**Regression Analysis – Social Networks**

***Version 1.0***

Contents

[INTRODUCTION: 3](#_Toc420352413)

[Data Collection & Data Preparation: 3](#_Toc420352414)

[Technical Architecture Diagram 4](#_Toc420352415)

[Data & Variable Description 6](#_Toc420352416)

[Modeling Strategy Process Guidelines 8](#_Toc420352417)

[Statistical Models and Linear Regression Analysis on each Model. 11](#_Toc420352418)

[Validation & Conclusion 15](#_Toc420352419)

[Appendix : 17](#_Toc420352420)



# INTRODUCTION:

Company X is a marketing research firm that collects and maintains data for Celebs across US and India.Celebs wants to reach out to maximum followers through Social Networks for promotional activities for their movies, product endorsements etc. Most of the Celebs for which company X maintains data are on twitter but not on Facebook, Company X will help celebs to predict the number of Facebook likes a celeb page will receive through linear regression model when a celeb creates a Facebook page. Company X collects celebrity data from Twitter and Facebook through Web Crawling and Data Mining Techniques and stores it in Mongo-DB as unstructured data and then apply data transformation techniques to convert unstructured data to structured data which can be used for further statistical modeling and regression analysis to build a model which determines the Facebook Likes a celebrity page is likely to receive predicted through linear regression equation.

# Data Collection & Data Preparation:

We started with selecting about 90 celebrities across United States and India across areas such as sports, entertainment and politics. After finalizing the celebrity list our next step was to decide the data sources from which the celebrity data will be collected from we came down to twitter and Facebook as the best sources of data. Python scripts were being written and executed in order to extract data from Facebook and Twitter. After performing the technical analysis on the nature of data and its format which will be getting from Facebook and twitter we decided to store the unstructured data in BSON format in Mongo-DB. Python scripts were written again to perform data mining on the unstructured data stored in Mongo-DB and then conversion of unstructured to structure data was performed and structure data was stored in SQL server. We have SQL tables for Facebook, twitter and also we have an aggregation table where the data for each celebrity will be stored at row level uniquely with all the parameters from Facebook and twitter. After this we will be using the aggregated data from SQL server to perform statistical modeling and regression analysis.

# Technical Architecture Diagram

Linear Regression and Statistical Modeling using R

RESULTS

Mongo DB Query for Statistics

Mongo DB

Python

Data Mining

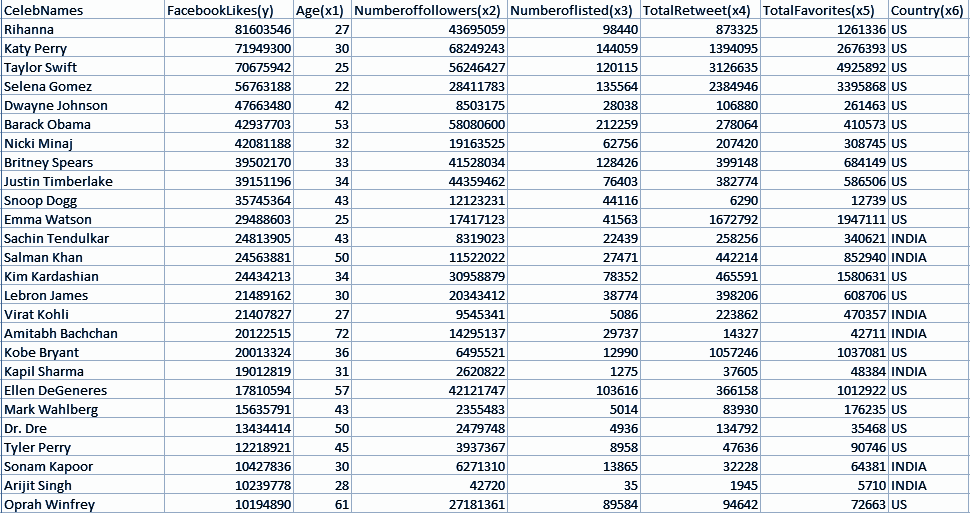
Celebrity List

**Following is the description of the Technical Architecture:**

* Select list of celebrities across **India** and **US** in the field of Entertainment, Sports & Politics.
* For each Celebrity perform Data Mining from Facebook and Twitter using Python script
* Insert for each celebrity one collection for Facebook and one for Twitter
* Execute Python scripts to extract statistical data for both Facebook and Twitter
* Extracted statistical data, needs to be stored in SQL database in order to perform aggregation of the statistics and convert multiple rows for each celebrity to one unique row, SQL database allows you to perform joins on the Celebrity name for the statistical parameters from Facebook and Twitter.
* Final data output from SQL database is a table which provides one row for each celebrity and has all the parameters required for statistical modeling.
* R-Programming language is used for Statistical Modeling and Linear Regression on the structured Celebrity data.

