Que. Predict Price of the computer

Ans.

> computer <- read\_csv("Computer\_Data.csv")

summary(computer)

X1 price speed hd ram

Min. : 1 Min. : 949 Min. : 25.00 Min. : 80.0 Min. : 2.000

1st Qu.:1566 1st Qu.:1794 1st Qu.: 33.00 1st Qu.: 214.0 1st Qu.: 4.000

Median :3130 Median :2144 Median : 50.00 Median : 340.0 Median : 8.000

Mean :3130 Mean :2220 Mean : 52.01 Mean : 416.6 Mean : 8.287

3rd Qu.:4694 3rd Qu.:2595 3rd Qu.: 66.00 3rd Qu.: 528.0 3rd Qu.: 8.000

Max. :6259 Max. :5399 Max. :100.00 Max. :2100.0 Max. :32.000

screen cd multi premium

Min. :14.00 Length:6259 Length:6259 Length:6259

1st Qu.:14.00 Class :character Class :character Class :character

Median :14.00 Mode :character Mode :character Mode :character

Mean :14.61

3rd Qu.:15.00

Max. :17.00

ads trend

Min. : 39.0 Min. : 1.00

1st Qu.:162.5 1st Qu.:10.00

Median :246.0 Median :16.00

Mean :221.3 Mean :15.93

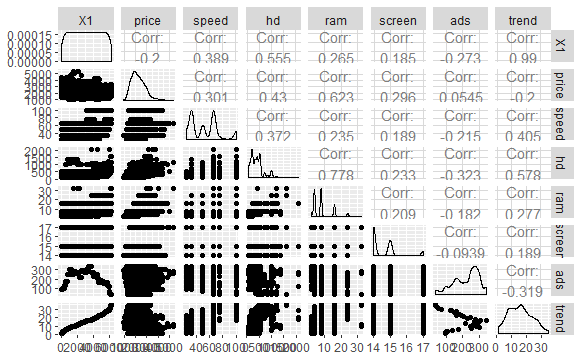
3rd Qu.:275.0 3rd Qu.:21.50

Max. :339.0 Max. :35.00

> sum(is.na(computer))

[1] 0

> ggpairs(computer1)



> cor(computer1)

X1 price speed hd ram

X1 1.0000000 -0.19992353 0.3885662 0.5554922 0.2653759

price -0.1999235 1.00000000 0.3009765 0.4302578 0.6227482

speed 0.3885662 0.30097646 1.0000000 0.3723041 0.2347605

hd 0.5554922 0.43025779 0.3723041 1.0000000 0.7777263

ram 0.2653759 0.62274824 0.2347605 0.7777263 1.0000000

screen 0.1846995 0.29604147 0.1890741 0.2328015 0.2089537

ads -0.2727142 0.05454047 -0.2152321 -0.3232220 -0.1816697

trend 0.9898184 -0.19998694 0.4054383 0.5777901 0.2768438

screen ads trend

X1 0.18469950 -0.27271422 0.9898184

price 0.29604147 0.05454047 -0.1999869

speed 0.18907412 -0.21523206 0.4054383

hd 0.23280153 -0.32322200 0.5777901

ram 0.20895374 -0.18166971 0.2768438

screen 1.00000000 -0.09391943 0.1886144

ads -0.09391943 1.00000000 -0.3185525

trend 0.18861444 -0.31855251 1.0000000

let us build the model to find the added variable plot function to identify whether variables are contributing for the output.

