Social Media Influence and Brand Engagement

Beth Hilbert, Gabrielle Peters and Katelynn Blalock

Dataset - http://archive.ics.uci.edu/ml/datasets/Facebook+metrics

Introduction

This research is focused on analyzing social media's role in influencing customers by measuring the impact of status updates and advertisements on Facebook for a particular cosmetic brand. The data used is from the UCI Machine Learning Repository and includes the Facebook Metrics Data Set (see note 1) containing performance metrics of a renowned cosmetic's brand Facebook page. The response variable in the linear regression will be "Page Total Likes". By making the assumption that the total number of page likes is a good representation of brand reputation and social media engagement with consumers this will allow for the analysis of other metics' impact on customer engagement via social media.

(Note 1) Our dataset is from the UCI Machine Learning Repository. It can be downloaded from http://archive.ics.uci.edu/ml/datasets/Facebook+metrics. Citation: (Moro et al., 2016) Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.

1. Dataset Exploration and Data Cleaning

Environment Setup

As a first step for the data exploration and cleaning we need to setup the RStudio environment by first installing the necessary packages and importing the related libraries. This includes car, ggplot2, mass and psych. Before importing the data, we clear the environment to remove an in memory variables. The last step in the setup is to read in the dataset and rename the variables for easy handling.

Exploratory Data Analysis

After the data is imported we analyze the data by reviewing the number of observations and the new variable names. There are 500 observations and 19 variables. The new variable names are as follows:

Page.Total.Likes
Type
Category
Post.Month
Post.Weekday
Post.Hour
Paid
Total.Reach
Total.Impressions
Engaged.Users
Consumers
Consumptions
Impressions.for.Users.with.Likes
Reach.by.Users.with.Likes
Users.with.Likes.and.Engagement
Comment
Like
Share
Total.Interactions

In reviewing the descriptive statistics, we found some variables are highly skewed with means and medians significantly different (such as Total.Reach). One of the variables in non-numeric (Type). Some of the variables only have a few discrete options (such as Category has 1, 2, 3).

We also reviewed the dataset for missing values and we observed that it does contains missing values, which could potentially cause problems within our model. The impact of missing values can be serious, leading to biased estimates, loss of information, decreased statistical power, increased standard errors, and weakened interpretation of findings. In order to understand the magnitude of missing values within our dataset, we ran a command to determine the percentage of missing values and determined the following variables had the corresponding percentage of missing values:

Paid .2%
Like .2%
Share .2%

Cleaning Data

As part of the data cleaning we evaluated two methods for handling the missing data values: 1) replacing missing values with column mean, or 2) replacing missing data with 0's.

The main reason for imputing values is to reduce bias due to missing values in order to maintain the sample size. This results in a potentially higher efficiency than deleting observations with missing values. This allows us to utilize the collected data in a complete dataset.

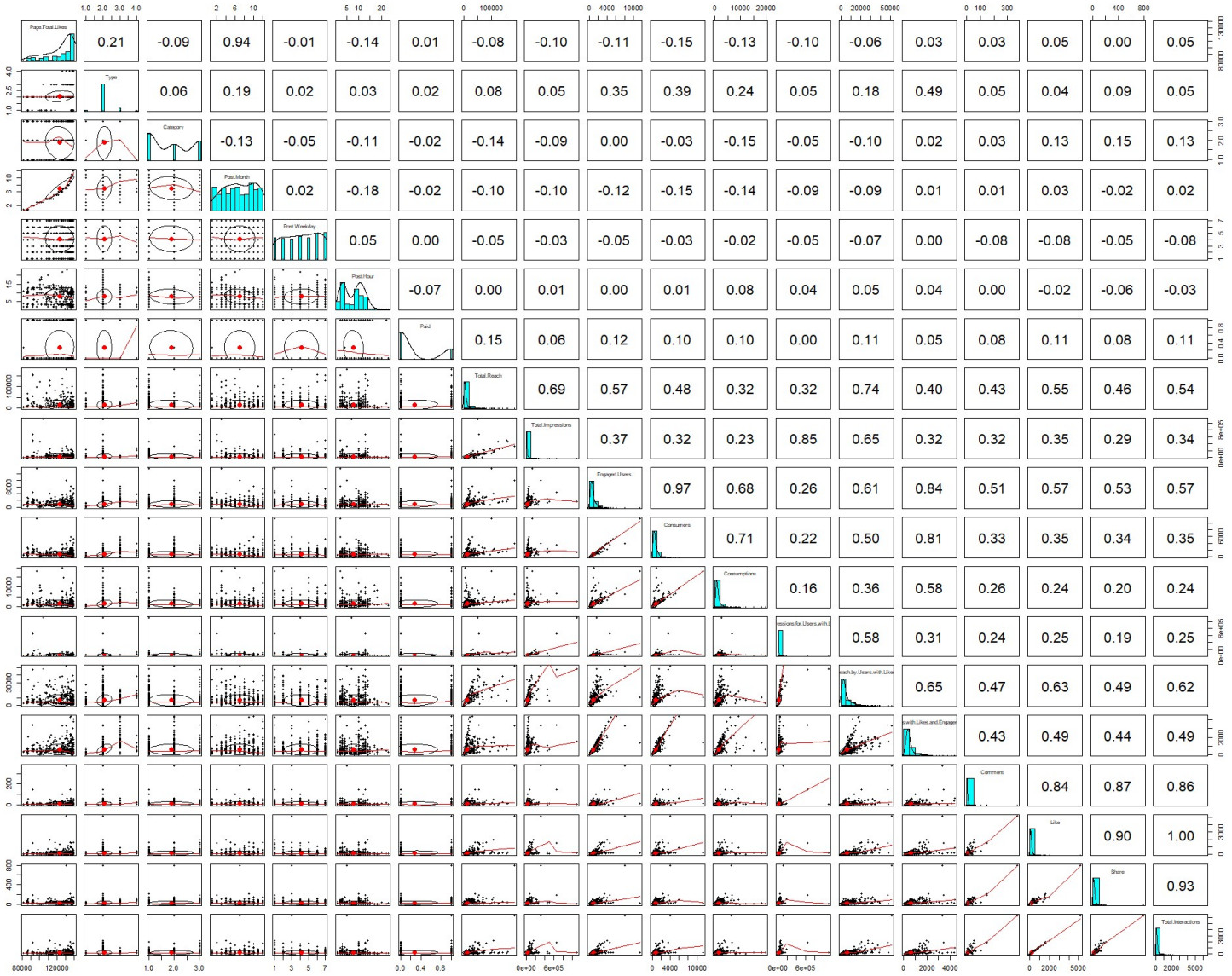
By replacing the missing values with the column mean average we run the risk of multicollinearity, which exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity is a problem because it undermines the statistical significance of an independent variable.

