-value:<2.2e-16
```

After removing the unusual observations also the fit is not that good and the variables are not also very explanatory for the variable price. Its clear from the adjusted r square value and the p values for the coefficients.

CHECKING THE NORMALITY

```
e=residuals(fit)
qqnorm(e)
par(new=T)
qqline(e,col="red")
```




```
shapiro.test(e)

##
##   Shapiro-Wilk normality test
##
##data:      e
##W=0.96143,p-value=5.778e-05
```

Even after removing the unusual observation also we don’t have normality of the dataset.

MODEL SELECTION:

We have seen that using all the variables also it is insufficient to explain the variability of price properly. So no subsets of these variables will be able to do that. So we don't check the AIC and Mallos cp for this case.

Conclusion:

In this dataset we see that none of the Gauss Markoff assumption holds good and the dataset contain many unusual observation. Even after removing the unusual observation we dont see any normality of the data set and the fitted model is also not that good. So we conclude that all these variables are unable to explain the variability of the variable price. And also we should not apply linear model to study the behaviour of this data. Rather we should think for some non linear model for explaining the dataset.