Prediction Model for Profit of Startup\_50 Data

#loading data  
startup\_50<-read.csv("E:\\assignments data\\50\_Startups.csv")  
#dataset details and EDA  
attach(startup\_50)  
colnames(startup\_50)

## [1] "R.D.Spend" "Administration" "Marketing.Spend" "State"   
## [5] "Profit"

head(startup\_50)

## R.D.Spend Administration Marketing.Spend State Profit  
## 1 165349.2 136897.80 471784.1 New York 192261.8  
## 2 162597.7 151377.59 443898.5 California 191792.1  
## 3 153441.5 101145.55 407934.5 Florida 191050.4  
## 4 144372.4 118671.85 383199.6 New York 182902.0  
## 5 142107.3 91391.77 366168.4 Florida 166187.9  
## 6 131876.9 99814.71 362861.4 New York 156991.1

dim(startup\_50)

## [1] 50 5

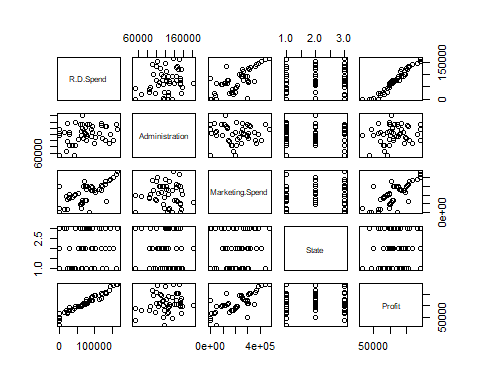
sum(is.null(startup\_50))

## [1] 0

library(plyr)  
startup\_50$State<-revalue(startup\_50$State, c("New York"="1", "California"="2", "Florida"="3"))  
startup\_50$State<-as.numeric(startup\_50$State)  
class(startup\_50$State)

## [1] "numeric"

# finding collinearity  
pairs(startup\_50)



# correlation matrix  
cor(startup\_50)

## R.D.Spend Administration Marketing.Spend State Profit  
## R.D.Spend 1.0000000 0.24195525 0.72424813 0.10468511 0.9729005  
## Administration 0.2419552 1.00000000 -0.03215388 0.01184720 0.2007166  
## Marketing.Spend 0.7242481 -0.03215388 1.00000000 0.07766961 0.7477657  
## State 0.1046851 0.01184720 0.07766961 1.00000000 0.1017963  
## Profit 0.9729005 0.20071657 0.74776572 0.10179631 1.0000000

# partial correlation  
library(corpcor)  
cor2pcor(cor(startup\_50))

## [,1] [,2] [,3] [,4] [,5]  
## [1,] 1.00000000 0.20880590 0.038920914 0.026971505 0.934484951  
## [2,] 0.20880590 1.00000000 -0.281913894 -0.013933444 -0.077271343  
## [3,] 0.03892091 -0.28191389 1.000000000 -0.001176456 0.237068057  
## [4,] 0.02697151 -0.01393344 -0.001176456 1.000000000 -0.002066896  
## [5,] 0.93448495 -0.07727134 0.237068057 -0.002066896 1.000000000

model<-lm(Profit~R.D.Spend+Administration+Marketing.Spend+State)  
summary(model)

##   
## Call:  
## lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +   
## State)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33504 -4736 90 6672 17338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.013e+04 6.885e+03 7.281 4.44e-09 \*\*\*  
## R.D.Spend 8.060e-01 4.641e-02 17.369 < 2e-16 \*\*\*  
## Administration -2.700e-02 5.223e-02 -0.517 0.608   
## Marketing.Spend 2.698e-02 1.714e-02 1.574 0.123   
## StateFlorida 1.988e+02 3.371e+03 0.059 0.953   
## StateNew York -4.189e+01 3.256e+03 -0.013 0.990   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9439 on 44 degrees of freedom  
## Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452   
## F-statistic: 169.9 on 5 and 44 DF, p-value: < 2.2e-16

# here Administation and Marketing.spend has greater than 0.05   
model.adm<-lm(Profit~Administration+Marketing.Spend)  
summary(model.adm)

##   
## Call:  
## lm(formula = Profit ~ Administration + Marketing.Spend)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -82155 -12168 2836 13650 56472   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.022e+04 1.770e+04 1.143 0.2589   
## Administration 3.237e-01 1.312e-01 2.468 0.0173 \*   
## Marketing.Spend 2.488e-01 3.005e-02 8.281 9.73e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 25710 on 47 degrees of freedom  
## Multiple R-squared: 0.6097, Adjusted R-squared: 0.5931   
## F-statistic: 36.71 on 2 and 47 DF, p-value: 2.496e-10

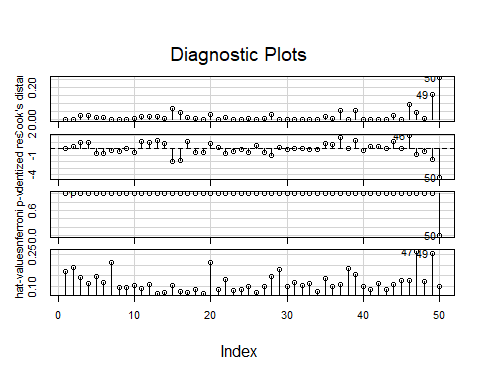
library(car)

