MLRAMAN

Contents

[Question 1 : 2](#_Toc19409962)

[BOXPLOT: 2](#_Toc19409963)

[Correlation 3](#_Toc19409964)

[PAIRS PLOT 3](#_Toc19409965)

[FITTING REGRESSION MODEL: 4](#_Toc19409966)

[Model 1: 4](#_Toc19409967)

[Model 2: 5](#_Toc19409968)

[Model 3: 5](#_Toc19409969)

[TABULATION: 6](#_Toc19409970)

[Conclusion: 6](#_Toc19409971)

[Question 2 : 7](#_Toc19409972)

[BOXPLOT: 7](#_Toc19409973)

[CORRELATION: 8](#_Toc19409974)

[PAIRS PLOT: 9](#_Toc19409975)

[FITTING REGRESSION MODEL: 9](#_Toc19409976)

[Model 1: 9](#_Toc19409977)

[Model 2: 10](#_Toc19409978)

[Model 3: 10](#_Toc19409979)

[Model 4: 11](#_Toc19409980)

[TABULATION: 12](#_Toc19409981)

[Conclusion: 12](#_Toc19409982)

[Question 3 : 13](#_Toc19409983)

[Summary: 13](#_Toc19409984)

[BOXPLOT: 13](#_Toc19409985)

[CORRELATION: 14](#_Toc19409986)

[FITTING REGRESSION MODEL: 15](#_Toc19409987)

[Model 1: 15](#_Toc19409988)

[Model 2: 17](#_Toc19409989)

[Tabulation 17](#_Toc19409990)

[CONCLUSION: 17](#_Toc19409991)

# Question 1 :

Prepare a prediction model for profit of 50\_startups data. Do transformations for getting better predictions of profit and make a table containing R² value for each prepared model.

**Answer:**

Variables available: “R.D.Spend”, Administration, “Marketing.Spend” , State Profit

Target Variable: Profit

Let’s have a look on the Summary Statistics of the Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| R.D.Spend | Administration | Marketing.Spend | State | Profit |
| Min.: 0 | Min.: 51283 | Min.: 0 | California:17 | Min.: 14681 |
| 1st Qu.: 39936 | 1st Qu.:103731 | 1st Qu.:129300 | Florida :16 | 1st Qu.: 90139 |
| Median: 73051 | Median :122700 | Median :212716 | New York :17 | Median :107978 |
| Mean.: 73722 | Mean :121345 | Mean :211025 |  | Mean :112013 |
| 3rd Qu.:101603 | 3rd Qu.:144842 | 3rd Qu.:299469 |  | 3rd Qu.:139766 |
| Max.:165349 | Max. :182646 | Max. :471784 |  | Max. :192262 |

Here we have 4 variables (including the Target Variable) of Continuous type. & one variable of Categorical type

i.e. “State”

## BOXPLOT:

Looking at the Boxplot we can say that no variables contains outliers.

Now lets move to the Correlation analysis between the continuous variables present in the data.

## Correlation

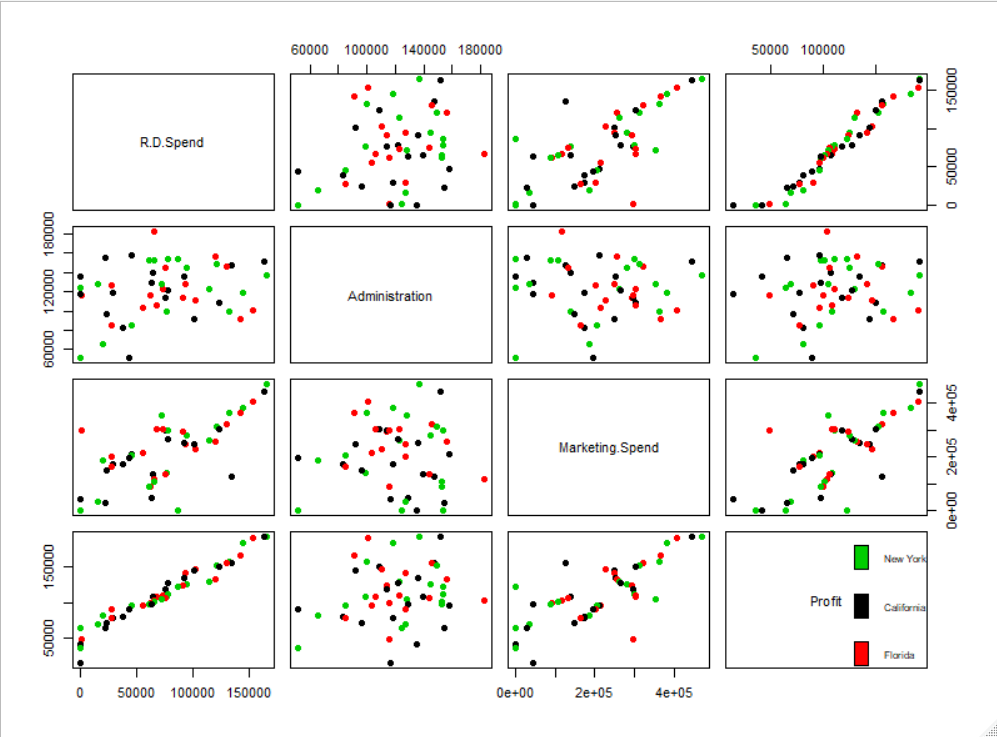
Let’s have a look on the correlation coefficient between the variables:



Seems like Except the pair (“Administrative” and “Marketing Spend”) all the variables are positively correlated with each other.

And there may be no collinearity problem in our Independent variables .

## PAIRS PLOT

Let’s Have a look on the Pairs Plot for visualizing the Scatterings of the data among themselves.