1. **Model 1**

> model.comp1 <- lm(price ~ .,data = computer1)

> summary(model.comp1)

Call:

lm(formula = price ~ ., data = computer1)

Residuals:

Min 1Q Median 3Q Max

-973.93 -194.75 -33.41 134.94 1959.60

Coefficients:

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

(Intercept) -155.64295 67.36445 -2.310 0.0209 \*

X1 0.11420 0.01621 7.047 2.03e-12 \*\*\*

speed 8.97725 0.20836 43.086 < 2e-16 \*\*\*

hd 0.73033 0.03094 23.604 < 2e-16 \*\*\*

ram 46.87530 1.18530 39.547 < 2e-16 \*\*\*

screen 126.32357 4.50429 28.045 < 2e-16 \*\*\*

ads 0.84856 0.05905 14.370 < 2e-16 \*\*\*

trend -73.77829 3.84792 -19.174 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

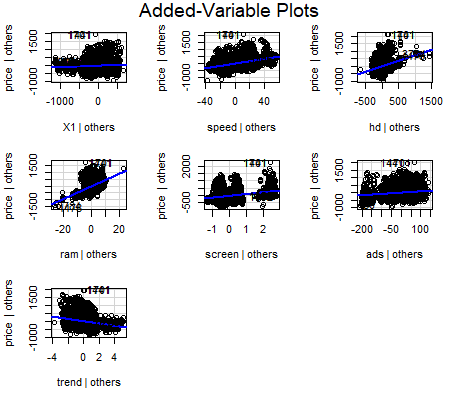
Residual standard error: 310.5 on 6251 degrees of freedom

Multiple R-squared: 0.7146, Adjusted R-squared: 0.7143

F-statistic: 2236 on 7 and 6251 DF, p-value: < 2.2e-16

**All the p values are significant but R sqrvalues are not upto the mark. So we will try to improve using some transformation techniques.**

> avPlots(model.comp)



1. **Model 2**

> model.comp2 <- lm(log(price) ~ .,data = computer1)

> summary(model.comp2)

Call:

lm(formula = log(price) ~ ., data = computer1)

Residuals:

Min 1Q Median 3Q Max

-0.56388 -0.08065 -0.00413 0.07157 0.53043

Coefficients:

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

(Intercept) 6.619e+00 2.972e-02 222.69 < 2e-16 \*\*\*

X1 4.805e-05 7.151e-06 6.72 1.98e-11 \*\*\*

speed 4.120e-03 9.193e-05 44.82 < 2e-16 \*\*\*

hd 3.243e-04 1.365e-05 23.75 < 2e-16 \*\*\*

ram 2.072e-02 5.230e-04 39.63 < 2e-16 \*\*\*

screen 5.549e-02 1.987e-03 27.92 < 2e-16 \*\*\*

ads 3.812e-04 2.605e-05 14.63 < 2e-16 \*\*\*

trend -3.226e-02 1.698e-03 -19.00 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.137 on 6251 degrees of freedom

Multiple R-squared: 0.7184, Adjusted R-squared: 0.7181

F-statistic: 2278 on 7 and 6251 DF, p-value: < 2.2e-16

**All variable p values are significant, Though R value improved as compared to previous model but not significant.**

1. **Model 3**

> summary(model.comp3)

Call:

lm(formula = sqrt(price) ~ ., data = computer1)

Residuals:

Min 1Q Median 3Q Max

-10.5927 -1.9700 -0.2385 1.5235 15.4014

Coefficients:

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

(Intercept) 21.9260322 0.6924475 31.665 < 2e-16 \*\*\*

X1 0.0011669 0.0001666 7.005 2.73e-12 \*\*\*

speed 0.0954305 0.0021417 44.558 < 2e-16 \*\*\*

hd 0.0076054 0.0003181 23.913 < 2e-16 \*\*\*

ram 0.4913381 0.0121839 40.327 < 2e-16 \*\*\*

screen 1.3120570 0.0463001 28.338 < 2e-16 \*\*\*

ads 0.0089269 0.0006070 14.707 < 2e-16 \*\*\*

trend -0.7661993 0.0395532 -19.371 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.191 on 6251 degrees of freedom

Multiple R-squared: 0.7224, Adjusted R-squared: 0.7221

F-statistic: 2324 on 7 and 6251 DF, p-value: < 2.2e-16

Still R sqr value not significant.

