University of Cincinnati

Predicting consumer's response on social media and evaluating its impact on brand building

BANA7038 -Data Analysis Methods-Section 002
Final Project

Ajmera, Himanshu – M12426133
Namani, Sriveni – M12624187
Sawant, Dipali – M12448427
Feb 22, 2018

Summary

Our data analysis study is focused on finding the relationships between online publications on social networks and the impact of such publications measured by users' interactions. We used a dataset of 500 posts published in a Facebook social network brand page in a complete year by a renowned cosmetic company for analysis.

The goal of this analysis is to implement a model that predicts the impact of posts using their characteristics and assessing the knowledge provided by the model in terms of which input features affect the impact metrics and how these input features influence each post, and hence supporting managers' decisions.

This data analysis provides insight on the social media engagement of a cosmetic company and provides predictions about the brand building. Manager will get help about current and future of their banding on social media, by using this data he/she can decide the promotions. Also, the manager can decide whether to post certain contents or not and if they want to post when to post, to catch maximum customer attention.

Dataset Used: *Data related to posts published on the Facebook's page of a renowned cosmetics brand*.

1. Data exploration and Data Cleaning

- **Introduction**

We used a dataset of 500 posts published in its Facebook social network brand page in a complete year by a renowned cosmetic company is considered for analysis.

The main goals of this study are as follows:

- Implementing a model that predicts the impact of posts using their characteristics
- Measuring the predictive value of the model when applied to output metric features, i.e., by evaluating the difference between the value predicted by the model and the real metric value
- Assessing the knowledge provided by the model in terms of which input features affect the impact metrics and how these input features influence each post, and hence supporting managers' decisions

- **Dataset Description**

  Data related to posts published on the Facebook's page of a renowned cosmetics brand.

The dataset contains 9 features(variables) for each of the posts. The description of each feature is listed below.

Input Features (Covariates):

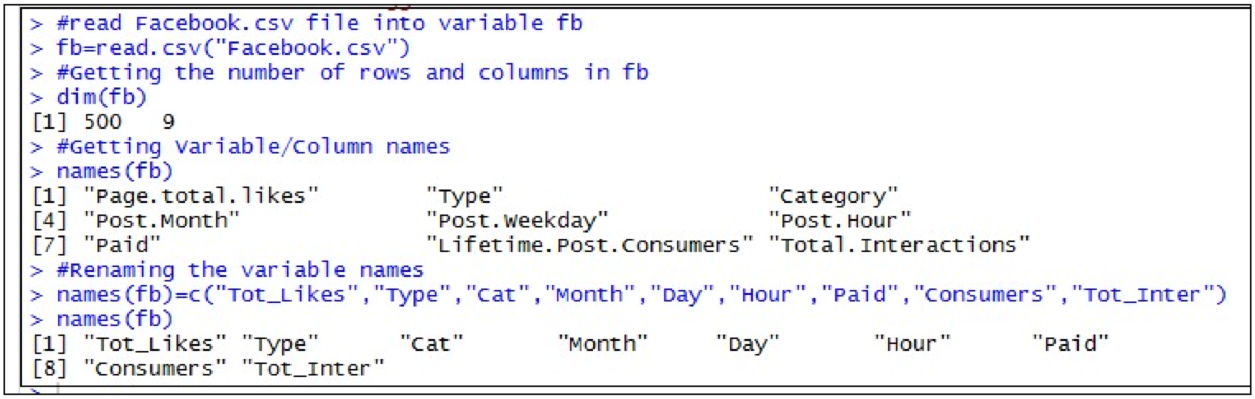
1. **Page total likes** - Number of people who have liked the company's page
2. **Type** - Type of content (Link, Photo, Status, Video).
3. **Category** – Provides the manual categorization according to the campaign to which the content posted is associated
4. **Post Month** - Month the post was published (January, February, March, …, December).
5. **Post Weekday** – Weekday the post was published (Sunday, Monday, …, Saturday)
6. **Post Hour** – Hour the post was published (0, 1, 2, 3, 4, …, 23).
7. **Paid** – If the company paid to Facebook for advertising (yes - 1, no - 0).
8. **Total Interactions** - The sum of “likes,” “comments,” and “shares” of the post.