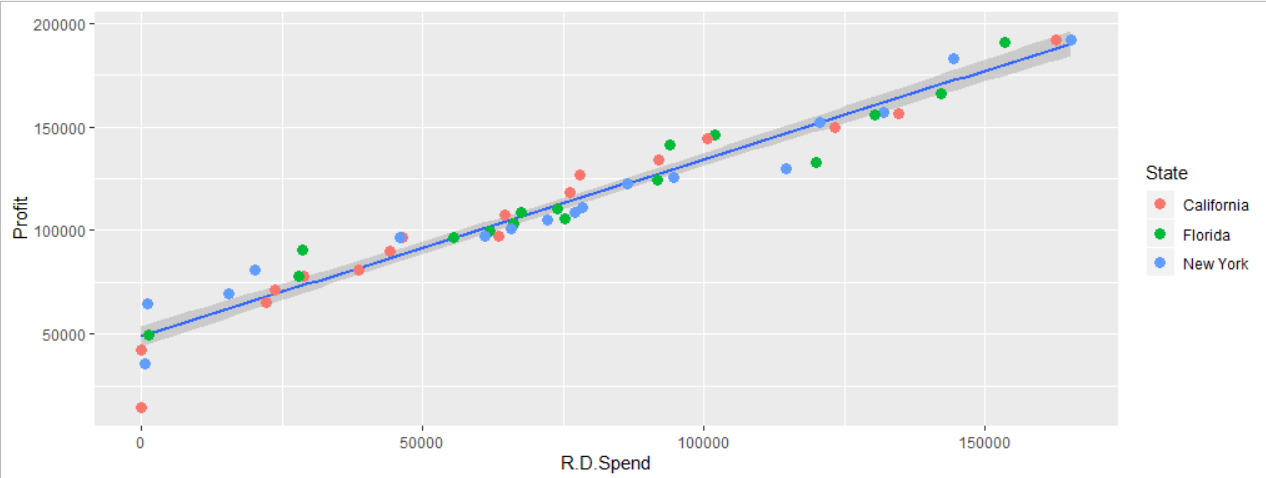
## FITTING REGRESSION MODEL:

### Model 1:

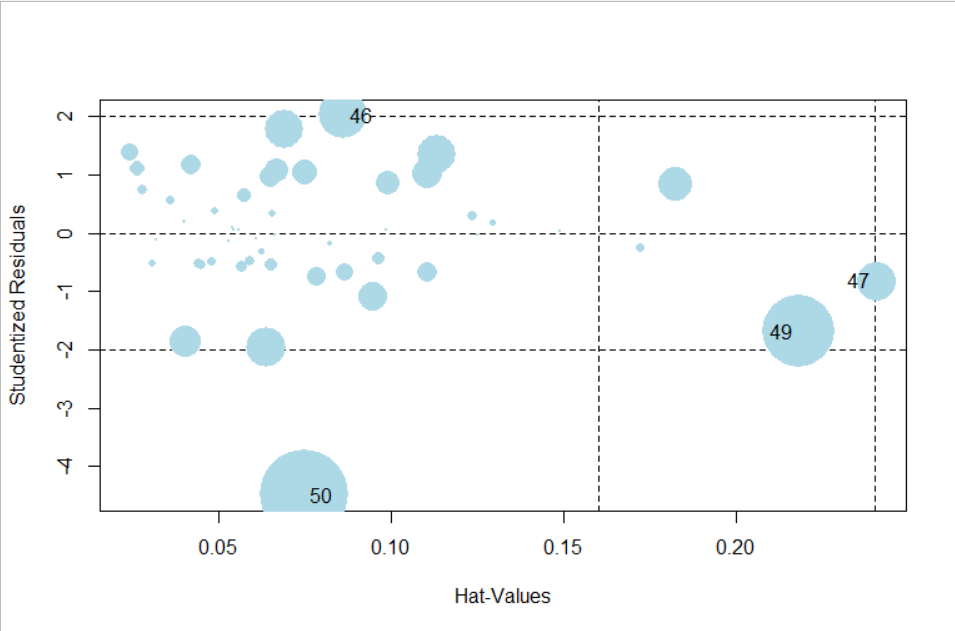
**model.S <- lm(Profit~R.D.Spend+Administration+Marketing.Spend)**

We get R² value as 0.9507. Which convey that 95% of variation in the “profit” is explained by the Independent variables in our model. Where we found that variable “Administration”, is not significant in our model, where the variable “Marketing Spend” is somewhat significant with significance level of 0.1 i.e. 90% confidence level.

#### **Now let’s have a look on the variable “State”**

Here in the above plot we can see clearly that, If we plot 3 simple linear regression, taking the variable “R.D.Spend” as independent and “Profit” as dependent. And plotting 3 regression lines for the 3 individual “State” data i.e. “California”, ”Florida” and “New York” than we can see that all the plots are overlapping with each other in the same confidence belt. The difference between the plots is negligible. Which may be considered as one regression line over there. As from above plot we can say that if we are not considering the state variable in our plot, then also we are getting the similar accuracy. So, we may not require to consider the “State” variable in our model.

#### Influence Plot:

Now we will look on the observations which are highly influencing our model fitting.

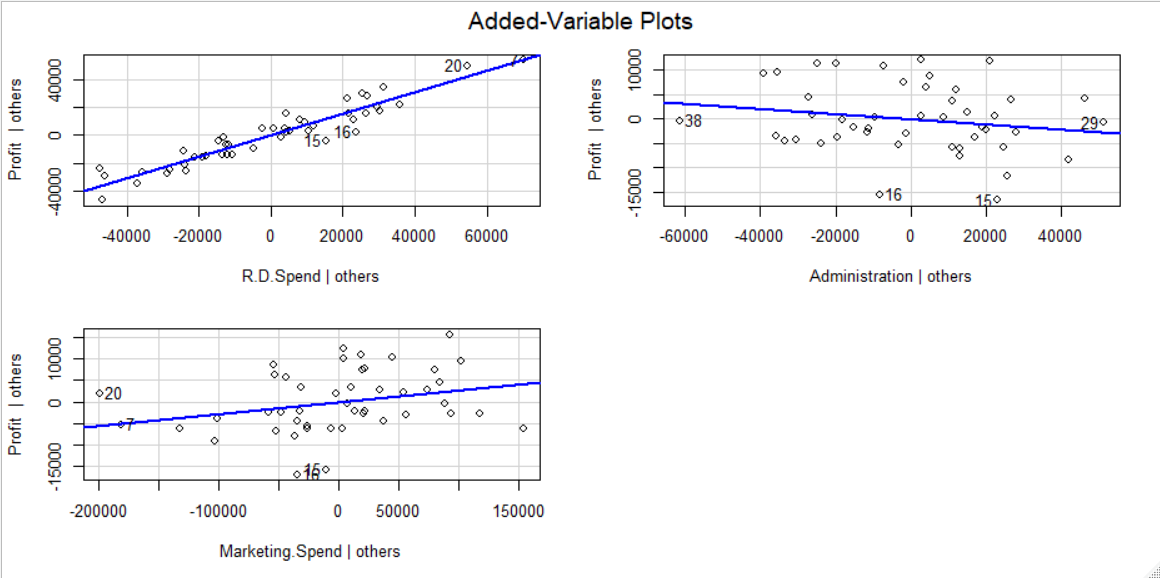
Here we can see that observation no 50.49,47,46 are our influence indexes. So, we may remove them for getting more accuracy in our model.

