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

Now, since ‘State’ is one of the predictor and is discrete in nature, it has been converted into 3 Dummy Variables corresponding to NY(New York), CA (California) and FL (Florida) and have values 1,0,0 to fit.

Y = Profit

X = R.D. Spend, Administration, Marketing Spend, NY, CA, FL

Step 1: Read the data set

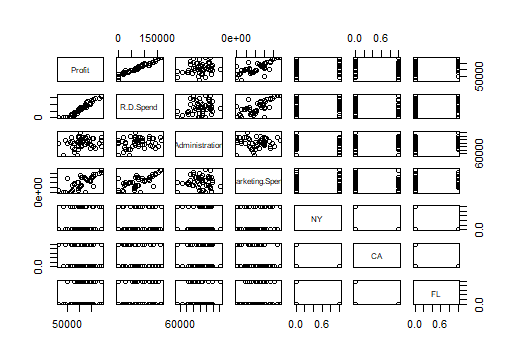
> su<-read.csv(file.choose())

> attach (su)

Step2: ScatterPlot of the data

> pairs (su)

Pairs help to plot multiple variables across each other



So, as it looks like:

Profit is having strong linear relationship with R.D.Spend

Moderate relationship with Marketing Spend

Weak linear relationship with Administration

No linear relationship with NY, CA or FL (dummy variables)

Step3: Co linearity

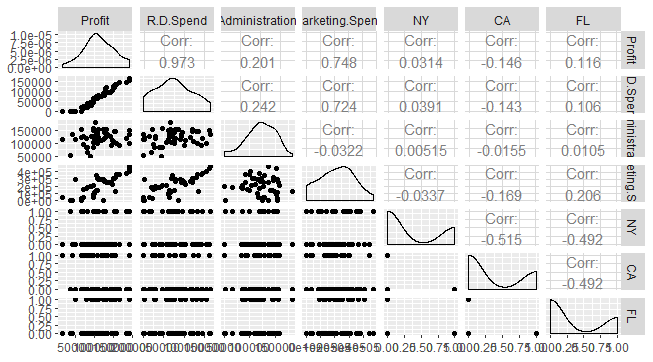
> cor (su)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Profit | R.D.Spend | Administration | Marketing. Spend | NY | CA | FL |
| Profit | 1 | 0.97290047 | 0.200716568 | 0.74776572 | 0.0313676 | -0.14583704 | 0.11624426 |
| R.D.Spend | 0.9729005 | 1 | 0.241955245 | 0.72424813 | 0.039068162 | -0.14316522 | 0.10571106 |
| Administration | 0.2007166 | 0.24195525 | 1 | -0.03215388 | 0.005145226 | -0.01547811 | 0.01049309 |
| Marketing. Spend | 0.7477657 | 0.72424813 | -0.032153875 | 1 | -0.0336698 | -0.16887523 | 0.20568545 |
| NY | 0.0313676 | 0.03906816 | 0.005145226 | -0.0336698 | 1 | -0.51515152 | -0.49236596 |
| CA | -0.145837 | -0.14316522 | -0.015478106 | -0.16887523 | -0.515151515 | 1 | -0.49236596 |
| FL | 0.1162443 | 0.10571106 | 0.010493089 | 0.20568545 | -0.492365964 | -0.49236596 | 1 |

Step4: Finding the ScatterPlot and Correlation Co-efficient altogether.

> ggpairs (su)

Ggpairs resides under a library (ggally)



So, from both these findings, it is evident that

1. Profit is having high co linearity with R.D. Spend
2. Profit is having moderate co linearity with Marketing. Spend

Also among the input parameters

1. R.D. Spend is having moderate co linearity with Marketing. Spend

Step5: Building the basic linear model with all input parameters

> msu<-lm (Profit~., data=su)

> summary (msu)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-33504 -4736 90 6672 17338

Coefficients: (1 not defined because of singularities)

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

(Intercept) 5.032e+04 7.252e+03 6.940 1.4e-08 \*\*\*

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

NY -2.407e+02 3.339e+03 -0.072 0.943

CA -1.988e+02 3.371e+03 -0.059 0.953

FL NA NA NA NA

---

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

So, it seems that only R.D. Spend is having significant co-efficient, contributing towards the prediction of model.

Also an error of ‘1 not defined because of singularities’ exist due to co linearity b/w predictors. In that case the dummy variables will be marked as n-1 for further implementations.

Step6: Now finding the model with Administration only to see its contribution

> msuadmin<-lm(Profit~Administration)

> summary(msuadmin)

Call:

lm(formula = Profit ~ Administration)

Residuals:

Min 1Q Median 3Q Max

-96072 -23426 -3564 25438 84870

Coefficients:

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

(Intercept) 7.697e+04 2.532e+04 3.040 0.00382 \*\*

Administration 2.887e-01 2.034e-01 1.419 0.16222

---

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

Residual standard error: 39900 on 48 degrees of freedom

Multiple R-squared: 0.04029, Adjusted R-squared: 0.02029

F-statistic: 2.015 on 1 and 48 DF, p-value: 0.1622

So, again it’s not significant.