## Loading required package: carData

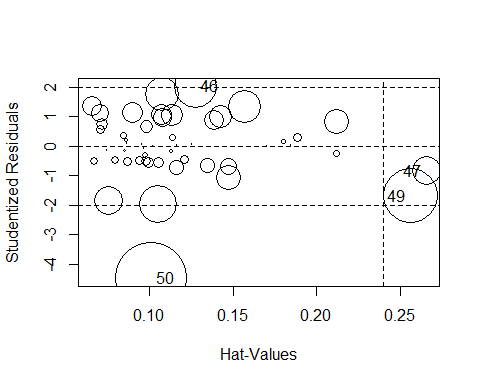
vif(model)

## GVIF Df GVIF^(1/(2\*Df))  
## R.D.Spend 2.495511 1 1.579719  
## Administration 1.177766 1 1.085249  
## Marketing.Spend 2.416797 1 1.554605  
## State 1.062673 2 1.015313

# vif<10 there is no multicollinearity  
influenceIndexPlot(model)



influencePlot(model)

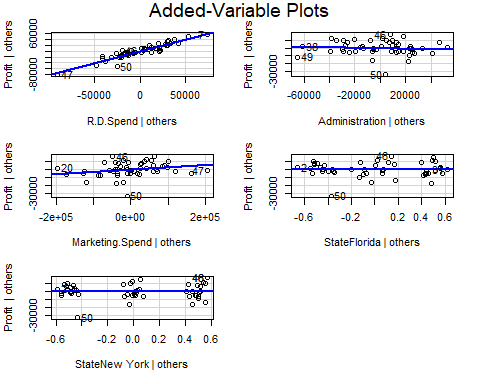


## StudRes Hat CookD  
## 46 2.0357210 0.1277290 0.09439478  
## 47 -0.8354542 0.2654200 0.04232333  
## 49 -1.6860294 0.2558868 0.15637613  
## 50 -4.4845939 0.1014896 0.26395944

# for better multiple R-square vakues removing influence variable   
model1<-lm(Profit~R.D.Spend+Administration+Marketing.Spend+State, data=startup\_50[-c(46,47,49,50)])  
summary(model1)

##   
## Call:  
## lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +   
## State, data = startup\_50[-c(46, 47, 49, 50)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33553 -4779 63 6595 17301   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.016e+04 7.322e+03 6.851 1.69e-08 \*\*\*  
## R.D.Spend 8.058e-01 4.576e-02 17.609 < 2e-16 \*\*\*  
## Administration -2.683e-02 5.160e-02 -0.520 0.606   
## Marketing.Spend 2.723e-02 1.663e-02 1.637 0.109   
## State -2.232e+01 1.610e+03 -0.014 0.989   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9334 on 45 degrees of freedom  
## Multiple R-squared: 0.9507, Adjusted R-squared: 0.9464   
## F-statistic: 217.2 on 4 and 45 DF, p-value: < 2.2e-16

avPlots(model)



#fitted multiple linear regression model  
model2<-lm(Profit ~.-Administration, data=startup\_50[-c(46,47,49,50)])  
summary(model2)

##   
## Call:  
## lm(formula = Profit ~ . - Administration, data = startup\_50[-c(46,   
## 47, 49, 50)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33655 -4633 -414 6473 17109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.700e+04 4.025e+03 11.677 2.35e-15 \*\*\*  
## R.D.Spend 7.966e-01 4.190e-02 19.013 < 2e-16 \*\*\*  
## Marketing.Spend 2.991e-02 1.569e-02 1.906 0.0628 .   
## State -1.076e+01 1.597e+03 -0.007 0.9947   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9260 on 46 degrees of freedom  
## Multiple R-squared: 0.9505, Adjusted R-squared: 0.9472   
## F-statistic: 294.1 on 3 and 46 DF, p-value: < 2.2e-16

#here multiple R-squared value is 0.9505 and Adjusted R-squared value is 0.9472  
#p value of marketing.spend is almost equal to 0.05  
predicted\_val<-predict(model2)  
predicted\_val

## 1 2 3 4 5 6 7 8   
## 192791.60 189787.25 181407.02 173432.01 171129.03 162869.75 158040.08 160456.90   
## 9 10 11 12 13 14 15 16   
## 152308.01 154355.59 135012.48 134650.17 129218.81 127823.34 150193.42 146022.78   
## 17 18 19 20 21 22 23 24   
## 117036.75 130819.12 128882.58 115805.85 116661.73 118373.53 114990.42 109886.09   
## 25 26 27 28 29 30 31 32   
## 112541.48 102623.47 110990.80 114967.86 103124.83 102429.48 99084.95 98303.51   
## 33 34 35 36 37 38 39 40   
## 98875.17 97600.37 90272.86 89765.18 75823.34 87984.18 68619.49 82934.86   
## 41 42 43 44 45 46 47 48   
## 75058.91 74112.96 70243.99 60378.27 65499.41 47817.31 56908.39 46985.10   
## 49 50   
## 47395.38 48336.14

par(mfrow=c(2,2))  
plot(model2)