### Model 2:

**model.S.2 <- lm(Profit~R.D.Spend+Administration+Marketing.Spend,data = df\_Startups)**

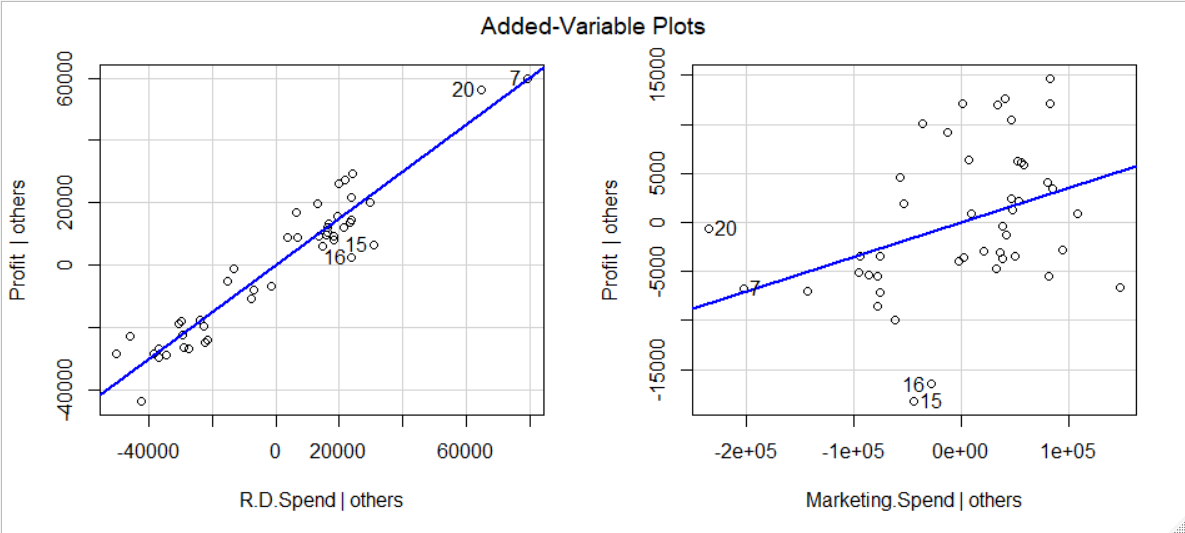
In our model 2 we remove the influencing observations. We get R² as 9626 with RMSE 6774 and

Correlation between actual and predicted = 0.9748282

We can see that still “Administrative” is insignificant in our model with level of significance (α) probability of error 0.21 i.e. 79% confidence. But incase of “Marketing Spend “we can say that its significance level is 0.06 i.e. approx. to 0.05 so we can relay on this variable to explain the model properly.

### Model 3:

**model.S.3 <- lm(Profit~R.D.Spend+Marketing.Spend,data = df\_Startups)**

In model 3 we get R² = 0.9612 and RMSE = 6899.99 and correlation between the actual and predicted is 0.9748121. In our model 3 we can see that After removing the variable “Administrative” we get probability of error (α) for considering the variable “MarketingSpend” is 0.01 i.e. less than 0.05, now we can say that it’s a significant variable in our model.

## TABULATION:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variable | R² | RMSE | Cor (Y, Predicted) |
| Model 1 | All variable except State | 0.9507 | 8855.344 | 0.975062 |
| Model 2 | Removed the Obs. # 50,49,47,46 | 0.9626 | 6774.245 | 0.9748282 |
| Model 3 | Removed variable “Administration: from model 2 | 0.9612 | 6899.99 | 0.9748121 |

## Conclusion:

So Here My best fit will be either Model 2, as having the least RMSE value as well as higher R² value as well as correlation between the Actual and predicted values.

But in certain case if we go for considering only the variable which are significant enough to explain our model properly. Then I might go for the Model 3, as in model 2 we are considering the insignificant variable “Administration”, so may be this is the possible reason for slight increase in our R² values, which may not matter as already in model 3, its explaining our 97% variation in our target variable. So, I may rely on my third model also.

# Question 2 :

Predict Price of the computer

Answer:

Available variable: "X", "price", "speed", "hd", "ram", "screen", "cd", "multi", "premium", "ads", "trend"

Target Variable : price

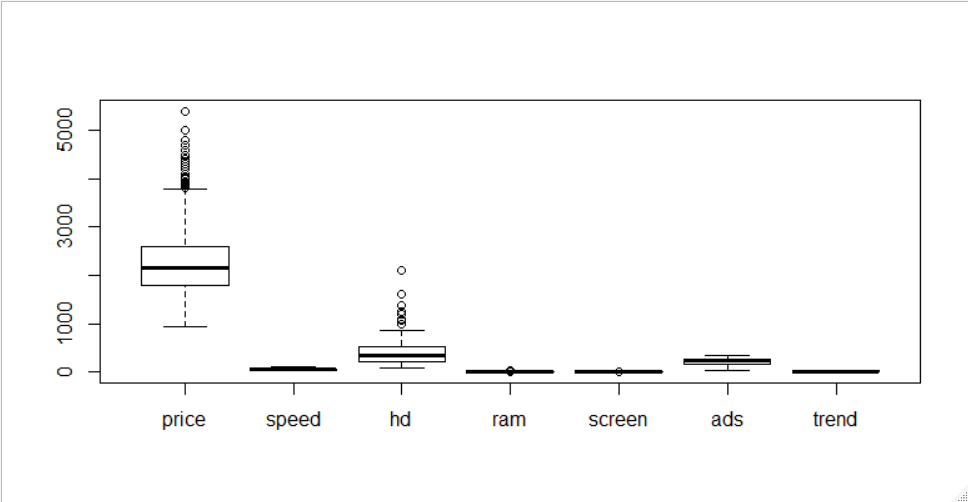
Lets have a look on the summary Statistics of the variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | price | speed | hd | ram | screen | ads | trend |
| Min. : 1 | Min. : 949 | Min. : 25.00 | Min. : 80.0 | Min. : 2.000 | Min. :14.00 | Min. : 39.0 | Min. : 1.00 |
| 1st Qu.:1566 | 1st Qu.:1794 | 1st Qu.: 33.00 | 1st Qu.: 214.0 | 1st Qu.: 4.000 | 1st Qu.:14.00 | 1st Qu.:162.5 | 1st Qu.:10.00 |
| Median :3130 | Median :2144 | Median : 50.00 | Median : 340.0 | Median : 8.000 | Median :14.00 | Median :246.0 | Median :16.00 |
| Mean :3130 | Mean :2220 | Mean : 52.01 | Mean : 416.6 | Mean : 8.287 | Mean :14.61 | Mean :221.3 | Mean :15.93 |
| 3rd Qu.:4694 | 3rd Qu.:2595 | 3rd Qu.: 66.00 | 3rd Qu.: 528.0 | 3rd Qu.: 8.000 | 3rd Qu.:15.00 | 3rd Qu.:275.0 | 3rd Qu.:21.50 |
| Max. :6259 | Max. :5399 | Max. :100.00 | Max. :2100.0 | Max. :32.000 | Max. :17.00 | Max. :339.0 | Max. :35.00 |

|  |  |  |
| --- | --- | --- |
| cd | multi | premium |
| no :3351 | no :5386 | no : 612 |
| yes:2908 | yes: 873 | yes:5647 |