Output Features (Response Variables):

9. Lifetime Post Consumers - The number of people who clicked anywhere in a post

- **Dataset Loading**

The dataset is named Facebook.csv, The code to read this .csv file using R into a variable "fb" and renaming the column names is as below.



```
> #read Facebook.csv file into variable fb
> fb=read.csv("Facebook.csv")
> #Getting the number of rows and columns in fb
> dim(fb)
[1] 500   9
> #Getting Variable/Column names
> names(fb)
[1] "Page.total.likes"          "Type"                      "Category"
[4] "Post.Month"                "Post.Weekday"              "Post.Hour"
[7] "Paid"                      "Lifetime.Post.Consumers"  "Total.Interactions"
> #Renaming the variable names
> names(fb)=c("Tot_Likes","Type","Cat","Month","Day","Hour","Paid","Consumers","Tot_Inter")
> names(fb)
[1] "Tot_Likes" "Type"      "Cat"       "Month"     "Day"       "Hour"      "Paid"
[8] "Consumers" "Tot_Inter"
```

- **Data Cleaning:**

  - **Checking for any missing values/null values**

    If there are any missing values in the dataset or there are NULL values, then the code to check them is as below:

```
#Checking for missing values/null values
> sum(is.na(fb))
[1] 0
> apply(fb,2, function(x){sum(is.na(x))})
Tot_Likes      Type       Cat     Month       Day      Hour      Paid Consumers Tot_Inter
        0         0         0         0         0         0         0         0         0
```

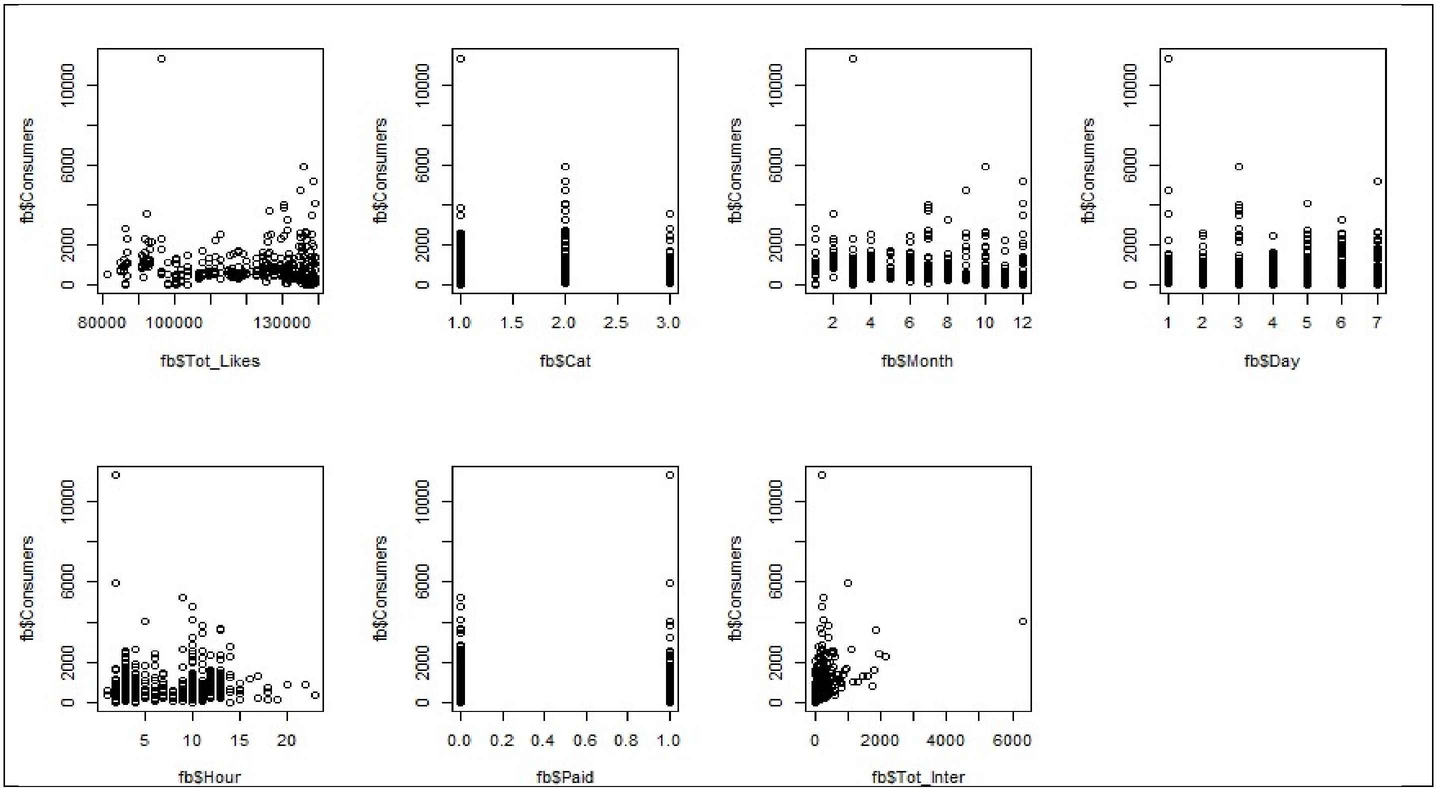
From the output, we observe that there are no missing/null values in the dataset.