# Data & Variable Description

**Following is the snapshot of the structured data on which the Regression will be performed.**

****

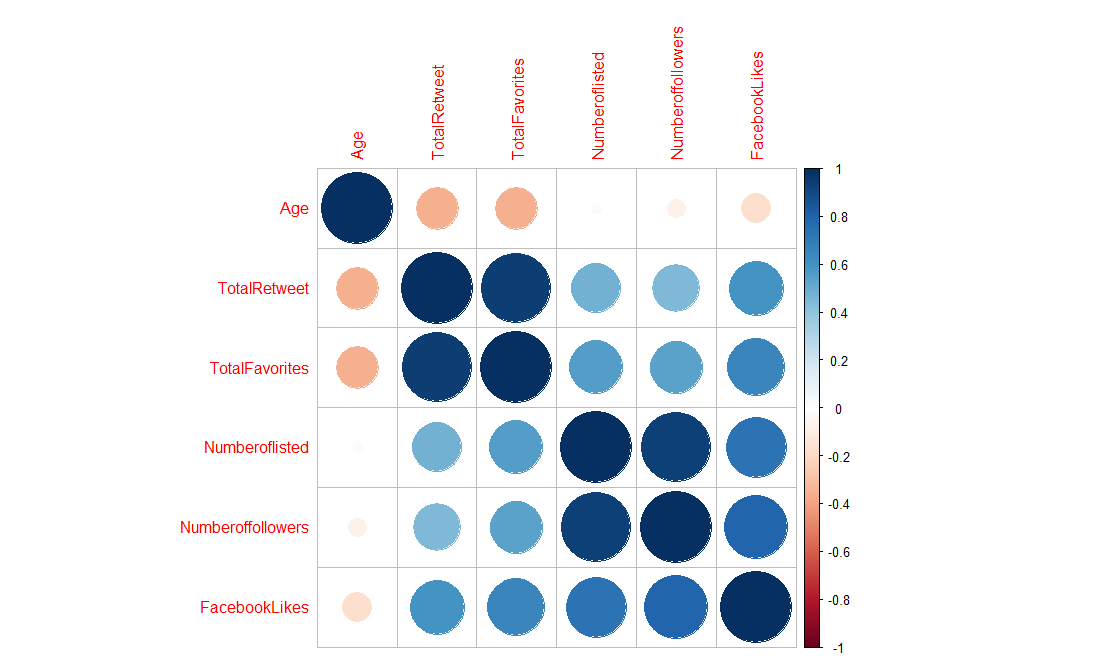
**Following is the description of the variables used for the Regression we have 6 independent variables and one dependent variable.**

|  |  |  |
| --- | --- | --- |
| **No** | **Parameter Name** | **Meaning** |
| **1** | **Facebook Likes (Y)** | **Count of the number of followers liking the Celebrity page on Facebook.** |
| **2** | **Age(X1)** | **Age of the Celebrity** |
| **3** | **Number of followers(X2)** | **Number of followers for the Celebrity on Twitter** |
| **4** | **Number of listed(X3)** | **Number of time the Celebrity is tagged by other users on Twitter** |
| **5** | **Total Retweets(X4)** | **Number of time the Celebrity Tweet is retweeted by other users on Twitter** |
| **6** | **Total Favorites(X5)** | **Number of likes each Celebrity Tweet receives** |
| **7** | **Country(X6)** | **US or India – Only Categorical Variable** |

# Modeling Strategy Process Guidelines

**Following are the steps and the guidelines we have used during the linear regression :**

1. We started by performing descriptive statistics on all the input and output variable and tried to understand correlation amongst all variables. Correlation was captured amongst independent variables and also between dependent and independent variable.

