######################Multi linear Regression###############################

**#Problem Statement:Predict Price of the computer**

**# Data : Computer\_Data.csv**

########################################################################

install.packages("psych")

library(psych)

# Diagnostic Plots

install.packages("car")

#to transform data from character to numeric

install.packages("plyr")

library(plyr)

library(car)

**##Step1 : Data Exploration**

ComputerDetails <- read.csv(file.choose()) #Computer\_Data.csv

ComputerData <- ComputerDetails[2:11]

ComputerData$cd <- as.numeric(revalue(ComputerData$cd,c("yes"=1, "no"=0)))

ComputerData$multi <- as.numeric(revalue(ComputerData$multi,c("yes"=1, "no"=0)))

ComputerData$premium <- as.numeric(revalue(ComputerData$premium,c("yes"=1, "no"=0)))

attach(ComputerData)

View(ComputerData)

summary(ComputerData)

#price speed hd ram screen cd

#Min. : 949 Min. : 25.00 Min. : 80.0 Min. : 2.000 Min. :14.00 Min. :1.000

#1st Qu.:1794 1st Qu.: 33.00 1st Qu.: 214.0 1st Qu.: 4.000 1st Qu.:14.00 1st Qu.:1.000

#Median :2144 Median : 50.00 Median : 340.0 Median : 8.000 Median :14.00 Median :1.000

#Mean :2220 Mean : 52.01 Mean : 416.6 Mean : 8.287 Mean :14.61 Mean :1.465

#3rd Qu.:2595 3rd Qu.: 66.00 3rd Qu.: 528.0 3rd Qu.: 8.000 3rd Qu.:15.00 3rd Qu.:2.000

#Max. :5399 Max. :100.00 Max. :2100.0 Max. :32.000 Max. :17.00 Max. :2.000

#multi premium ads trend

#Min. :1.000 Min. :1.000 Min. : 39.0 Min. : 1.00

#1st Qu.:1.000 1st Qu.:2.000 1st Qu.:162.5 1st Qu.:10.00

#Median :1.000 Median :2.000 Median :246.0 Median :16.00

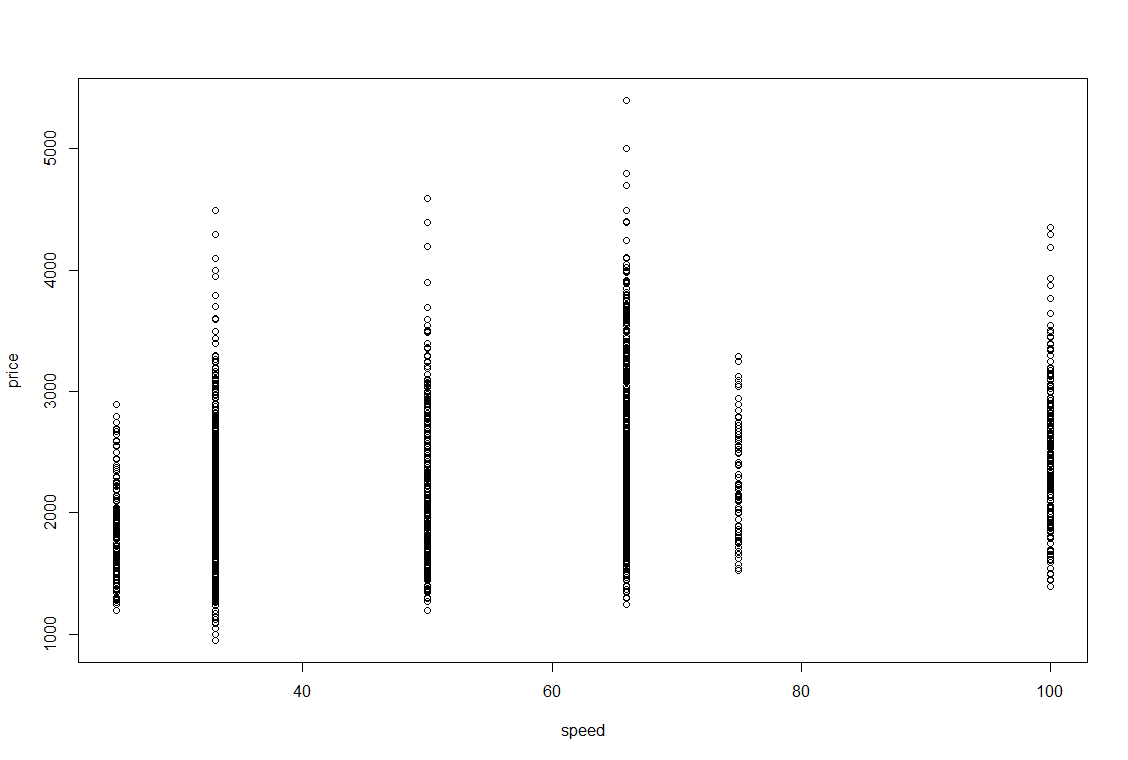
#Mean :1.139 Mean :1.902 Mean :221.3 Mean :15.93

#3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:275.0 3rd Qu.:21.50

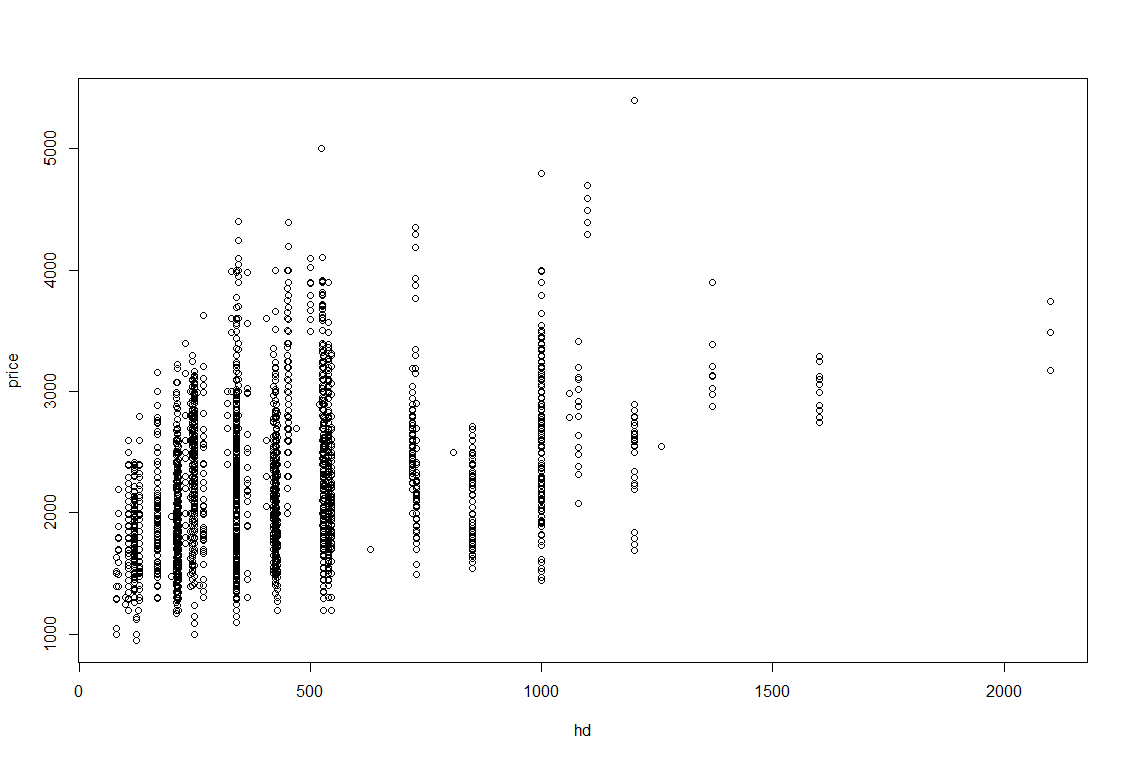
#Max. :2.000 Max. :2.000 Max. :339.0 Max. :35.00

