Facebook metrics Data Set

Goutham Maiyalagan (71620025), Hardik Gupta (71620027), Praveen Bhandari (71620051), Soumya Mallick (71620071)

8 April 2017

Business Objective

One of the most important factor while weighing the success of a facebook cosmetic page is how many users are engaged with the page. It is not just Page Likes but over the time how many people have appreciated or not appreciated the content that is uploaded on the page. Engagement with any post is an important factor measuring how much the content uploaded is affecting and reaching people. Based on this, we have chosen "Lifetime people who have liked your page and engaged with your post" as our dependent variable and our goal is to predict this based on the other regressors. We would like to examine which are those factors affecting the success of the post by the people who have liked the page. We would like to examine what kind of post (link,video, photo, status), time. day, month and other various factors receives maximum or minimum engagement. Based on this analysis, marketers can choose what kind of content and at what particular time can get the maximum level of engagement and hence better marketing.

Dependent Variable

1. Lifetime people who have liked your page and engaged with your post

Independent Variables:

1. Post Hour
2. Post Weekday
3. Post month
4. Type
5. Category
6. Paid
7. Page total likes
8. Comments
9. Likes
10. Shares
11. Total interactions

Loading libraries and Initialisation

```r
library(GGally)
library(ggplot2)
library(car)
library(MASS)
library(corrplot)
library(ggcorrplot)
library(perturb)
library(caTools)
library(qpcR)
options(scipen = 1000)
```

Preliminary Data Analysis

Inspecting the data set

```
fb.raw <- read.csv("F:/BIG DATA/ISB/Assignments/Term 2/Statistical Analysis 2/Project/data/Facebook.csv")
summary(fb.raw)
str(fb.raw)
```

```
## Page.total.likes    Type      Category     Post.Month
## Min.   : 81370   Link  : 22   Min.   :1.00   Min.   : 1.000
## 1st Qu.:112676   Photo :426   1st Qu.:1.00   1st Qu.: 4.000
## Median :129600   Status: 45   Median :2.00   Median : 7.000
## Mean   :123194   Video :  7   Mean   :1.88   Mean   : 7.038
## 3rd Qu.:136393                3rd Qu.:3.00   3rd Qu.:10.000
## Max.   :139441                Max.   :3.00   Max.   :12.000
##
##  Post.Weekday    Post.Hour         Paid        Lifetime.Post.Total.Reach
## Min.   :1.00   Min.   : 1.00   Min.   :0.0000   Min.   :   238
## 1st Qu.:2.00   1st Qu.: 3.00   1st Qu.:0.0000   1st Qu.:  3315
## Median :4.00   Median : 9.00   Median :0.0000   Median :  5281
## Mean   :4.15   Mean   : 7.84   Mean   :0.2786   Mean   : 13903
## 3rd Qu.:6.00   3rd Qu.:11.00   3rd Qu.:1.0000   3rd Qu.: 13168
## Max.   :7.00   Max.   :23.00   Max.   :1.0000   Max.   :180480
##                                NA's   :1
## Lifetime.Post.Total.Impressions Lifetime.Engaged.Users
## Min.   :    570                 Min.   :    9.0
## 1st Qu.:   5695                 1st Qu.:  393.8
## Median :   9051                 Median :  625.5
## Mean   :  29586                 Mean   :  920.3
## 3rd Qu.:  22086                 3rd Qu.: 1062.0
## Max.   :1110282                 Max.   :11452.0
##
## Lifetime.Post.Consumers Lifetime.Post.Consumptions
## Min.   :    9.0         Min.   :    9.0
## 1st Qu.:  332.5         1st Qu.:  509.2
## Median :  551.5         Median :  851.0
## Mean   :  798.8         Mean   : 1415.1
## 3rd Qu.:  955.5         3rd Qu.: 1463.0
## Max.   :11328.0         Max.   :19779.0
##
## Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
## Min.   :    567
## 1st Qu.:   3970
## Median :   6256
## Mean   :  16766
## 3rd Qu.:  14860
## Max.   :1107833
##
## Lifetime.Post.reach.by.people.who.like.your.Page
## Min.   :   236
```

```
## 1st Qu.: 2182
## Median : 3417
## Mean   : 6585
## 3rd Qu.: 7989
## Max.   :51456
##
## Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
## Min.   :   9.0
## 1st Qu.: 291.0
## Median : 412.0
## Mean   : 610.0
## 3rd Qu.: 656.2
## Max.   :4376.0
##
##    comment            like            share        Total.Interactions
## Min.   :  0.000   Min.   :   0.0   Min.   :  0.00   Min.   :   0.0
## 1st Qu.:  1.000   1st Qu.:  56.5   1st Qu.: 10.00   1st Qu.:  71.0
## Median :  3.000   Median : 101.0   Median : 19.00   Median : 123.5
## Mean   :  7.482   Mean   : 177.9   Mean   : 27.27   Mean   : 212.1
## 3rd Qu.:  7.000   3rd Qu.: 187.5   3rd Qu.: 32.25   3rd Qu.: 228.5
## Max.   :372.000   Max.   :5172.0   Max.   :790.00   Max.   :6334.0
##                   NA's   :1        NA's   :4
## 'data.frame':    500 obs. of  19 variables:
## $ Page.total.likes                                         : int  139441 139441 139441 139441 139441 139441 139441 139441 139
441 139441 ...
## $ Type                                                     : Factor w/ 4 levels "Link","Photo",..: 2 3 2 2 2 3 2 2 3 2 ...
## $ Category                                                 : int  2 2 3 2 2 2 3 3 2 3 ...
## $ Post.Month                                               : int  12 12 12 12 12 12 12 12 12 12 ...
## $ Post.Weekday                                             : int  4 3 3 2 2 1 1 7 7 6 ...
## $ Post.Hour                                                : int  3 10 3 10 3 9 3 9 3 10 ...
## $ Paid                                                     : int  0 0 0 1 0 0 1 1 0 0 ...
## $ Lifetime.Post.Total.Reach                                : int  2752 10460 2413 50128 7244 10472 11692 13720 11844 4694 .
..
## $ Lifetime.Post.Total.Impressions                          : int  5091 19057 4373 87991 13594 20849 19479 24137 22538 8
668 ...
## $ Lifetime.Engaged.Users                                   : int  178 1457 177 2211 671 1191 481 537 1530 280 ...
## $ Lifetime.Post.Consumers                                  : int  109 1361 113 790 410 1073 265 232 1407 183 ...
## $ Lifetime.Post.Consumptions                               : int  159 1674 154 1119 580 1389 364 305 1692 250 ...
## $ Lifetime.Post.Impressions.by.people.who.have.liked.your.Page      : int  3078 11710 2812 61027 6228 16034 15432 1
9728 15220 4309 ...
## $ Lifetime.Post.reach.by.people.who.like.your.Page         : int  1640 6112 1503 32048 3200 7852 9328 11056 7912
2324 ...
## $ Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post: int  119 1108 132 1386 396 1016 379 422 12
50 199 ...
## $ comment                                                  : int  4 5 0 58 19 1 3 0 0 3 ...
## $ like                                                     : int  79 130 66 1572 325 152 249 325 161 113 ...
## $ share                                                    : int  17 29 14 147 49 33 27 14 31 26 ...
## $ Total.Interactions                                       : int  100 164 80 1777 393 186 279 339 192 142 ...
```

Handling Missing Vaues

Few missing values were observed in columns paid, like and share. Since these rows constitue only 1% of the entire dataset, we've removed them.

```
fb.raw <- fb.raw[-c(which(is.na(fb.raw$Paid))),]
fb.raw <- fb.raw[-c(which(is.na(fb.raw$share))),]
```

