Regression

Multi Linear Regression

Assignment 1

BP: 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.

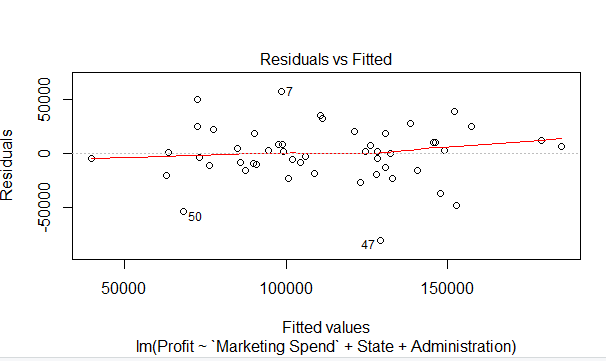
1. Now, from the given data we can observe that there is qualitative data after analysing the data we can eliminate the qualitative data as it doesn’t have any effect on the output.  
   This can be done by using dummies and dummy packages.
2. Now we have to create a multi linear model for this , by the formula :

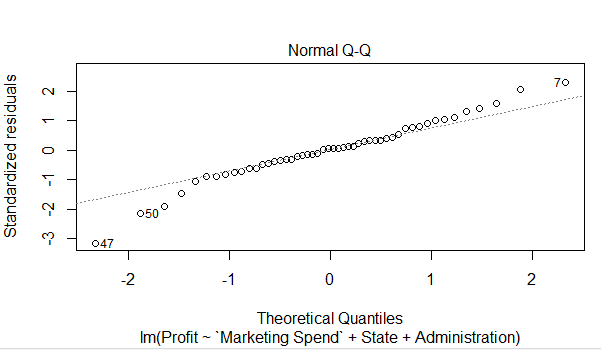
m2 <- lm(Profit ~ `Marketing Spend` + State + Administration)

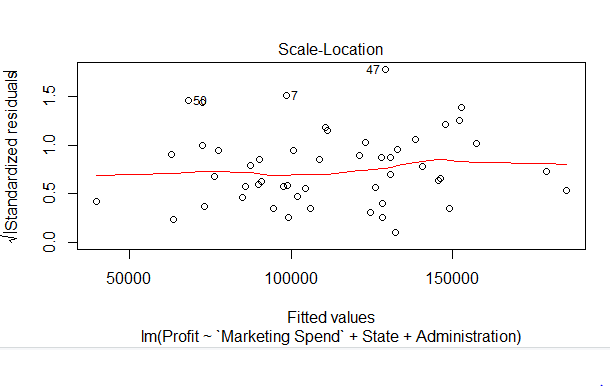
On using summary (m2) we get,

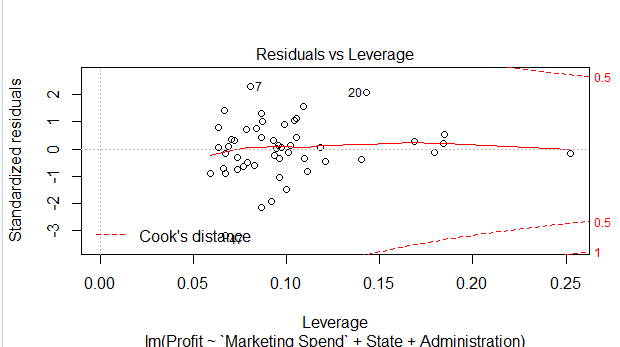
Now and value of R2= 0.6097

EDA of the data is as follows:





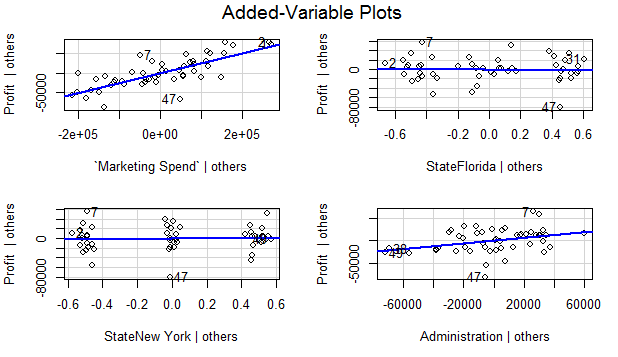




1. Now we have to try and build model , we can use variance inflation models,

If vif > 10 collinearity is good, but from the given values we can see that collinearity is on lower side, we cannot remove all, but can eliminate at least the least one.  
So let’s check which factor is effecting the correlation to bring it down.

Now on plotting added variable plots, we get as follows



From the following plots we can assume that, ads , multi , cd has been pulling down the value of correlation . But we cannot judge directly like that, so we use another method of regression called as Akaike information criterion model .  
  
on doing so we get as follows,

Profit ~ `Marketing Spend` + State + Administration

Df Sum of Sq RSS AIC

- State 2 2.6942e+08 3.1068e+10 1018.4

<none> 3.0799e+10 1021.9

- Administration 1 4.0290e+09 3.4827e+10 1026.1

- Marketing Spend 1 4.3775e+10 7.4573e+10 1064.2

Step: AIC=1018.37

Profit ~ `Marketing Spend` + Administration

Df Sum of Sq RSS AIC

<none> 3.1068e+10 1018.4

- Administration 1 4.0256e+09 3.5094e+10 1022.5

- Marketing Spend 1 4.5330e+10 7.6398e+10 1061.4

Call:

lm(formula = Profit ~ `Marketing Spend` + Administration)

Coefficients:

(Intercept) `Marketing Spend` Administration

2.022e+04 2.488e-01 3.237e-01

According to AIC regression, it is suggesting to remove the quantitative data column which is trying to reduce the collinearity.

The value of R2 is = 0.6131.

This is increased by a very small value.

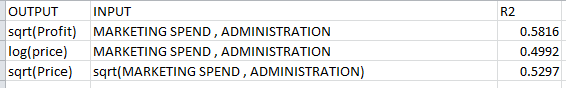
Now from the Plots we can observe that 47th entry is trying to reduce the collinearity, so let’s remove that and try to form the model

On doing so we get the value of R2 =0.683.

We can observe that it has increased by a great value .

To increase the value of R2 we can try some transformations.

The value before and after transformation are:

   
  
So we can conclude that after transformations, it’s getting even worse so, better not to apply transformations.