Profit analysis of 50 startups using regression model

BY:K S S Nikhil

20BCE1307

Abstract

This project aims in studying the “Startup logistics” dataset. In this study we will be analyzing the dataset and predicting the profit using the linear regression and the Poisson regression model and will also be suggesting the better model for practical purposes. We have 5 variables namely administration, marketing.Spend, RD.Spend, state and profit. The variables administration, marketing.spend and RD.spend represent the money spent on the administrative , marketing and research department respectively. The startups were in three different states namely, New York, Florida and California. The first 4 variables and independent variable and profit is the dependent variable.

Here we have data for 50 startups in various states. We will visualize all the data provided using different plots and graphs for better understanding of the data provided. Then we study the independent variables and find which best shows the variation of profit of the startups using the concept of correlation. We then create the best regression models and find the difference between the actual and predicted values and suggest the better model for practical purposes. The study is done using the programming language R.

Objective of the work / Proposed Work

We will be studying the different costs spent to begin a startup and analyze the profits gained using the dataset given.

Need and importance

A start-up is a young company founded by one or more entrepreneurs to develop a unique product or service and bring it to market. A startup is a necessity in various places, especially in developing countries as it helps in economic growth of the country. Its also helps in creating jobs and decreases the unemployment percentage in the country.

Using this analysis different entrepreneurs can plan their finances to initiate a start-up.

Correlation and Regression

Correlation:

In a bivariate distribution we have to find out the if there is any correlation or covariance between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables are deviate in the same direction, that is, if the increase (or decrease) in one result in a corresponding increase (or decrease) in the other, correlation is said to be positive. But, if they are constantly deviate in the opposite directions, that is if increase (or decrease) in one result in corresponding decrease (or increase) in the other, correlation is said to be negative.

Regression:

Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and co-founders are known as predictors or independent variables. The dependent variable is shown by “y” and independent variables are shown by “x” in regression analysis.

The different types of regression models used in this project are :-

(i)Linear regression:- **Linear regression** is a linear approach to computing the relationship between the dependent and one or more independent variables. If the regression has one independent variable, then it is known as a simple linear regression. If it has more than one independent variable, then it is known as multiple linear regression.

A linear regression line has an equation of the form Y = a + bX, where X is the independent variable and Y is the dependent variable.

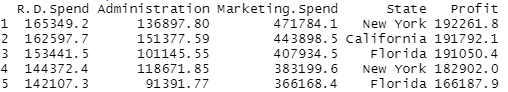
(ii)Poisson regression: When the response variable is rare event over the large population, the Poisson regression model is more appropriate. While logarithm of the response variable is linked to a linear function of independent variables.

The the Poisson regression model can be expressed as

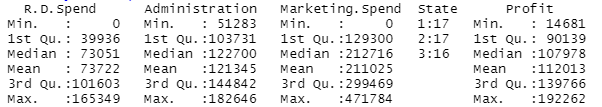
ln(y)= a + bX1 + cX2 + ……

Experimental Analysis

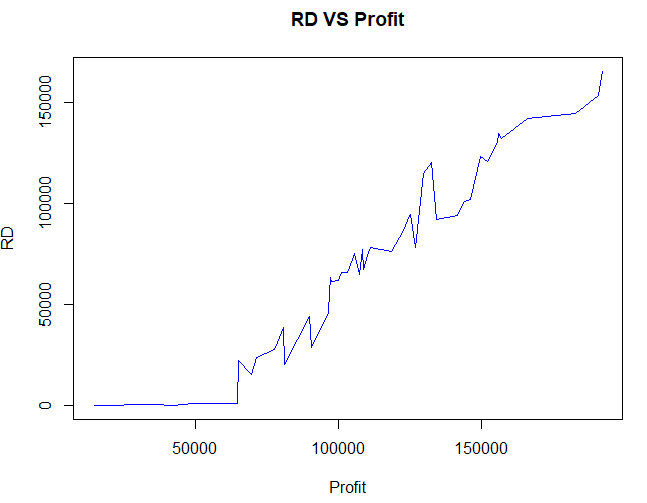
->Displaying the initial rows of the data set