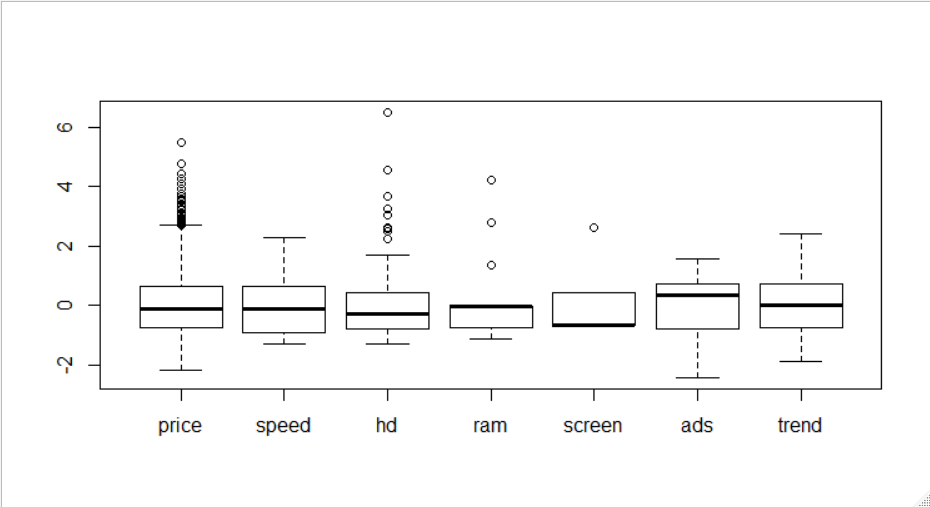
All except cd, multi, premium are discrete type. Where as cd, multi, premium is of factor type.

## BOXPLOT:



We can see there is lots of outlier in the variable price and hd.

Lets make the plot unitless and scale free

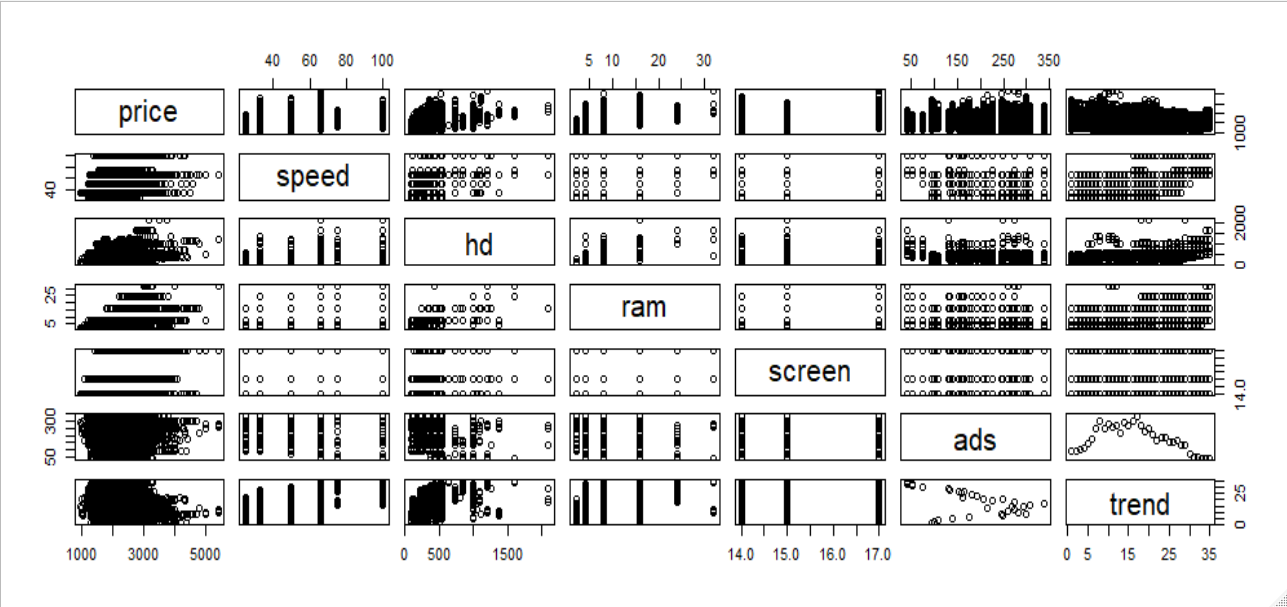


Now we have a clear view in each and every variable here

## CORRELATION:

None of the variables are strongly correlated as seen from the data bars.

