Prepare a prediction model for profit of 50\_startups data.

Do transformations for getting better predictions of profit and

make a table containing R^2 value for each prepared model.

R&D Spend -- Research and devolop spend in the past few years

Administration -- spend on administration in the past few years

Marketing Spend -- spend on Marketing in the past few years

State -- states from which data is collected

Profit -- profit of each state in the past few years

a=read.csv(file.choose())

> View(a)

> str(a)

'data.frame': 50 obs. of 5 variables:

$ R.D.Spend : num 165349 162598 153442 144372 142107 ...

$ Administration : num 136898 151378 101146 118672 91392 ...

$ Marketing.Spend: num 471784 443899 407935 383200 366168 ...

$ State : chr "New York" "California" "Florida" "New York" ...

$ Profit : num 192262 191792 191050 182902 166188 ...

> #creating dummy variables for state

> a$State <-as.integer(factor(a$State,levels = c('New York','California','Florida'),labels=c(1,2,3)))

> View(a)

> summary(a)#1st moment business decision

R.D.Spend Administration Marketing.Spend State Profit

Min. : 0 Min. : 51283 Min. : 0 Min. :1.00 Min. : 14681

1st Qu.: 39936 1st Qu.:103731 1st Qu.:129300 1st Qu.:1.00 1st Qu.: 90139

Median : 73051 Median :122700 Median :212716 Median :2.00 Median :107978

Mean : 73722 Mean :121345 Mean :211025 Mean :1.98 Mean :112013

3rd Qu.:101603 3rd Qu.:144842 3rd Qu.:299469 3rd Qu.:3.00 3rd Qu.:139766

Max. :165349 Max. :182646 Max. :471784 Max. :3.00 Max. :192262

> attach(a)

> class(a)

[1] "data.frame"

> library(psych)

> describe(a)#2nd ,3rd & 4th moment business decision

vars n mean sd median trimmed mad min

R.D.Spend 1 50 73721.62 45902.26 73051.08 72883.87 47053.05 0.00

Administration 2 50 121344.64 28017.80 122699.80 123127.46 31868.81 51283.14

Marketing.Spend 3 50 211025.10 122290.31 212716.24 211200.78 128222.13 0.00

State 4 50 1.98 0.82 2.00 1.98 1.48 1.00

Profit 5 50 112012.64 40306.18 107978.19 111727.66 39346.94 14681.40

max range skew kurtosis se

R.D.Spend 165349.2 165349.2 0.15 -0.89 6491.56

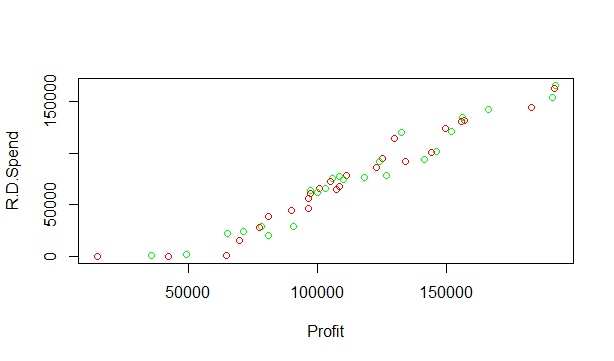
Administration 182645.6 131362.4 -0.46 -0.04 3962.32

Marketing.Spend 471784.1 471784.1 -0.04 -0.81 17294.46

State 3.0 2.0 0.04 -1.54 0.12

Profit 192261.8 177580.4 0.02 -0.29 5700.15

> plot(Profit,R.D.Spend,col=c("green","red"))



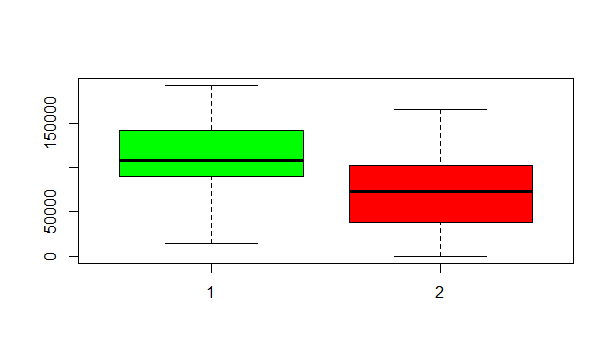
>plot(Profit,Marketing.Spend,col=c("green","red"))