Justification for using the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors within the dataset.

By replacing the missing values with 0's, we will tend to underestimate the standard errors and overestimate the level of precision. Thus, a single imputation of "0" gives us more apparent power than the dataset in reality.

Scatter Plot, Histogram, and Correlation Coefficient

We only visualized data with numeric variables; however, we could convert non-numeric variables in order for us to make interpretations on the full data set. Three functions are included in this one matrix. The lower left shows pairwise combinations of continuous variables in scatterplots. The histogram down the diagonal shows the data distribution of each variables. The correlation coefficients in upper right show variables that might be related. This can help pinpoint variables that may have similar correlations to our dataset.



2. Initial Model Building

This initial model is looking at all 19 variables within the dataset, with Page.Total.Likes as the chosen response variable. We plan to look at each variable and determine which ones are significant to build a final, reliable model.

```
Call:
lm(formula = Page.Total.Likes ~ Post.Month + Post.Weekday + Post.Hour +
    Paid + Total.Reach + Total.Impressions + Engaged.Users +
    Consumers + Consumptions + Impressions.for.Users.with.Likes +
    Reach.by.Users.with.Likes + Users.with.Likes.and.Engagement +
    Comment + Like + Share + Total.Interactions, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-13403.2  -4201.6    284.2   4557.3  10618.0

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      9.086e+04  1.037e+03  87.643  < 2e-16 ***
Post.Month                       4.590e+03  8.131e+01  56.447  < 2e-16 ***
Post.Weekday                    -1.738e+02  1.205e+02  -1.441  0.15010
Post.Hour                        6.090e+01  5.732e+01   1.062  0.28855
Paid                             8.902e+02  5.489e+02   1.622  0.10552
Total.Reach                      1.006e-01  4.697e-02   2.142  0.03268 *
Total.Impressions               -3.139e-02  1.678e-02  -1.871  0.06197 .
Engaged.Users                   -1.786e+01  7.746e+00  -2.306  0.02156 *
Consumers                        1.515e+01  7.743e+00   1.957  0.05097 .
Consumptions                     7.309e-02  1.780e-01   0.411  0.68151
Impressions.for.Users.with.Likes 2.364e-02  1.859e-02   1.272  0.20413
Reach.by.Users.with.Likes       -1.055e-01  1.001e-01  -1.054  0.29240
Users.with.Likes.and.Engagement  3.013e+00  1.015e+00   2.968  0.00314 **
Comment                          1.860e+01  3.749e+01   0.496  0.62010
Like                             1.401e+01  2.590e+01   0.541  0.58869
Share                           -3.599e+01  3.199e+01  -1.125  0.26124
Total.Interactions               2.833e+00  2.626e+01   0.108  0.91415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5377 on 483 degrees of freedom
Multiple R-squared:  0.8943,    Adjusted R-squared:  0.8908
F-statistic: 255.4 on 16 and 483 DF,  p-value: < 2.2e-16
```