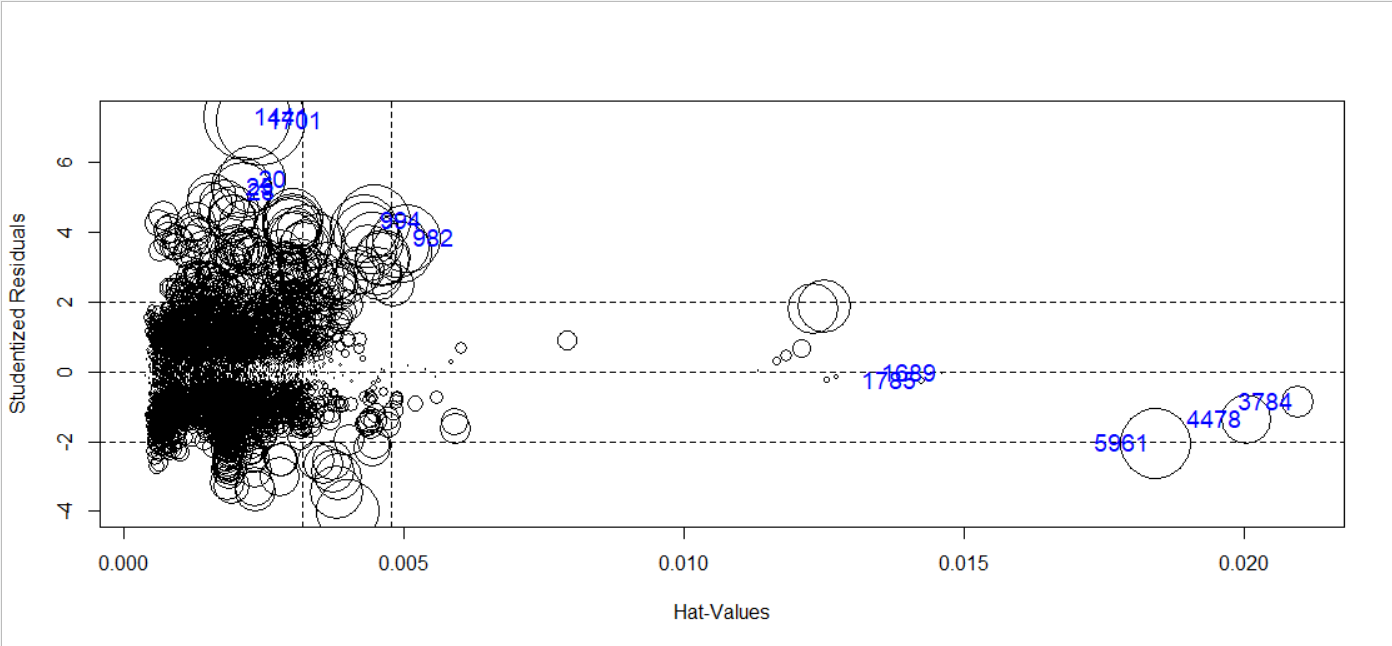
## PAIRS PLOT:



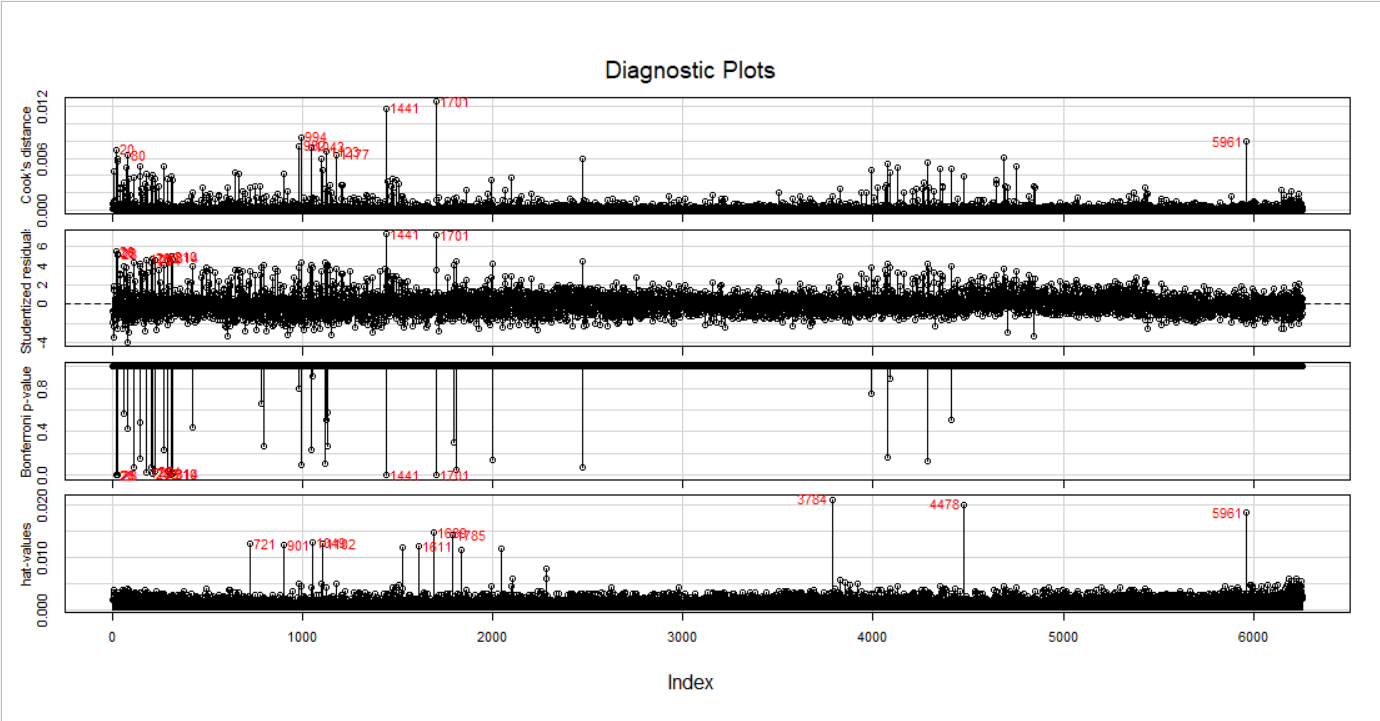
## FITTING REGRESSION MODEL:

### Model 1:

model\_Comp\_1 <-lm(price~speed+hd+ram+screen+cd+multi+premium+ads+trend,data = df\_comp)

In model 1 I simply consider all the variables and fit the model and come of with no insignificant variables, lots of influencing indexes, with coefficient of determination 0.7756 , RMSR value 275.1298 and finally correlation between the predicted and actual value to be 0.8806631, which was not that much bad for me.

From above plot we can see the dispersion of the points.

Looking at this plot we can say that there are pretty large numbers of influencing observations in our model.

Than I make the data scale free and unit less for performing my next model.

### Model 2:

**df\_comp2 <- data.frame(scale(log(Comp[,-c(1,2,7,8,9)])),"price" = df\_comp$price,"cd" = df\_comp$cd,"premium" = df\_comp$premium,"multi" = df\_comp$multi)**

**model\_Comp\_2 <- lm(price~.,data=df\_comp2)**

In model 2 I make the whole data (which columns are numerical) to unitless and scale free i.e. standardized the data, along with log transformation. Here I come up with coefficient of determination 0.7426, which was less than the previous model. So, I think better removing the influencing index from my data for my next model.

