**Multi-Linear Regression**

**Example- Computer Dataset**

**Target Variable is Price**

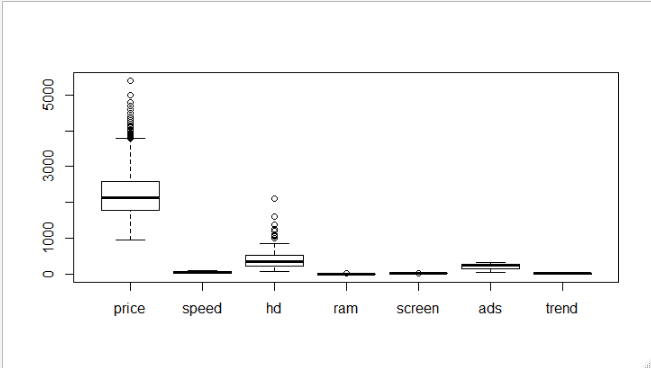
**Summary 🡺**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

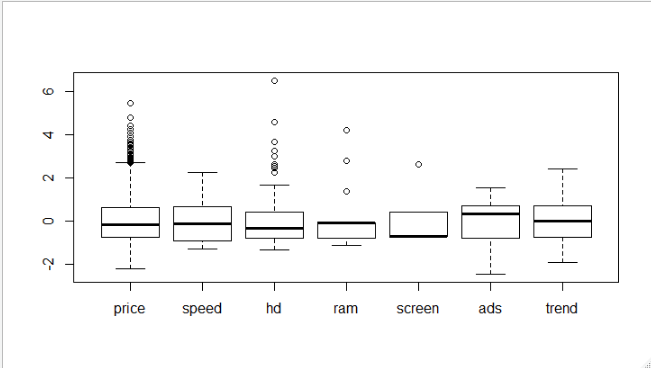
**From the above summary cd, multi and premium are factor type and rest all are in discrete type.**

**Box Plot 🡺**

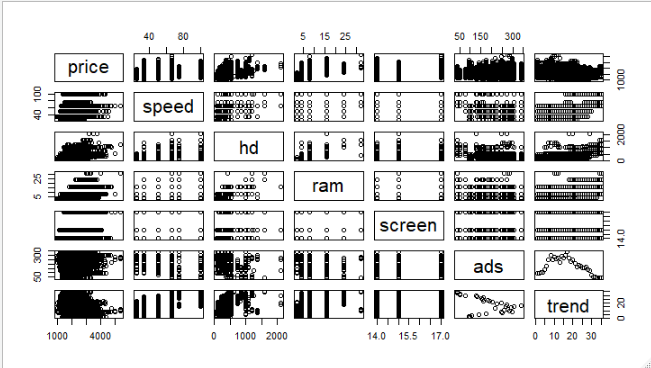


**From the above plot, so many outliers are present in variable price and hd.**

**Unitless and scale free box plot**



**Pairs Plot 🡺**



**Correlation 🡺**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **price** | **speed** | **hd** | **ram** | **screen** | **ads** | **trend** |
| **price** | **1** | **0.300976459** | **0.430257794** | **0.622748245** | **0.296041474** | **0.054540473** | **-0.199986935** |
| **speed** | **0.300976459** | **1** | **0.372304101** | **0.234760496** | **0.189074122** | **-0.21523206** | **0.405438333** |
| **hd** | **0.430257794** | **0.372304101** | **1** | **0.777726299** | **0.23280153** | **-0.323222005** | **0.577790128** |
| **ram** | **0.622748245** | **0.234760496** | **0.777726299** | **1** | **0.20895374** | **-0.181669713** | **0.276843843** |
| **screen** | **0.296041474** | **0.189074122** | **0.23280153** | **0.20895374** | **1** | **-0.093919429** | **0.188614445** |
| **ads** | **0.054540473** | **-0.21523206** | **-0.323222005** | **-0.181669713** | **-0.093919429** | **1** | **-0.318552508** |
| **trend** | **-0.199986935** | **0.405438333** | **0.577790128** | **0.276843843** | **0.188614445** | **-0.318552508** | **1** |