  - **Checking for any outliers**

    The outliers in the given dataset can be found by drawing the plots of each covariate with the response and observing for any data point which is outside the selected range of the plot.

```
#plotting each regressor with the response variable
#Here Consumers is our response variable
par(mfrow=c(2,4))
plot(fb$Tot_Likes,fb$Consumers)
plot(fb$Cat,fb$Consumers)
plot(fb$Month,fb$Consumers)
plot(fb$Day,fb$Consumers)
plot(fb$Hour,fb$Consumers)
plot(fb$Paid,fb$Consumers)
plot(fb$Tot_Inter,fb$Consumers)
```

Output plot:



```
#Removing an oulier where fb$Consumers is maximum as it is influencing the whole plot
> fbnew=fb[-c(which(fb$Consumers==max(fb$Consumers))),]
> dim(fbnew)
[1] 499   9
```

From the output, we can see that, the outlier with maximum Consumer value is deleted and this new dataset is now stored in the variable fbnew.

2. Data Visualization and Modelling

Building a Linear Regression model for the given set for the response variable Consumers with all the covariates. The code is as below:

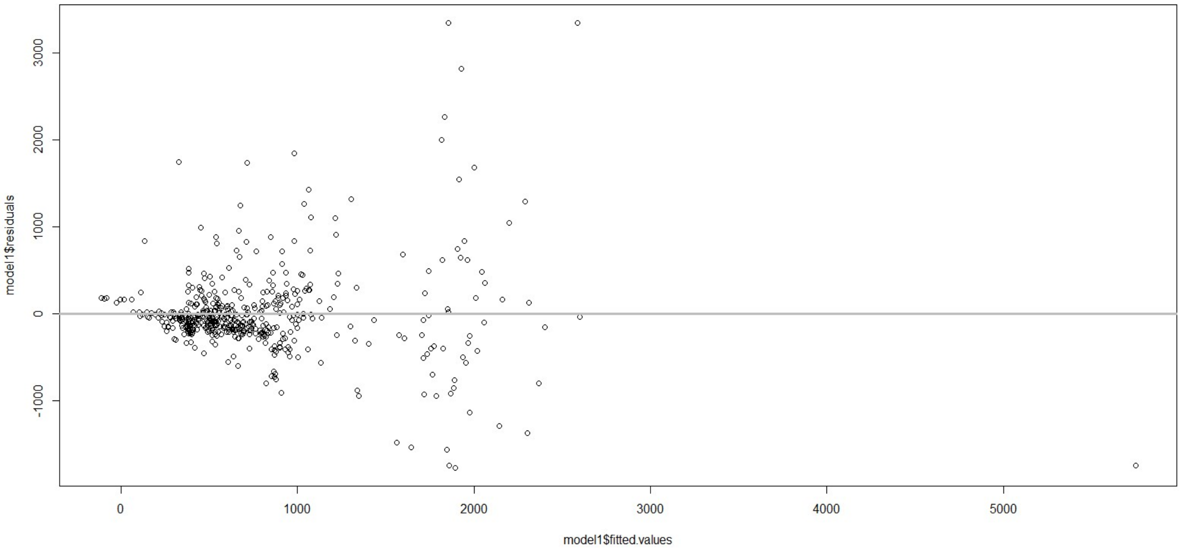
```
#Basic multiple regression taking all regressors into consideration
> model1=lm(Consumers~Tot_Likes+Type+Cat+Month+Day+Hour+Paid+Tot_Inter,data
=fbnew)
> summary(model1)

Call:
lm(formula = Consumers ~ Tot_Likes + Type + Cat + Month + Day + 
    Hour + Paid + Tot_Inter, data = fbnew)

Residuals:
    Min      1Q  Median      3Q     Max 
-1771.2  -184.6   -60.6   111.6  3348.7 

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.291e+03  3.970e+02   3.250 0.001233 ** 
Tot_Likes   -6.576e-03  4.336e-03  -1.517 0.130025    
TypePhoto    3.931e+02  1.181e+02   3.327 0.000944 ***
TypeStatus   1.835e+03  1.408e+02  13.030  < 2e-16 ***
TypeVideo    1.306e+03  2.281e+02   5.724 1.82e-08 ***
Cat         -1.086e+02  2.885e+01  -3.763 0.000189 ***
Month       -3.370e+01  2.148e+01  -1.569 0.117331    
Day          2.365e+00  1.155e+01   0.205 0.837796    
Hour         2.330e+00  5.536e+00   0.421 0.673944    
Paid         5.424e+01  5.257e+01   1.032 0.302700    
Tot_Inter    8.379e-01  6.216e-02  13.480  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 517.1 on 488 degrees of freedom
Multiple R-squared:  0.5299,	Adjusted R-squared:  0.5203 
F-statistic:    55 on 10 and 488 DF,  p-value: < 2.2e-16
```