### Model 3:

**influ\_comp <- as.integer(rownames(influencePlot(model\_Comp\_2,id = list(n=20,col="blue"))))**

**df\_comp3 <- df\_comp2[-c(influ\_comp),]#head(df\_comp2)**

**model\_Comp\_3 <- lm(price~.,data=df\_comp3)**

In model 3 I removed 20 influencing observations. And fit my model again. Now I come up with slight improvement in my coefficient of determination as 0.7508, RMSE as 281.3819 and finally correlation between the actual and predicted value as 0.8664

As we now our data set still contains lots of influencing index over there with count 291. So I thought removing 3% of my data in my model 4.

Have a look on this influencing index of model 3 , where I come up with more than 200 influencing index for Cook’s distance

### Model 4:

**influencing\_obs <- which(rowSums(influence.measures(model\_Comp\_1)$is.inf) > 0);influencing\_obs # These are the influencing observations**

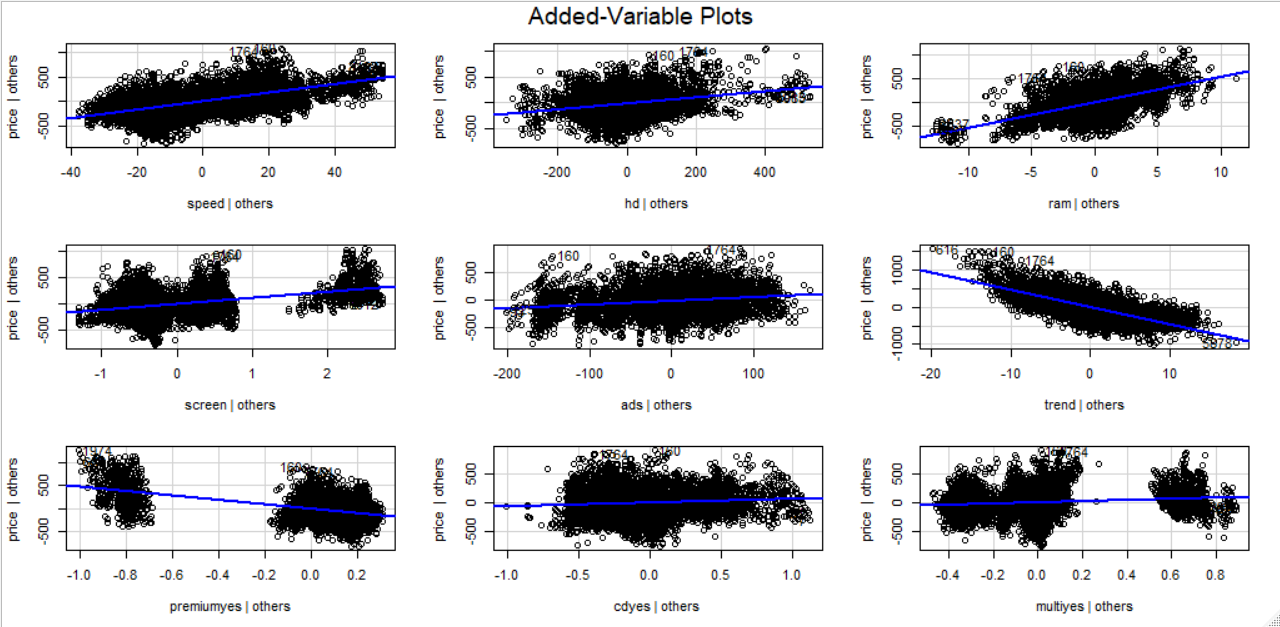
Here I get 294 influencing Index

**influence\_obs <- as.integer(rownames(influencePlot(model\_Comp\_1,id=list(n=90,col="red"))))**

Here I get 186 influencing Index

**df\_Comp\_scale <- data.frame(df\_comp[,-c(6,7,8)],"premium"=df\_comp$premium,"cd"=df\_comp$cd,"multi"=df\_comp$multi)**

**df\_Comp\_scale <- df\_Comp\_scale[-c(influence\_obs),]**

**model\_Comp\_4 <- lm(price~.,data=df\_Comp\_scale)**

Added Variable Plot

In my 4th model I prefer not to do any transformation with our data as it seems useless (observing the above two models) so in this model I prefer to focus more on my influencing observations and come up with 186 influencing index (defaulters) i.e. approx. 3% of my data set. So I build the model without those influencing factors but I predict my model with all the data set given to me. Now I come up with coefficient of determination as 0.804, RMSE is 238 and finally correlation between the actual and predicted values is 0.879

## TABULATION:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model No | Modeled with | Predicted with | Transformation | R² | RMSE | cor |
| 1 | All Observations | All Observations | NA | 0.7756 | 275.1298 | 0.880663 |
| 2 | All Observations | All Observations | NA | 0.7426 | - | - |
| 3 | 99.4% data | 99.4% data | log, standard scalar | 0.7508 | 281.3819 | 0.86647 |
| 4 | 97.02% data | All Observations | NA | 0.804 | 238.0004 | 0.879204 |

## Conclusion:

It’s obvious that I will go for my 4th model which is explaining 80% of variation in our target variable due to the observations, along with least RMSE value among all fitted models.

# Question 3 :

Consider only the below columns and prepare a prediction model for predicting Price.

Corolla<-Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

Answer:

Available variable: "Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight"

Target Variable : price

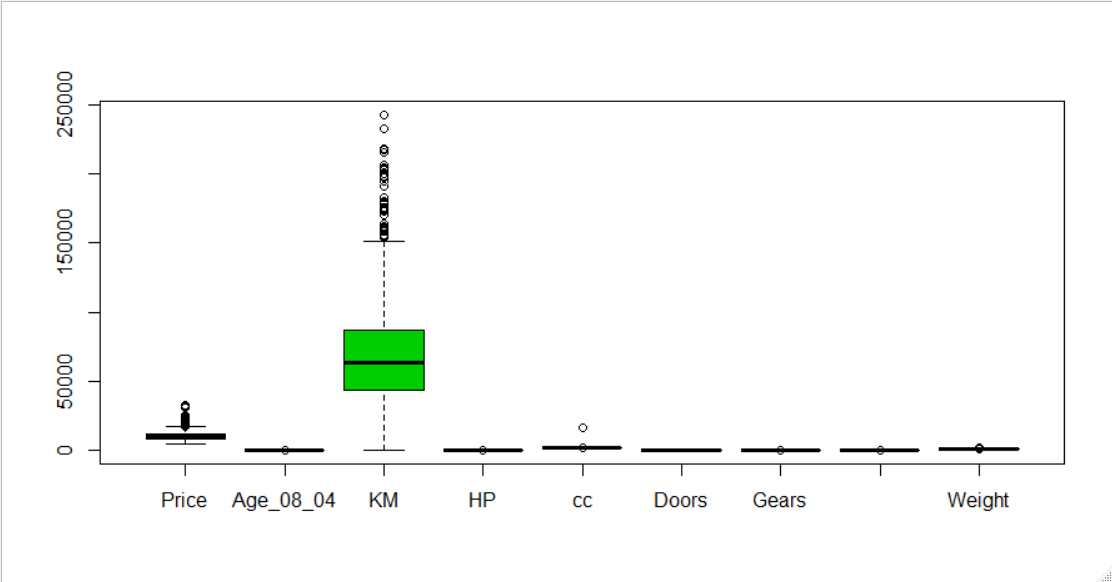
Lets have a look on the summary Statistics of the variables