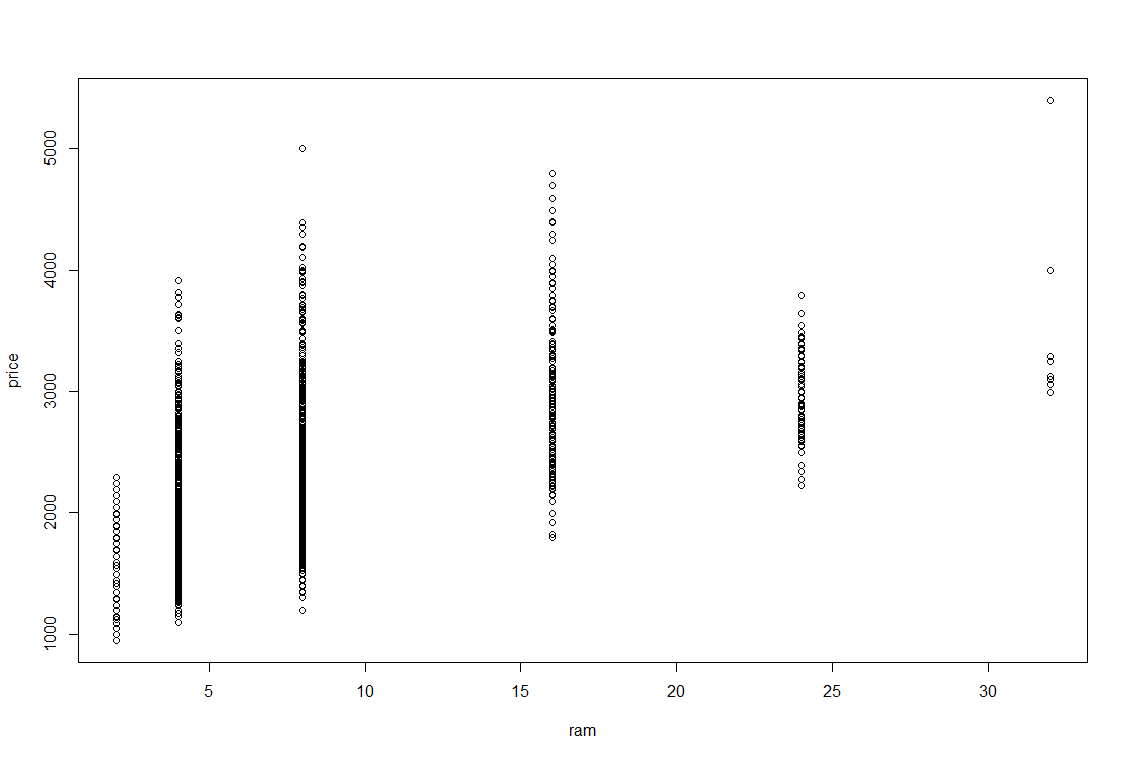
**#Data Visualization**

plot(speed,price)

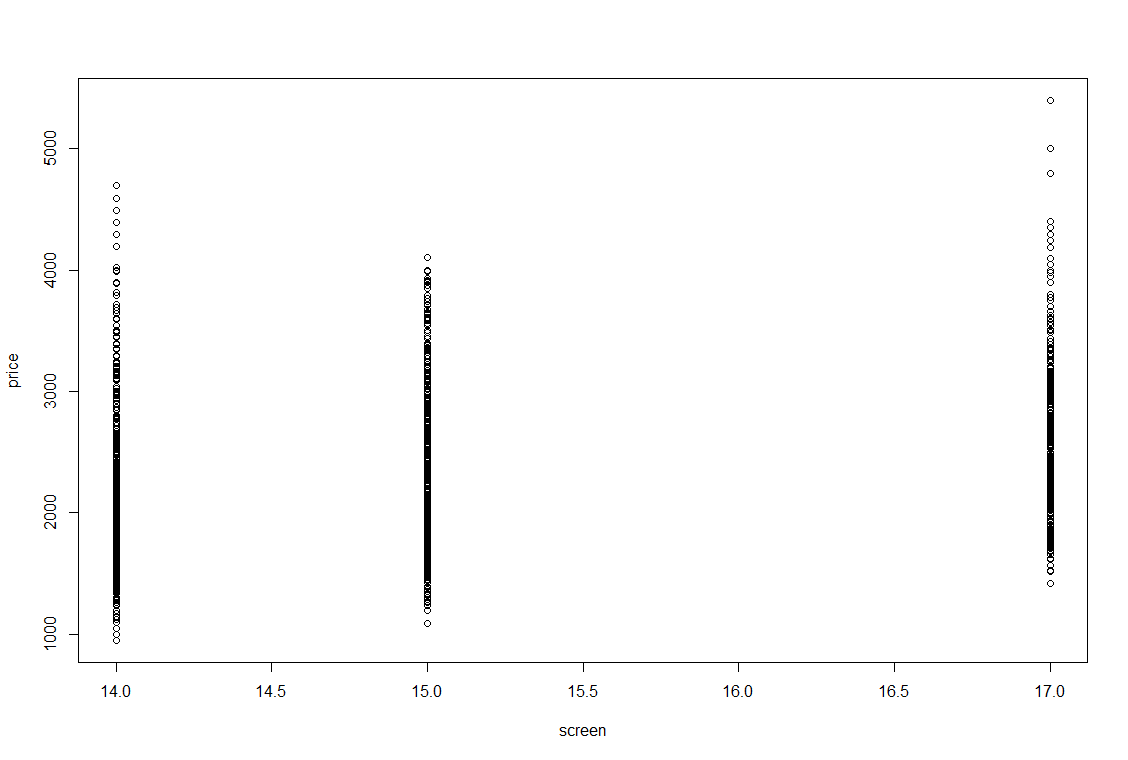


plot(hd,price)

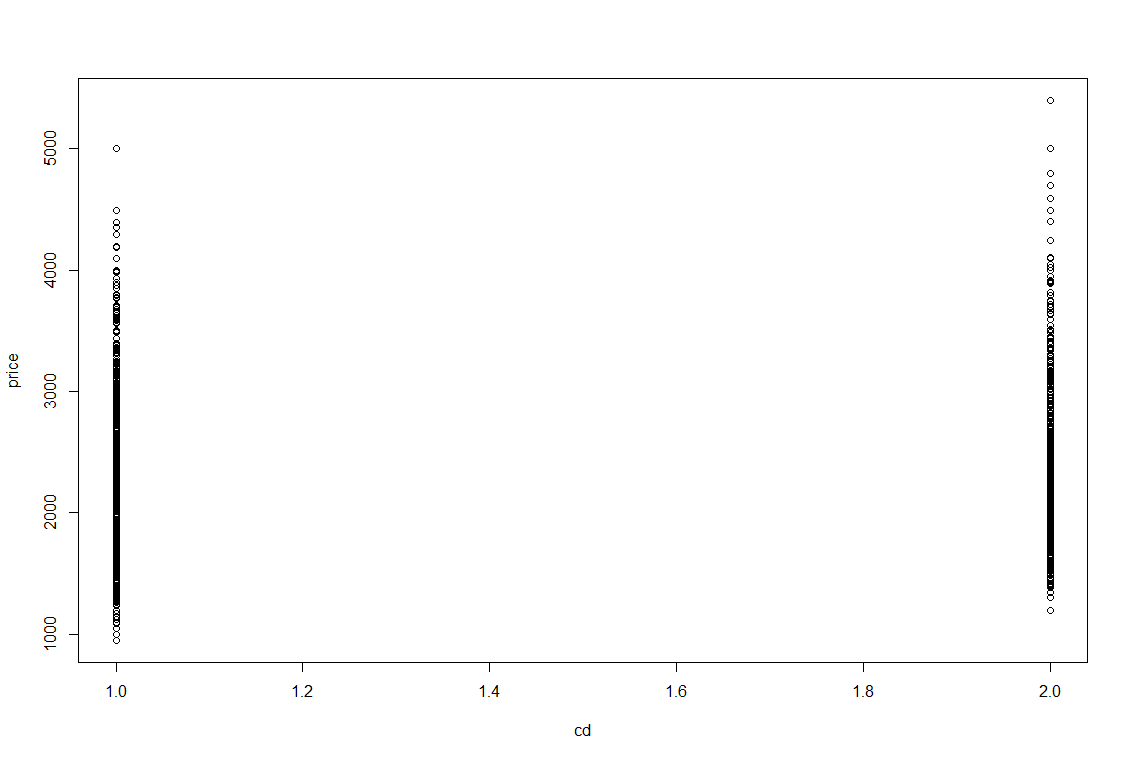


plot(ram,price) 