Model Checking

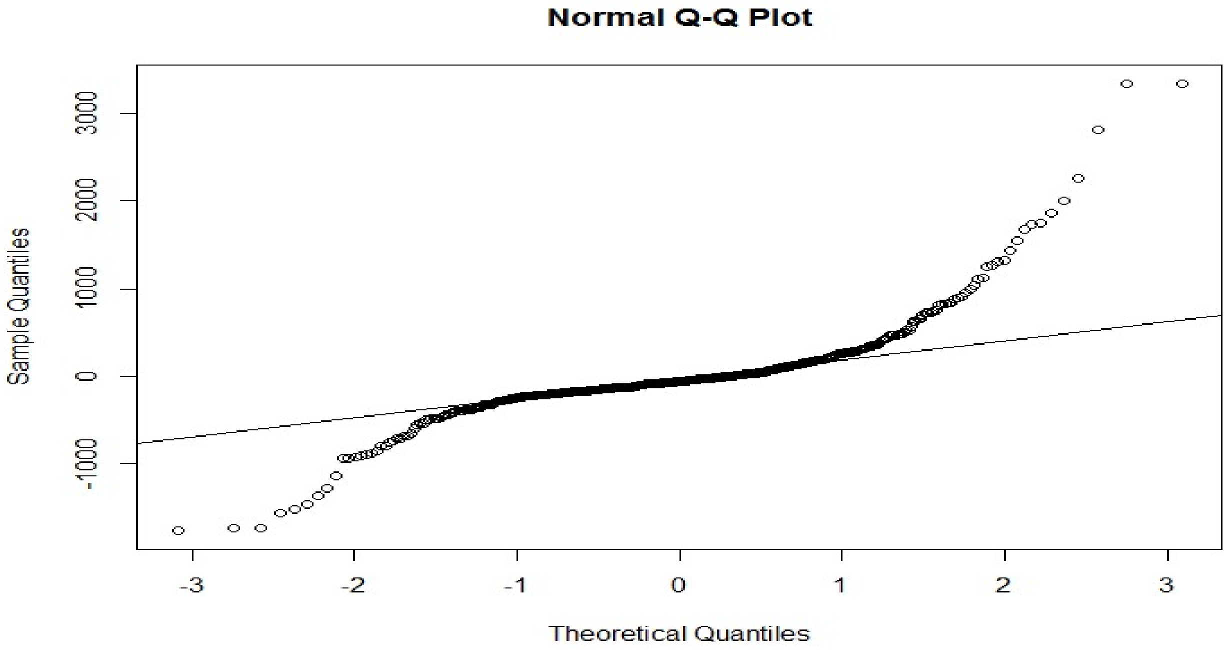
Now, for checking this model1, we plot the residual values against the fitted values and draw the qqplots to check for the variance.

```
#Plotting the residual vs fitted values
plot(model1$fitted.values,model1$residuals)
abline(h=0,col="grey",lwd=3)
```



From the above plot when checked for equal variance, the pattern for this residual plot is in funnel shaped. So, it is not as expected to be.

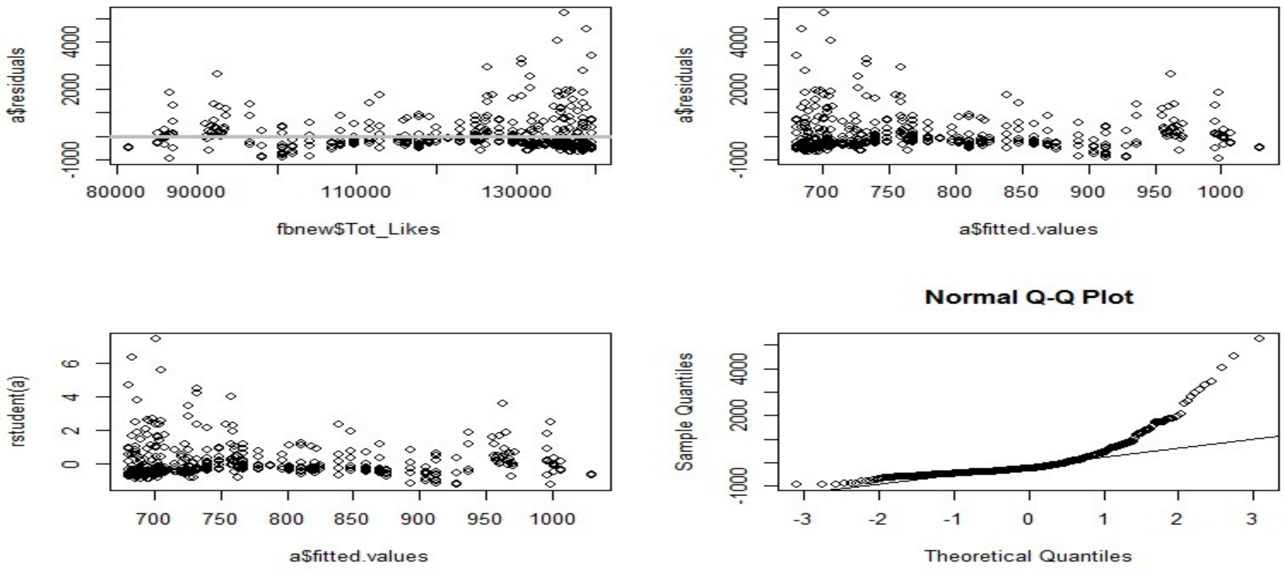
```
#Plotting the qqplot
qqnorm(model1$residuals)
qqline(model1$residuals)
```



From the qq plot, to observe the normality assumption, the plot is not ideal and instead heavy tailed distribution is observed. So, it is not as expected to be.