## Summary:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Price | Age\_08\_04 | KM | HP | cc | Doors | Gears | Quarterly\_Tax | Weight |
| Min. : 4350 | Min. : 1.00 | Min. : 1 | Min. : 69.0 | Min. : 1300 | Min. :2.000 | Min. :3.000 | Min. : 19.00 | Min. :1000 |
| 1st Qu.: 8450 | 1st Qu.:44.00 | 1st Qu.: 43000 | 1st Qu.: 90.0 | 1st Qu.: 1400 | 1st Qu.:3.000 | 1st Qu.:5.000 | 1st Qu.: 69.00 | 1st Qu.:1040 |
| Median : 9900 | Median :61.00 | Median : 63390 | Median :110.0 | Median : 1600 | Median :4.000 | Median :5.000 | Median : 85.00 | Median :1070 |
| Mean :10731 | Mean :55.95 | Mean : 68533 | Mean :101.5 | Mean : 1577 | Mean :4.033 | Mean :5.026 | Mean : 87.12 | Mean :1072 |
| 3rd Qu.:11950 | 3rd Qu.:70.00 | 3rd Qu.: 87021 | 3rd Qu.:110.0 | 3rd Qu.: 1600 | 3rd Qu.:5.000 | 3rd Qu.:5.000 | 3rd Qu.: 85.00 | 3rd Qu.:1085 |
| Max. :32500 | Max. :80.00 | Max. :243000 | Max. :192.0 | Max. :16000 | Max. :5.000 | Max. :6.000 | Max. :283.00 | Max. :1615 |

Price, “Age\_08\_04”, “KM” have significant difference in their mean and median. So, we can say they may contain outliers.

## BOXPLOT:



This Plot cant convey as much as we desire.

So lets go for a scale free plot

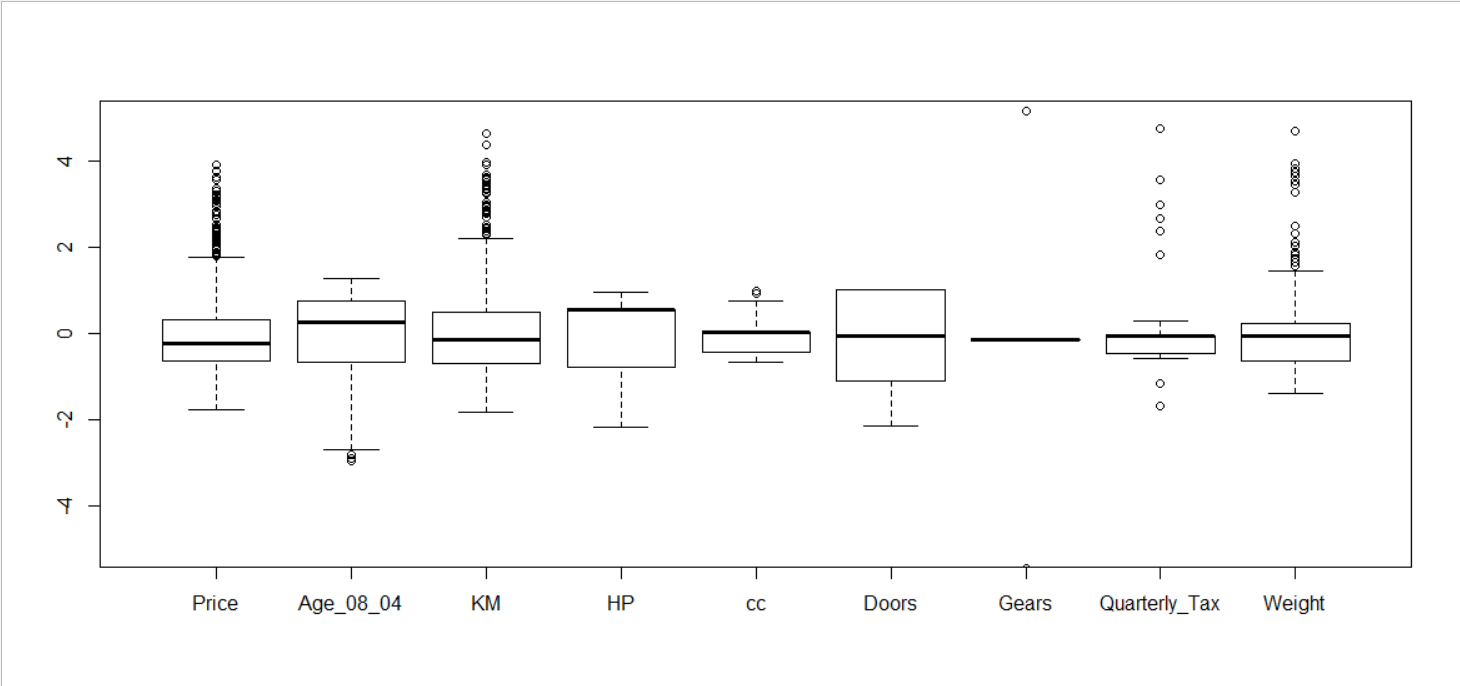
Even if I have considered only y-limit from -4 to 4 , but originally there is lot more points beyond this plot limit. So, from this plot we can see that Weight , Quarterly\_Tax also contains outliers in a larger amount.