Converting Categorical Variables to factor and inspection of factor levels

```
fb.raw$Post.Hour <- as.factor(fb.raw$Post.Hour)
fb.raw$Post.Weekday <- as.factor(fb.raw$Post.Weekday)
fb.raw$Post.Month <- as.factor(fb.raw$Post.Month)
fb.raw$Type <- as.factor(fb.raw$Type)
fb.raw$Category <- as.factor(fb.raw$Category)
fb.raw$Paid <- as.factor(fb.raw$Paid)

table(fb.raw$Type)
table(fb.raw$Post.Month)
table(fb.raw$Paid)
table(fb.raw$Category)
table(fb.raw$Post.Weekday)
table(fb.raw$Post.Hour)

##
##   Link  Photo Status  Video
##     22    421     45      7
##
##  1  2  3  4  5  6  7  8  9 10 11 12
## 24 26 36 50 37 49 52 34 35 57 45 50
##
##   0   1
## 356 139
##
##   1   2   3
## 211 129 155
##
##  1  2  3  4  5  6  7
## 68 66 64 71 66 80 80
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   4  39 105  34  13  15  13  11  29  77  44  29  52  13   6   1   3   3
##  19  20  22  23
##   1   1   1   1
```

We observe that Post.Hour variable has only one record for levels 16,19,20,21,22,23. This will create problems while splitting the data. Hence, we remove these records containing single values.

```
fb.raw <- fb.raw[-which(fb.raw$Post.Hour == 16),]
fb.raw <- fb.raw[-which(fb.raw$Post.Hour == 19),]
fb.raw <- fb.raw[-which(fb.raw$Post.Hour == 20),]
fb.raw <- fb.raw[-which(fb.raw$Post.Hour == 22),]
fb.raw <- fb.raw[-which(fb.raw$Post.Hour == 23),]

fb.raw$Post.Hour <- factor(fb.raw$Post.Hour)
table(fb.raw$Post.Hour)
rownames(fb.raw) <- NULL
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  17  18
##   4  39 105  34  13  15  13  11  29  77  44  29  52  13   6   3   3
```

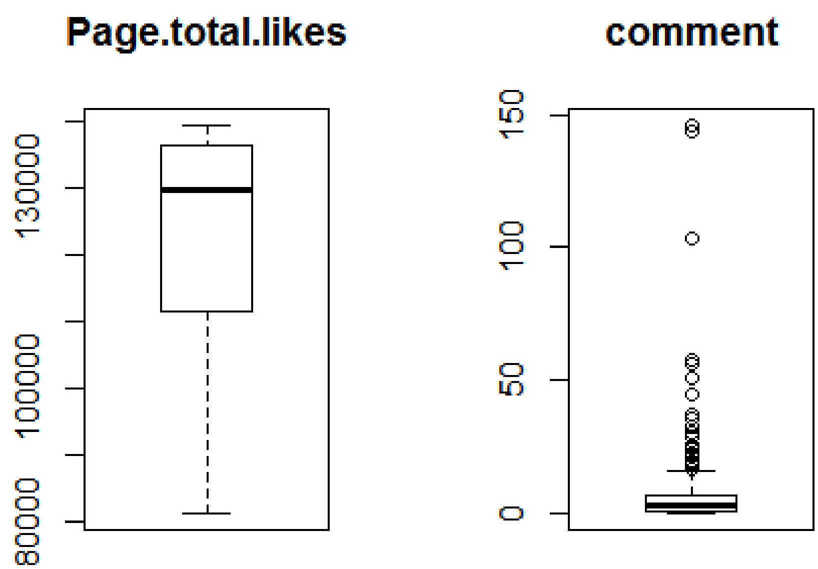
Training and test data classification

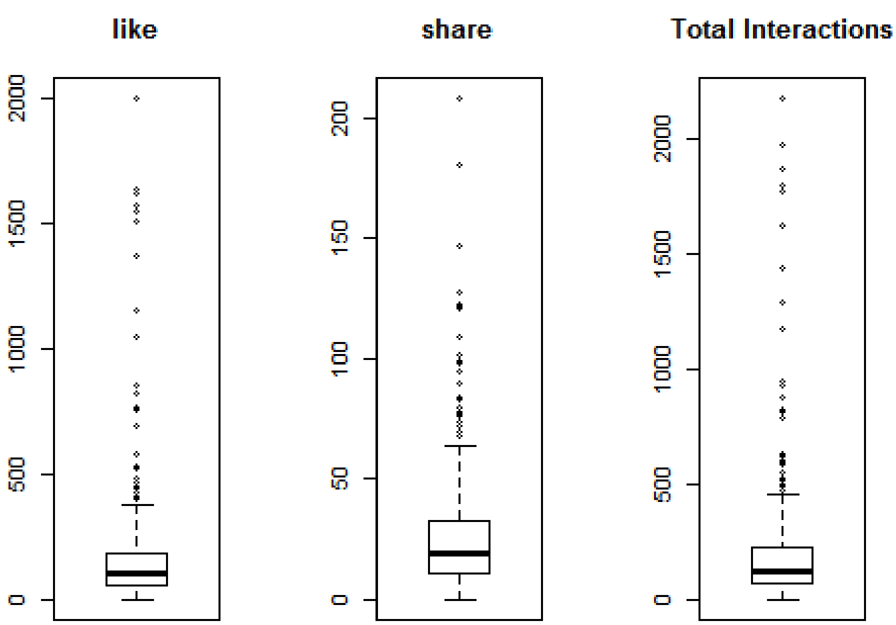
We divide the data into training and test data sets in a ratio of 80:20

```
set.seed(55)
spl = sample.split(fb.raw$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post, SplitRatio = 0.8)
Train = subset(fb.raw, spl==TRUE)
Test = subset(fb.raw, spl==FALSE)

dim(Train)
dim(Test)

## [1] 392  19
## [1] 98 19
```

Inspecting Independent and Dependent variables

Independent variables





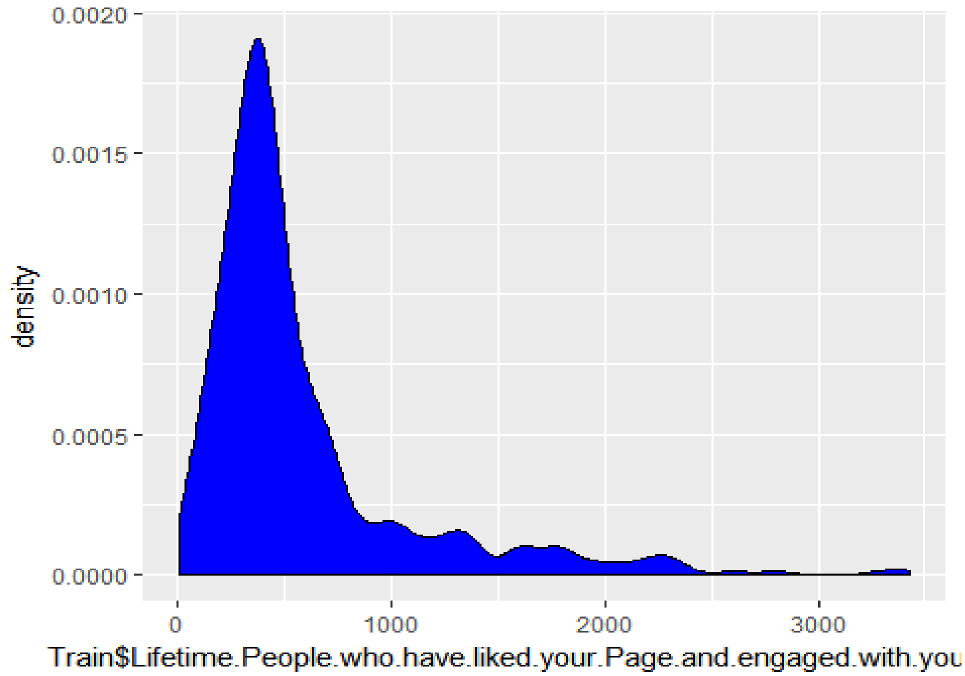
Our observations:

1. Numerous outliers in the variables such as comment, share, like, total interactions.

2. The variables are heavily right skewed which could suggest a need for transformation.

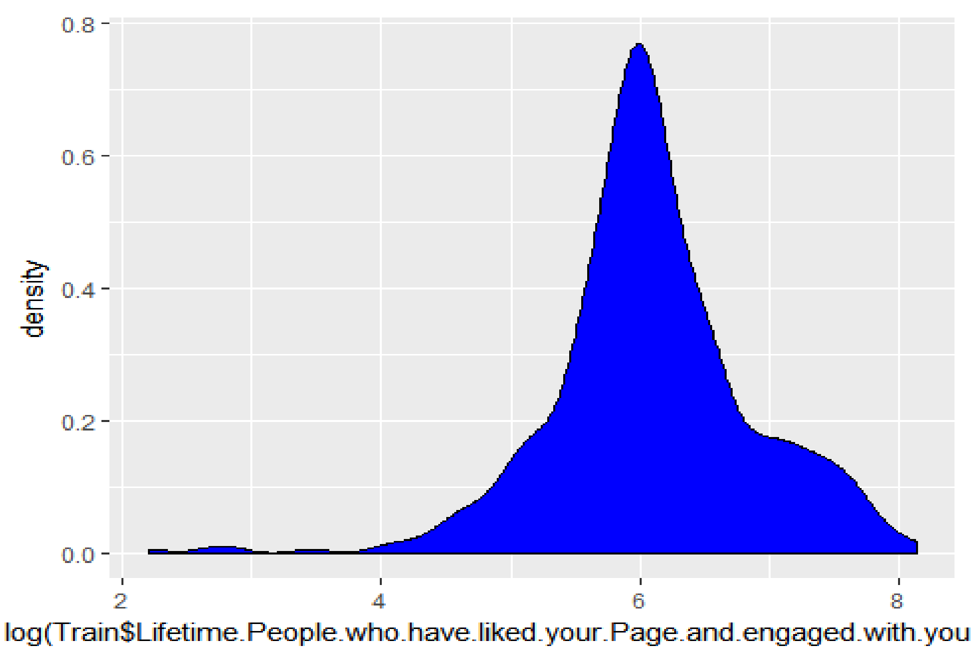
Dependent variable - Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post

```
ggplot(Train, aes(x=Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)) +
 geom_density(fill="blue")
```



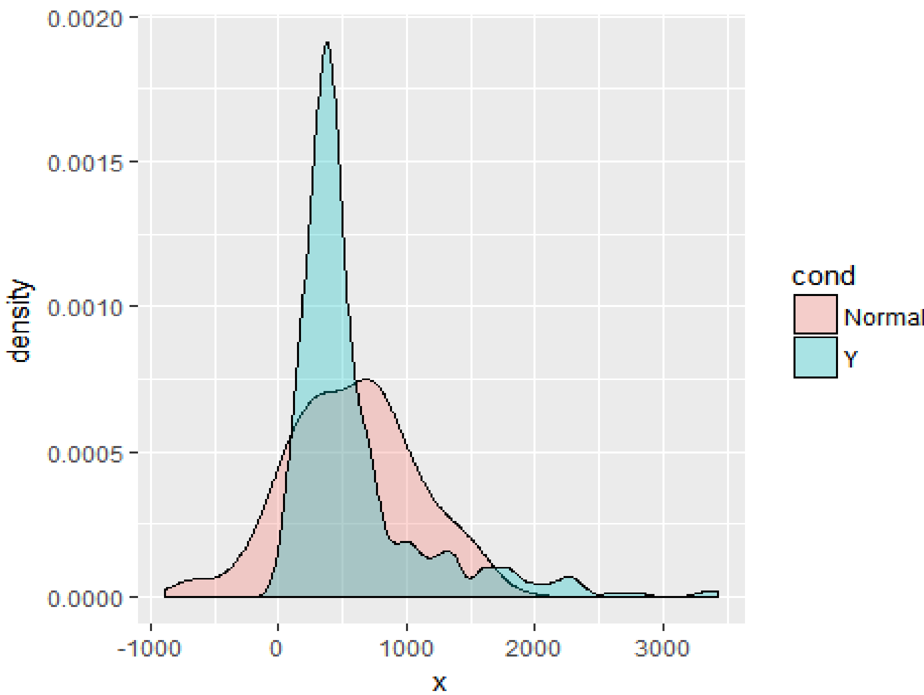
The dependent variable looks heavily right skewed. We can try a log transformation.

```
ggplot(Train, aes(x=log(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post))) +
 geom_density(fill="blue")
```

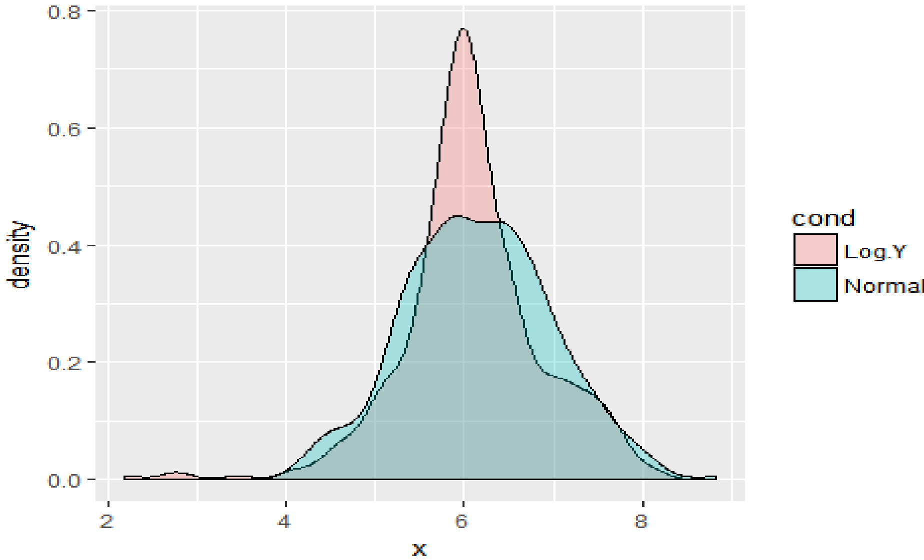


We compare the distribution of the dependent variable and its log transformation with a normal distribution of same mean and standard deviation.

```
norm<-rnorm(392, mean=mean(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post),
        sd=sd(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post))
dat <- data.frame(cond = factor(rep(c("Y","Normal"), each=392)),
            x = c(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post,norm))
ggplot(dat, aes(x, fill=cond)) + geom_density(alpha=.3)
```