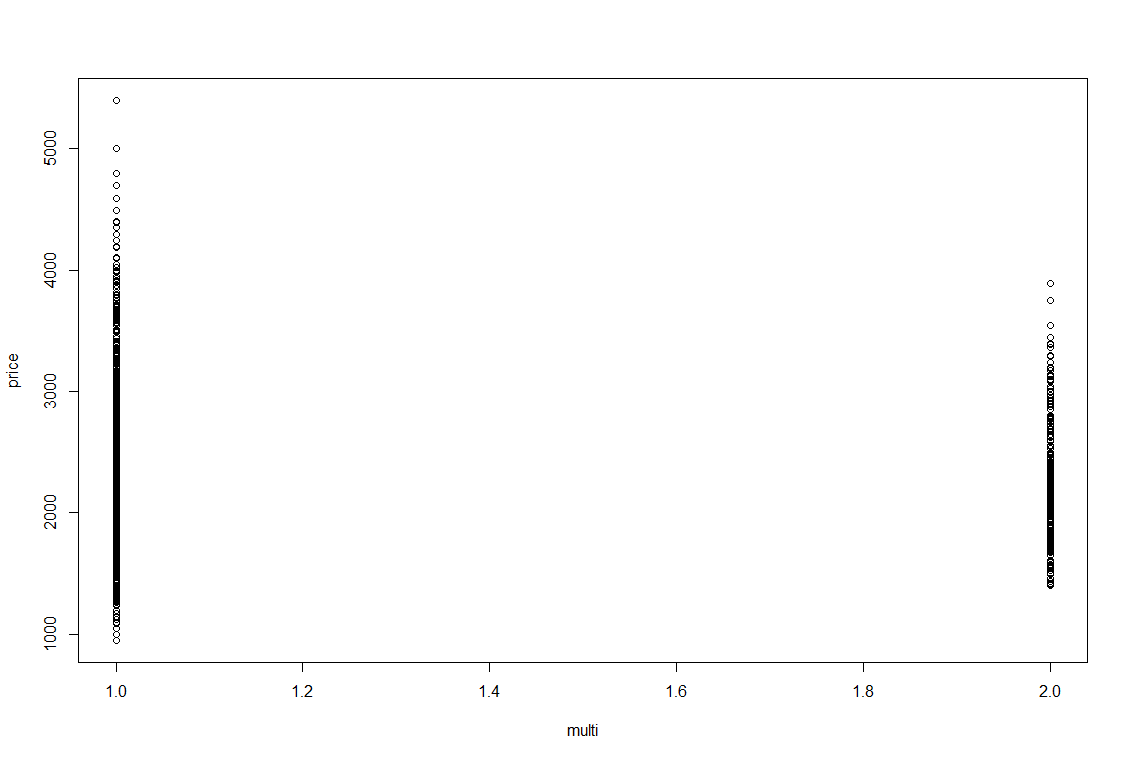
+plot(screen,price)



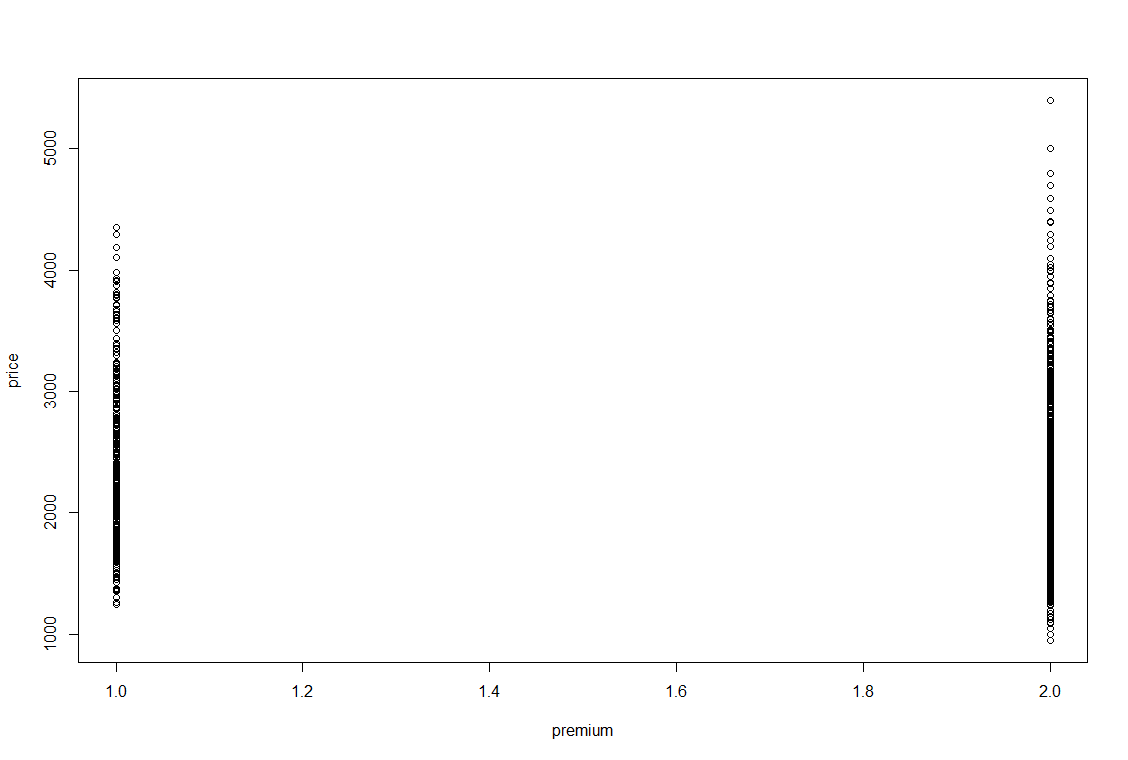
plot(cd,price)



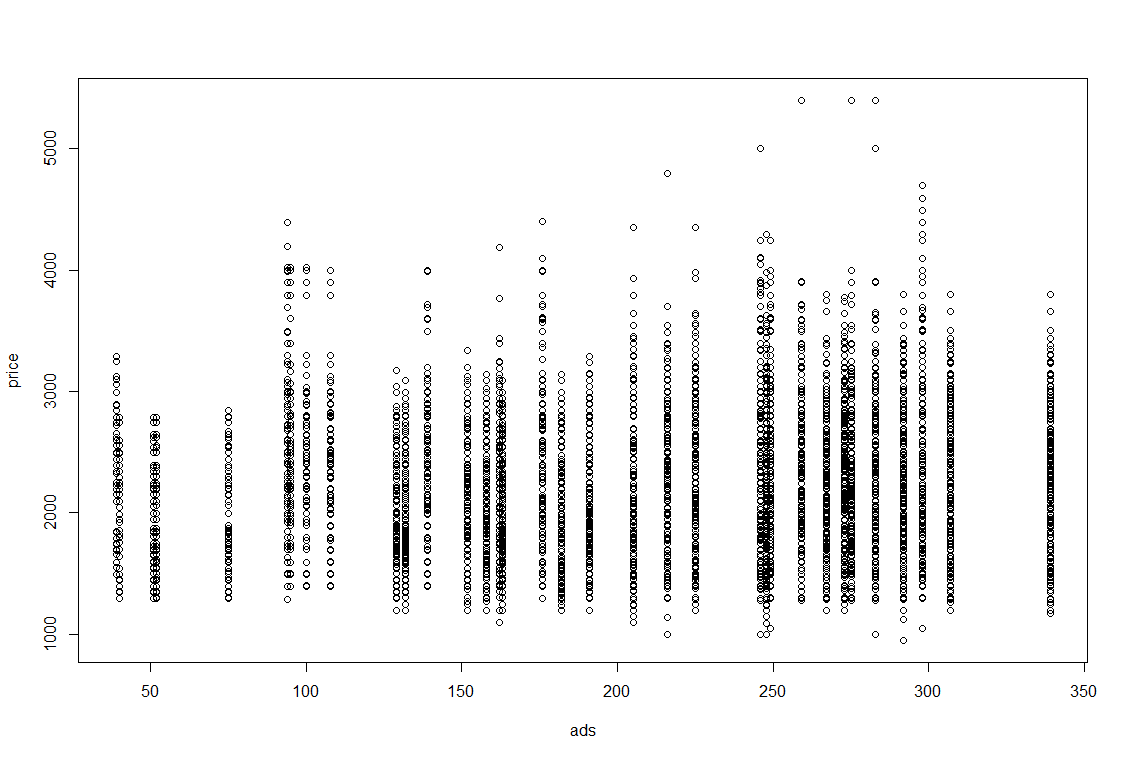
plot(multi,price)