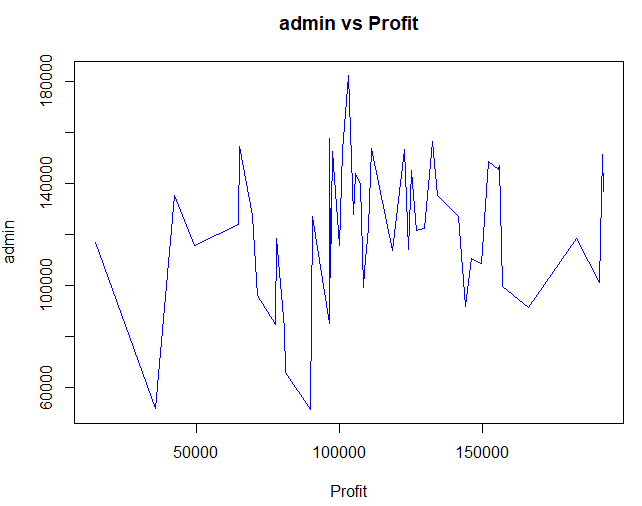


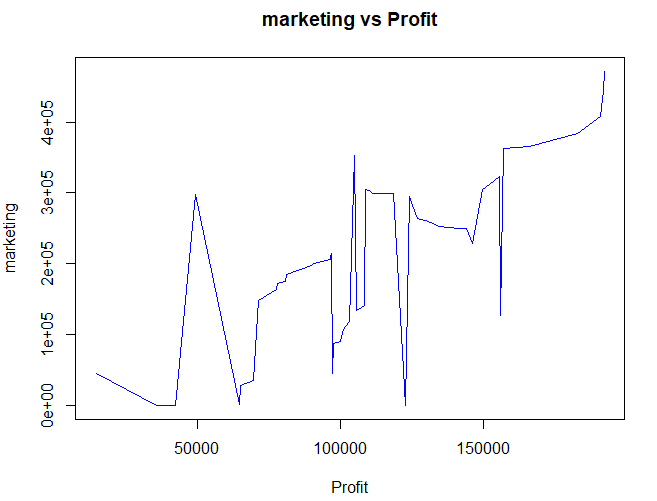
->Summary of the dataset



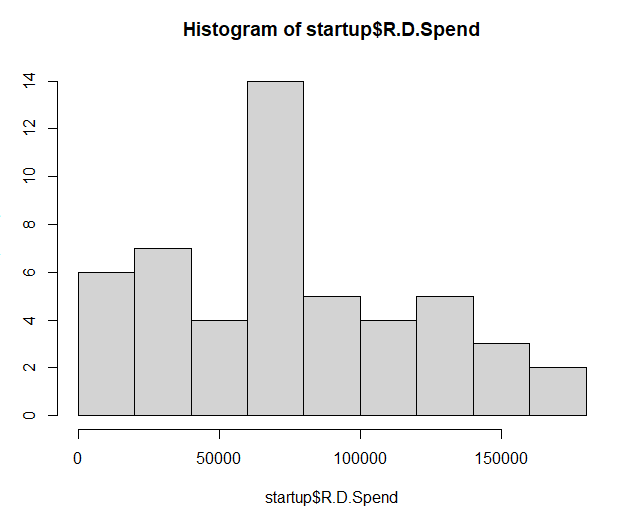
Visualization of data using plots

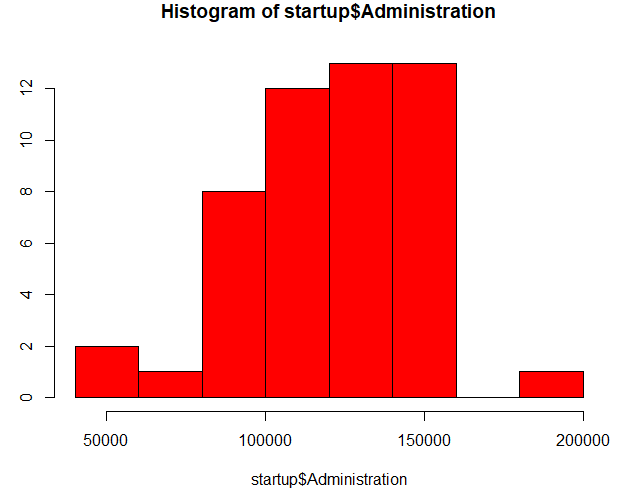


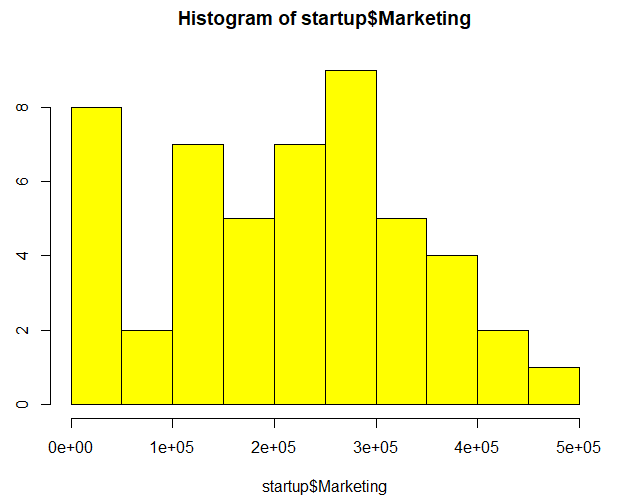


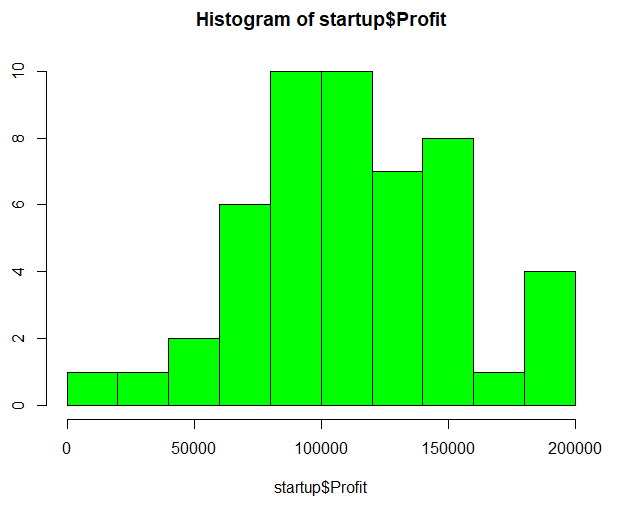


->Using Histograms