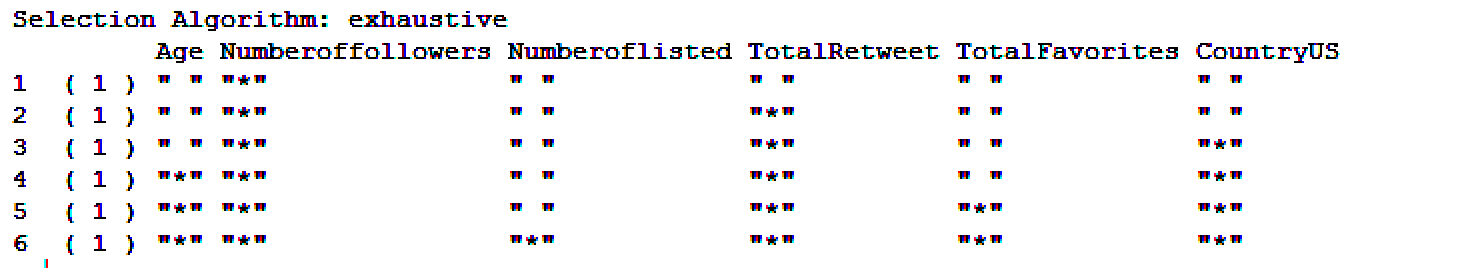
****

**The values highlighted in yellow indicate which variable mostly correlated with another one.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameters** | **Age** | **TotalRetweet** | **TotalFavorites** | **Numberoflisted** | **Numberoffollowers** | **FacebookLikes** |
| **Age** | **1** | **-0.354395** | **-0.352621** | **-0.02301995** | **-0.072425** | **-0.1742424** |
| **TotalRetweet** | **-0.3544** | **1** | **0.946708** | **0.479598** | **0.4478789** | **0.5958577** |
| **TotalFavorites** | **-0.3526** | **0.946708** | **1** | **0.55796684** | **0.5357931** | **0.6583138** |
| **Numberoflisted** | **-0.023** | **0.479598** | **0.557967** | **1** | **0.9378948** | **0.7314523** |
| **Numberoffollowers** | **-0.0724** | **0.447879** | **0.535793** | **0.93789477** | **1** | **0.7998733** |
| **FacebookLikes** | **-0.1742** | **0.595858** | **0.658314** | **0.73145231** | **0.7998733** | **1** |

1. After Correlation analysis we were determining which technique such as forward, backward, Stepwise or Exhaustive Search to use for building models. As the number of input variables were only 6 we decided to go ahead with Exhaustive Search using LEAPS package in R to generate 2^6 – 1 = 63 models and then select best 6 subsets of linear models out of 63 .

Following are the top 6 models from LEAPS Package

****

In the above chart \* indicate that this parameter should be used and “ “ indicate that this parameter should not be used.

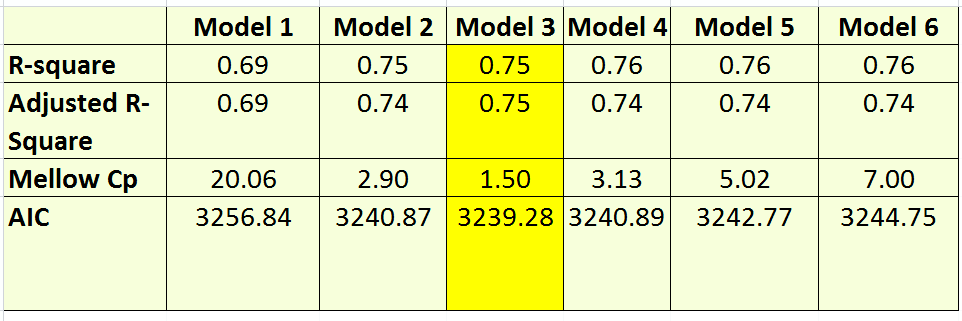
Model with 1 variable – X1: Numberoffollowers

Model with 2 variables– X1: Numberoffollowers

X2: Total Retweet

Similarly for Model with 3, 4, 5 and 6 variables can be interpreted from the above table.

1. All the Models we have built are with confidence interval of 95 % with margin of error as 5 %.
2. Out of the 6 subsets of linear model we had to determine which is the best full model to start the regression with, in order to determine the best full model to start we compare statistical measures such as AIC,Mallows’s CP, R-Square, Adjusted R-Square and also for few models we used VIF also as a measure.

