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.

**Ans :**

**R Code :**

## Multi Linear Regression

########## 50 Startups Data Set #########

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

startup1 <- read.csv('D:\\Data Science\\Excelr\\Assignments\\Assignment\\Multi Linear Regression\\50\_Startups.csv')

startup <- startup1[,-c(4)]

#Model Building

#Regression Model and Summary

model.startup<-lm(Profit~.,data = startup)

summary(model.startup)

car::vif(model.startup)

#Diagnostic Plots:

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

plot(model.startup)

#Residuals vs Regressors

library(car)

residualPlots(model.startup)

#Added Variable Plots

avPlots(model.startup)

#QQ plots of studentized residuals

qqPlot(model.startup)

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

####Iteration 1

#Remove 50th observation

startup1<-startup[-50,]

model1<-lm(Profit~.,data = startup1)

summary(model1)

car::vif(model1)

plot(model1)

residualPlots(model1)

qqPlot(model1)

influenceIndexPlot(model1)

#iteration2

startup2<-startup[-c(50,49,47),]

model2<-lm(Profit~.,data = startup2)

summary(model2)

car::vif(model2)

plot(model2)

residualPlots(model2)

qqPlot(model2)

influenceIndexPlot(model2)

#iteration3

startup3<-startup[-c(50,49,47,46,15),]

model3<-lm(Profit~.,data = startup3)

summary(model3)

car::vif(model3)

plot(model3)

residualPlots(model3)

qqPlot(model3)

influenceIndexPlot(model3)

#iteration4

startup4<-startup[-c(50,49,47,46,15,37,28),]

model4<-lm(Profit~.,data = startup4)

summary(model4)

car::vif(model4)

plot(model4)

residualPlots(model4)

qqPlot(model4)

influenceIndexPlot(model4)

#iteration5

startup5<-startup[-c(50,49,47,46,15,37,28,20),]

model5<-lm(Profit~.,data = startup5)

summary(model5)

car::vif(model5)

plot(model5)

residualPlots(model5)

qqPlot(model5)

influenceIndexPlot(model5)

**Results :**

> #Regression Model and Summary

> model.startup<-lm(Profit~.,data = startup)

> summary(model.startup)

Call:

lm(formula = Profit ~ ., data = startup)

Residuals:

Min 1Q Median 3Q Max

-33534 -4795 63 6606 17275

Coefficients:

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

(Intercept) 5.012e+04 6.572e+03 7.626 1.06e-09 \*\*\*

R.D.Spend 8.057e-01 4.515e-02 17.846 < 2e-16 \*\*\*

Administration -2.682e-02 5.103e-02 -0.526 0.602

Marketing.Spend 2.723e-02 1.645e-02 1.655 0.105

---

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

Residual standard error: 9232 on 46 degrees of freedom

Multiple R-squared: 0.9507, Adjusted R-squared: 0.9475

F-statistic: 296 on 3 and 46 DF, p-value: < 2.2e-16

> car::vif(model.startup)

R.D.Spend Administration Marketing.Spend

2.468903 1.175091 2.326773

> #Residuals vs Regressors

> residualPlots(model.startup)

Test stat Pr(>|Test stat|)

R.D.Spend -0.7210 0.4746

Administration -0.4176 0.6782

Marketing.Spend -0.3680 0.7146

Tukey test -0.6503 0.5155

> #QQ plots of studentized residuals

> qqPlot(model.startup)

[1] 46 50

> ####Iteration 1

> #Remove 50th observation

> startup1<-startup[-50,]

> model1<-lm(Profit~.,data = startup1)

> summary(model1)

Call:

lm(formula = Profit ~ ., data = startup1)

Residuals:

Min 1Q Median 3Q Max

-16692.8 -4862.3 -993.9 6135.9 14468.8

Coefficients:

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

(Intercept) 5.238e+04 5.543e+03 9.451 2.96e-12 \*\*\*

R.D.Spend 7.830e-01 3.825e-02 20.470 < 2e-16 \*\*\*

Administration -2.220e-02 4.287e-02 -0.518 0.6071

Marketing.Spend 2.523e-02 1.382e-02 1.825 0.0746 .