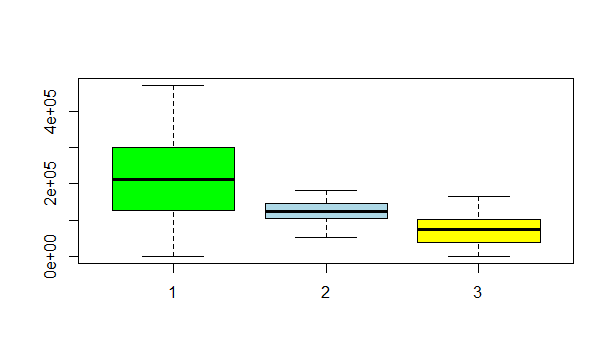
> plot(Profit,Administration,col=c("green","red"))



> boxplot(Profit,R.D.Spend,col=c("green","red"))



> boxplot(Marketing.Spend,Administration,R.D.Spend,col=c("green","lightblue","yellow"))



> pairs(a,col="red")

> cor(a)

R.D.Spend Administration Marketing.Spend State Profit

R.D.Spend 1.00000000 0.241955245 0.72424813 0.037929529 0.97290047

Administration 0.24195525 1.000000000 -0.03215388 0.003025813 0.20071657

Marketing.Spend 0.72424813 -0.032153875 1.00000000 0.137777487 0.74776572

State 0.03792953 0.003025813 0.13777749 1.000000000 0.04847097

Profit 0.97290047 0.200716568 0.74776572 0.048470972 1.00000000

>library(corpcor)

>cor2pcor(cor(a[,-c(4)]))

[,1] [,2] [,3] [,4]

[1,] 1.00000000 0.20852619 0.03890336 0.93477127

[2,] 0.20852619 1.00000000 -0.28192506 -0.07725021

[3,] 0.03890336 -0.28192506 1.00000000 0.23707116

[4,] 0.93477127 -0.07725021 0.23707116 1.00000000

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

> summary(model) #R^2=95.07

Call:

lm(formula = Profit ~ Marketing.Spend + R.D.Spend + Administration +

State, data = a)

Residuals:

Min 1Q Median 3Q Max

-33546 -4709 109 6638 17374

Coefficients:

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

(Intercept) 4.993e+04 7.183e+03 6.951 1.2e-08 \*\*\*

Marketing.Spend 2.703e-02 1.687e-02 1.602 0.116

R.D.Spend 8.061e-01 4.588e-02 17.569 < 2e-16 \*\*\*

Administration -2.699e-02 5.164e-02 -0.523 0.604

State 1.185e+02 1.649e+03 0.072 0.943

---

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

Residual standard error: 9334 on 45 degrees of freedom

Multiple R-squared: 0.9508, Adjusted R-squared: 0.9464

F-statistic: 217.2 on 4 and 45 DF, p-value: < 2.2e-16

> rmse <- mean(model$residuals^2)^.5 #8854.34

> pred <- predict(model,a)

> cor(pred,Profit) #0.9750

[1] 0.9750649

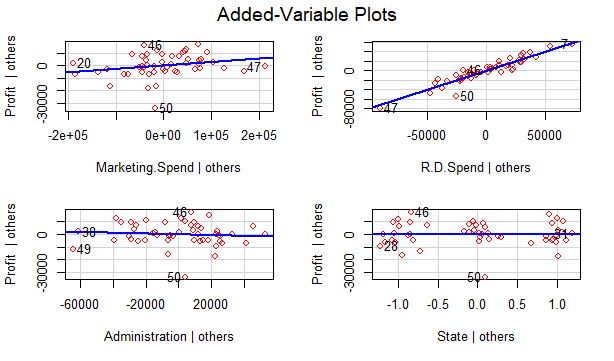
> library(car)

> vif(model)

Marketing.Spend R.D.Spend Administration State

2.392965 2.494486 1.177595 1.029971

> avPlots(model,col="red")



> influencePlot(model,col="red")