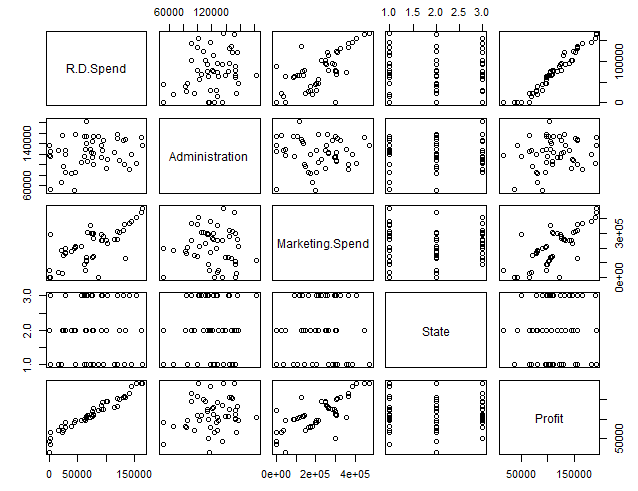






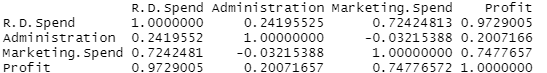


->A complete plot of the dataset

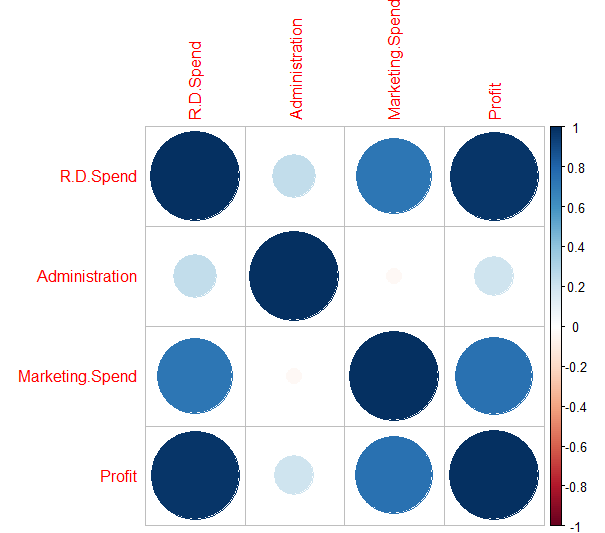


->Now, we find the correlation between all the independent variables

And profit (dependent variable)



->From the correlation matrix we see that high correlation