---

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

Residual standard error: 7754 on 45 degrees of freedom

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

F-statistic: 372.8 on 3 and 45 DF, p-value: < 2.2e-16

> car::vif(model1)

R.D.Spend Administration Marketing.Spend

2.377754 1.175172 2.239957

> qqPlot(model1)

[1] 15 49

> #iteration2

> startup2<-startup[-c(50,49,47),]

> model2<-lm(Profit~.,data = startup2)

> summary(model2)

Call:

lm(formula = Profit ~ ., data = startup2)

Residuals:

Min 1Q Median 3Q Max

-15877 -4779 -1557 6163 12489

Coefficients:

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

(Intercept) 5.823e+04 5.905e+03 9.860 1.32e-12 \*\*\*

R.D.Spend 7.642e-01 4.086e-02 18.704 < 2e-16 \*\*\*

Administration -5.319e-02 4.424e-02 -1.202 0.2358

Marketing.Spend 2.509e-02 1.480e-02 1.696 0.0972 .

---

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

Residual standard error: 7289 on 43 degrees of freedom

Multiple R-squared: 0.9614, Adjusted R-squared: 0.9587

F-statistic: 357.1 on 3 and 43 DF, p-value: < 2.2e-16

> car::vif(model2)

R.D.Spend Administration Marketing.Spend

2.708100 1.231713 2.684378

> qqPlot(model2)

[1] 15 16

> #iteration3

> startup3<-startup[-c(50,49,47,46,15),]

> model3<-lm(Profit~.,data = startup3)

> summary(model3)

Call:

lm(formula = Profit ~ ., data = startup3)

Residuals:

Min 1Q Median 3Q Max

-16563 -5312 -1416 6091 12968

Coefficients:

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

(Intercept) 5.504e+04 5.545e+03 9.926 1.82e-12 \*\*\*

R.D.Spend 7.795e-01 3.802e-02 20.500 < 2e-16 \*\*\*

Administration -3.996e-02 4.124e-02 -0.969 0.3383

Marketing.Spend 2.701e-02 1.372e-02 1.969 0.0557 .

---

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

Residual standard error: 6725 on 41 degrees of freedom

Multiple R-squared: 0.967, Adjusted R-squared: 0.9646

F-statistic: 400.8 on 3 and 41 DF, p-value: < 2.2e-16

> car::vif(model3)

R.D.Spend Administration Marketing.Spend

2.508026 1.214146 2.509105

> qqPlot(model3)

16 37

15 36

> #iteration4

> startup4<-startup[-c(50,49,47,46,15,37,28),]

> model4<-lm(Profit~.,data = startup4)

> summary(model4)

Call:

lm(formula = Profit ~ ., data = startup4)

Residuals:

Min 1Q Median 3Q Max

-16722 -5304 -1090 5080 11663

Coefficients:

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

(Intercept) 5.426e+04 5.300e+03 10.238 1.31e-12 \*\*\*

R.D.Spend 7.834e-01 3.881e-02 20.188 < 2e-16 \*\*\*

Administration -3.981e-02 3.983e-02 -0.999 0.324

Marketing.Spend 2.878e-02 1.397e-02 2.061 0.046 \*

---

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

Residual standard error: 6338 on 39 degrees of freedom

Multiple R-squared: 0.9717, Adjusted R-squared: 0.9695

F-statistic: 445.7 on 3 and 39 DF, p-value: < 2.2e-16

> car::vif(model4)

R.D.Spend Administration Marketing.Spend

2.840626 1.273056 2.840575

> qqPlot(model4)

13 16

13 15

> #iteration5

> startup5<-startup[-c(50,49,47,46,15,37,28,20),]

> model5<-lm(Profit~.,data = startup5)

> summary(model5)

Call:

lm(formula = Profit ~ ., data = startup5)

Residuals:

Min 1Q Median 3Q Max

-16249.8 -5184.1 -180.2 4319.6 11846.7

Coefficients:

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

(Intercept) 5.349e+04 5.297e+03 10.100 2.59e-12 \*\*\*

R.D.Spend 7.650e-01 4.123e-02 18.554 < 2e-16 \*\*\*

Administration -3.814e-02 3.956e-02 -0.964 0.3412

Marketing.Spend 3.721e-02 1.540e-02 2.416 0.0206 \*

---

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

Residual standard error: 6291 on 38 degrees of freedom

Multiple R-squared: 0.9728, Adjusted R-squared: 0.9706

F-statistic: 452.7 on 3 and 38 DF, p-value: < 2.2e-16

> car::vif(model5)