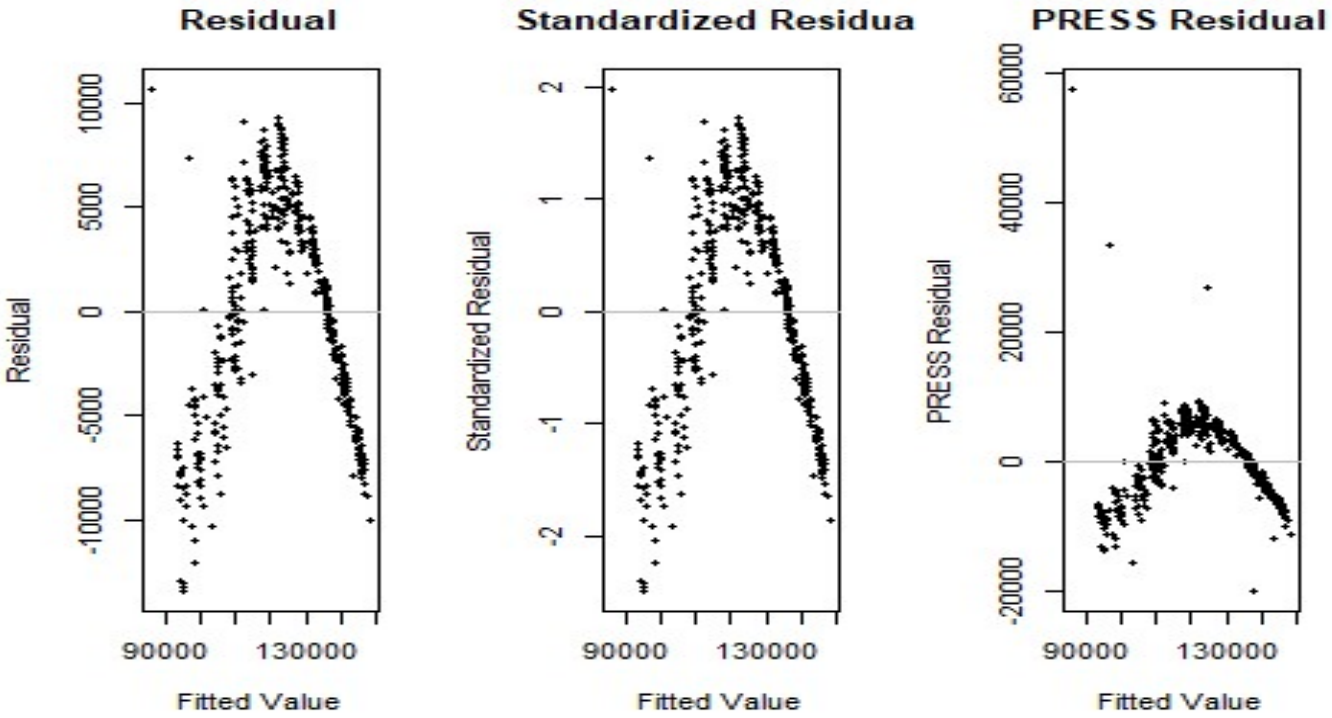
3. Model Adequacy Checking

Model adequacy is evaluated by evaluating residuals. This checks the fit of the model to the observed data. These are referred to as the LINE assumptions. Gross violations of these assumptions can result in an unstable model in which different samples result in totally different models.

The LINE assumptions are:
- Linear relationship
- Independent errors
- Normally distributed errors
- Equal variance in errors