Step7: Now finding the model with Marketing Spend only to see its contribution

> msumrktspnd<-lm(Profit~Marketing.Spend)

> summary(msumrktspnd)

Call:

lm(formula = Profit ~ Marketing. Spend)

Residuals:

Min 1Q Median 3Q Max

-83739 -18802 4925 15879 64642

Coefficients:

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

(Intercept) 6.000e+04 7.685e+03 7.808 4.29e-10 \*\*\*

Marketing. Spend 2.465e-01 3.159e-02 7.803 4.38e-10 \*\*\*

---

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

Residual standard error: 27040 on 48 degrees of freedom

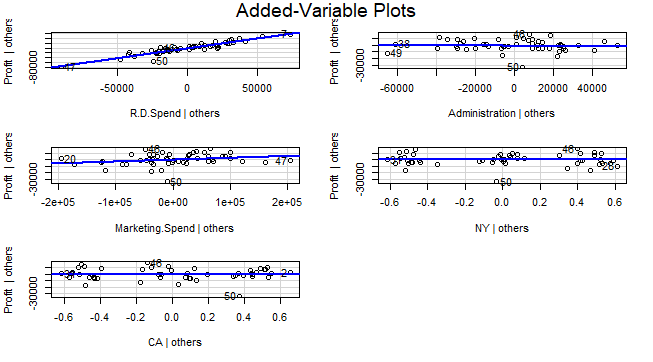
Multiple R-squared: 0.5592, Adjusted R-squared: 0.55

F-statistic: 60.88 on 1 and 48 DF, p-value: 4.381e-10

Now, the Marketing. Spend is significant. So, it is supportive that Administration is not contributing much towards the model prediction.

Step8: Also to double confirm

> avPlots(msu)



So, it supports that none of the predictors (Administration, CA, and NY) are significant or contributing towards the prediction of model except R.D. Spend and Marketing. Spend.

Step9:So, building a new model with R.D. Spend and Marketing. Spend

> msuf<-lm(Profit~R.D.Spend+Marketing.Spend)

> summary(msuf)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-33645 -4632 -414 6484 17097

Coefficients:

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

(Intercept) 4.698e+04 2.690e+03 17.464 <2e-16 \*\*\*

R.D.Spend 7.966e-01 4.135e-02 19.266 <2e-16 \*\*\*

Marketing. Spend 2.991e-02 1.552e-02 1.927 0.06 .

---

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

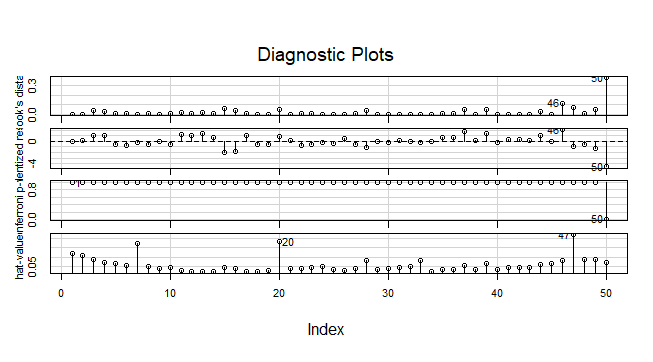
Residual standard error: 9161 on 47 degrees of freedom

Multiple R-squared: 0.9505, Adjusted R-squared: 0.9483

F-statistic: 450.8 on 2 and 47 DF, p-value: < 2.2e-16

Step10: Finding the most influence indicator

> influenceIndexPlot(msuf,id.n=3)



So, as it looks like 50 is the most common influence indicator/outlier in the data set

Step11: Building a model (excluding the influence indicator)

msufei<-lm(Profit~R.D.Spend+Marketing.Spend, data = su[-50,])

> summary(msufei)

Call:

lm(formula = Profit ~ R.D.Spend + Marketing. Spend, data = su[-50,

])

Residuals:

Min 1Q Median 3Q Max

-17224.8 -4753.4 -960.8 5375.2 14313.0

Coefficients:

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

(Intercept) 4.979e+04 2.342e+03 21.261 <2e-16 \*\*\*

R.D.Spend 7.754e-01 3.503e-02 22.136 <2e-16 \*\*\*

Marketing. Spend 2.745e-02 1.304e-02 2.104 0.0408 \*

---

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

Residual standard error: 7692 on 46 degrees of freedom

Multiple R-squared: 0.9611, Adjusted R-squared: 0.9594

F-statistic: 568 on 2 and 46 DF, p-value: < 2.2e-16

So, finally we have got a basic model (without any transformations) where we have the significant co-efficients contributing towards the prediction of the model.

Step12: Finding mean of the residuals

> mean(msufei$residuals)

[1] 9.030418e-14

It should be closer to zero.

Step13: Finding the RMSE

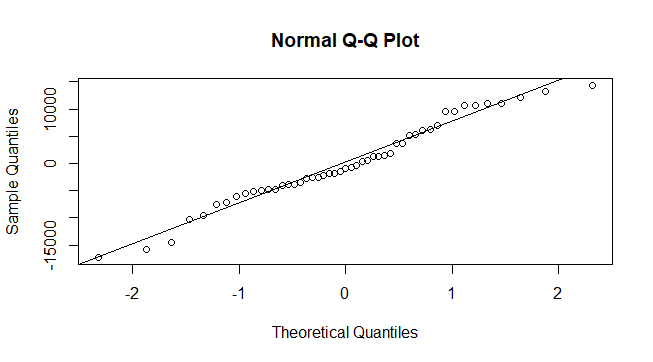
> sqrt(mean((msufei$residuals)^2))

[1] 7452.7

Step14:

> qqnorm(msufei$residuals)

> qqline(msufei$residuals)



The final model would be:

So, the final model will be: Y = 4.979e+04 + 7.754e-01 (R.D. Spend)+ 2.745e-02 (Marketing. Spend) Similarly different transformations are applied and the best model is identified as:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Input** | **Output** | **Cor** | **R^2** | **RMSE** | **Model** |
| Multiple Linear Regression | R.D. Spend, Marketing. Spend | Profit | 0.85 | 0.96 | 7452.7 | Profit = 4.979e+04 + 7.754e-01 (R.D. Spend)+ 2.745e-02 (Marketing. Spend) |