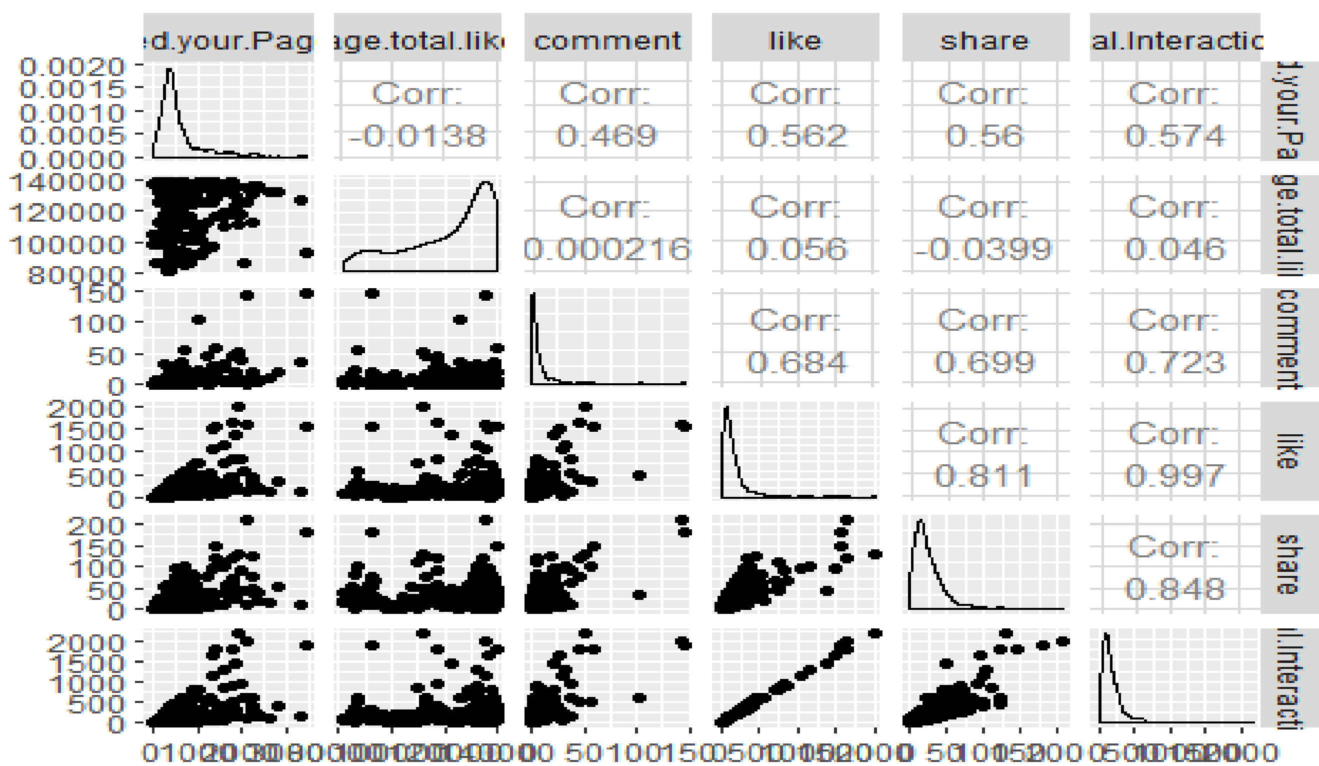

```
lnorm<-rnorm(392, mean=mean(log(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)),
        sd=sd(log(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)))
dat <- data.frame(cond = factor(rep(c("Log.Y","Normal"), each = 392)),
            x = c(log(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post),lnorm))
ggplot(dat, aes(x, fill=cond)) + geom_density(alpha=.3)
```



We see that the log transformed variable fits better and is close to a normal distribution.

Correlation and Scatter Plot Matrices

```r
mcor <- round(cor(Train[,-c(2:15)]),2)
#corrplot(mcor, method="number")
ggpairs(Train[,c(15,1,16,17,18,19)])
```



There is positive correlation between like, comment, share and Interactions. Output variable is also positively correlated with these variables

---

Initial Model Fitting and Basic Diagnostics

```
model1 <- lm(Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~ Page.total.likes + Type +
           Category + Post.Month + Post.Weekday + Post.Hour + Paid + comment + like + share + Total.Interactions,
          data = Train)
summary(model1)

##
## Call:
## lm(formula = Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~
##     Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##        Post.Hour + Paid + comment + like + share + Total.Interactions,
##     data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1155.67  -118.15   -27.51    96.44  1512.02
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate  Std. Error t value            Pr(>|t|)
## (Intercept)    954.744738  895.187521   1.067            0.286923
## Page.total.likes -0.008567   0.010318  -0.830            0.406955
## TypePhoto      227.459347   76.251157   2.983            0.003056
## TypeStatus    1357.917125   93.447005  14.531 < 0.0000000000000002
## TypeVideo      616.481246  132.599953   4.649   0.0000047383955262
## Category2     -149.932447   44.297961  -3.385            0.000794
## Category3     -218.192484   38.975208  -5.598   0.0000000439766722
## Post.Month2    126.462219  108.630579   1.164            0.245161
## Post.Month3    -33.751720  170.510401  -0.198            0.843203
## Post.Month4    219.239631  264.415267   0.829            0.407589
## Post.Month5    179.397449  340.422424   0.527            0.598540
## Post.Month6    349.130412  413.895657   0.844            0.399516
## Post.Month7    349.156893  459.720903   0.759            0.448069
## Post.Month8    349.988023  493.029878   0.710            0.478259
## Post.Month9    248.108362  516.427543   0.480            0.631222
## Post.Month10   340.225118  530.084176   0.642            0.521405
## Post.Month11     2.167494  543.194575   0.004            0.996819
## Post.Month12   123.088485  557.414200   0.221            0.825362
## Post.Weekday2  -50.567424   54.858125  -0.922            0.357279
## Post.Weekday3   30.947710   57.044417   0.543            0.587808
## Post.Weekday4 -122.230736   55.815764  -2.190            0.029195
## Post.Weekday5  -99.591998   55.365700  -1.799            0.072916
## Post.Weekday6    3.361021   53.404223   0.063            0.949854
## Post.Weekday7   47.803946   52.766971   0.906            0.365592
## Post.Hour2      53.716314  150.057379   0.358            0.720581
```

```
## Post.Hour3          35.921411  142.863466   0.251          0.801623
## Post.Hour4          48.732997  150.394801   0.324          0.746107
## Post.Hour5          39.240175  163.815985   0.240          0.810829
## Post.Hour6        -133.457987  159.923453  -0.835          0.404565
## Post.Hour7         -54.898219  165.775584  -0.331          0.740723
## Post.Hour8        -101.990891  169.431532  -0.602          0.547593
## Post.Hour9          74.063873  151.048827   0.490          0.624209
## Post.Hour10         36.776937  143.496322   0.256          0.797877
## Post.Hour11          4.160009  146.535610   0.028          0.977368
## Post.Hour12        105.687337  150.962630   0.700          0.484339
## Post.Hour13         80.059622  145.723432   0.549          0.583087
## Post.Hour14        128.513781  164.977206   0.779          0.436521
## Post.Hour15        -77.534519  196.823711  -0.394          0.693875
## Post.Hour17        218.285313  222.302375   0.982          0.326817
## Post.Hour18         30.727714  249.134584   0.123          0.901911
## Paid1               45.979266   32.402695   1.419          0.156795
## comment              3.454755    1.511999   2.285          0.022920
## like                 0.846200    0.107449   7.875  0.0000000000000437
## share                2.917569    1.147526   2.542          0.011439
## Total.Interactions         NA          NA      NA                NA
##
## (Intercept)
## Page.total.likes
## TypePhoto        **
## TypeStatus       ***
## TypeVideo        ***
## Category2        ***
## Category3        ***
## Post.Month2
## Post.Month3
## Post.Month4
## Post.Month5
## Post.Month6
## Post.Month7
## Post.Month8
## Post.Month9
## Post.Month10
## Post.Month11
## Post.Month12
## Post.Weekday2
## Post.Weekday3
## Post.Weekday4    *
## Post.Weekday5    .
## Post.Weekday6
## Post.Weekday7
## Post.Hour2
## Post.Hour3
## Post.Hour4
## Post.Hour5
## Post.Hour6
## Post.Hour7
## Post.Hour8
## Post.Hour9
```

```
## Post.Hour10
## Post.Hour11
## Post.Hour12
## Post.Hour13
## Post.Hour14
## Post.Hour15
## Post.Hour17
## Post.Hour18
## Paid1
## comment         *
## like          ***
## share          *
## Total.Interactions
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270 on 348 degrees of freedom
## Multiple R-squared:  0.7592, Adjusted R-squared:  0.7295
## F-statistic: 25.52 on 43 and 348 DF,  p-value: < 0.00000000000000022
```

Interpretation from Model-1

- R-Squared for the model is 69% which indicates that the model initially fits just well.
- Few of the regressors are insignifcant and these need to be analysed and removed
- Regressor Total.Interactions has coefficient values as NA. This is possibly because Total.Interactions is linearly related to the other variables (from correlation matrix we observe that correlation between like and Total Interactions is 1).