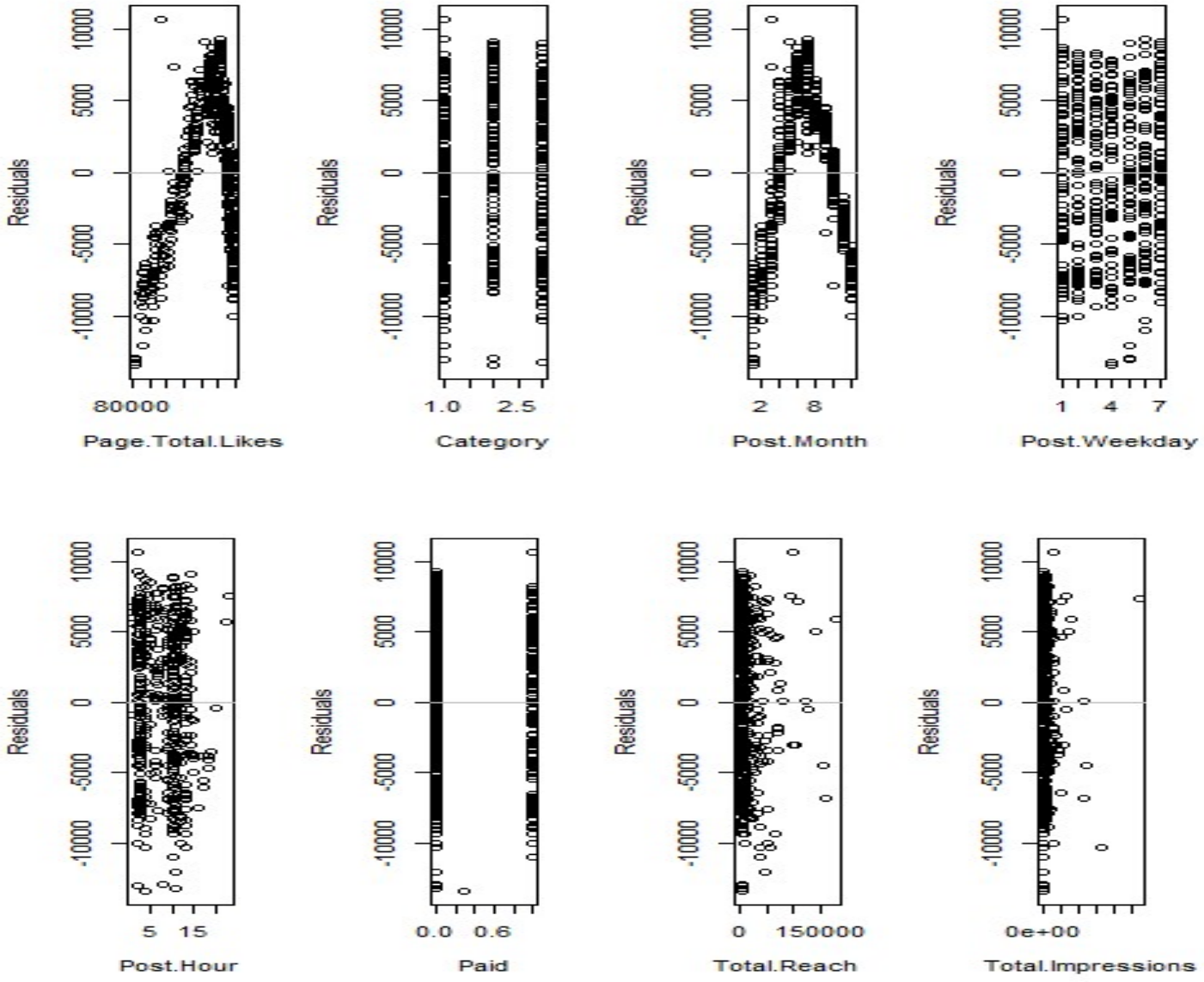
Different Types of Residuals

Residuals measure the variability in the response not explained by the model. There are several ways to measure residuals. For most of our analysis we used the definition of residual as the observed data minus the fitted data (plot1). But here we also show two results of scaling residuals (which are helpful in identifying outliers and extreme values). Standardized Residuals (plot2) shows the residual standardized by its standard deviation. PRESS residuals (plot3) shows how well the model may perform in predicting new data.



Checking Linearity and Equal Variance

We reevaluate all variables in the dataset to see if they are viable options to remain covariates in our final model. To do this we plot each covariate against the residuals in order to determine if and how they need to be transformed. Each are expected to be linear and equal variance. After reviewing these plots we can see that the Post.Month needs to be transformed because of its nonlinearity, it violates the linearity of errors assumption.