So, now checking the model of individual covariates with the response variable and drawing the residual plot and qq plots.

```
#Implementing the linear regression for each of the input vs y and plotting
the residual vs fitted values
a=lm(fbnew$Consumers~fbnew$Tot_Likes)

plot(fbnew$Tot_Likes,a$residuals)
abline(h=0,col="grey",lwd=3)

plot(a$fitted.values,a$residuals)
plot(a$fitted.values,rstudent(a))

#Checking the qqplot
qqnorm(a$residuals)
qqline(a$residuals)
```



So, from the above plots, we observe the same result, the regression model with one covariate to the response, there is no normal distribution and is of unequal variance.

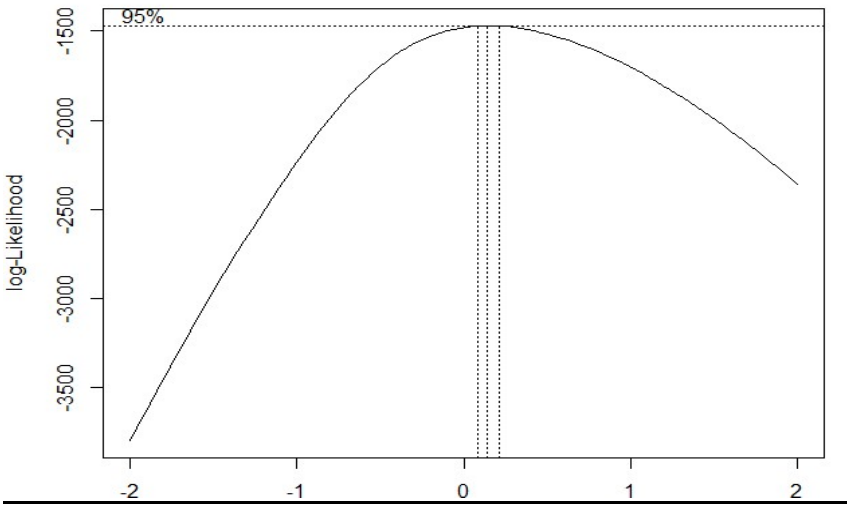
Similarly, when the same is done with other covariates with respect to response, the result is same, no normal distribution and unequal variance.

3. Re-exploration of data

- **Transformation of variables:**

Since, the models are not as expected, we need to transform the variables.
Boxcox plot for the model gives the λ value through which we can determine what transformation to do on each of the variables.

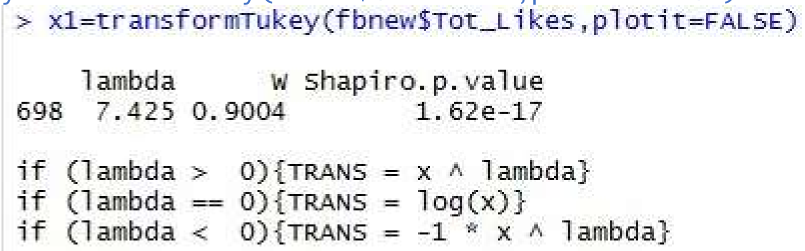
```
#Doing boxcox to know the λ value
library(car)
require(MASS)
boxcox(a)
```



We get the same box-cox plot for each of the regressors if taken individually with response. Since λ =! 1, we need to transform the variables accordingly.

Code to calculate λ value for each variable and applying this λ value to transform

```
install.packages("rcompanion")
library(rcompanion)

#Code to get λ values
x1=transformTukey(fbnew$Tot_Likes,plotit=FALSE)
x2=transformTukey(fbnew$Cat,plotit=FALSE)
x3=transformTukey(fbnew$Month,plotit=FALSE)
x4=transformTukey(fbnew$Day,plotit=FALSE)
x5=transformTukey(fbnew$Hour,plotit=FALSE)
x6=transformTukey(fbnew$Paid,plotit=FALSE)
x7=transformTukey(fbnew$Tot_Inter,plotit=FALSE)
y=transformTukey(fbnew$Consumers,plotit=FALSE)
```



```
> x1=transformTukey(fbnew$Tot_Likes,plotit=FALSE)

    lambda      W Shapiro.p.value
698  7.425 0.9004        1.62e-17

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```

From the above output of x1, the λ value is 7.425 and since it is greater than 0, the transformation of this x1 would be x1^7.425. Below is the code to perform transformations on all the variables similarly.