R.D.Spend Administration Marketing.Spend

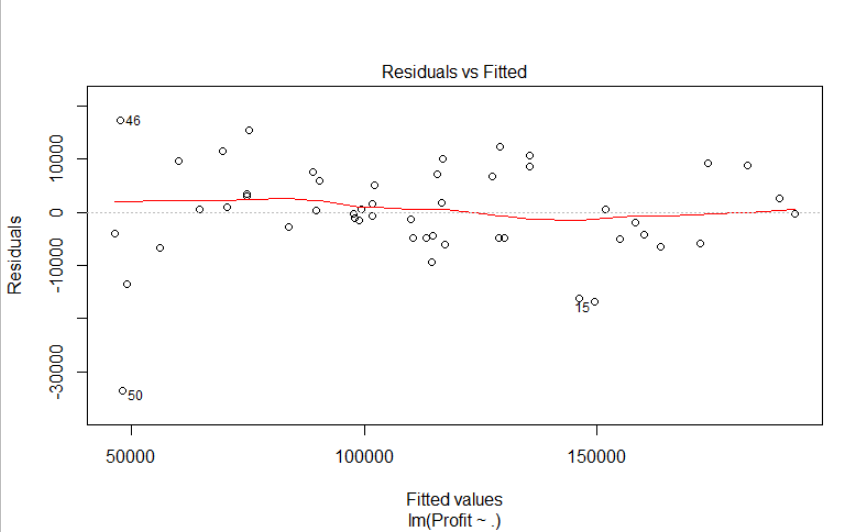
3.252322 1.234354 3.212198

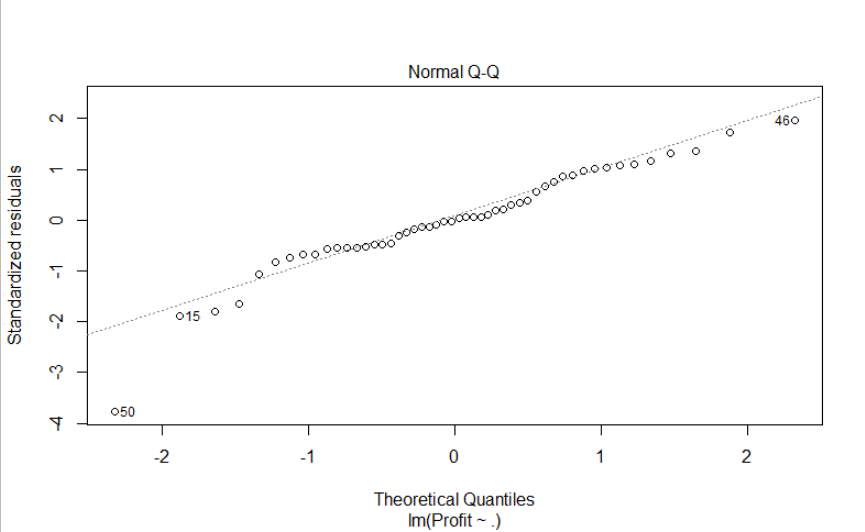
> qqPlot(model5)