Residuals
10000
5000
0
-5000
-10000
0
8000
Engaged.Users

Residuals
10000
5000
0
-5000
-10000
0
8000
Consumers

Residuals
10000
5000
0
-5000
-10000
0
15000
Consumptions

Residuals
10000
5000
0
-5000
-10000
0e+00
pressions.for.Users.with

Residuals
10000
5000
0
-5000
-10000
0
40000
Reach.by.Users.with.Likes.and.Enga

Residuals
10000
5000
0
-5000
-10000
0
3000
Users.with.Likes.and.Enga

Residuals
10000
5000
0
-5000
-10000
0
200
Comment

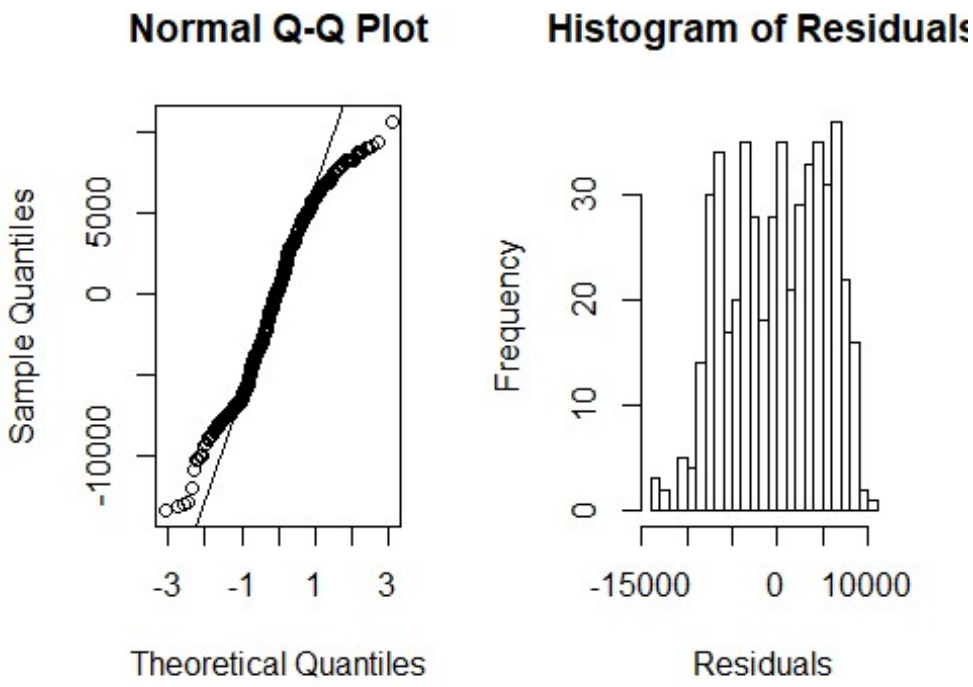
Residuals
10000
5000
0
-5000
-10000
0
4000
Like

Residuals
10000
5000
0
-5000
-10000
0
600
Share

Residuals
10000
5000
0
-5000
-10000
0
4000
Total.Interactions

Check Normality Assumption

To check this assumption we plot each covariate against the residual. We use the QQ plot to check the normality assumption. On the plot we plot the residuals for the model and show the QQ line to assist in visualizing the normal distribution. The plot is shown below. From this we can see that the residual plot appears to be a light tailed distribution. The histogram below confirms this.

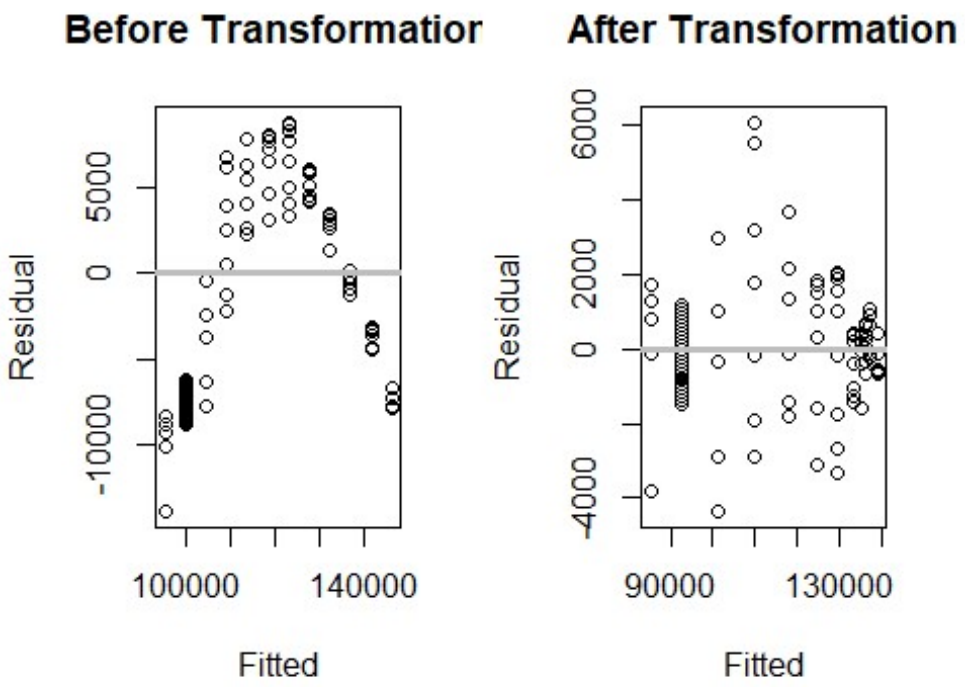


4. Transformation

Initially only Posts.Month needed transforming. We tried a number of different transformations but found that the most successful transformation was a higher order ploynomial transformation. Part of the transformations with higher order involves centering the variable's data in attempt to remove multicollinearity. We also reviewed the variance inflation factor of the variable. Sometimes, centering the regressor variables can minimize or eliminate at least some of the ill-conditioning that may be present in a polynomial model.