```
#Now the transform variables depending on the λ values
t1=fbnew$Tot_Likes^7.425
t2=fbnew$Cat^0.775
t3=fbnew$Month^0.925
t4=fbnew$Day^0.975
t5=fbnew$Hour^0.85
t6=fbnew$Paid^0.025
t7=fbnew$Tot_Inter^0.25
ty=fbnew$Consumers^0.15
```

So, here, t1, t2…t7 are the transformed covariances and ty is the transformed response variable.

Building a regression model for this transform variable

```
#Multiple Linear regression with the transformed values
m=lm(ty~t1+t2+t3+t4+t5+t6+t7)
summary(m)
```

- **Checking for multicollinearity:**

Before doing the regression model, we check for multicollinearity of the transformed covariances. To do this, we use the Variance Inflation factor VIF to check for multicollinearity. Below is the code:

```
#Checking for Multicollinearity
> vif(m)
       t1        t2        t3        t4        t5        t6        t7 
35.050092  1.144043 35.000491  1.015985  1.061767  1.023984  1.142918 
```

From, the output, we observe that VIF of t1 & t3 are high, so to simplify our regression model, we eliminate these covariances and build the regression model.

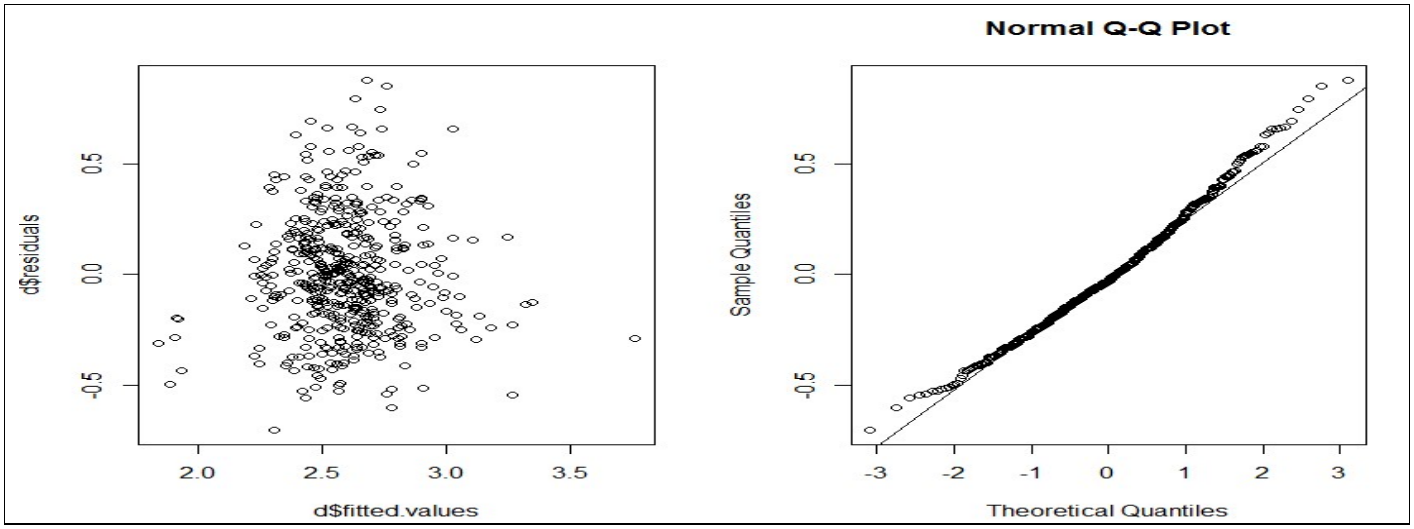
4. Re-modeling

Building a multiple regression model for these refined transformed variables, below is the code.

```
#Removing t1,t3 due to high VIF, Final regression is as below
> d=lm(ty~t2+t4+t5+t6+t7)
> summary(d)

Call:
lm(formula = ty ~ t2 + t4 + t5 + t6 + t7)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7046 -0.1816 -0.0308  0.1640  0.8791 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.858416   0.064331  28.889  < 2e-16 ***
t2          -0.066144   0.022142  -2.987  0.00296 ** 
t4           0.008490   0.006337   1.340  0.18092    
t5           0.012545   0.004421   2.838  0.00473 ** 
t6           0.013281   0.027181   0.489  0.62534    
t7           0.215987   0.012850  16.809  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2685 on 492 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.377,	Adjusted R-squared:  0.3706 
F-statistic: 59.53 on 5 and 492 DF,  p-value: < 2.2e-16
```

Now we will check the residual plots and QQ plots for transformed model