between RD spend-Profit and a moderate correlation between Marketing Spend-Profit

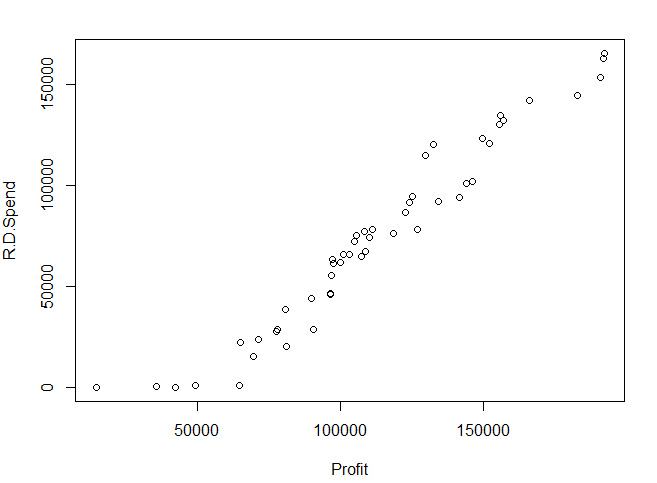


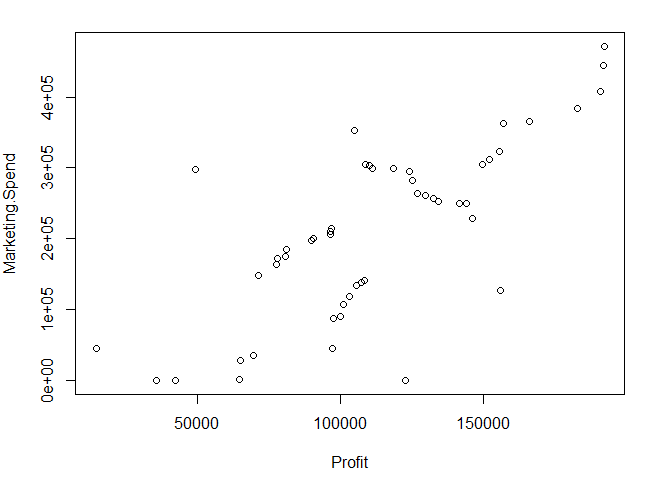
In the above graph we see a huge dark circle for RD Spend-Profit and a moderate circle for Marketing Spend-Profit. Darker and bigger the circle, greater is the correlation