****

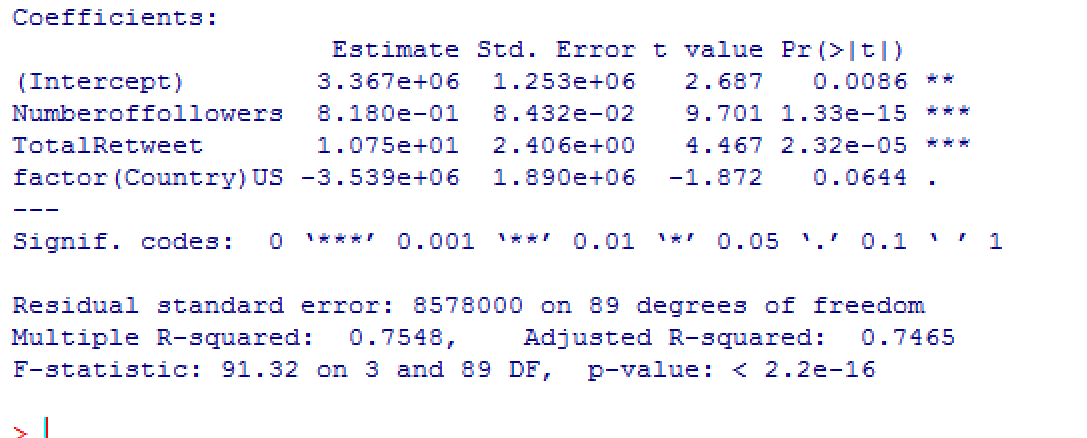
As per the above table Model 3 is the best full Model to start with 3 variables for the Regression Analysis as it has the lease AIC, Mallows’s CP and optimal R-Square and Adjusted R-Square.

1. While building the model we have taken care of the basic principle of any modeling technique which is the assurance that there is no violation of Constant Variance, Normality, Linearity and Independence for any models which were built.
2. We have analyzed the problem via the visual way from the residual plot and normal QQ plot but in order to confirm our analysis we have also ran Breusch-Pagan (BP) test for Heteroscedasticity on most of the best models we have.
3. Appropriate transformation techniques have been applied on both X and Y in order to ensure no variance violation happens.
4. At the end we have performed validation in 3 stages for the final linear equation built via regression, validation test is performed on the sample training data and test data to calculate the overall accuracy and the error rate for the final model along with that a prediction is also done for the celebs who are on Twitter and not on Facebook and if they join what will be approximate number of fan followers the Celeb Facebook page will receive.

# Statistical Models and Linear Regression Analysis on each Model.

**Model 3 : Analysis**

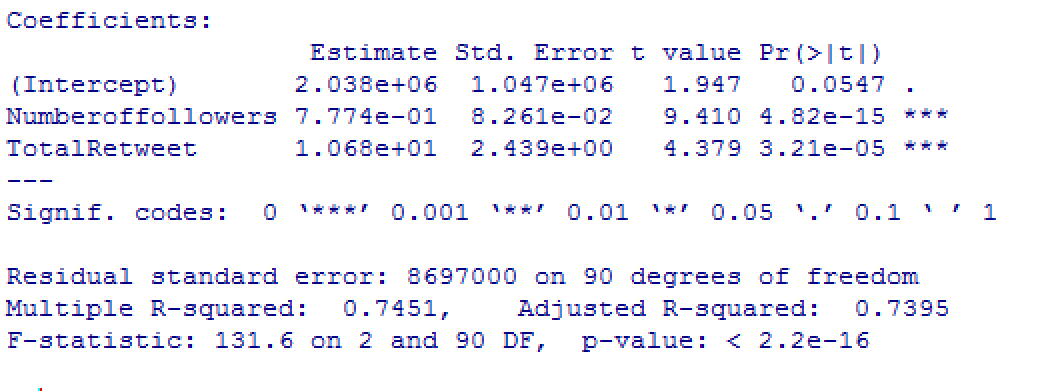
**FacebookLikes ~ Numberoffollowers+TotalRetweet+Country**