## CORRELATION:



From the above Tabulation we can see, only Age\_08\_04 and Price is highly (negatively) correlated with correlation coefficient -0.876

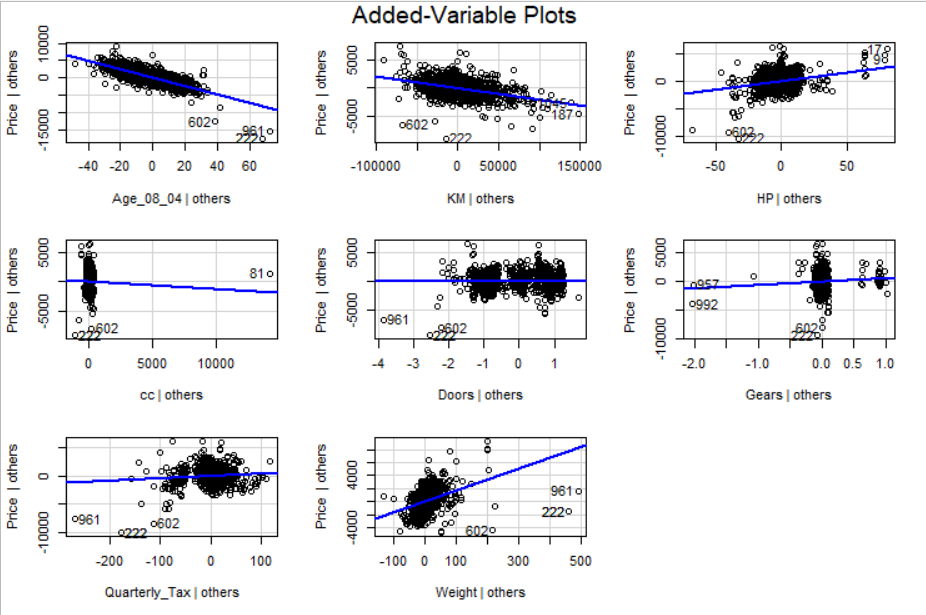
Its useless to go for pairs plot as we know that maximum variables are discrete with less unique values so those data may not convey much for finding out relationships, i.e. seen by our correlation table with a mare look.



## FITTING REGRESSION MODEL:

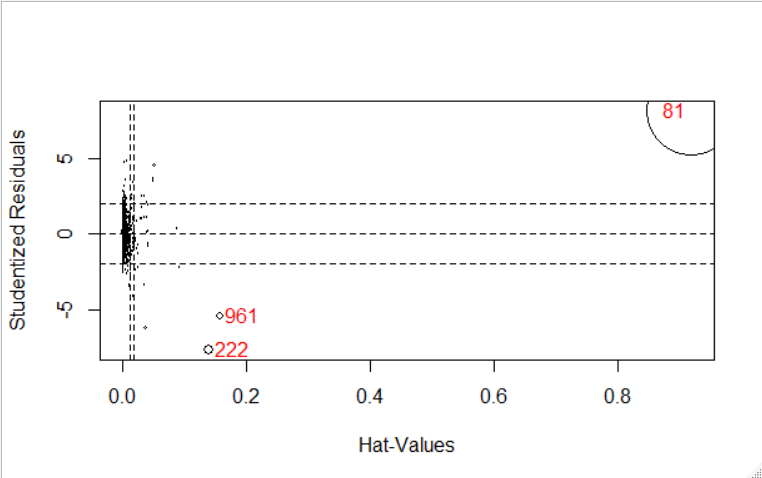
### Model 1:

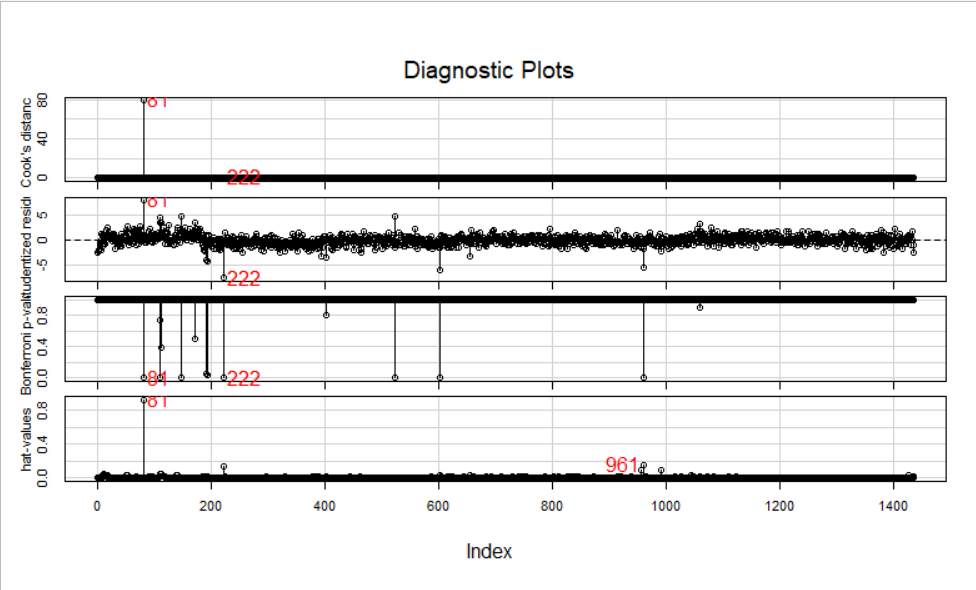
**model\_T\_1 <- lm(Price~.,data = Corolla)**

This is my simple model, considering all the observation as well as all the variables. In this model I get coefficient of determination as 0.8698, RMSE as 1338.25 and finally correlation between the actual and predicted as 0.929 .

From the above model we find that the variable cc and Doors are insignificant. But looking at the plot we can also say that Gears may also be insignificant for our model.

Influence Plot:



It seems like observation 81,961, and 222 are influencing more in our model.

### Model 2:

In our second model we are removing the three influencing observations.

df\_Corola <- Corolla[-c(influence\_index),]

model\_T\_2 <- lm(Price~.,data = df\_Corola)

In our second model we come up with all significant variables. We can see our R² value little bit increased i.e. 0.8852 with RMSE decreased 1227.474 and finally correlation between the actual and predicted value is found out to be 0.9408425

##### Added Variable Plot:

From this plot we can see that now cc is showing somewhat significant behavior, but even if Doors and Gear are not showing that much significance, but we may consider them as per our |t| statistics value in our model.

## Tabulation

|  |  |  |  |
| --- | --- | --- | --- |
| Model No | R² | RMSE | Cor |
| 1 | 0.8698 | 1338.25 | 0.929 |
| 2 | 0.8852 | 1227.474 | 0.9408425 |

## CONCLUSION:

Its Obvious that I will go for my second model with higher R² value i.e. this model can able to explain 88% of variation in the price with the help of all the given independent variable in the data set. This is enough for predicting price. This model also with lower RMSE and higher Correlation between the actual and predicted variable as compare to the previous model.