1. **Model 4**

> model.comp4 <- lm(log(price) ~ log(X1) + price + speed+ hd + ram + screen + log(ads) + trend,data = computer1)

> summary(model.comp4)

Call:

lm(formula = log(price) ~ log(X1) + price + speed + hd + ram +

screen + log(ads) + trend, data = computer1)

Residuals:

Min 1Q Median 3Q Max

-0.47040 -0.01348 0.01461 0.02753 0.05674

Coefficients:

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

(Intercept) 6.666e+00 1.159e-02 575.092 < 2e-16 \*\*\*

log(X1) 5.469e-03 1.946e-03 2.811 0.004957 \*\*

price 4.193e-04 1.726e-06 243.005 < 2e-16 \*\*\*

speed 3.632e-04 3.253e-05 11.166 < 2e-16 \*\*\*

hd 1.780e-05 4.397e-06 4.050 5.19e-05 \*\*\*

ram 1.067e-03 1.808e-04 5.900 3.82e-09 \*\*\*

screen 2.525e-03 6.517e-04 3.874 0.000108 \*\*\*

log(ads) -1.447e-03 2.360e-03 -0.613 0.539607

trend -1.990e-03 2.834e-04 -7.020 2.45e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04234 on 6250 degrees of freedom

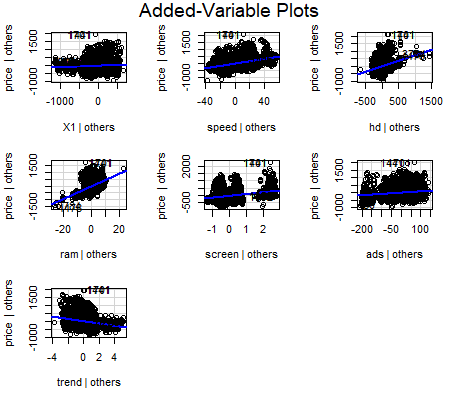
Multiple R-squared: 0.9731, Adjusted R-squared: 0.9731

F-statistic: 2.826e+04 on 8 and 6250 DF, p-value: < 2.2e-16

**Now the R sqr value is greatly improved and all p values are significant but ads is not significant.**

**As from the avPlot function it is clear that ads variable is not making contributions towards op variable as their seems no relation from the plot.**

> avPlots(model.comp)



1. **Model 5**

> model.comp5 <- lm(log(price) ~ log(X1) + price + speed+ hd + ram + screen + trend,data = computer1)

> summary(model.comp5)

Call:

lm(formula = log(price) ~ log(X1) + price + speed + hd + ram +

screen + trend, data = computer1)

Residuals:

Min 1Q Median 3Q Max

-0.47057 -0.01359 0.01459 0.02746 0.05708

Coefficients:

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

(Intercept) 6.664e+00 1.106e-02 602.543 < 2e-16 \*\*\*

log(X1) 4.485e-03 1.101e-03 4.075 4.66e-05 \*\*\*

price 4.193e-04 1.720e-06 243.729 < 2e-16 \*\*\*

speed 3.641e-04 3.250e-05 11.202 < 2e-16 \*\*\*

hd 1.818e-05 4.354e-06 4.175 3.02e-05 \*\*\*

ram 1.066e-03 1.808e-04 5.897 3.89e-09 \*\*\*

screen 2.526e-03 6.517e-04 3.877 0.000107 \*\*\*

trend -1.863e-03 1.936e-04 -9.621 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04234 on 6251 degrees of freedom

Multiple R-squared: 0.9731, Adjusted R-squared: 0.9731

F-statistic: 3.231e+04 on 7 and 6251 DF, p-value: < 2.2e-16

> sqrt(mean(model.comp2$residuals)^2)

[1] 2.800684e-19

**All P values are significant. R sqr value is significant. And RMSE is also significant.**

**Model Plots:**

> plot(model.comp2)