Model-2

Observing our model1, we build model2 by removing Total.Interactions.

```
model2 <- lm(Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~ Page.total.likes + Type +
           Category + Post.Month + Post.Weekday + Post.Hour + Paid + comment + like + share,
          data = Train)
summary(model2)
```

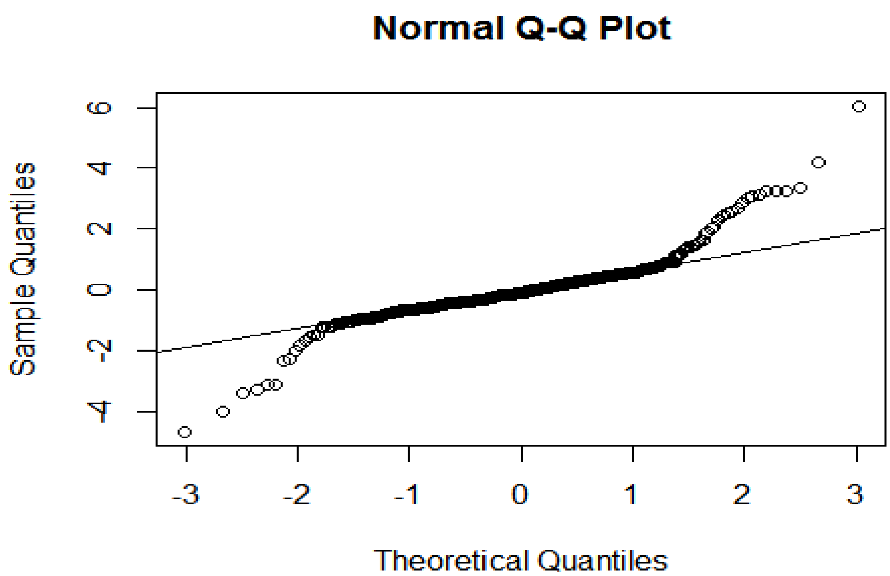
```
##
## Call:
## lm(formula = Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~
##     Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##       Post.Hour + Paid + comment + like + share, data = Train)
##
## Residuals:
##      Min      1Q   Median      3Q     Max
## -1155.67  -118.15   -27.51   96.44  1512.02
##
## Coefficients:
##                  Estimate  Std. Error t value           Pr(>|t|)
## (Intercept)      954.744738  895.187521   1.067          0.286923
## Page.total.likes   -0.008567    0.010318  -0.830          0.406955
## TypePhoto        227.459347   76.251157   2.983          0.003056 **
## TypeStatus      1357.917125   93.447005  14.531 < 0.0000000000000002 ***
## TypeVideo        616.481246  132.599953   4.649   0.0000047383955262 ***
```

```
## Category2       -149.932447   44.297961  -3.385           0.000794 ***
## Category3       -218.192484   38.975208  -5.598   0.0000000439766722 ***
## Post.Month2      126.462219  108.630579   1.164           0.245161
## Post.Month3      -33.751720  170.510401  -0.198           0.843203
## Post.Month4      219.239631  264.415267   0.829           0.407589
## Post.Month5      179.397449  340.422424   0.527           0.598540
## Post.Month6      349.130412  413.895657   0.844           0.399516
## Post.Month7      349.156893  459.720903   0.759           0.448069
## Post.Month8      349.988023  493.029878   0.710           0.478259
## Post.Month9      248.108362  516.427543   0.480           0.631222
## Post.Month10     340.225118  530.084176   0.642           0.521405
## Post.Month11       2.167494  543.194575   0.004           0.996819
## Post.Month12     123.088485  557.414200   0.221           0.825362
## Post.Weekday2    -50.567424   54.858125  -0.922           0.357279
## Post.Weekday3     30.947710   57.044417   0.543           0.587808
## Post.Weekday4   -122.230736   55.815764  -2.190           0.029195 *
## Post.Weekday5    -99.591998   55.365700  -1.799           0.072916 .
## Post.Weekday6      3.361021   53.404223   0.063           0.949854
## Post.Weekday7     47.803946   52.766971   0.906           0.365592
## Post.Hour2        53.716314  150.057379   0.358           0.720581
## Post.Hour3        35.921411  142.863466   0.251           0.801623
## Post.Hour4        48.732997  150.394801   0.324           0.746107
## Post.Hour5        39.240175  163.815985   0.240           0.810829
## Post.Hour6      -133.457987  159.923453  -0.835           0.404565
## Post.Hour7       -54.898219  165.775584  -0.331           0.740723
## Post.Hour8      -101.990891  169.431532  -0.602           0.547593
## Post.Hour9        74.063873  151.048827   0.490           0.624209
## Post.Hour10       36.776937  143.496322   0.256           0.797877
## Post.Hour11        4.160009  146.535610   0.028           0.977368
## Post.Hour12      105.687337  150.962630   0.700           0.484339
## Post.Hour13       80.059622  145.723432   0.549           0.583087
## Post.Hour14      128.513781  164.977206   0.779           0.436521
## Post.Hour15      -77.534519  196.823711  -0.394           0.693875
## Post.Hour17      218.285313  222.302375   0.982           0.326817
## Post.Hour18       30.727714  249.134584   0.123           0.901911
## Paid1             45.979266   32.402695   1.419           0.156795
## comment            3.454755    1.511999   2.285           0.022920 *
## like               0.846200    0.107449   7.875   0.0000000000000437 ***
## share              2.917569    1.147526   2.542           0.011439 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270 on 348 degrees of freedom
## Multiple R-squared:  0.7592, Adjusted R-squared:  0.7295
## F-statistic: 25.52 on 43 and 348 DF,  p-value: < 0.00000000000000022
```

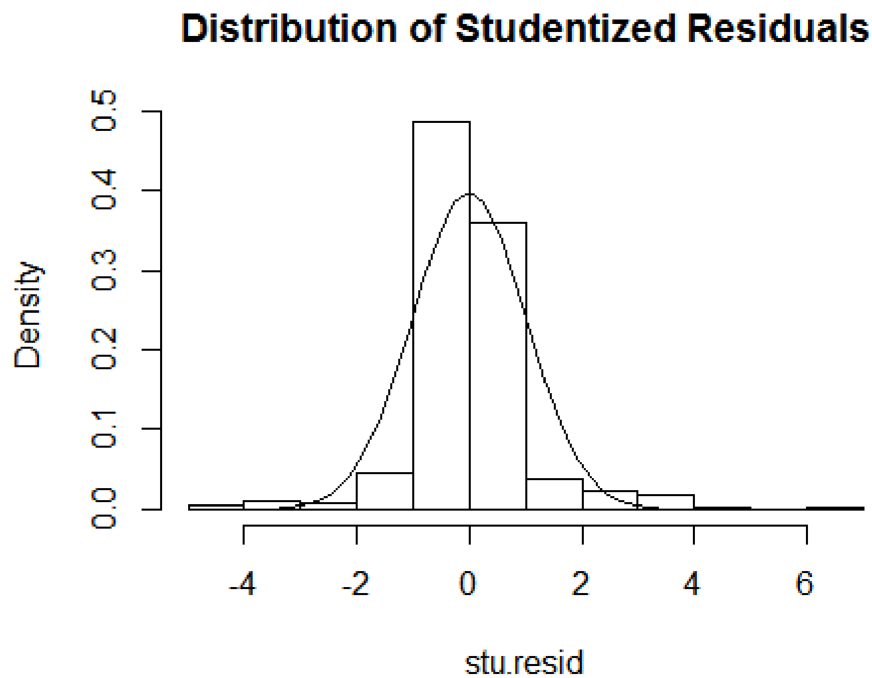
We obtain a model with R-Squared value of 0.7592