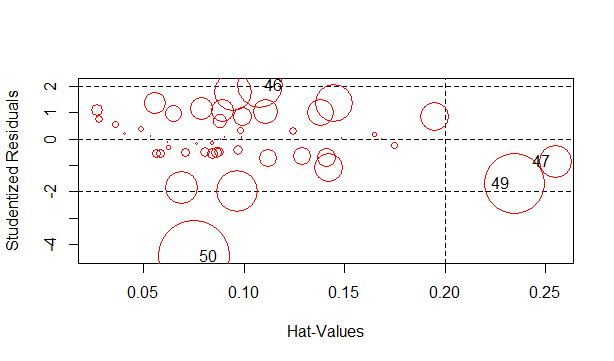
StudRes Hat CookD

46 2.0386686 0.10785323 0.09390326

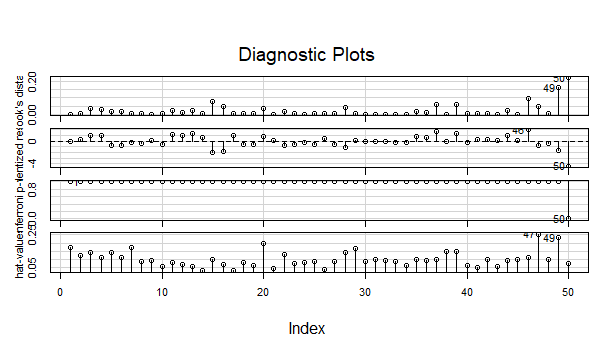
47 -0.8357182 0.25494143 0.04811938

49 -1.6749535 0.23451409 0.16526587

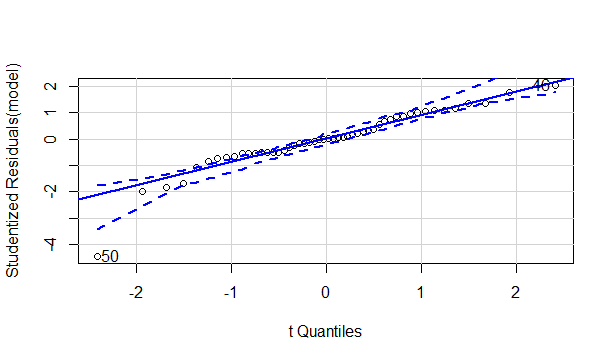
50 -4.4496953 0.07508785 0.22675297



> influenceIndexPlot(model)



> qqPlot(model)



[1] 46 50

> vif(model)

Marketing.Spend R.D.Spend Administration State

2.392965 2.494486 1.177595 1.029971

> #modelinfluence plot

> model2<- lm(Profit~R.D.Spend+Marketing.Spend+Administration+State,data = a[-c(46,50),])

> summary(model2) #r^2=0.9635

Call:

lm(formula = Profit ~ R.D.Spend + Marketing.Spend + Administration +

State, data = a[-c(46, 50), ])

Residuals:

Min 1Q Median 3Q Max

-17359.4 -4950.2 -479.9 6078.2 13770.5

Coefficients:

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

(Intercept) 5.019e+04 5.905e+03 8.500 9.39e-11 \*\*\*

R.D.Spend 7.922e-01 3.773e-02 20.995 < 2e-16 \*\*\*

Marketing.Spend 2.712e-02 1.372e-02 1.976 0.0546 .

Administration -2.669e-02 4.192e-02 -0.637 0.5277

State 6.589e+02 1.353e+03 0.487 0.6288

---

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

Residual standard error: 7565 on 43 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9601

F-statistic: 284 on 4 and 43 DF, p-value: < 2.2e-16

> rmse2 <- mean(model2$residuals^2)^.5 #7160

> pred2 <- predict(model2,a)

> cor(pred2,a$Profit) #0.9748

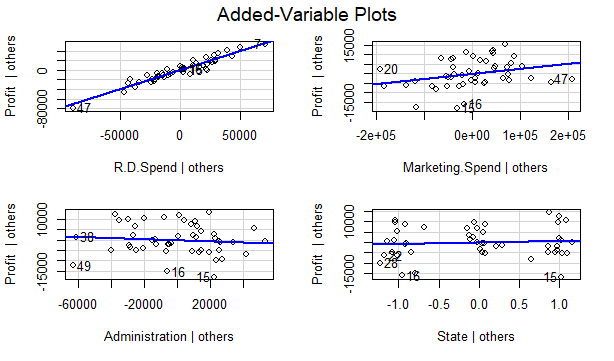
[1] 0.974998

> vif(model2)