We can visualize the transformation by creating a model with the original Post.Month values and a second with the transformed Post.Month values and plotting both against residuals.

```
# centering
dataset$Post.Month.Centered = dataset$Post.Month - mean(dataset$Post.Month)

# transforming
lm(dataset$Page.Total.Likes ~ dataset$Post.Month.Centered+I(dataset$Post.Month.Centered^2
)+I(dataset$Post.Month.Centered^3)+I(dataset$Post.Month.Centered^4))
```



5. Variable Selection

Variance inflation factors (VIF) are very useful in determining if multicollinearity is present. Multicollinearity occurs when independent variables are strongly correlated and can give very wrong estimates for the betas (intercepts and slopes). The square root of VIF indicates how much larger the standard error is compared with what it would be if that variable were uncorrelated.

After reviewing the VIF results, we see there are 12 variables with a VIF score higher than 10. We proceeded to do backwards elimination of variables, eliminating the highest VIF score each time. After this process we eliminated Post.Month, Total.Impressions, Engaged.Users and Total.Interactions.

Initial VIF

dataset$Post.Month.Centered	I(dataset$Post.Month.Centered^2)	I(dataset$Post.Month.Centered^3)
6.886723	12.692696	8.572386
I(dataset$Post.Month.Centered^4)	dataset$Post.Weekday	dataset$Post.Hour
15.145175	1.044128	1.103117
dataset$Paid	dataset$Total.Reach	dataset$Total.Impressions
1.052023	19.989041	29.360192
dataset$Engaged.Users	dataset$Consumers	dataset$Consumptions
1038.950206	831.744644	2.202810
dataset$Impressions.for.Users.with.Likes	dataset$Reach.by.Users.with.Likes	dataset$Users.with.Likes.and.Engagement
21.584520	10.272602	6.767960
dataset$Comment	dataset$Like	dataset$Share
10.903958	1220.083972	32.025621
dataset$Total.Interactions		
1725.805011		

Final VIF

dataset$Post.Hour	dataset$Paid	dataset$Total.Reach
1.035778	1.042742	3.193144
dataset$Consumers	dataset$Consumptions	dataset$Impressions.for.Users.with.Likes
4.996422	2.087754	1.722414
dataset$Reach.by.Users.with.Likes	dataset$Users.with.Likes.and.Engagement	dataset$Comment
6.001376	5.170489	4.605599
dataset$Like	dataset$Share	
8.382640	7.791800	

6. Re-modeling

Standardized Regression Coefficients

To begin remodeling we start by comparing the influence of each of the variables. It's often difficult to directly compare regression coefficients due to possible varying dimensions, so we scale them. Dimensionless regression coefficients are referred to as standardized regression coefficients. This allows us to compare the relative strength of the coefficients. We do this by converting the units of each coefficient to a standard unit of measure and viewing the summary.

```
Call:
lm(formula = Page.Total.Likes ~ Post.Hour + Paid + Total.Reach +
    Consumers + Consumptions + Impressions.for.Users.with.Likes +
    Reach.by.Users.with.Likes + Users.with.Likes.and.Engagement +
    Comment + Like + Share, data = dataset_standard)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7707 -0.5682  0.3185  0.6764  3.8655

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.807e-15  4.115e-02   0.000 1.000000
Post.Hour                       -1.543e-01  4.192e-02  -3.681 0.000258 ***
Paid                             1.779e-02  4.206e-02   0.423 0.672609
Total.Reach                      1.470e-01  7.361e-02   1.997 0.046413 *
Consumers                       -5.451e-01  9.208e-02  -5.920 6.07e-09 ***
Consumptions                    -6.009e-02  5.952e-02  -1.010 0.313201
Impressions.for.Users.with.Likes -5.297e-02  5.406e-02  -0.980 0.327623
Reach.by.Users.with.Likes       -3.350e-01  1.009e-01  -3.320 0.000969 ***
Users.with.Likes.and.Engagement  6.487e-01  9.367e-02   6.925 1.38e-11 ***
Comment                          1.532e-01  8.840e-02   1.733 0.083669 .
Like                             3.500e-01  1.193e-01   2.935 0.003493 **
Share                           -4.444e-01  1.150e-01  -3.865 0.000126 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9202 on 488 degrees of freedom
Multiple R-squared:  0.1719,    Adjusted R-squared:  0.1533
F-statistic: 9.212 on 11 and 488 DF,  p-value: 4.478e-15
```

Partial F Test

We used the standardized regression coefficient to determine which coefficients were potentially insignificant. We used a progression of partial F tests to eliminate these insignificant variables from the model.

To further reduce our model we used these results from the standardized regression coefficients to determine which variables to include in a partial F-test. Our partial F-test tested a subset of variables to see if the slope parameter is significant.