```
#Checking the residual plots & qq-plots
plot(d$fitted.values,d$residuals)
qqnorm(d$residuals)
qqline(d$residuals)
```



From, the above result, we get the satisfactory results.

5. Model Selection:

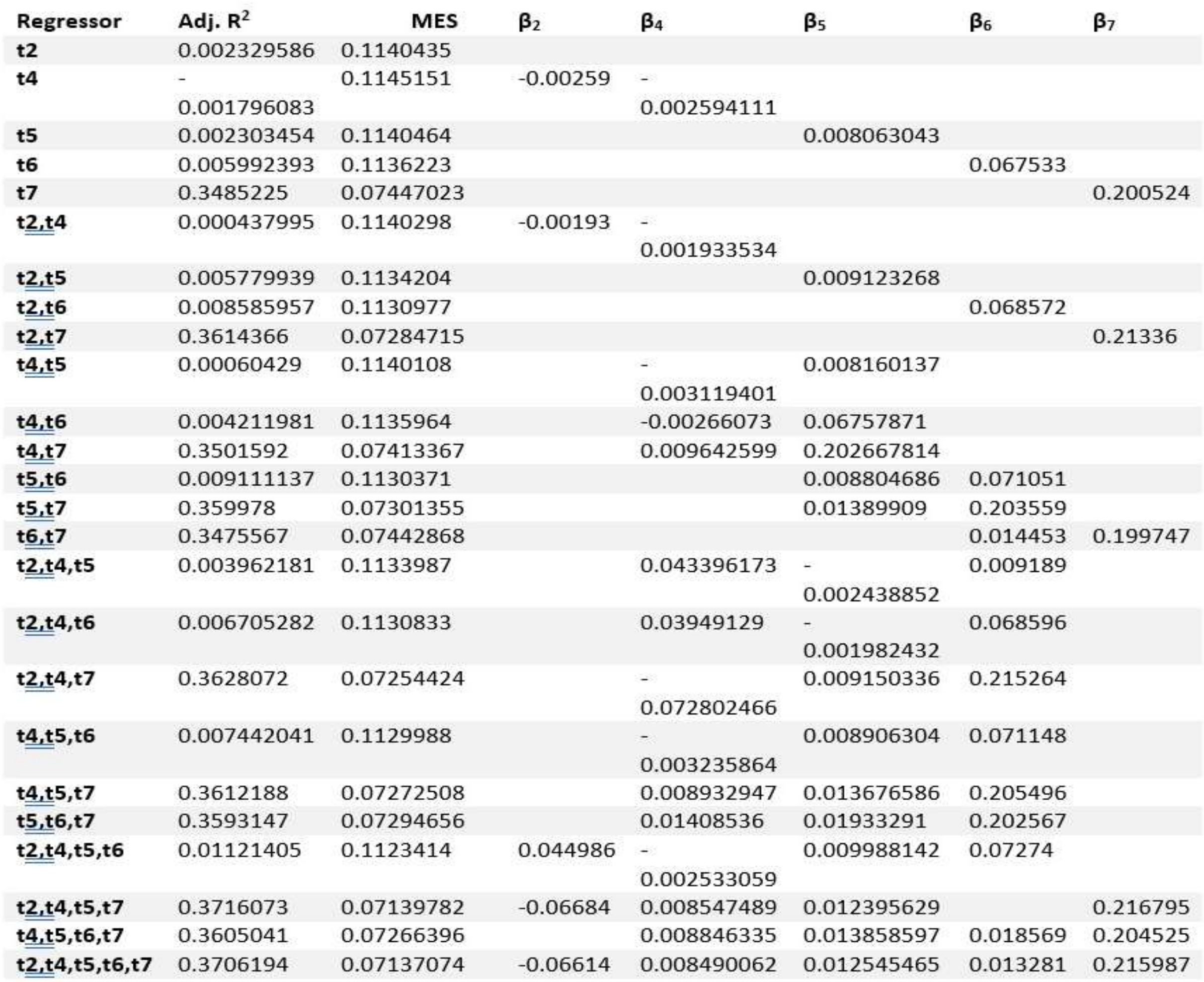
To select a suitable model for the analysis, we get the summary of all regressors. Summary includes the Adjusted R square value, Mean Sum of squares of residuals, the estimated slope values.

Below is the code:

```
###Model Selection
##Getting the summary of all regressors
> d=lm(ty~t2)
> summary(d)$coef[,1]
(Intercept)          t2
 2.53183161  0.03875706
> summary(d)$adj.r.square
[1] 0.002329586
> ssres=sum((y-d$fitted.values)^2)
> Mres=ssres/(499-1-1)
> Mres
[1] 0.1140435
```

Similarly, we will calculate the summary for all the regressors. And below is the report of it.

Model checking Finalizing the model:



Regressor	Adj. R^2	MES	β_2	β_4	β_5	β_6	β_7
t2	0.002329586	0.1140435					
t4	-0.001796083	0.1145151	-0.00259	-0.002594111			
t5	0.002303454	0.1140464			0.008063043		
t6	0.005992393	0.1136223				0.067533	
t7	0.3485225	0.07447023					0.200524
t2,t4	0.000437995	0.1140298	-0.00193	-0.001933534			
t2,t5	0.005779939	0.1134204			0.009123268		
t2,t6	0.008585957	0.1130977				0.068572	
t2,t7	0.3614366	0.07284715					0.21336
t4,t5	0.00060429	0.1140108		-0.003119401	0.008160137		
t4,t6	0.004211981	0.1135964		-0.00266073	0.06757871		
t4,t7	0.3501592	0.07413367		0.009642599	0.202667814		
t5,t6	0.009111137	0.1130371			0.008804686	0.071051	
t5,t7	0.359978	0.07301355			0.01389909	0.203559	
t6,t7	0.3475567	0.07442868				0.014453	0.199747
t2,t4,t5	0.003962181	0.1133987		0.043396173	-0.002438852	0.009189	
t2,t4,t6	0.006705282	0.1130833		0.03949129	-0.001982432	0.068596	
t2,t4,t7	0.3628072	0.07254424		-0.072802466	0.009150336	0.215264	
t4,t5,t6	0.007442041	0.1129988		-0.003235864	0.008906304	0.071148	
t4,t5,t7	0.3612188	0.07272508		0.008932947	0.013676586	0.205496	
t5,t6,t7	0.3593147	0.07294656		0.01408536	0.01933291	0.202567	
t2,t4,t5,t6	0.01121405	0.1123414	0.044986	-0.002533059	0.009988142	0.07274	
t2,t4,t5,t7	0.3716073	0.07139782	-0.06684	0.008547489	0.012395629		0.216795
t4,t5,t6,t7	0.3605041	0.07266396		0.008846335	0.013858597	0.018569	0.204525
t2,t4,t5,t6,t7	0.3706194	0.07137074	-0.06614	0.008490062	0.012545465	0.013281	0.215987

From the above report calculated, we observe that for the regression model, t2, t4, t5, t7 has the highest Adjusted R Square value and suitable slope values.

Now, building the regression model for these covariates with the response:

```
#From checking the values, we select t2, t4, t5, t7 as our covariates
> dnew=lm(ty~t2+t4+t5+t7)
> summary(dnew)

Call:
lm(formula = ty ~ t2 + t4 + t5 + t7)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.70688 -0.18439 -0.02653  0.16485  0.87541 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.861125   0.063978  29.090  < 2e-16 ***
t2          -0.066837   0.022055  -3.030  0.00257 ** 
t4           0.008547   0.006324   1.352  0.17714    
t5           0.012396   0.004401   2.816  0.00505 ** 
t7           0.216795   0.012717  17.047  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.268 on 494 degrees of freedom
Multiple R-squared:  0.3767,	Adjusted R-squared:  0.3716 
F-statistic: 74.62 on 4 and 494 DF,  p-value: < 2.2e-16
```

6. Prediction:

We compute the confidence interval of the model above and the predictions using confint() function and predict() function.

```
#Getting the confidence interval of the model and the predictions
> confint(dnew)
                   2.5 %      97.5 %
(Intercept)  1.735421392  1.98682833
t2          -0.110170779 -0.02350254
t4          -0.003878360  0.02097334
t5           0.003748153  0.02104311
t7           0.191807899  0.24178170
> head((predict(dnew,fbnew,interval = c("pred"),level = 0.95,type="response"))^(1/0.15))
        fit       lwr      upr
1  445.9853  91.55470 1601.707
2  639.0075 144.16664 2156.571
3  351.5893  67.32887 1320.645
4 2633.0438 800.41136 7229.050
5  870.5034 210.91830 2800.224
6  640.4405 144.10212 2165.281
```

Making the predictions for various other values, below is the code:
Here, t2 - transformation of Category; t4 – transformation of Post Weekday; t5- transformation of Post Hour; t7- transformation of Total Interaction.
From the above output, we get the prediction values of the response, i.e., Lifetime Post Consumers when different other values of inputs are given.

Conclusion:

We are predicting social media performance metrics and evaluation of the impact on brand building for data extracted from Facebook page of a cosmetic company. After analyzing our dataset Facebook.csv, we found that the Lifetime Post Consumers model will be best suitable for predicting the customer response to the advertisement posts. We have built a regression model with few of the input features (Category of the post, Post Weekday, Post Hour and Total Interactions) and predicted the response (Lifetime Post Consumers). We found that there is need of applying a transformation, hence have transformed variables, ignored few input features which are not significant.

By using the prediction of lifetime post consumers, managers can decide when to post an advertisement and on which particular day or hour there are huge consumer responses. And thus, this prediction model will help managers to evaluate online branding of different cosmetic products.