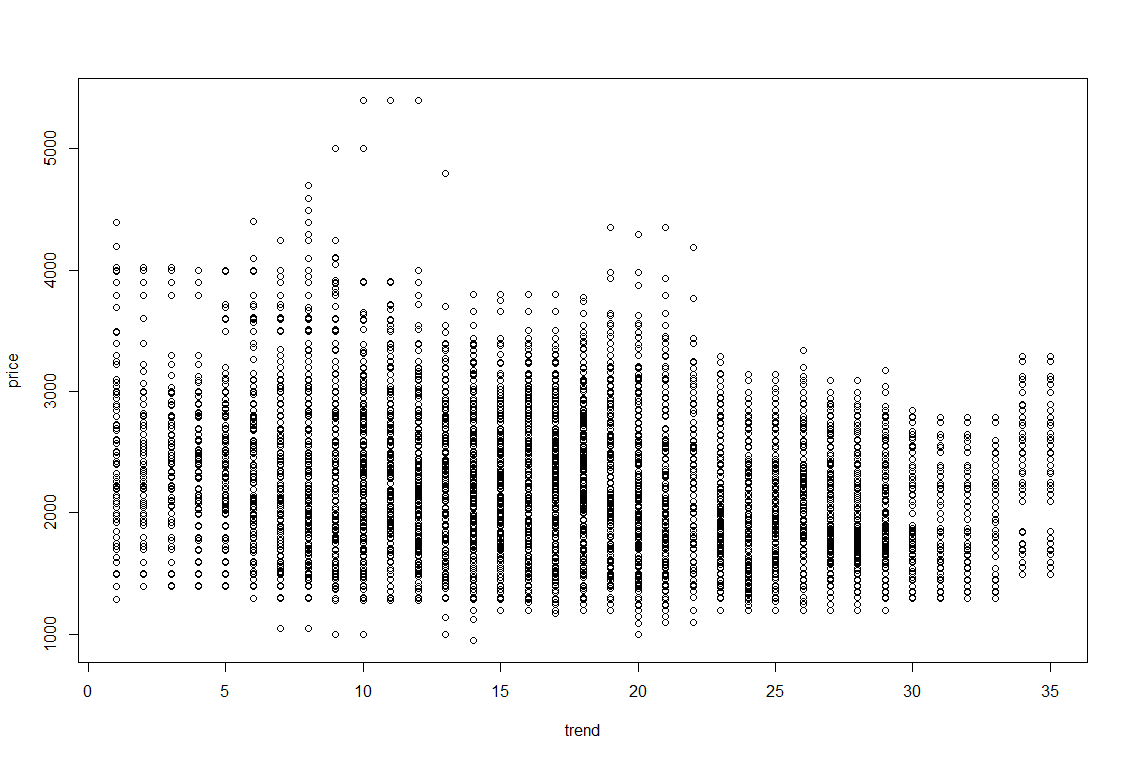
plot(premium,price)



plot(ads,price)

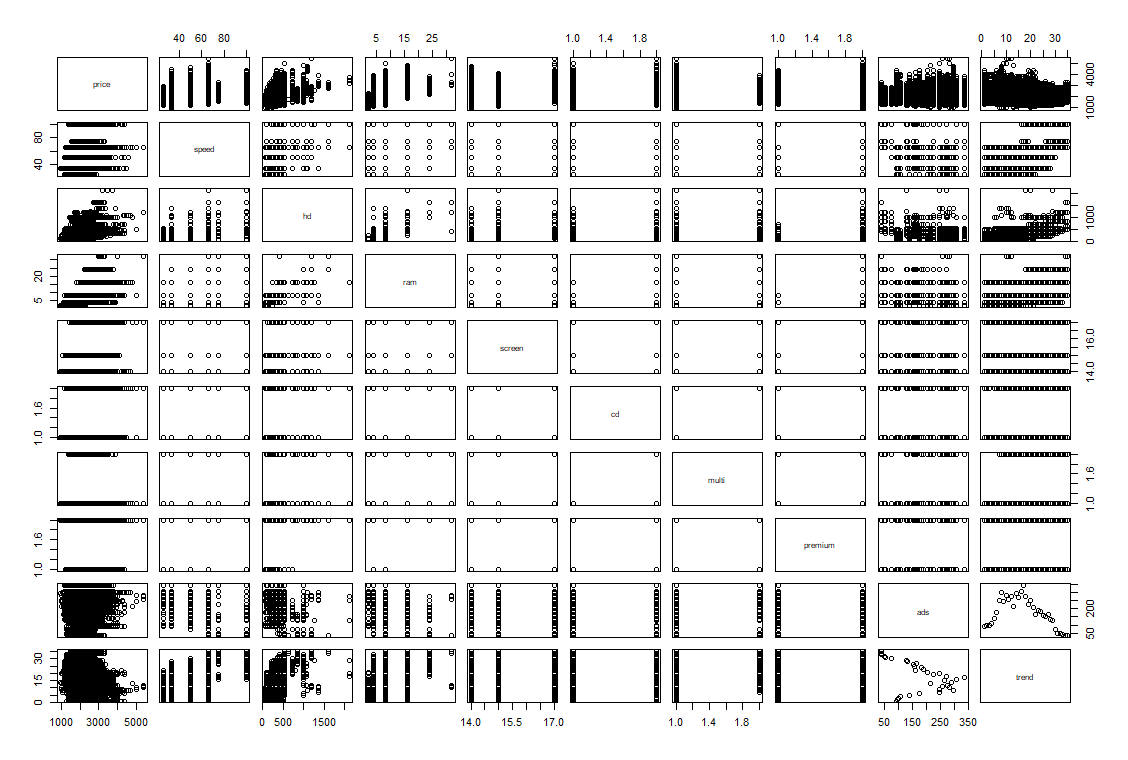


plot(trend,price)



#Scatter plot for all pairs of variables.

pairs(ComputerData)



#Correlation coefficient -strength and direction of correlation

cor(ComputerData)

#price speed hd ram screen cd multi premium

#price 1.00000000 0.30097646 0.43025779 0.62274824 0.296041474 0.19734334 -0.016651388 -0.08069636

#speed 0.30097646 1.00000000 0.37230410 0.23476050 0.189074122 0.25825980 0.084171934 0.11420791

#hd 0.43025779 0.37230410 1.00000000 0.77772630 0.232801530 0.50357041 0.092804830 0.19692359

#ram 0.62274824 0.23476050 0.77772630 1.00000000 0.208953740 0.43850441 0.045496894 0.19714459

#screen 0.29604147 0.18907412 0.23280153 0.20895374 1.000000000 0.12948766 -0.001740414 0.01874522

#cd 0.19734334 0.25825980 0.50357041 0.43850441 0.129487662 1.00000000 0.432179298 0.21607660

#multi -0.01665139 0.08417193 0.09280483 0.04549689 -0.001740414 0.43217930 1.000000000 0.12477474

#premium -0.08069636 0.11420791 0.19692359 0.19714459 0.018745223 0.21607660 0.124774741 1.00000000

#ads 0.05454047 -0.21523206 -0.32322200 -0.18166971 -0.093919429 -0.06109108 -0.030394260 -0.15202274

#trend -0.19998694 0.40543833 0.57779013 0.27684384 0.188614445 0.44578018 0.210907431 0.04210738

# ads trend

#price 0.05454047 -0.19998694

#speed -0.21523206 0.40543833

#hd -0.32322200 0.57779013

#ram -0.18166971 0.27684384

#screen -0.09391943 0.18861444

#cd -0.06109108 0.44578018

#multi -0.03039426 0.21090743

#premium -0.15202274 0.04210738

#ads 1.00000000 -0.31855251

#trend -0.31855251 1.00000000

#Building the linear regression model

computerModel <- lm(price ~ speed+hd+ram+screen+cd+multi+premium+ads+trend)

summary(computerModel)

#Call:

# lm(formula = price ~ speed + hd + ram + screen + cd + multi +

# premium + ads + trend)

#Residuals:

# Min 1Q Median 3Q Max

#-1093.77 -174.24 -11.49 146.49 2001.05

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) 651.97219 64.49224 10.109 < 2e-16 \*\*\*

# speed 9.32028 0.18506 50.364 < 2e-16 \*\*\*

# hd 0.78178 0.02761 28.311 < 2e-16 \*\*\*

# ram 48.25596 1.06608 45.265 < 2e-16 \*\*\*

# screen 123.08904 3.99950 30.776 < 2e-16 \*\*\*

# cd 60.91671 9.51559 6.402 1.65e-10 \*\*\*

# multi 104.32382 11.41268 9.141 < 2e-16 \*\*\*

# premium -509.22473 12.34225 -41.259 < 2e-16 \*\*\*

# ads 0.65729 0.05132 12.809 < 2e-16 \*\*\*

# trend -51.84958 0.62871 -82.470 < 2e-16 \*\*\*

#Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 275.3 on 6249 degrees of freedom

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

#F-statistic: 2399 on 9 and 6249 DF, p-value: < 2.2e-16

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

computerModelS <- lm(price ~ speed)

summary(computerModelS)

**#Multiple R-squared: 0.09059, Adjusted R-squared: 0.09044**

computerModel1 <- lm(price ~ speed+hd)

summary(computerModel1)

**#Multiple R-squared: 0.2081, Adjusted R-squared: 0.2079**

computerModel2<- lm(price ~ speed+hd+ram)

summary(computerModel2)

**#Multiple R-squared: 0.4335, Adjusted R-squared: 0.4332**

computerModel3<- lm(price ~ speed+hd+screen)

summary(computerModel3)

**#Multiple R-squared: 0.2425, Adjusted R-squared: 0.2421**

computerModel4<- lm(price ~ speed+hd+screen+cd)

summary(computerModel4)

**#Multiple R-squared: 0.2439, Adjusted R-squared: 0.2434**

computerModel5<- lm(price ~ speed+hd+screen+cd+multi)

summary(computerModel5)

**#Multiple R-squared: 0.2462, Adjusted R-squared: 0.2456**

computerModel6<- lm(price ~ speed+hd+screen+cd+multi+premium)

summary(computerModel6)