If the resulting p-value is greater than .05 we fail to reject the null hypothesis. This means that after adjusting for other regressors not in the partial F-test in the linear regression, these regressors in the subset are not significant as their slopes are not significantly different than zero. We eliminated the following as the partial F-test showed that the p-values of the following are greater than .05 and were thus eliminated from the model: Comment (0.08367) Paid (0.2045) Impressions.for.Users.with.Likes (0.2938) Consumptions (0.3682) Total.Reach (0.09598)

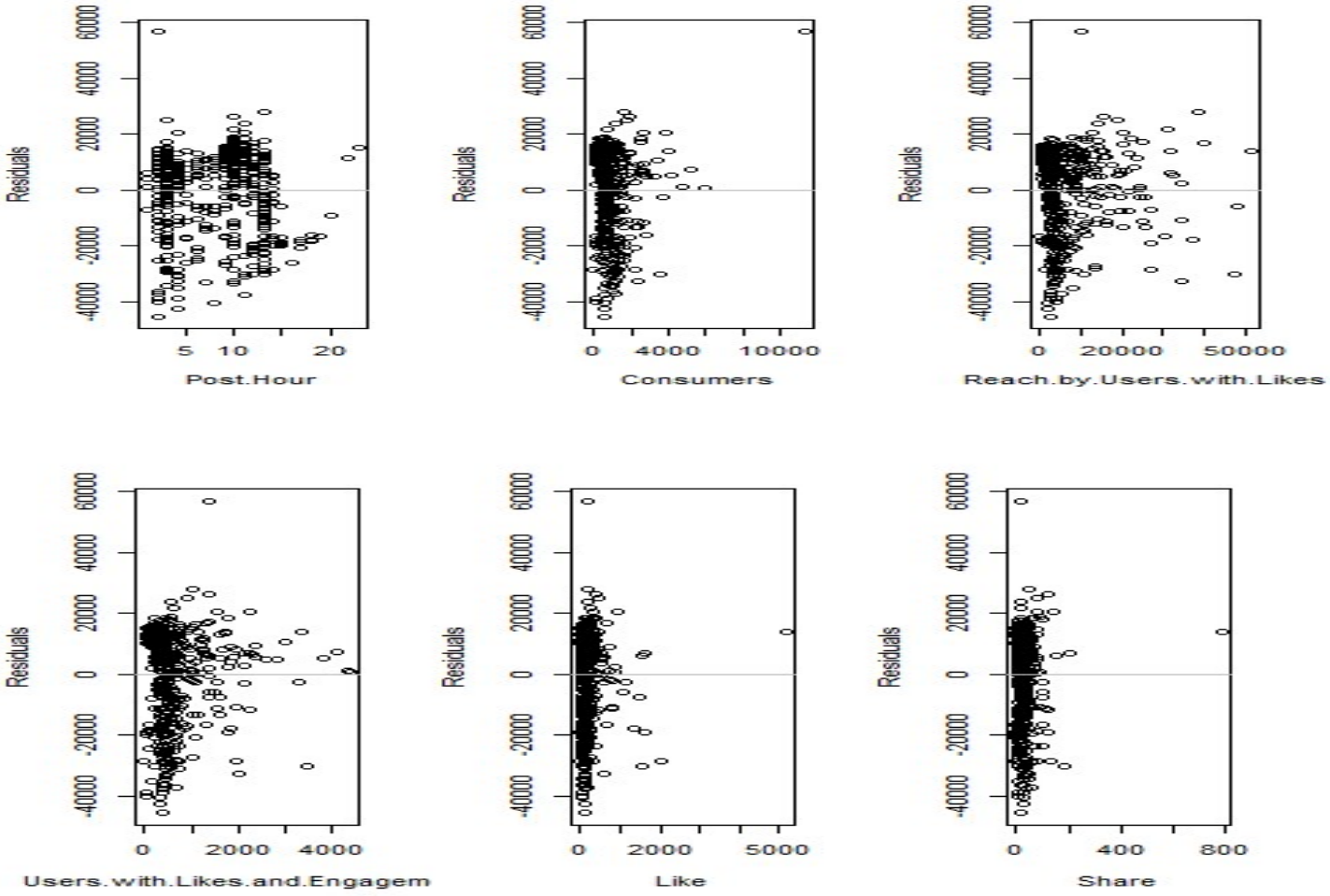
After eliminating these variables from the standardized model, we can see from the results of the summary that all remaining variables are significant.

