Que. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Output Y | Input X | Model | R^2 |
| Profit | R&D Spend, Administration , Marketing Spend | lm(Profit ~ . , data = startup) | 0.9614 |
| Profit | R&D Spend, Administration , Marketing Spend | lm(Profit ~ sqrt(`R&D Spend`) + sqrt(Administration) + sqrt(`Marketing Spend`),data = startup1) | 0.88 |
| log(Profit) | R&D Spend, Administration , Marketing Spend | lm(log(Profit) ~ sqrt(`R&D Spend`) + log(Administration) + sqrt(`Marketing Spend`),data = startup1) | 0.9476 |
| Profit | R&D Spend , Marketing Spend | lm(Profit ~ . , data = startup) | 0.9585 |

Ans:

> X50\_Startups <- read\_csv("50\_Startups.csv")

> summary(X50\_Startups)

R&D Spend Administration Marketing Spend State Profit

Min. : 0 Min. : 51283 Min. : 0 Length:50 Min. : 14681

1st Qu.: 39936 1st Qu.:103731 1st Qu.:129300 Class :character 1st Qu.: 90139

Median : 73051 Median :122700 Median :212716 Mode :character Median :107978

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

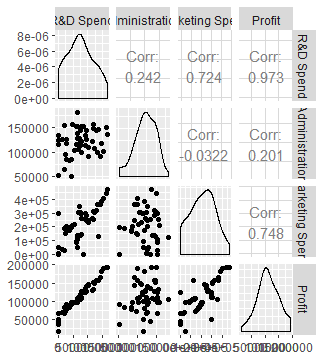
3rd Qu.:101603 3rd Qu.:144842 3rd Qu.:299469 3rd Qu.:139766

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

State isn’t relevant data, so we will drop the column.

> startup <- Startups[c("R&D Spend", 'Administration', "Marketing Spend", "Profit")]

> ggpairs(startup)



> cor(startup)

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

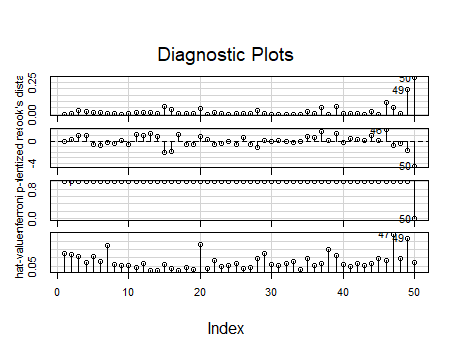
**From above plot and correlation function input variable R&D spend and Administration shows a good correlation.**

**We will begin with building a model to perform necessary transformation.**

**1. Trial 1**

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

> influenceIndexPlot(model.startup)



> influencePlot(model.startup)

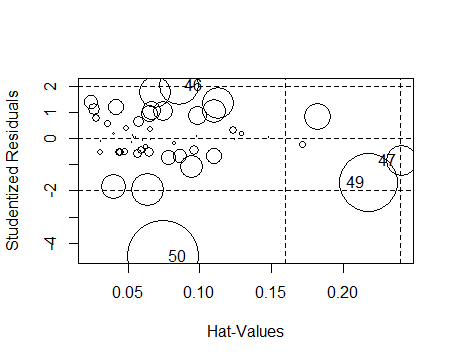
StudRes Hat CookD

46 2.0220730 0.08617007 0.09032342

47 -0.8268684 0.24060165 0.05453034

49 -1.6861241 0.21801940 0.19052744

50 -4.4961657 0.07477116 0.28808229



**From above Influence plot observation no. 50, 49 and 47 are influencing the entire model relationship. So we will discard these values and build new model.**

**2. Trial 2**

> startup1 <- startup[-c(50,49,47),]

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

> summary(model.startup1)

Call:

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

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

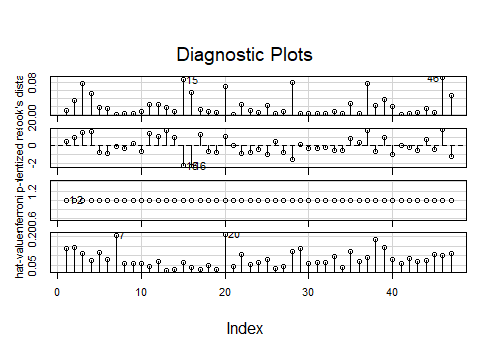
> vif(model.startup1)

`R&D Spend` Administration `Marketing Spend`

2.708100 1.231713 2.684378

**All the vif values are significant.**

> influenceIndexPlot(model.startup1)



> influencePlot(model.startup1)