****

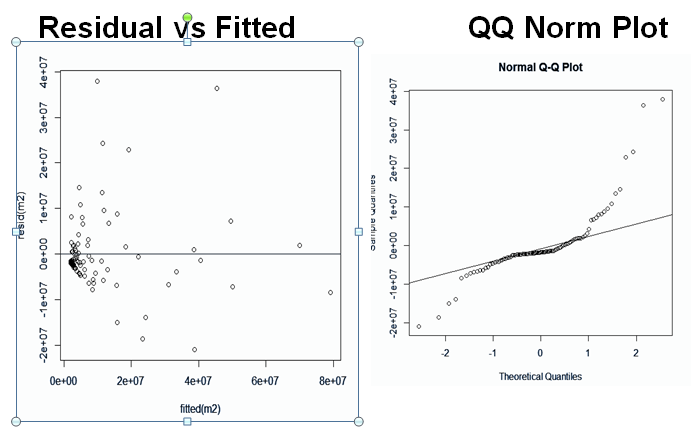
**Conclusion: Country is not a significant predictor as the P value(0.0644) is > than Alpha (0.05) and it’s not contributing significantly to predict Facebook Likes(Y).**

**Model 2 : Analysis**

***FacebookLikes ~ Numberoffollowers+TotalRetweet***

****

**VIF: 3.92 (<10) which indicates that multi-co-linearity amongst all the variables is not over fitting the model.**



* **Quantile plot has heavy tail**
* **Non constant variance in residual vs fitted**

**Breusch- Pagan Test : Test of Heteroscedasticity**

* **H0 : Homoscedasticity**
* **H1 : Homoscedasticity**

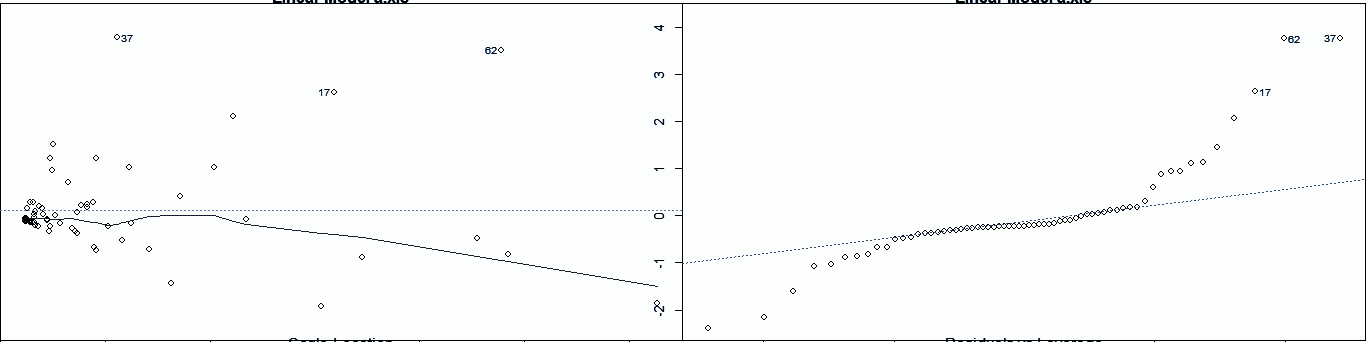
**Decision Rule: After running BP test if P value less than Alpha (0.05) Reject H0 else Fail to Reject H0**