Observing the residual plots and checking for Normality

```
residuals <- rstandard(model2)
qqnorm(residuals)
qqline(residuals)
```



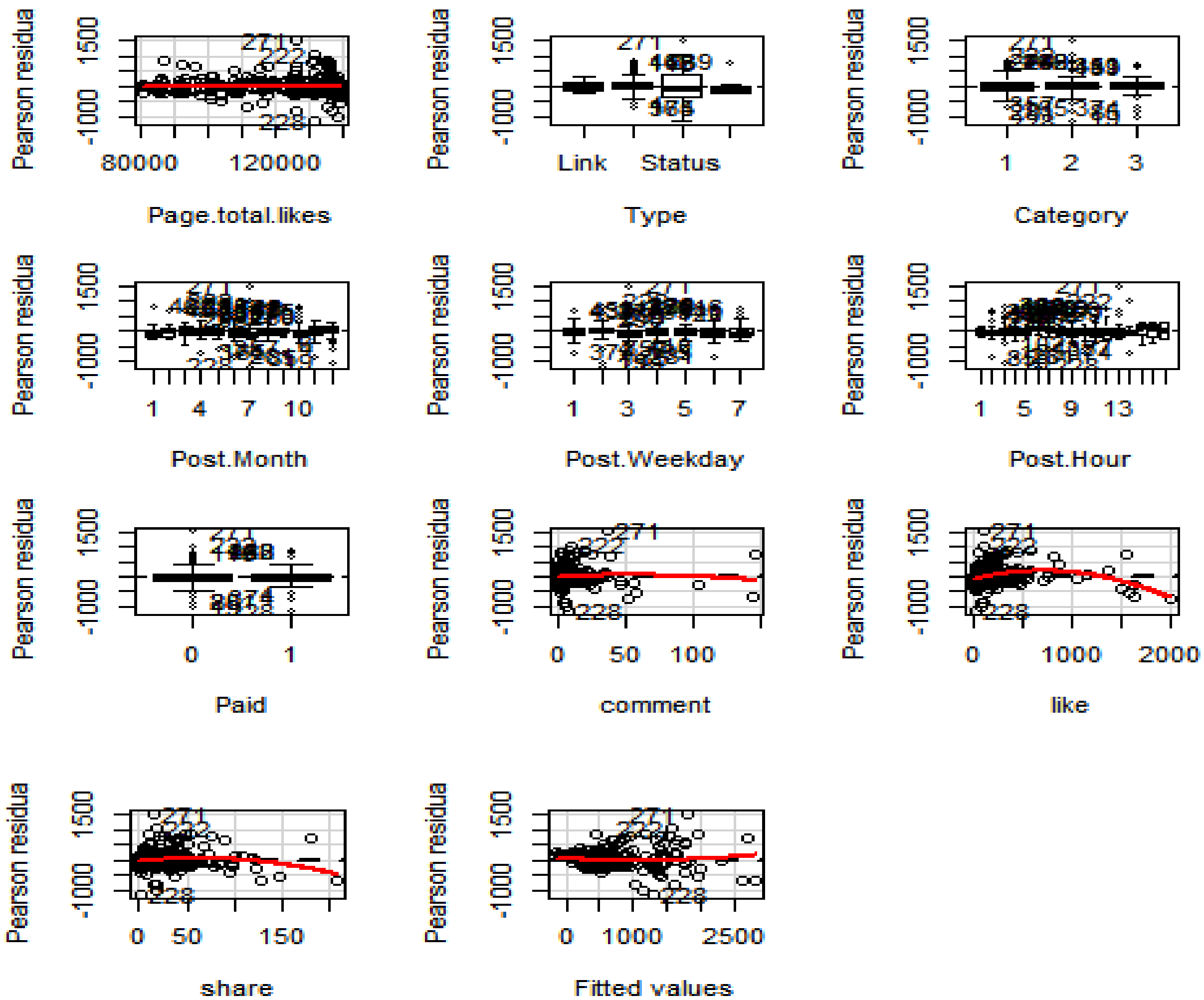

```
stu.resid <- studres(model2)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(-3.5, 7,length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Observing the above plots shows that the model fits just well with the data, however the histogram is distorted

Residuals plot with Fitted values and other Regressors