**#Multiple R-squared: 0.273, Adjusted R-squared: 0.2723**

computerModel7<- lm(price ~ speed+hd+screen+cd+multi+premium+ads)

summary(computerModel7)

**#Multiple R-squared: 0.3179, Adjusted R-squared: 0.3171**

computerModel8<- lm(price ~ speed+hd+screen+cd+multi+premium+ads+trend)

summary(computerModel8)

**#Multiple R-squared: 0.702, Adjusted R-squared: 0.7016**

### Partial Correlation matrix - Pure correlation between the variables

install.packages("corpcor")

library(corpcor)

cor2pcor(cor(ComputerData))

#[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]

#[1,] 1.0000000 0.537326713 0.33716508 0.49690857 0.36279673 0.080719005 0.11486991 -0.46269692

#[2,] 0.5373267 1.000000000 -0.10433213 -0.28438702 -0.11491585 -0.008311233 -0.07381036 0.29280827

#[3,] 0.3371651 -0.104332126 1.00000000 0.43156783 -0.10045528 0.087223167 -0.12420636 0.20637502

#[4,] 0.4969086 -0.284387018 0.43156783 1.00000000 -0.11960849 0.104933080 -0.10279078 0.25151338

#[5,] 0.3627967 -0.114915854 -0.10045528 -0.11960849 1.00000000 -0.024623811 -0.07426269 0.14359050

#[6,] 0.0807190 -0.008311233 0.08722317 0.10493308 -0.02462381 1.000000000 0.40003932 0.15563350

#[7,] 0.1148699 -0.073810359 -0.12420636 -0.10279078 -0.07426269 0.400039320 1.00000000 0.10912193

#[8,] -0.4626969 0.292808267 0.20637502 0.25151338 0.14359050 0.155633504 0.10912193 1.00000000

#[9,] 0.1599485 -0.140381854 -0.20186341 -0.04815123 -0.06683674 0.164413496 -0.04983936 -0.04691126

#[10,] -0.7219155 0.504953018 0.55265421 0.15896659 0.30166432 0.214237661 0.14157426 -0.42328259

#[,9] [,10]

#[1,] 0.159948478 -0.721915473

#[2,] -0.140381854 0.504953018

#[3,] -0.201863412 0.552654205

#[4,] -0.048151226 0.158966590

#[5,] -0.066836740 0.301664320

#[6,] 0.164413496 0.214237661

#[7,] -0.049839360 0.141574261

#[8,] -0.046911259 -0.423282587

#[9,] 1.000000000 -0.001856949

#[10,] -0.001856949 1.000000000

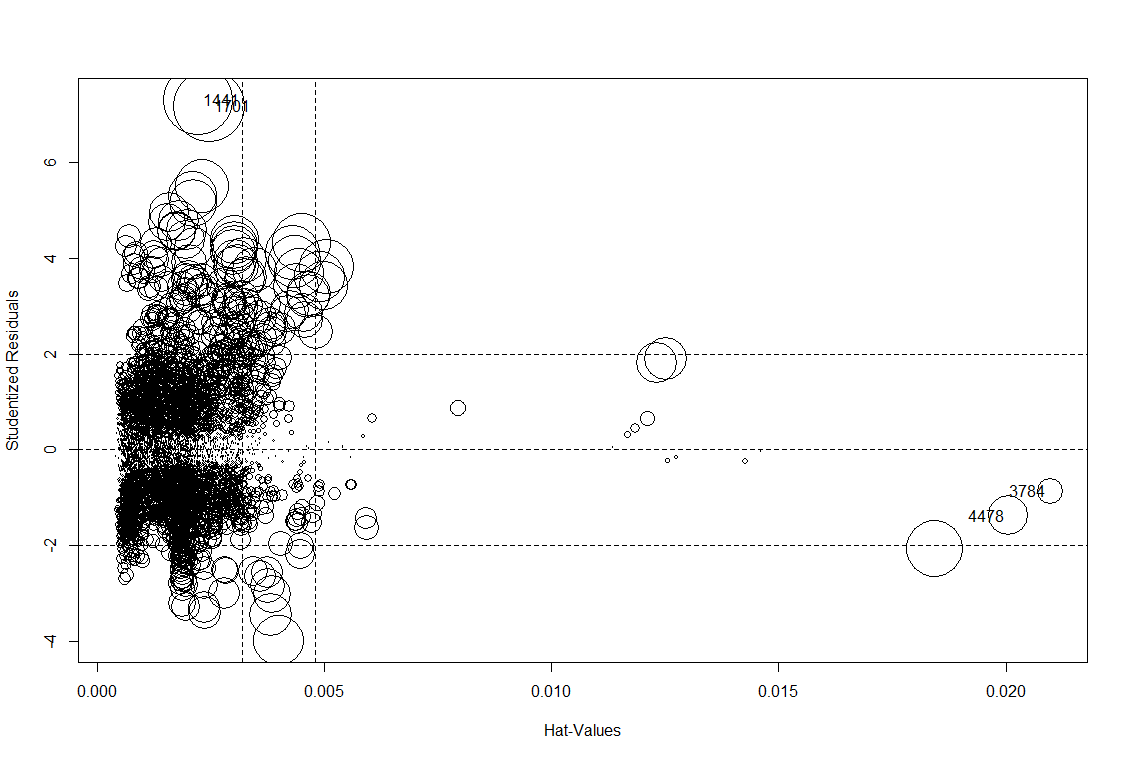
#deletion diagnostics for identifying the influential variable.

influence.measures(computerModel)

influenceIndexPlot(computerModel) # Index Plots of the influence measures

influencePlot(computerModel)# A user friendly representation of the above

?influencePlot



#Data frame with the hat values, Studentized residuals and Cook's distance of the identified points.

# StudRes Hat CookD

#1441 7.3058529 0.002228075 0.011819949

#1701 7.1838002 0.002464463 0.012647347

#3784 -0.8667018 0.020972880 0.001609237

#4478 -1.3795547 0.020060286 0.003895407

#Regression model after deleting the 1441 & 1701 observation

computerModel <- lm(price ~ speed+hd+ram+screen+cd+multi+premium+ads+trend, data=ComputerData[-c(1440,1701),])

summary(computerModel)

**#Multiple R-squared: 0.7766, Adjusted R-squared: 0.7763**

#Logarthimic transformation

model.computerDataLog <- lm(price ~ log(speed)+log(hd)+log(ram)+log(screen)+log(cd)+log(multi)+log(premium)

+log(ads)+log(trend), data=ComputerData[-c(1440,1701),] )

summary(model.computerDataLog)

**## Adjusted R-squared: 0.7431**

##Exponential Transformation

model.computerDataExp <-lm(log(price) ~ speed+hd+ram+screen+cd+multi+premium+ads+trend, data=ComputerData[-c(1440,1701),])

summary(model.computerDataExp)

**#Adjusted R-squared: 0.7831**

##Quadratic Transformation

model.computerDataQuad<-lm(price ~ speed+hd+ram+screen+cd+multi+premium+ads+trend

+I(speed^2)+I(hd^2)+I(ram^2)+I(screen^2)+I(cd^2)+I(multi^2)+I(premium^2)

+I(ads^2)+I(trend^2), data=ComputerData[-c(1440,1701),])

summary(model.computerDataQuad)

**#Multiple R-squared: 0.8044, Adjusted R-squared: 0.804**

#Poly model transformation

FinalModel<-lm(price ~ speed+hd+ram+screen+cd+multi+premium+ads+trend

+I(speed^2)+I(hd^2)+I(ram^2)+I(screen^2)+I(cd^2)+I(multi^2)+I(premium^2)

+I(ads^2)+I(trend^2) +I(speed^3)+I(hd^3)+I(ram^3)+I(screen^3)+I(cd^3)+I(multi^3)+I(premium^2)

+I(ads^3)+I(trend^3), data=ComputerData[-c(1440,1701),])

summary(FinalModel)

#Call:

# lm(formula = price ~ speed + hd + ram + screen + cd + multi +

# premium + ads + trend + I(speed^2) + I(hd^2) + I(ram^2) +

# I(screen^2) + I(cd^2) + I(multi^2) + I(premium^2) + I(ads^2) +

# I(trend^2) + I(speed^3) + I(hd^3) + I(ram^3) + I(screen^3) +

# I(cd^3) + I(multi^3) + I(premium^2) + I(ads^3) + I(trend^3),

# data = ComputerData[-c(1440, 1701), ])

#Residuals:

# Min 1Q Median 3Q Max

#-1049.27 -161.77 -21.93 133.37 1872.73

#Coefficients: (6 not defined because of singularities)

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) 1.088e+04 8.794e+02 12.371 < 2e-16 \*\*\*