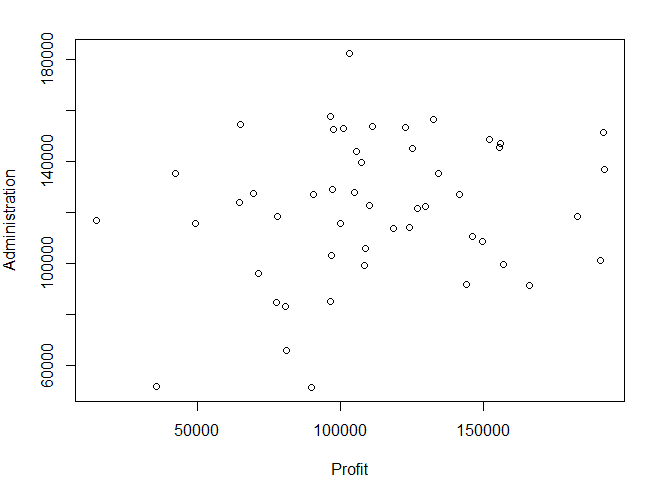
The above graph gives us a rough idea about the relation between all the variables

We now observe the scatter plots to arrive at a final conclusion on which independent variables effects Profit the most for our regression analysis.

Scatter Plots

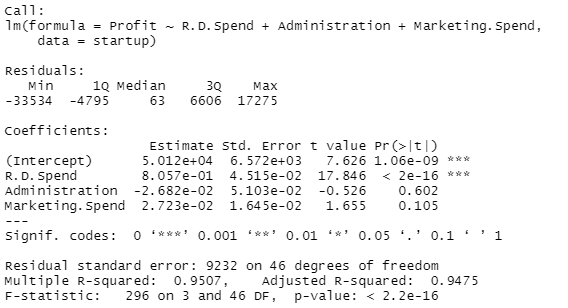






I have formed a regression model using all the dependent variables and profit and examined the summary.

Multiple Linear regression:

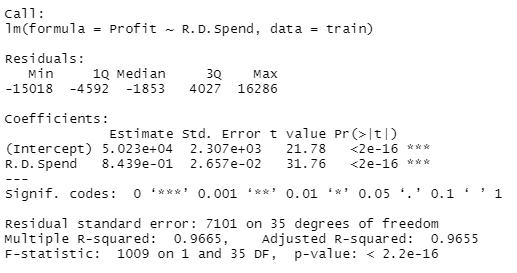


From the above plots and summary ,we come to a conclusion that RD Spend Gives the best linear graph. Hence RD spend is used for our regression models.

->I have now split the data into a train set and a test set.

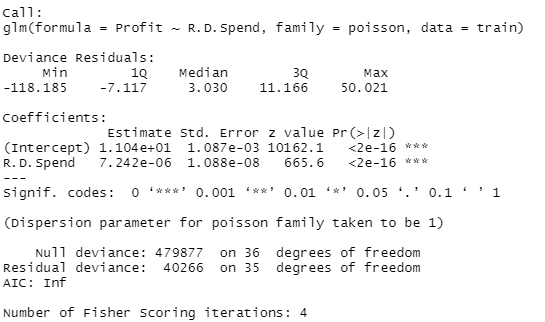
I have created one regression model for linear and another model for poisson using the trainset.

Summary of Linear model



We have a Adjusted R-squared of 0.966.5.Therefore, this means that the this model shows 96.65% variation of profit.

Summary of Poisson model



Now, I predicted the profits with both the modules using a particular value of RD Spend and compared them with the actual value in the dataset.

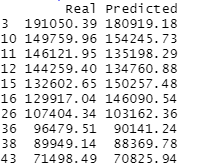
When RD Spend = 101913.08 , the actual profit is 146121.95

Using the linear model, we get profit as 136236.5

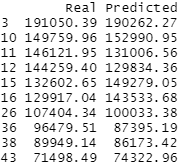
Using the poison model, we get profit as 131006.6

To get a clearer idea, I used the test set and found the actual and predicted values using both the models.

For linear model



For poison model



Findings and Results

From the above analysis, we have found that profit is most dependent of RD spend. When we predicted a profit value using both the models,the linear model has produced a value closer to the actual value than the poison value.