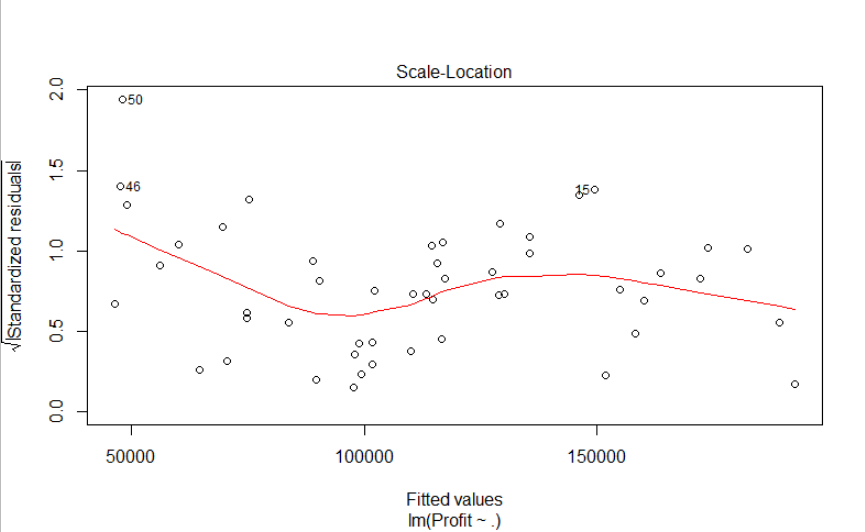
13 16

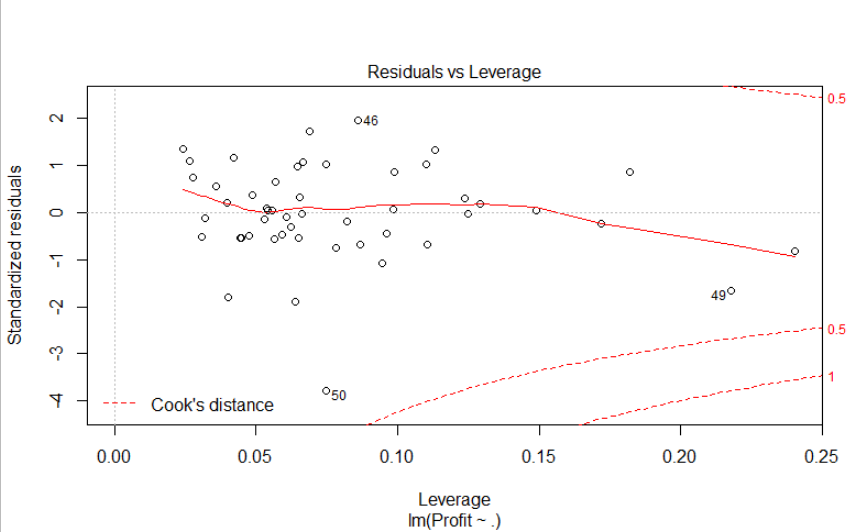
13 15

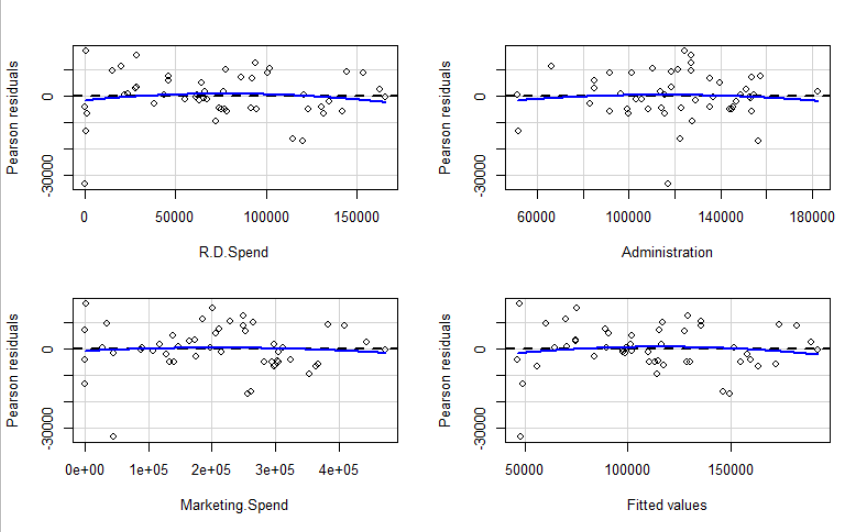
**Plots :**

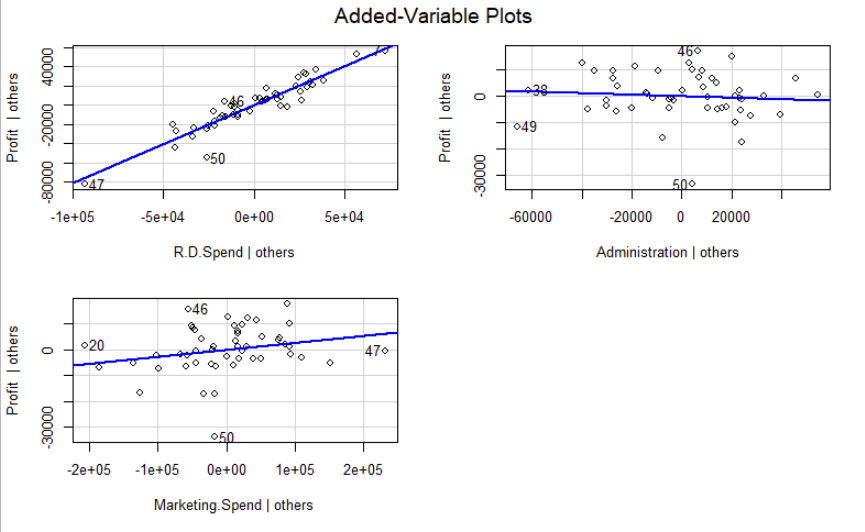


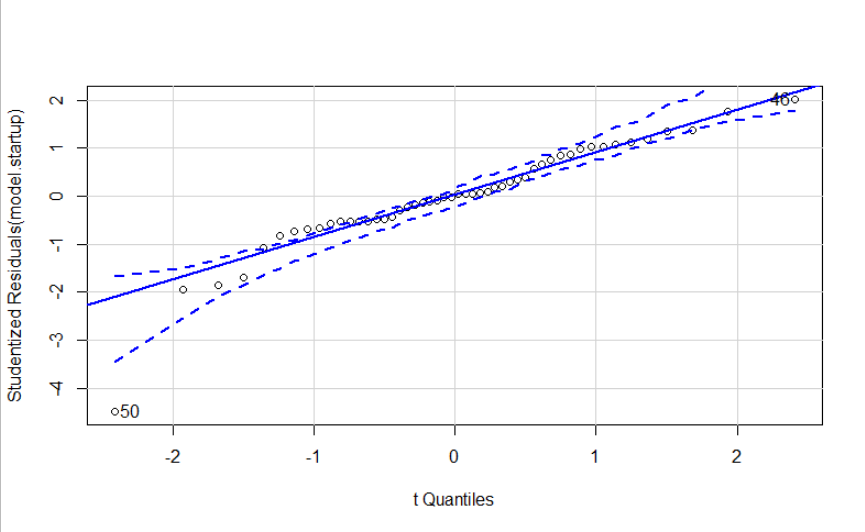


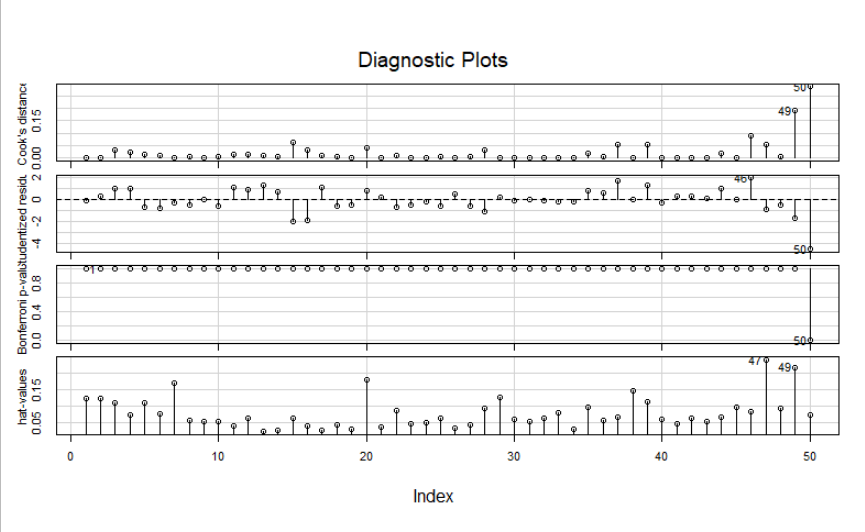












**Inference :**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No. | Model No. | Multiple R-squared | Adjusted R-squared |
| 1 | Model.startup | 0.9507 | 0.9475 |
| 2 | Model1 | 0.9613 | 0.9587 |
| 3 | Model2 | 0.9614 | 0.9587 |
| 4 | Model3 | 0.967 | 0.9646 |
| 5 | Model4 | 0.9717 | 0.9695 |
| 6 | Model5 | 0.9728 | 0.9706 |

R.D.Spend and Marketing.spend were some of the important features found.