# speed 6.668e+01 4.654e+00 14.326 < 2e-16 \*\*\*

# hd 2.550e+00 1.140e-01 22.375 < 2e-16 \*\*\*

# ram 2.931e+01 6.629e+00 4.422 9.96e-06 \*\*\*

# screen -1.334e+03 1.147e+02 -11.629 < 2e-16 \*\*\*

# cd 4.507e+01 8.924e+00 5.051 4.52e-07 \*\*\*

# multi 9.674e+01 1.046e+01 9.253 < 2e-16 \*\*\*

# premium -5.452e+02 1.150e+01 -47.422 < 2e-16 \*\*\*

# ads -3.694e+00 9.620e-01 -3.840 0.000124 \*\*\*

# trend -7.062e+00 8.107e+00 -0.871 0.383707

#I(speed^2) -9.069e-01 8.003e-02 -11.331 < 2e-16 \*\*\*

# I(hd^2) -2.181e-03 1.504e-04 -14.502 < 2e-16 \*\*\*

# I(ram^2) 1.401e+00 4.861e-01 2.882 0.003963 \*\*

# I(screen^2) 4.698e+01 3.726e+00 12.610 < 2e-16 \*\*\*

# I(cd^2) NA NA NA NA

#I(multi^2) NA NA NA NA

#I(premium^2) NA NA NA NA

#I(ads^2) 1.383e-02 4.955e-03 2.791 0.005264 \*\*

# I(trend^2) -1.858e+00 4.641e-01 -4.004 6.31e-05 \*\*\*

# I(speed^3) 4.302e-03 4.191e-04 10.264 < 2e-16 \*\*\*

# I(hd^3) 6.804e-07 5.842e-08 11.648 < 2e-16 \*\*\*

# I(ram^3) -2.775e-02 1.090e-02 -2.545 0.010939 \*

# I(screen^3) NA NA NA NA

#I(cd^3) NA NA NA NA

#I(multi^3) NA NA NA NA

#I(ads^3) -1.784e-05 8.070e-06 -2.211 0.027070 \*

# I(trend^3) 1.362e-02 7.904e-03 1.724 0.084789 .

# Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

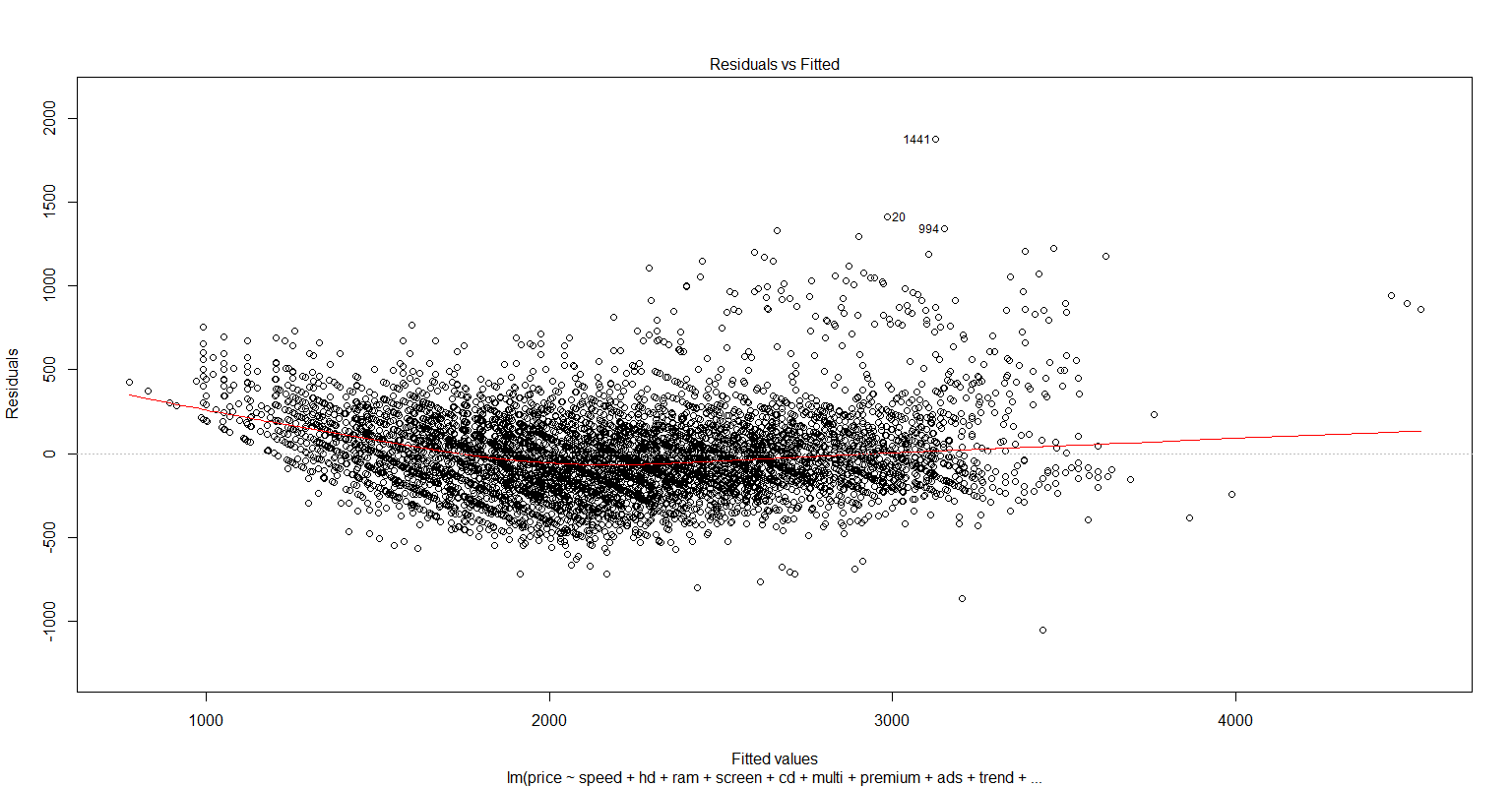
#Residual standard error: 251.4 on 6236 degrees of freedom

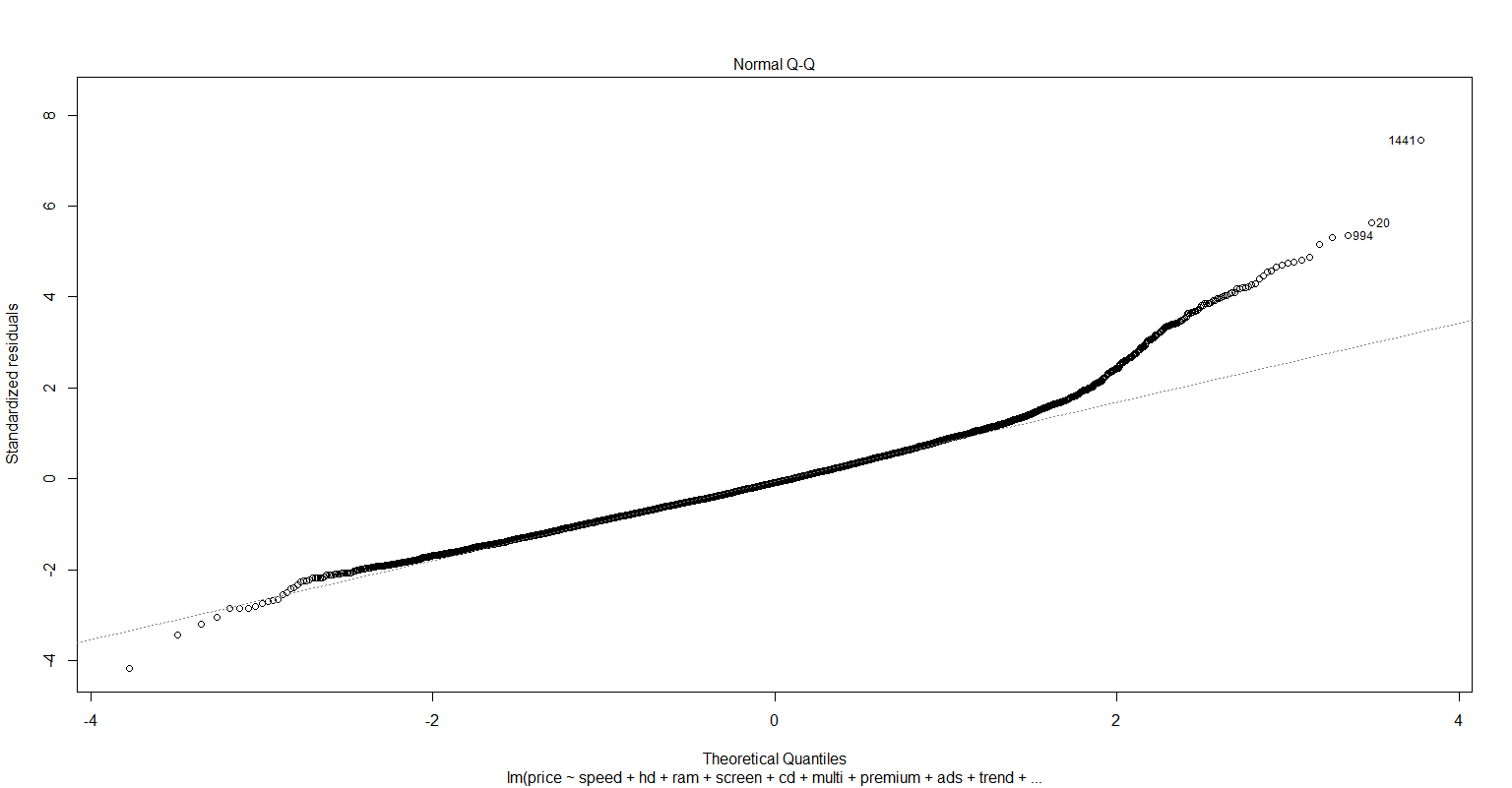
#Multiple R-squared: 0.8126, Adjusted R-squared: 0.812