But the difference between both the predicted values is small so I made a dataframe with the actual values and the predicted values for both the models using the test set. It is observed that in both the models the predicted values are inconsistent. But comparatively it is seen that the linear model is a better regression model than the poison model for the above data set.

Conclusion

The R studio is a great tool for data analysis and data visualization. It has many different features which helps us in understanding the dataset better. Correlation and regression is a very helpful concept used to find relationships between various data and predicting values using regression models. It can be used in various fields. We can see that the above model was inconsistent and this is due to smaller data set. With more data we can create a more consistent model for predicting a value.

Overall this project was a great learning curve and helped me understand the subject better

References

1)https://www.kaggle.com/karthickveerakumar/startup-logistic-regression

2) <https://rpubs.com/ID_Tech/S1>

3)<http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software> (Finding the correlation plot and matrix)

Appendix (Code)

#installing the reguired packages for the study

library(corrplot)

library(mltools)

#importing the data csv

startup<-read.csv("C:\\Users\\NikhilK\\OneDrive\\Desktop\\50\_Startups.csv")

#displaying the data set

Startup

#reading the initial 8 rows of the dataset

head(startup, 5)

#to check wether there are any missing values

cat("Contains missing values : ", any(is.na(startup)))

startup$State<-factor(startup$State,levels=c('New York','California','Florida'),labels=c(1,2,3))

startup

summary(startup)

#vizualization of data

#using graphs

plot(startup[,5],startup[,1],type="l",main="RDVSProfit",xlab="Profit",ylab="RD",col="blue")

#with admin

plot(startup[,5],startup[,2],type="l",main="adminvsProfit",xlab="Profit",ylab="admin",col="blue")

#with marketing

plot(startup[,5],startup[,3],type="l",main="marketingvsProfit",xlab="Profit",ylab="marketing",col="blue")

#using histograhm

#with rd

hist(startup$R.D.Spend)

#with admin

hist(startup$Administration, col="blue")

#with marketing

hist(startup$Marketing, col="yellow")

#with profit

hist(startup$Profit, col="green")

#plot all the dataset together

plot(startup)

#correlation of dependent variable with the other independent variables

cor(startup$R.D.Spend,startup$R.D.Spend)

cor(startup$Profit,startup$Administration)

cor(startup$Profit,startup$Marketing)

cor(startup[sapply(startup, is.numeric)])

#plot the correlation

corrplot(cor(startup[sapply(startup, is.numeric)]))

#scatter diagrams

plot(R.D.Spend ~ Profit, data = startup)

plot(Marketing.Spend ~ Profit, data = startup)

plot(Administration ~ Profit, data = startup)

#setting up the models

regr <- lm(Profit ~ R.D.Spend + Administration + Marketing.Spend, data = startup)

summary(regr)

startup1 <- subset(startup, select = c(Profit, R.D.Spend))

startup1

#splitting the data into train and test data

set.seed(420)

smp\_siz = floor(0.75\*nrow(startup1))

train\_ind = sample(seq\_len(nrow(startup1)),size = smp\_siz)

train =startup1[train\_ind,]

test=startup1[-train\_ind,]

cat("Training set size is ", length(train$Profit), "\n")

cat("Testing set size is ", length(test$Profit))

test

#Linear Model

model <- lm(Profit ~ R.D.Spend, data = train)

summary(model)

confint(model)

a=data.frame(R.D.Spend=101913.08)

result=predict(model,a)

result

predicteddata <- (predict(model,test))

bound <- cbind(test$Profit, predicteddata)

colnames(bound) = c('Real', 'Predicted')

predicted <- as.data.frame(bound)

head(predicted, n = 10)

#poisson Model

model2<-glm(formula = Profit ~ R.D.Spend, data = train,family = poisson)

summary(model2)

confint(model2)

b=data.frame(R.D.Spend= 101913.08)

result1=exp(predict(model2,b))

result1

predicteddata <- exp((predict(model,test)))

bound <- cbind(test$Profit, predicteddata)

colnames(bound) = c('Real', 'Predicted')

predicted <- as.data.frame(bound)

head(predicted, n = 10)