Multiple Linear Regression

Dataset : 50\_Startups.csv

startup<-read.csv("50\_Startups.csv",stringsAsFactors = FALSE)

startup<-startup[-4]

pairs(startup)

#Correlation Matrix:###

cor(startup)

#Regression Model and Summary

model.startup<-lm(Profit~R.D.Spend+Administration+Marketing.Spend,data = startup)

summary(model.startup)

Output:

R.D.Spend Administration Marketing.Spend Profit

R.D.Spend 1.0000000 0.24195525 0.72424813 0.9729005

Administration 0.2419552 1.00000000 -0.03215388 0.2007166

Marketing.Spend 0.7242481 -0.03215388 1.00000000 0.7477657

Profit 0.9729005 0.20071657 0.74776572 1.0000000

|  |  |  |  |
| --- | --- | --- | --- |
|  | Multiple-R^2 | Adjusted R-squared | Significant Term |
| Profit ~ Marketing.Spend | 0.5592 | 0.55 |  |
| Profit ~ R.D.Spend + Marketing.Spend | 0.9505 | 0.9483 | R.D.spend |
| Profit~. | 0.9507 | 0.9475 | R.D.spend |

install.packages("car")

library(car)

#Variance Inflation Factor - Multi collinearity values

car::vif(model.startup)

Output:

Coefficients:

(Intercept) R.D.Spend Marketing.Spend

4.698e+04 7.966e-01 2.991e-02

VIF of R.D.Spend is high. Therefore, it is highly correlated.

##Subset selection

install.packages("MASS")

library(MASS)

stepAIC(model.startup)

Start: AIC=916.88

Profit ~ R.D.Spend + Administration + Marketing.Spend

Df Sum of Sq RSS AIC

- Administration 1 2.3539e+07 3.9444e+09 915.18

<none> 3.9209e+09 916.88

- Marketing.Spend 1 2.3349e+08 4.1543e+09 917.77

- R.D.Spend 1 2.7147e+10 3.1068e+10 1018.37

Step: AIC=915.18

Profit ~ R.D.Spend + Marketing.Spend

Df Sum of Sq RSS AIC

<none> 3.9444e+09 915.18

- Marketing.Spend 1 3.1165e+08 4.2560e+09 916.98

- R.D.Spend 1 3.1149e+10 3.5094e+10 1022.46

Call:

lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = startup)

Coefficients:

(Intercept) R.D.Spend Marketing.Spend

4.698e+04 7.966e-01 2.991e-02

AIC VALUE WHICH IS LESS WILL BE PROCEEDED.

install.packages("car")

library(car)

#Variance Inflation Factor

car::vif(model.pf)

#Diagnostic Plots:

#Residual Plots, QQ-Plos, Std. Residuals vs Fitted

plot(model.pf)

#Residuals vs Regressors

residualPlots(model.pf)

#Added Variable Plots

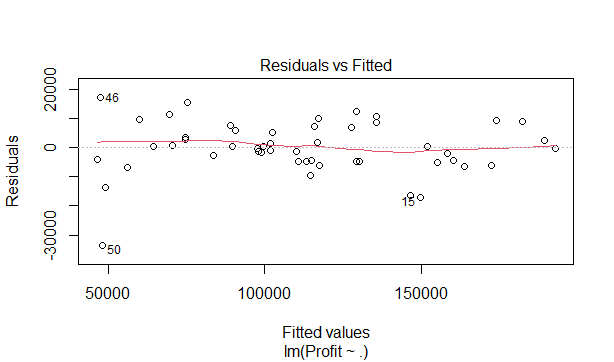
avPlots(model.pf)

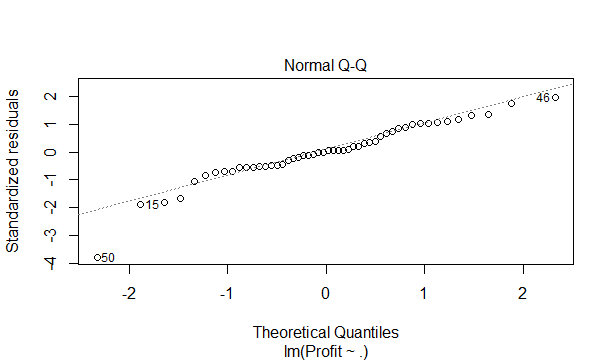
#QQ plots of studentized residuals

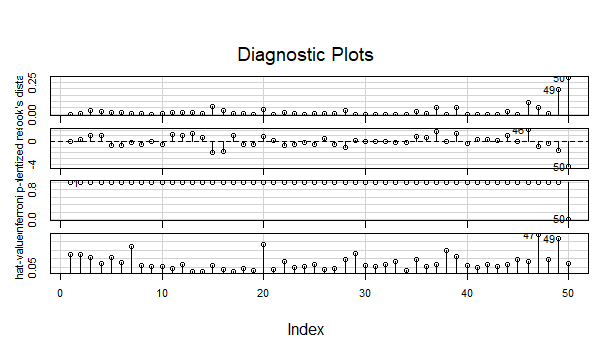
qqPlot(model.pf)

#Deletion Diagnostics

influenceIndexPlot(model.pf) # Index Plots of the influence measures







From this 50th observation has to be removed. Now after removing this , the model obtained

Multiple R-squared: 0.9613, Adjusted R-squared: 0.9587

df2<-newdf[-c(20,47,49),]

model2<-lm(Profit~.,data = df2)

summary(model2)

After removing 20,47,49th Observations, Multiple R-squared: 0.9623, Adjusted R-squared: 0.9597