**From the above table it is seen that none of the variables are strongly correlated.**

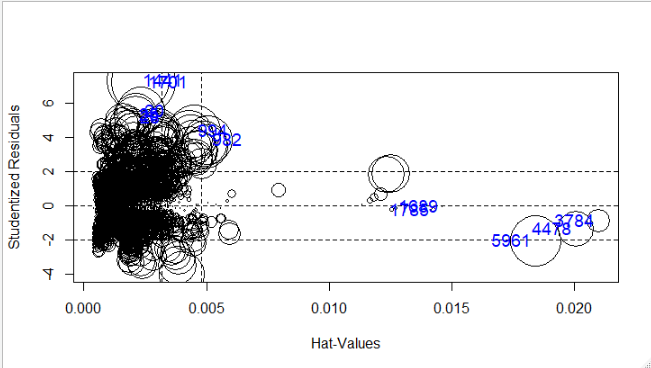
**Model-1 🡺**

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

Multiple R-squared: 0.7756, Adjusted R-squared: 0.7752

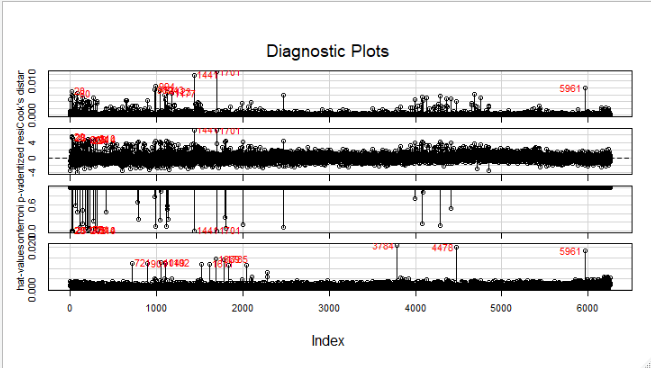
Correlation is 0.8806631

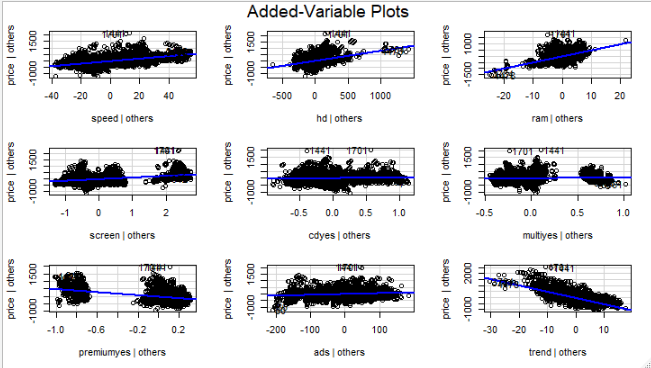
rmse = 275.1298



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

**Large number of influencing observations available in our model.**





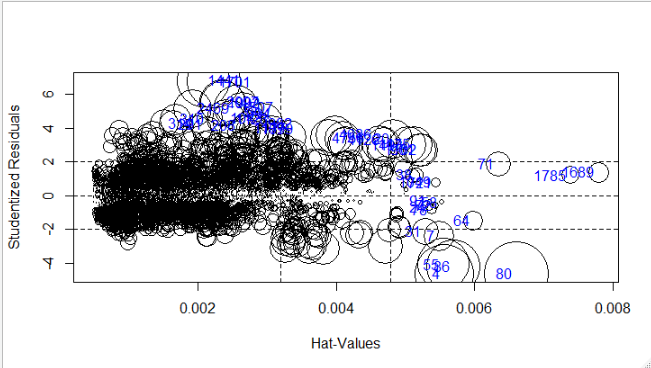
**Now we will make data scale free and unitless for next model.**

**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)

Multiple R-squared: 0.7426, Adjusted R-squared: 0.7422



**After standardizing the whole data and log transformation we are getting coefficient of determination 0.742 which is less than previous model.**

**So, we are removing influencing index for our next model.**

**Model-3 🡺**

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

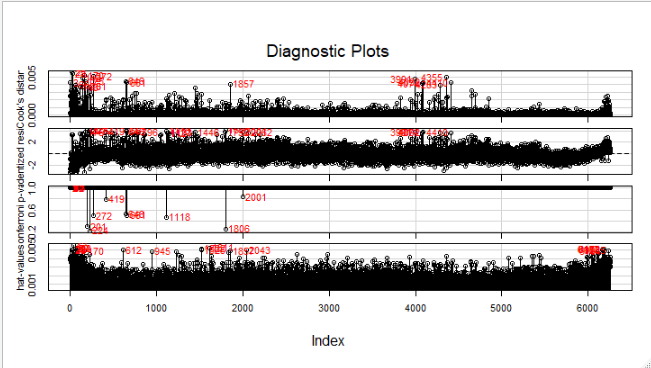
df\_comp3 <- df\_comp2[-c(influ\_comp),]#head(df\_comp2)

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

Multiple R-squared: 0.7508, Adjusted R-squared: 0.7504

Correlation is 0.8664749

rmse = 281.3819



**In mode-3 we have removed 20 influencing observations , so there is only slight improvement coefficient of determination, RMSE, and correlation .**

**But our dataset contain still more influencing index with count 291, so now we will remove 3% of data in our next model.**

**Model-4 🡺**

nfluencing\_obs <- length(which(rowSums(influence.measures(model\_Comp\_1)$is.inf) > 0));influencing\_obs

# These are the influencing observations

[1] 294

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

> length(influence\_obs)

[1] 186

df\_Comp\_scale <- data.frame(df\_comp[,-c(6,7,8)],"premium"=df\_comp$premium,"cd"=df\_comp$cd,"multi"=df\_comp$multi)#,"cd"=df\_comp3$cd,"multi"=df\_comp3$multi

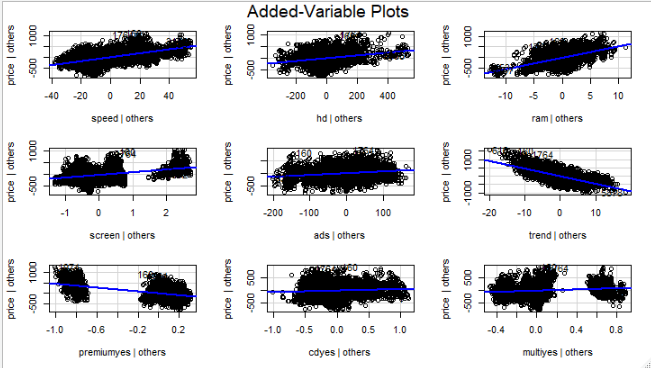
> df\_Comp\_scale <- df\_Comp\_scale[-c(influence\_obs),]

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

Multiple R-squared: 0.804, Adjusted R-squared: 0.8037

Correlation is 0.879

rmse = 238.0004



**In model-4 without influencing factor we are getting good results with less RMSE compared with other models.**

**Comparison 🡺**

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

**From the above comparison we can infer that Model-4 is good model with 80% of variation in our target variable due to observations along with least RMSE.**