StudRes Hat CookD

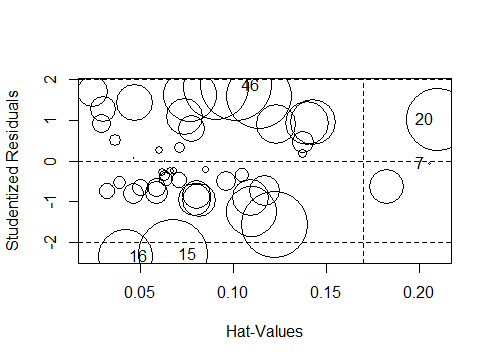
7 -0.05397278 0.20538367 0.0001927024

15 -2.29357874 0.06790622 0.0871740802

16 -2.33792344 0.04188306 0.0541137175

20 1.03054686 0.21014766 0.0705388088

46 1.85902293 0.10199893 0.0928337321



**From above influence plots it is clear that there are no influencing outliers.**

**So we will finalize our model for transformation technique.**

**Transformation Trial:**

**1. Trial 1**

> model.startup1 <- lm(Profit ~ sqrt(`R&D Spend`) + sqrt(Administration) + sqrt(`Marketing Spend`),data = startup1)

> summary(model.startup1)

Call:

lm(formula = Profit ~ sqrt(`R&D Spend`) + sqrt(Administration) +

sqrt(`Marketing Spend`), data = startup1)

Residuals:

Min 1Q Median 3Q Max

-14597 -10148 -3771 6188 35819

Coefficients:

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

(Intercept) 23858.068 19216.265 1.242 0.221

sqrt(`R&D Spend`) 359.489 33.101 10.860 6.65e-14 \*\*\*

sqrt(Administration) -18.592 53.001 -0.351 0.727

sqrt(`Marketing Spend`) 9.864 18.496 0.533 0.597

---

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

Residual standard error: 12850 on 43 degrees of freedom

Multiple R-squared: 0.88, Adjusted R-squared: 0.8717

F-statistic: 105.2 on 3 and 43 DF, p-value: < 2.2e-16

**P values not significant.**

**Trial 2**

> model.startup1 <- lm(log(Profit) ~ sqrt(`R&D Spend`) + log(Administration) + sqrt(`Marketing Spend`),data = startup1)

> summary(model.startup1)

Call:

lm(formula = log(Profit) ~ sqrt(`R&D Spend`) + log(Administration) +

sqrt(`Marketing Spend`), data = startup1)

Residuals:

Min 1Q Median 3Q Max

-0.113839 -0.051537 -0.006665 0.052785 0.278643

Coefficients:

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

(Intercept) 11.0488886 0.5981008 18.473 <2e-16 \*\*\*

sqrt(`R&D Spend`) 0.0033286 0.0001960 16.984 <2e-16 \*\*\*

log(Administration) -0.0303801 0.0512144 -0.593 0.556

sqrt(`Marketing Spend`) 0.0001040 0.0001091 0.953 0.346

---

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

Residual standard error: 0.07617 on 43 degrees of freedom

Multiple R-squared: 0.9476, Adjusted R-squared: 0.9439

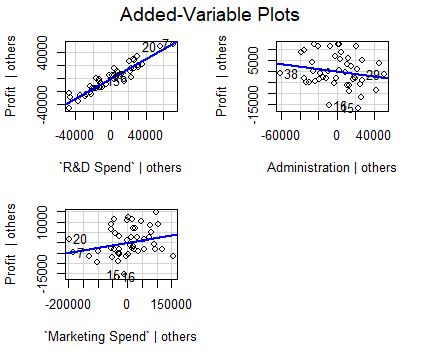
F-statistic: 259 on 3 and 43 DF, p-value: < 2.2e-16

**P values not significant.**

As p values are not significant after several trials checking for added variable plot,

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

> avPlot(model.startup1)



**As observed in plot the administration have very less relationship towards output variable. So we will build the final model discarding Administration column.**

**Final Model:**

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

> summary(model.startup3)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-16056 -4174 -1210 4959 12536

Coefficients:

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

(Intercept) 5.225e+04 2.532e+03 20.638 <2e-16 \*\*\*

`R&D Spend` 7.253e-01 3.983e-02 18.209 <2e-16 \*\*\*

`Marketing Spend` 3.672e-02 1.514e-02 2.426 0.0196 \*

---

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

Residual standard error: 7270 on 42 degrees of freedom

Multiple R-squared: 0.9585, Adjusted R-squared: 0.9566

F-statistic: 485.4 on 2 and 42 DF, p-value: < 2.2e-16

**The p values as well as R sqr values are significant. So final Model is ready to deploy.**