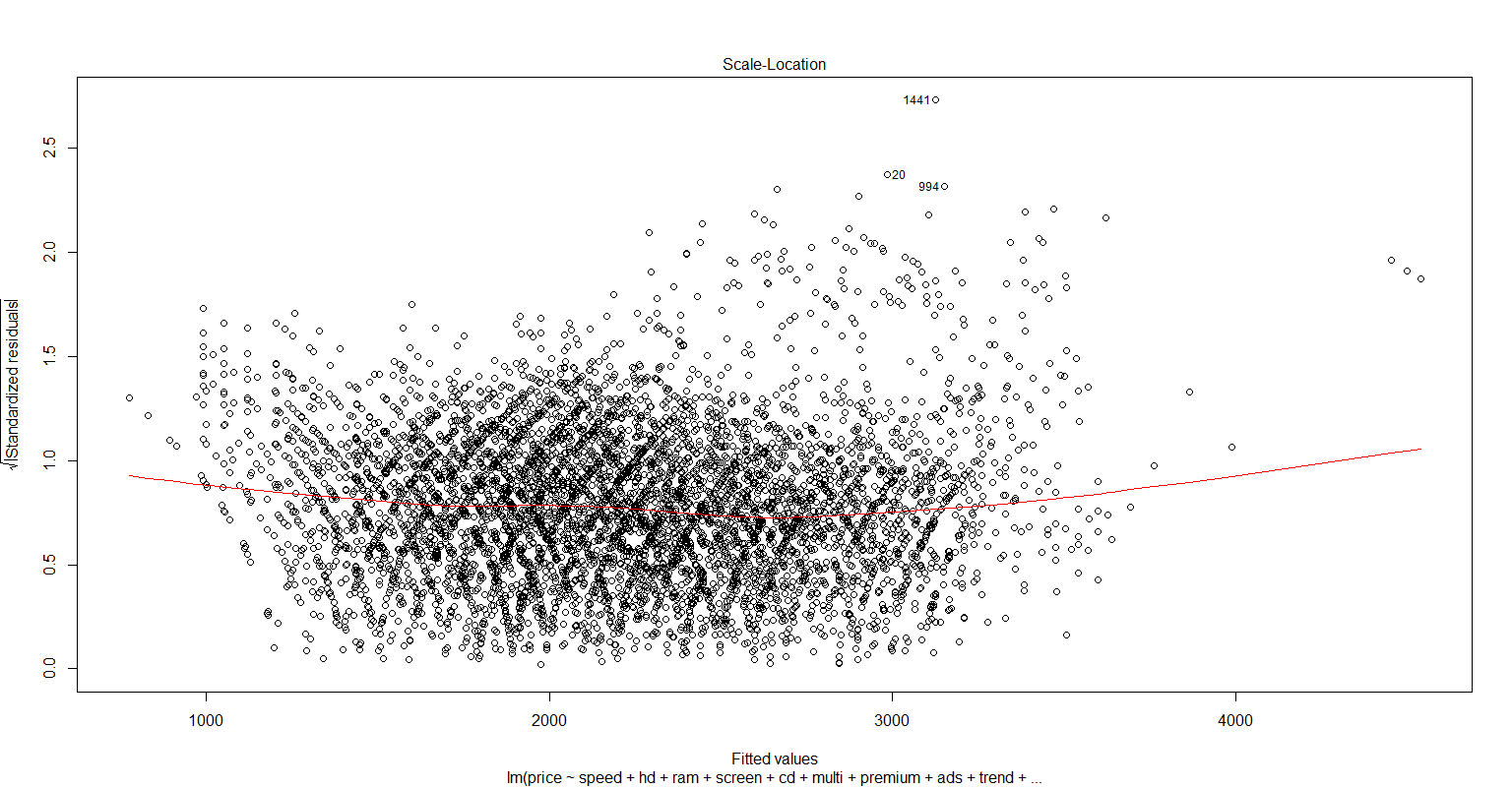
#F-statistic: 1352 on 20 and 6236 DF, p-value: < 2.2e-16

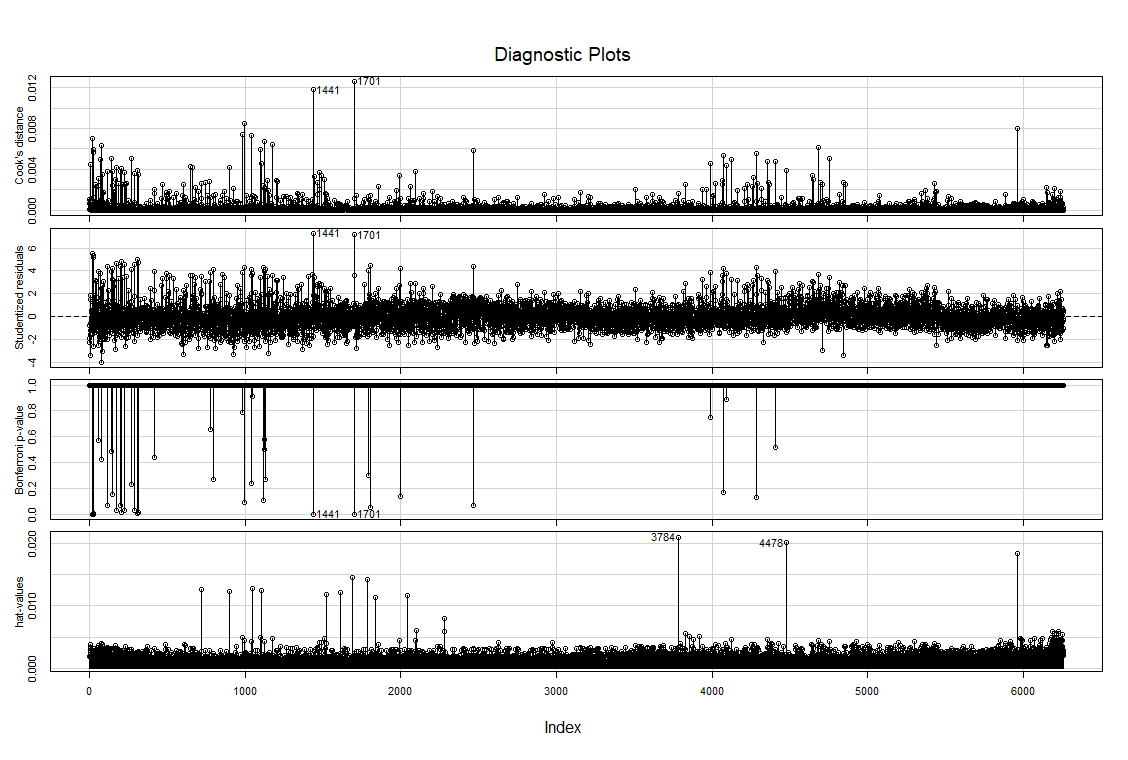
**#Multiple R-squared: 0.8126, Adjusted R-squared: 0.812**

plot(FinalModel)