```
residualPlots(model2,id.n=3)
```



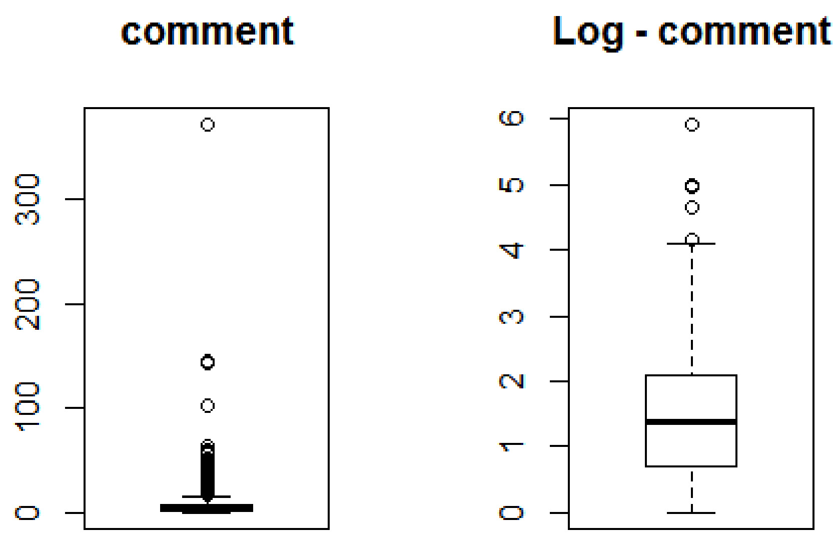

```
##                  Test stat Pr(>|t|)
## Page.total.likes    -1.738    0.083
## Type                    NA       NA
## Category                NA       NA
## Post.Month              NA       NA
## Post.Weekday            NA       NA
## Post.Hour               NA       NA
## Paid                    NA       NA
## comment             -1.107    0.269
## like                -6.418    0.000
## share               -3.911    0.000
## Tukey test           1.730    0.084
```

Observing the residual plots, we perform the following Transformation

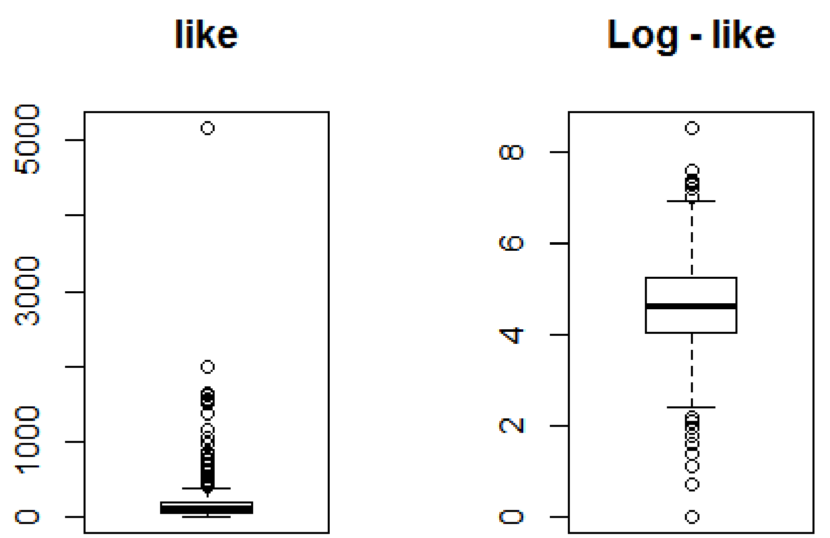
- Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post - Logarithimic transformation (Since skewed to right)
- comment - Logarithimic transformation (Since skewed to right)
- like - Logarithimic transformation (Since skewed to right)
- share - Logarithimic transformation (Since skewed to right)

Transformations

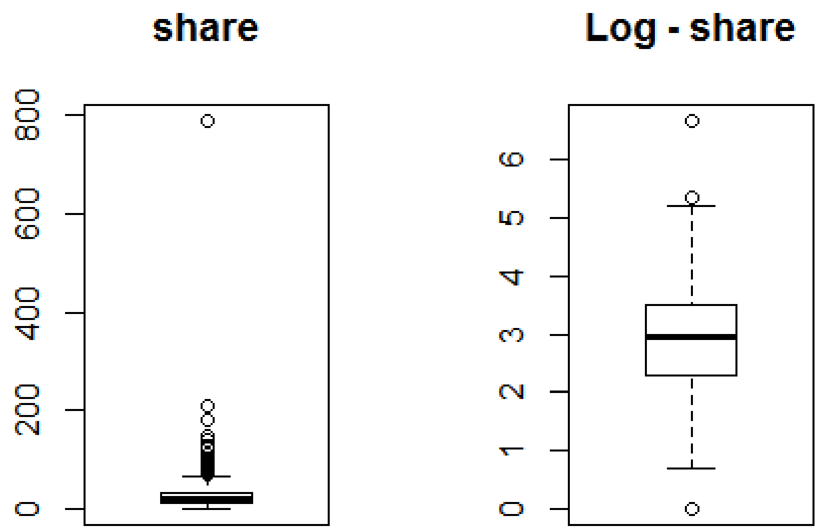
```
log.comment <- log(fb.raw$comment+1)
par(mfrow=c(1, 2))
boxplot(fb.raw$comment, main = "comment")
boxplot(log.comment, main = "Log - comment")
```




```
log.like <- log(fb.raw$like)
par(mfrow=c(1, 2))
boxplot(fb.raw$like, main = "like")
boxplot(log.like, main = "Log - like")
```




```
log.share <- log(fb.raw$share)
par(mfrow=c(1, 2))
boxplot(fb.raw$share, main = "share")
boxplot(log.share, main = "Log - share")
```

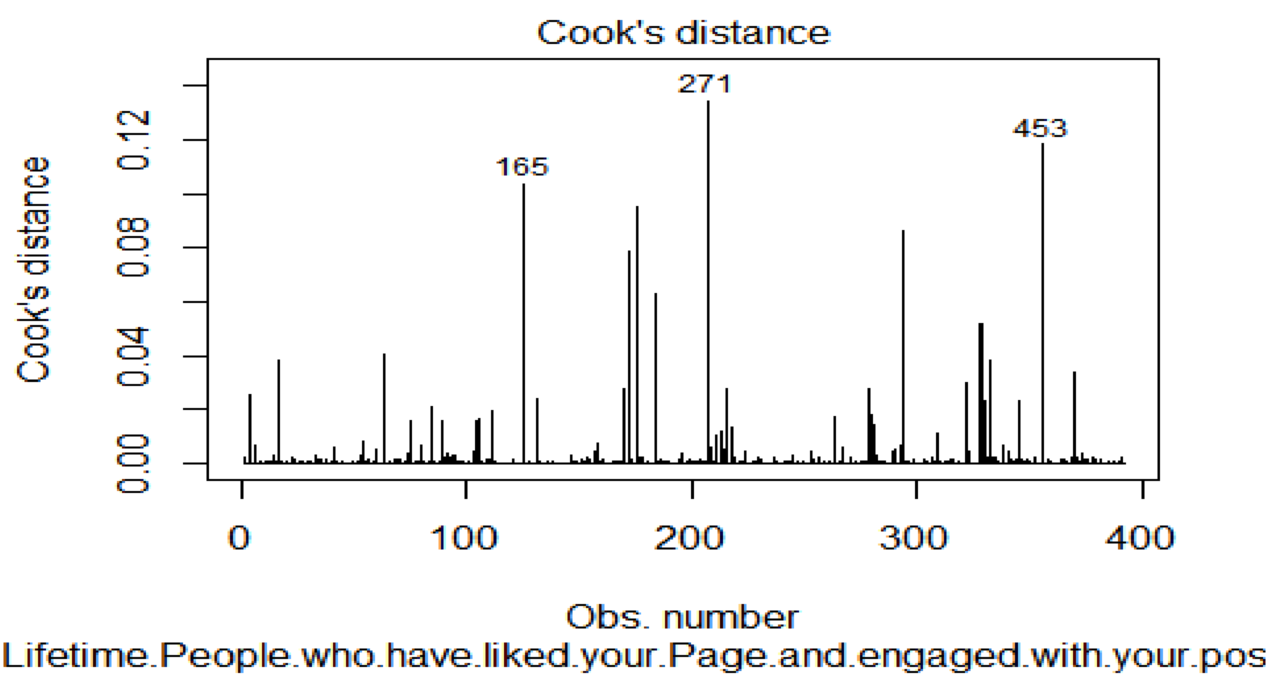


The data fits better after performing the Transformations

Checking for Influential Observations/ Deletion Diagnostics

Analysing the influential variables using Cook's Distance

```
cutoff <- 4/((nrow(Train)-length(model2$coefficients)-2))
plot(model2, which=4, cook.levels=cutoff)
```



We observe that observation 165, 271, 453 have very large Cook's distance. Next we check whether their deletion affects our model or not

Model 3 - Running the model by removing the influential observation

```
Train_1 <- Train[-which(row.names(Train) == 165),]
Train_1 <- Train[-which(row.names(Train) == 271),]
Train_1 <- Train[-which(row.names(Train) == 453),]

model3 <- lm(Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~ Page.total.likes + Type +
           Category + Post.Month + Post.Weekday + Post.Hour + Paid + comment + like + share,
          data = Train_1)
summary(model3)

##
## Call:
## lm(formula = Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post ~
##     Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##       Post.Hour + Paid + comment + like + share, data = Train_1)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -1154.49  -121.36  -20.14   96.83  1588.99
##
## Coefficients:
##                   Estimate  Std. Error t value           Pr(>|t|)
## (Intercept)      979.893246  883.002947   1.110           0.267885
## Page.total.likes   -0.009042    0.010178  -0.888           0.374978
## TypePhoto        234.211100   75.238761   3.113           0.002006 **
## TypeStatus      1357.742362   92.171600  14.731 < 0.0000000000000002 ***
## TypeVideo        622.679768  130.803880   4.760  0.00000284140458740 ***
## Category2       -151.254369   43.695225  -3.462           0.000604 ***
## Category3       -229.001027   38.585027  -5.935  0.00000000712402789 ***
## Post.Month2       90.192873  107.720226   0.837           0.403007
## Post.Month3      -39.791807  168.193312  -0.237           0.813119
## Post.Month4      226.477543  260.815758   0.868           0.385807
## Post.Month5      199.358204  335.831593   0.594           0.553149
## Post.Month6      370.545998  408.299068   0.908           0.364753
## Post.Month7      371.731593  453.498886   0.820           0.412952
## Post.Month8      365.152890  486.322815   0.751           0.453255
## Post.Month9      273.866462  509.439909   0.538           0.591209
## Post.Month10     356.036789  522.871620   0.681           0.496373
## Post.Month11      20.988919  535.811635   0.039           0.968776
## Post.Month12     147.275881  549.856011   0.268           0.788978
## Post.Weekday2    -31.073339   54.436661  -0.571           0.568494
## Post.Weekday3     56.359015   56.799717   0.992           0.321771
## Post.Weekday4    -97.481703   55.571539  -1.754           0.080285 .
## Post.Weekday5    -80.842389   54.910096  -1.472           0.141855
## Post.Weekday6     22.113006   52.986426   0.417           0.676693
## Post.Weekday7     63.105540   52.256618   1.208           0.228020
## Post.Hour2        64.326821  148.044855   0.435           0.664189
## Post.Hour3        41.882330  140.925366   0.297           0.766495
```