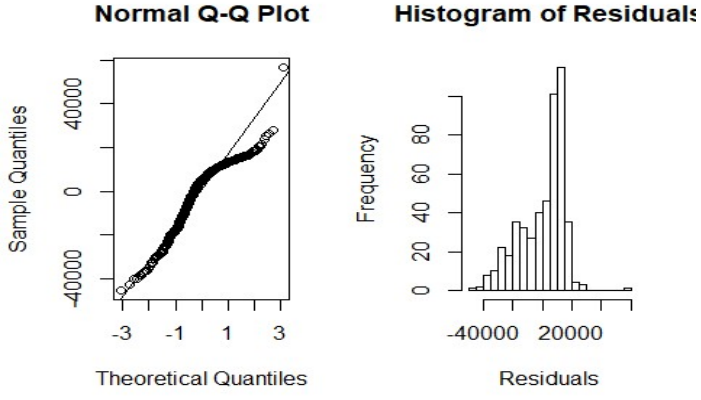
```
Call:
lm(formula = Page.Total.Likes ~ Post.Hour + Consumers + Reach.by.Users.with.Likes +
    Users.with.Likes.and.Engagement + Like + Share, data = dataset_standard)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7884 -0.6463  0.3237  0.7094  3.4845

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.775e-15  4.133e-02   0.000 1.000000
Post.Hour                     -1.548e-01  4.167e-02  -3.715 0.000227 ***
Consumers                     -4.980e-01  7.223e-02  -6.895 1.65e-11 ***
Reach.by.Users.with.Likes     -2.661e-01  6.436e-02  -4.134 4.19e-05 ***
Users.with.Likes.and.Engagement  5.577e-01  8.198e-02   6.803 2.98e-11 ***
Like                          4.297e-01  1.140e-01   3.768 0.000184 ***
Share                         -3.468e-01  1.017e-01  -3.409 0.000704 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9243 on 493 degrees of freedom
Multiple R-squared:  0.156,     Adjusted R-squared:  0.1457
F-statistic: 15.19 on 6 and 493 DF,  p-value: 5.479e-16
```

7. Model Adequacy Re-checking

Check Residuals

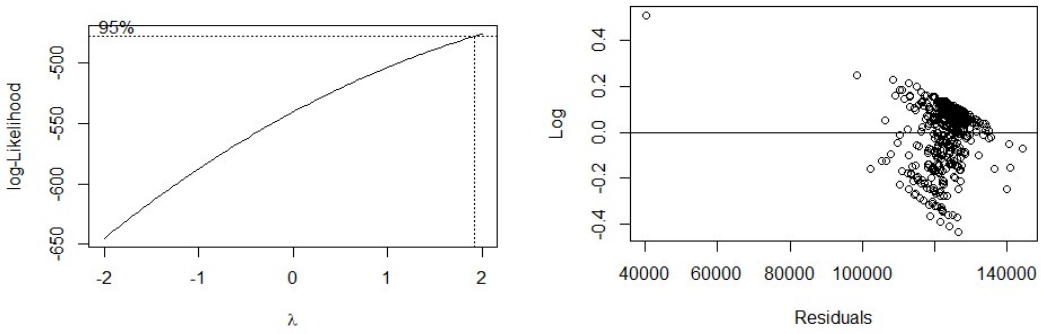
Before finalizing our model, we need to recheck residuals to be sure that the LINE assumptions have not been violated with this new model. First by reevaluating the remaining variables in the dataset by plotting each covariate against the residuals in the new model to ensure that each are linear and have equal variance. Then we rerun the QQ plot to check the normality assumption. From this we can see that the normal distribution appears to have a negative skew thus we must transform the predictor variable.





Transform Regressor to Adjust for Non-Normal Distribution

We determined from the QQ plot that there was a slight negative skew and ran a boxcox power transformation on the new model and found that the fit would not be improved from a log transformation. To prove this we ran the log transformation on the predictor and found the results to be similar after transformation.



8. Conclusion - Finalizing the Model

In our final model our response is Page.Total.Likes and our covariates are Post.Hour, Consumers, Reach.by.Users.with.Likes, Users.with.Likes.and.Engagement, Like and Share.

Page.Total.Likes = 1.295e+05 + (-5.765e+02 * Post.Hour) + (-9.184e+00 * Consumers) + (-5.636e-01 * Reach.by.Users.with.Likes) + (1.481e+01 * Users.with.Likes.and.Engagement) + (2.164e+01 * Like) + (-1.330e+02 * Share)

```
Call:
lm(formula = Page.Total.Likes ~ Post.Hour + Consumers + Reach.by.Users.with.Likes +
    Users.with.Likes.and.Engagement + Like + Share, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-45375 -10518   5267  11544  56702

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.295e+05  1.576e+03  82.153  < 2e-16 ***
Post.Hour                      -5.765e+02  1.552e+02  -3.715 0.000227 ***
Consumers                      -9.184e+00  1.332e+00  -6.895 1.65e-11 ***
Reach.by.Users.with.Likes      -5.636e-01  1.363e-01  -4.134 4.19e-05 ***
Users.with.Likes.and.Engagement  1.481e+01  2.177e+00   6.803 2.98e-11 ***
Like                            2.164e+01  5.743e+00   3.768 0.000184 ***
Share                          -1.330e+02  3.900e+01  -3.409 0.000704 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15040 on 493 degrees of freedom
Multiple R-squared:  0.156,     Adjusted R-squared:  0.1457
F-statistic: 15.19 on 6 and 493 DF,  p-value: 5.479e-16
```