R.D.Spend Marketing.Spend Administration State

2.290271 2.167464 1.180137 1.023687

> avPlots(model2)



> modelrd <- lm(Profit~R.D.Spend+Marketing.Spend,data = a[-c(46,50),])

> summary(modelrd) #R^2=0.963

Call:

lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = a[-c(46,

50), ])

Residuals:

Min 1Q Median 3Q Max

-17299.7 -4489.1 -786.7 5708.1 13813.8

Coefficients:

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

(Intercept) 4.834e+04 2.378e+03 20.325 <2e-16 \*\*\*

R.D.Spend 7.812e-01 3.404e-02 22.949 <2e-16 \*\*\*

Marketing.Spend 3.064e-02 1.273e-02 2.408 0.0202 \*

---

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

Residual standard error: 7448 on 45 degrees of freedom

Multiple R-squared: 0.963, Adjusted R-squared: 0.9614

F-statistic: 585.7 on 2 and 45 DF, p-value: < 2.2e-16

> rmserd <- mean(modelrd$residuals^2)^.5 #7211

> predrd <- predict(modelrd,a)

> cor(predrd,Profit) #0.9747

[1] 0.9749069

> #transformation method

> #exp model

> log=log(a[,5])

> a2 <- data.frame(log,a[,-5])

> View(a2)

> head(a2)

log R.D.Spend Administration Marketing.Spend State

1 12.16661 165349.2 136897.80 471784.1 1

2 12.16417 162597.7 151377.59 443898.5 2

3 12.16029 153441.5 101145.55 407934.5 3

4 12.11671 144372.4 118671.85 383199.6 1

5 12.02087 142107.3 91391.77 366168.4 3

6 11.96394 131876.9 99814.71 362861.4 1

> model3 <- lm(log~.,data = a2)

> summary(model3) #r^2 0.7619

Call:

lm(formula = log ~ ., data = a2)

Residuals:

Min 1Q Median 3Q Max

-1.29544 -0.06047 0.06779 0.10920 0.23271

Coefficients:

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

(Intercept) 1.082e+01 1.815e-01 59.651 < 2e-16 \*\*\*

R.D.Spend 8.147e-06 1.159e-06 7.029 9.24e-09 \*\*\*

Administration 2.959e-07 1.305e-06 0.227 0.822

Marketing.Spend 3.069e-07 4.261e-07 0.720 0.475

State 8.568e-03 4.167e-02 0.206 0.838

---

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

Residual standard error: 0.2358 on 45 degrees of freedom

Multiple R-squared: 0.7619, Adjusted R-squared: 0.7408

F-statistic: 36.01 on 4 and 45 DF, p-value: 1.714e-13

> rmse3 <- mean(model3$residuals^2)^.5 #0.2237

> pred3 <- predict(model3,a2)

> cor(pred3,a2$log) #0.87

[1] 0.8728934

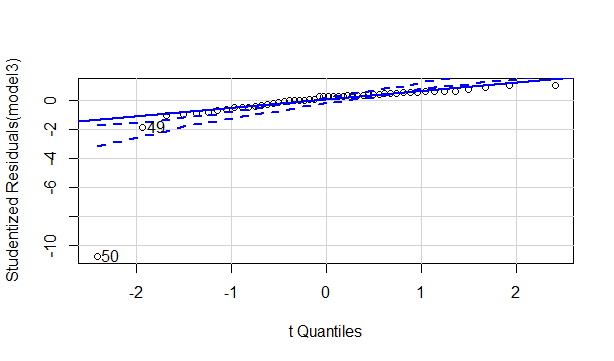
> vif(model3)

R.D.Spend Administration Marketing.Spend State

2.494486 1.177595 2.392965 1.029971

> qqPlot(model3)

[1] 49 50



**model 2 have higher R^2 value and accuracy**

**model 3 is best model with less rmse**

> plot(model2)