**Results: P Value: 0.009837 < Alpha(0.05)**

**Conclusion: Reject H0**

**Model 1:Linear Regression (Single Variable) *FacebookLikes(Y) ~ Numberoffollowers(X2)***

**Residual vs Fitted Plot QQ NORM**

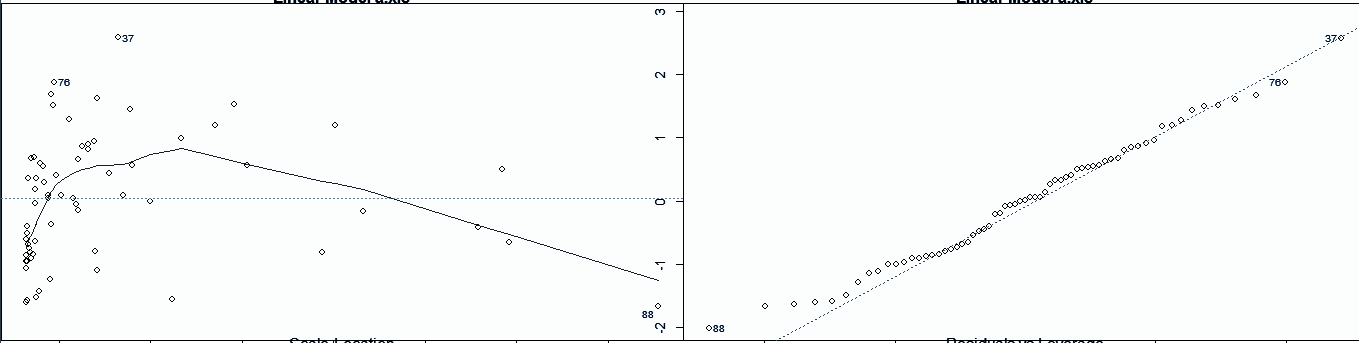
******

* **Quantile plot has heavy tail**
* **Non constant variance in residual vs fitted**

**Model 1.1 Transform Y to Y^0.25**

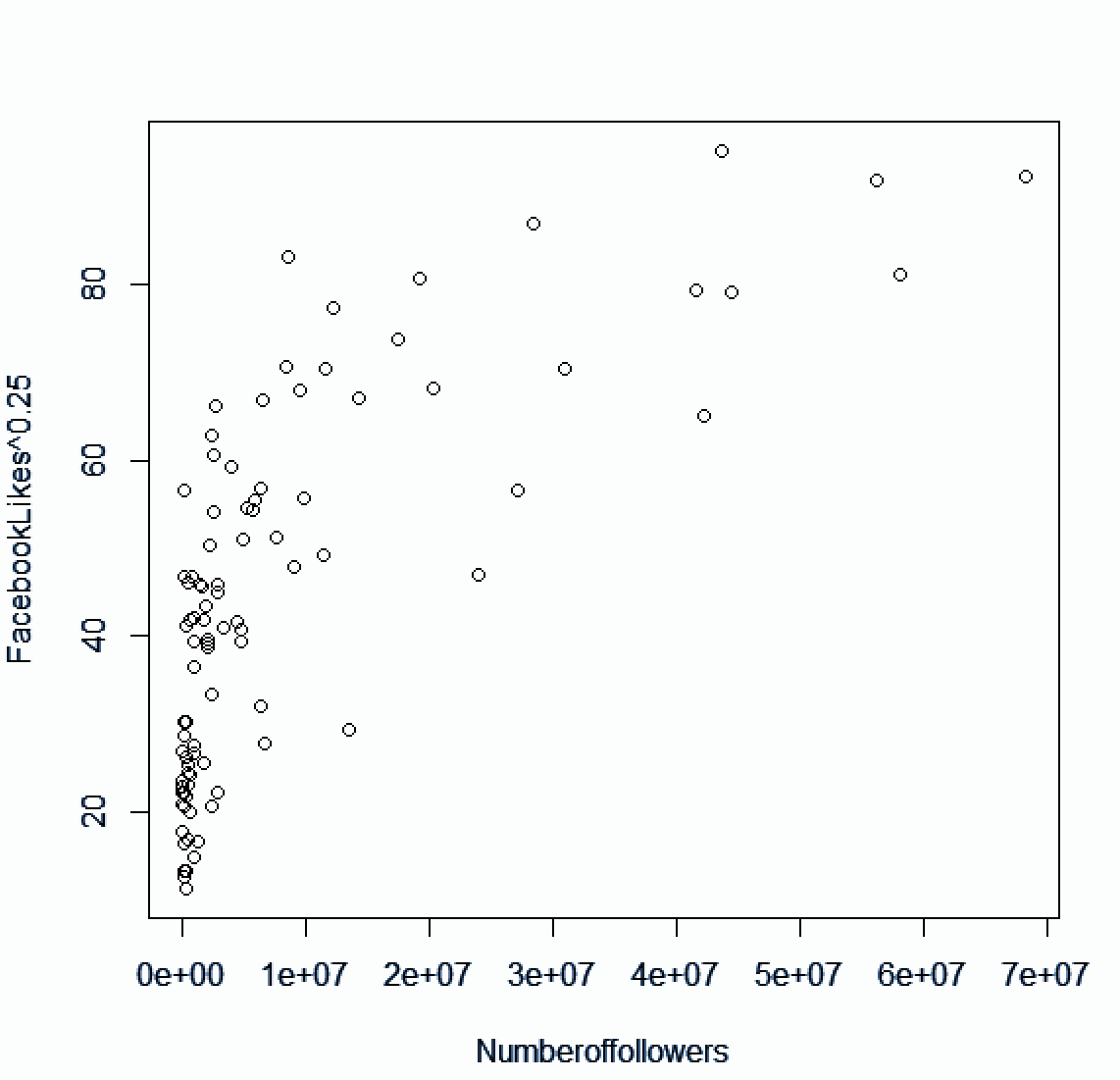
**Using the Box-Cox method we found that Y needs to transformed to Y^0.25. The below shows the Residual vs Fitted plot and Quantile Plot of Model 1.1**