## There are some outliers.

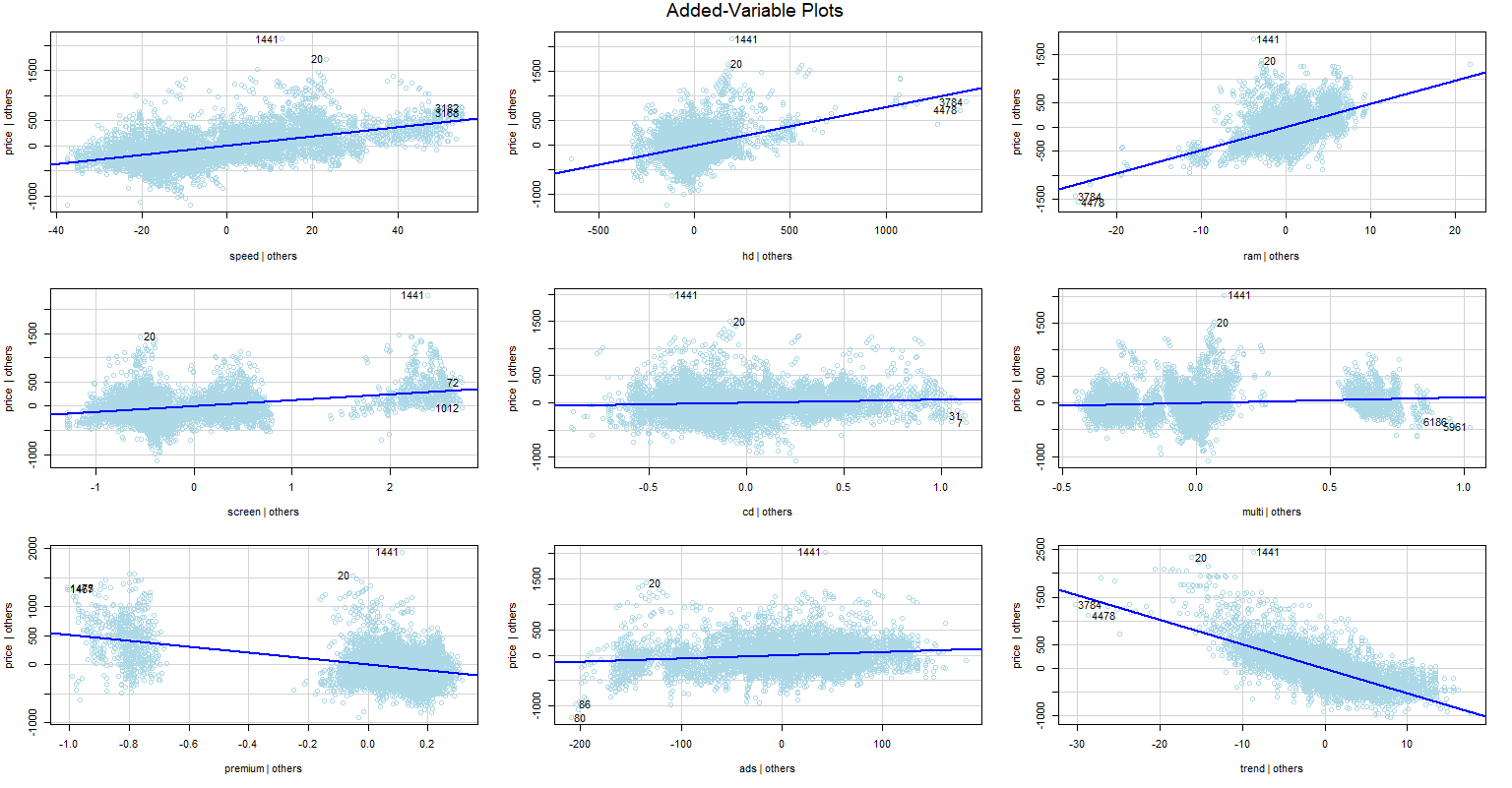
### Variance Inflation Factors

vif(computerModel) # VIF is > 10 => collinearity

#speed hd ram screen cd multi premium ads trend

#1.265351 4.208798 2.975865 1.081701 1.859724 1.290632 1.109381 1.217196 2.023754

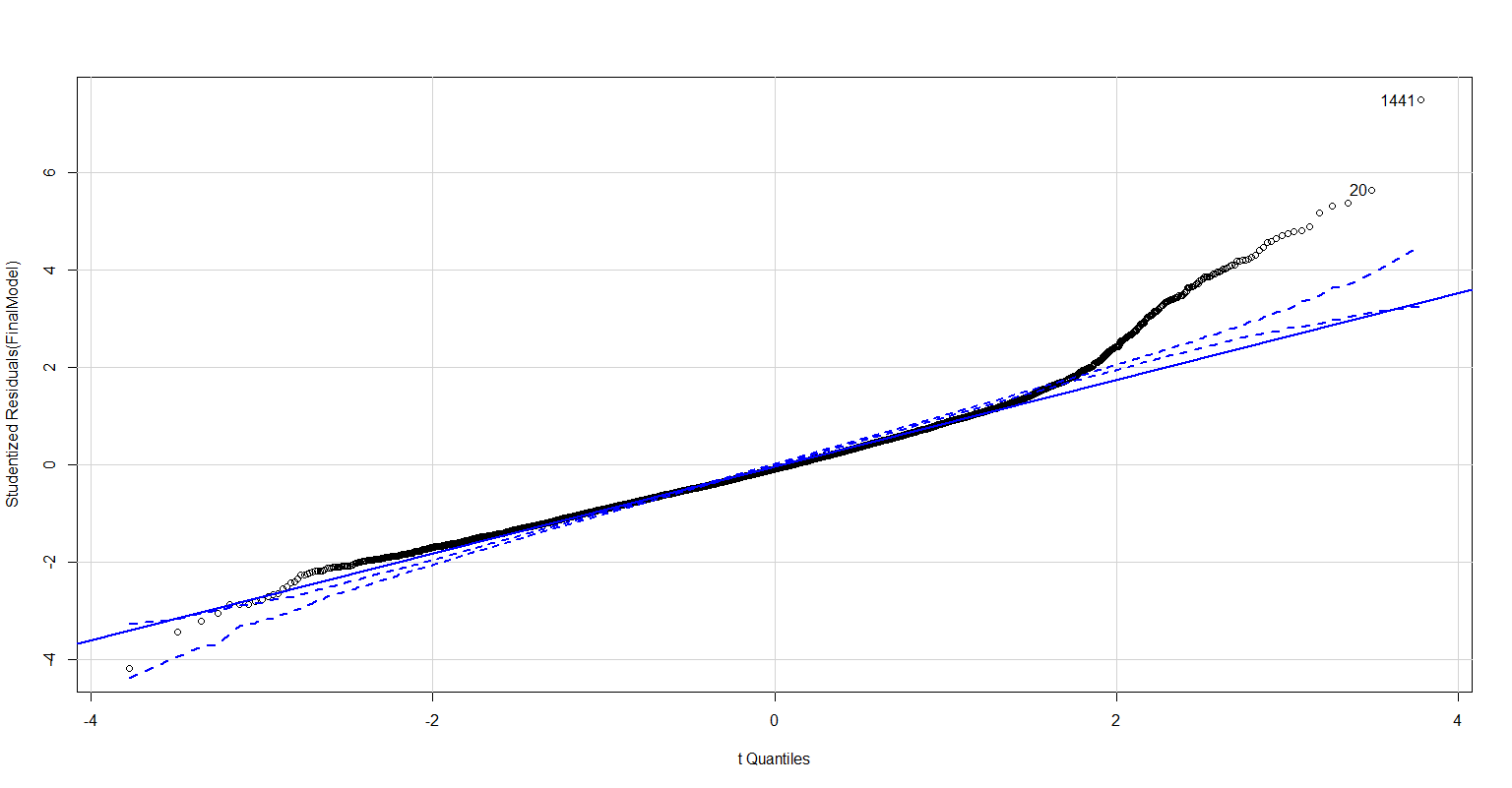
avPlots(computerModel, id.n=2, id.cex=0.8, col="lightblue")



install.packages("lattice")

library("lattice")

qqPlot(FinalModel)



## qqplot of studentized residuals, identified outliers with 1441,20 observation

library(MASS)

stepAIC(computerModel)

#Start: AIC=70264.61

#price ~ speed + hd + ram + screen + cd + multi + premium + ads + trend

# Df Sum of Sq RSS AIC

#<none> 469894852 70265

#- cd 1 2974407 472869259 70302

#- multi 1 6432377 476327228 70348

#- ads 1 12310139 482204991 70424

#- hd 1 60266517 530161369 71018

#- screen 1 70591458 540486310 71138

#- premium 1 129146521 599041373 71782

#- ram 1 156268216 626163068 72059

#- speed 1 191906458 661801310 72405

#- trend 1 513102713 982997564 74881

#Call:

# lm(formula = price ~ speed + hd + ram + screen + cd + multi +

# premium + ads + trend, data = ComputerData[-c(1440, 1701),])

#Coefficients:

# (Intercept) speed hd ram screen cd multi premium

#666.7050 9.3106 0.7787 48.4096 122.1142 59.6160 105.1261 -509.3978

#ads trend

#0.6539 -51.7387

**#Conclusion: Hence the Poly model transformation model prices 81.2% of the variation in price using explanatory variables**

**#speed + hd + ram + screen + cd + multi + premium + ads + trend**

**#ram is most significant with t-value of 22.375, followed by hd with t-value 14.32**

**#The least significant variable is trend.**