**Residual Vs Fitted Quantile Plot**

****

* **The quantile plot looks much better it was in the previous models.**
* **The Residual plot displays a little curvilinear nature.**

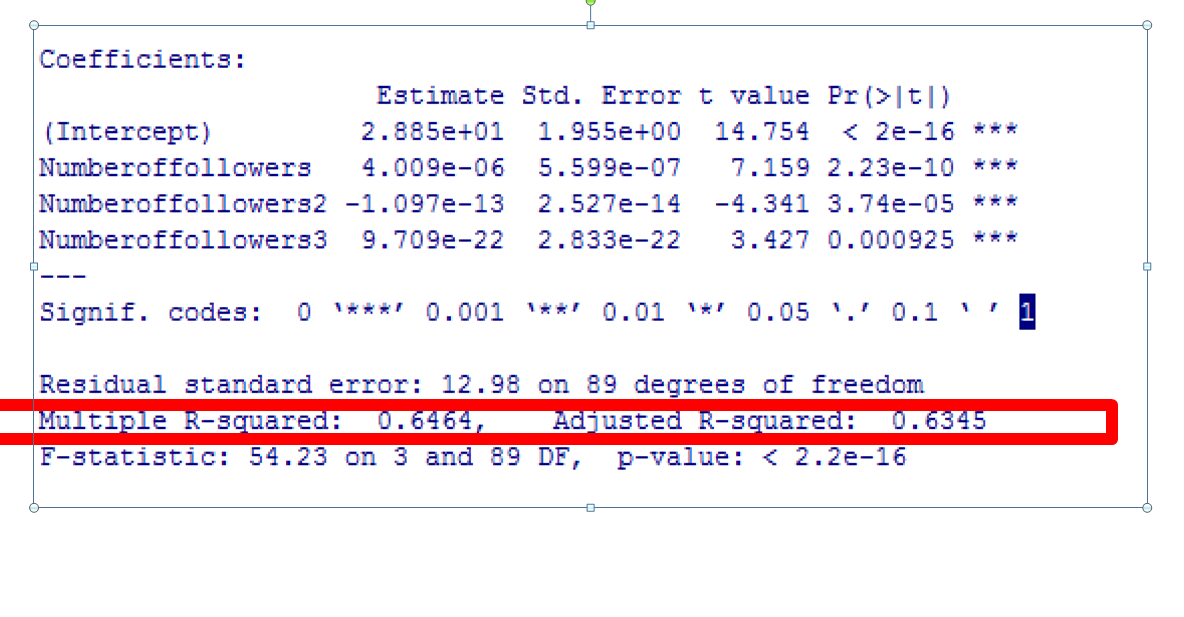
**Scatterplot (FacebookLikes ^0.25 ~ Numberoffollowers)**

****

* **The scatter plot shows a curvilinear relationship between FacebookLikes^0.25 and Numberoffollowers axis.**

***Model 1.2 : Transforming X to Polynomial form of X***

**FacebookLikes^0.25 ~ Numberoffollowers + Numberoffollowers^2 + Numberoffollowers^3**



# Validation & Conclusion

**Based on the final equation of the linear model which as follow x` = 28.85**

**+ (4.009e-06) \* (Numberoffollowers)**

**- (1.097e-13) \* (Numberoffollowers)^2**

**+ (9.709e-22) \* (Numberoffollowers)^3**

**Validation Results:**

Validation on Sample training data:

US Celeb – **Barack Obama**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook**  **Likes** | **Numberoffollowers** | **Actual Transformed Y** | **Estimated from Model Transformed Y** | **Error** |
| **43 million** | **58 million** | **81.86** | **80.94** | **2.38%** |

India Celeb – **Salman Khan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook**  **Likes** | **Numberoffollowers** | **Actual Transformed Y** | **Estimated from Model Transformed Y** | **Error** |
| **24 million** | **11.5 million** | **70.40** | **61.96** | **13.62%** |

**Validation of Sample Test Data**

India Celeb – **Hrithik Roshan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook**  **Likes** | **Numberoffollowers** | **Actual Transformed Y** | **Estimated from Model Transformed Y** | **Error** |
| **14.8 million** | **9.6 million** | **62.07** | **58.02** | **7%** |

US Celeb - **Chris Pratt**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facebook**  **Likes** | **Numberoffollowers** | **Actual Transformed Y** | **Estimated from Model Transformed Y** | **Error** |
| **11.8 million** | **1.9 million** | **32.98** | **36.14** | **8.76%** |

**Prediction:**

Following are the celebrities who are on twitter and not on Facebook and if they join Facebook with the linear regression equation we are 95 % confident with the predicted number of Facebook like the below celebs will receive with 5 % Margin of error.

**Alia Bhatt**

|  |  |  |
| --- | --- | --- |
| **Number Of Followers in Twitter** | **Estimated (Y^0.25)** | **Expected Facebook likes** |
| **4.6 million** | **45.24** | **4.2 million** |

**Suresh Raina**

|  |  |  |
| --- | --- | --- |
| **Number Of Followers in Twitter** | **Estimated (Y^0.25)** | **Expected Facebook likes** |
| **3.8 million** | **42.84** | **3.37 million** |

# Appendix

**R-Codes used for the project for reference.**

***Import Data into R***

dataset <- read.xlsx("C:/Chanchal/MyClasses/Advanced Stats - STT863/Project/CelebsData.xls", sheetIndex=1)

**Generate Scatter Plot Matrix**

pairs(~FacebookLikes+Listed+Age+Numberoffollowers+Numberoflisted+TotalRetweet+TotalFavorites+sharecount,data=dataset, main="Simple Scatterplot Matrix",col="red")

***Using Leaps Package and Generate metrics of models such as R-square, Mellow Cp etc***

library(leaps)

test = regsubsets((FacebookLikes) ~ Age+Numberoffollowers+Numberoflisted+TotalRetweet+TotalFavorites+(Country),data=dataset)

stest=summary(test)

Rp2=stest$rsq

Rap2=stest$adjr2

C\_p=stest$cp

AIC(<linear Model)

***Different Models***

Model 1 : One vairable

m1=lm(FacebookLikes ~ Numberoffollowers,data=dataset )

m11 =lm(FacebookLikes^0.25 ~ Numberoffollowers,data=dataset )

Numberoffollowers2=dataset$Numberoffollowers^2

Numberoffollowers3=Numberoffollowers2\*dataset$Numberoffollowers

Numberoffollowers4=Numberoffollowers3\*dataset$Numberoffollowers

m11 = lm(FacebookLikes^0.25 ~ Numberoffollowers+Numberoffollowers2,data=dataset )

m12 = lm(FacebookLikes^0.25 ~ Numberoffollowers+Numberoffollowers2+Numberoffollowers3,data=dataset )

m13 =lm(FacebookLikes^0.25 ~ Numberoffollowers+Numberoffollowers2+Numberoffollowers3+Numberoffollowers4,data=dataset )

Model 2

m2=lm(FacebookLikes ~ Numberoffollowers+TotalRetweet,data=dataset )

m21=lm((FacebookLikes)^0.25 ~ Numberoffollowers+TotalRetweet,data=dataset )

m22=lm(log(FacebookLikes) ~ Numberoffollowers+TotalRetweet,data=dataset )

boxcox(m2,lambda=seq(-5,+5,seq=1/10))

Model 3

m3=lm(FacebookLikes ~ Numberoffollowers+TotalRetweet+factor(Country),data=dataset )

***Plot Residuals***

Plot(resid(m1) ~ fitted(m1))

***Quantitle Plot***

 qqnorm(resid(m1))

qqline(resid(m1))