```
## Post.Hour4       54.549914  148.352784   0.368           0.713318
## Post.Hour5       36.203896  161.582793   0.224           0.822844
## Post.Hour6     -125.373910  157.760085  -0.795           0.427324
## Post.Hour7      -64.261417  163.538039  -0.393           0.694602
## Post.Hour8     -103.563375  167.119721  -0.620           0.535865
## Post.Hour9       75.169980  148.987606   0.505           0.614203
## Post.Hour10      46.799197  141.570965   0.331           0.741168
## Post.Hour11       4.996563  144.535831   0.035           0.972443
## Post.Hour12     123.841242  149.005615   0.831           0.406479
## Post.Hour13      75.075797  143.742588   0.522           0.601799
## Post.Hour14     146.056212  162.813866   0.897           0.370301
## Post.Hour15     -68.626094  194.156449  -0.353           0.723960
## Post.Hour17     231.095535  219.303238   1.054           0.292721
## Post.Hour18      44.291468  245.769238   0.180           0.857088
## Paid1            47.757982   31.965070   1.494           0.136067
## comment           1.000286    1.669528   0.599           0.549468
## like              0.865210    0.106142   8.151  0.00000000000000662 ***
## share             3.125749    1.133653   2.757           0.006138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.3 on 347 degrees of freedom
## Multiple R-squared:  0.747,  Adjusted R-squared:  0.7156
## F-statistic: 23.83 on 43 and 347 DF,  p-value: < 0.00000000000000022
```

Removing influential observation did not affect the model.

We will now perform transformation on the dependent variable and few of the independent variable by observing the residual plots from model2

Model 4

```r
Train$log.comment <- log(Train$comment+1)
Train$log.like <- log(Train$like+1)
Train$log.share <- log(Train$share+1)
Train$log.Y <- log(Train$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)


model4 <- lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + Post.Hour + Paid +
             log.comment + log.like + log.share,
           data = Train)
summary(model4)
```

```
##
## Call:
## lm(formula = log.Y ~ Page.total.likes + Type + Category + Post.Month +
##     Post.Weekday + Post.Hour + Paid + log.comment + log.like +
##     log.share, data = Train)
```

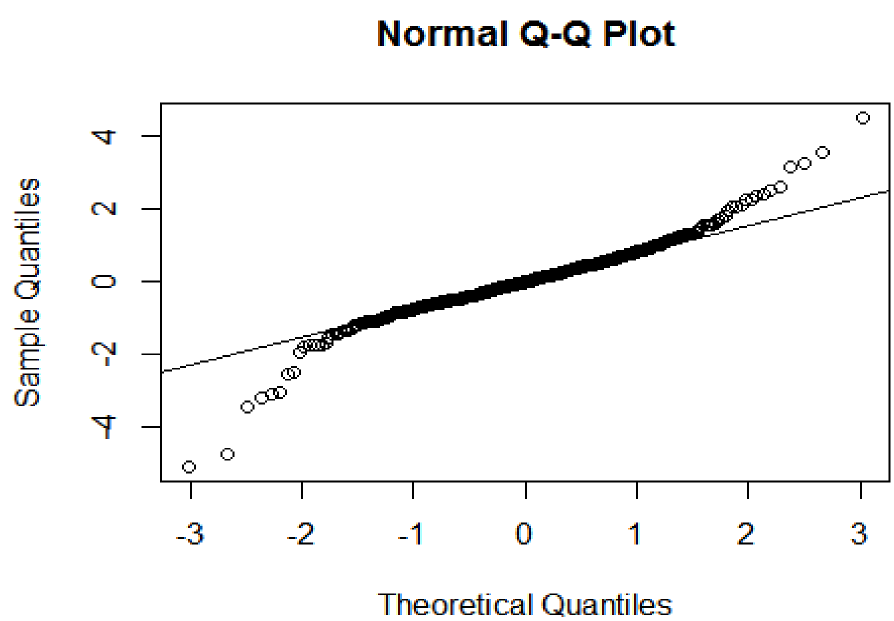
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66463 -0.17884 -0.00784  0.16947  1.52990
##
## Coefficients:
##                    Estimate  Std. Error t value           Pr(>|t|)
## (Intercept)      5.63617267  1.17924468   4.779    0.00000259710058 ***
## Page.total.likes -0.00002615  0.00001361  -1.922            0.0555 .
## TypePhoto        0.72273554  0.10117026   7.144    0.00000000000535 ***
## TypeStatus       1.88254673  0.12357322  15.234 < 0.0000000000000002 ***
## TypeVideo        1.16823945  0.17350435   6.733    0.00000000006868 ***
## Category2       -0.37343372  0.06013821  -6.210    0.00000000151377 ***
## Category3       -0.46347493  0.05349838  -8.663 < 0.0000000000000002 ***
## Post.Month2      0.24003961  0.14356663   1.672            0.0954 .
## Post.Month3      0.08495988  0.22529287   0.377            0.7063
## Post.Month4      0.66017136  0.34845928   1.895            0.0590 .
## Post.Month5      0.68025287  0.44834042   1.517            0.1301
## Post.Month6      1.10816601  0.54493317   2.034            0.0428 *
## Post.Month7      0.99057102  0.60574692   1.635            0.1029
## Post.Month8      1.06189838  0.64983387   1.634            0.1031
## Post.Month9      0.97436139  0.68147967   1.430            0.1537
## Post.Month10     1.25257545  0.69919500   1.791            0.0741 .
## Post.Month11     0.48087932  0.71593271   0.672            0.5022
## Post.Month12     0.80144306  0.73429080   1.091            0.2758
## Post.Weekday2   -0.03735547  0.07215373  -0.518            0.6050
## Post.Weekday3    0.03801590  0.07574590   0.502            0.6161
## Post.Weekday4   -0.17431760  0.07387508  -2.360            0.0188 *
## Post.Weekday5   -0.13138854  0.07325130  -1.794            0.0737 .
## Post.Weekday6    0.01290201  0.07019559   0.184            0.8543
## Post.Weekday7    0.12866271  0.06954814   1.850            0.0652 .
## Post.Hour2       0.05399090  0.19869242   0.272            0.7860
## Post.Hour3      -0.04366312  0.18901613  -0.231            0.8174
## Post.Hour4       0.06906503  0.19909683   0.347            0.7289
## Post.Hour5       0.08245327  0.21682085   0.380            0.7040
## Post.Hour6      -0.21916498  0.21151035  -1.036            0.3008
## Post.Hour7       0.06553770  0.21959985   0.298            0.7655
## Post.Hour8      -0.17697778  0.22443548  -0.789            0.4309
## Post.Hour9       0.01653453  0.19999619   0.083            0.9342
## Post.Hour10     -0.00389007  0.18984246  -0.020            0.9837
## Post.Hour11     -0.07121409  0.19400241  -0.367            0.7138
## Post.Hour12      0.21477356  0.19968052   1.076            0.2829
## Post.Hour13      0.06254419  0.19289912   0.324            0.7460
## Post.Hour14      0.14692064  0.21777779   0.675            0.5004
## Post.Hour15      0.51324958  0.26210354   1.958            0.0510 .
## Post.Hour17      0.24352324  0.29456466   0.827            0.4090
## Post.Hour18      0.29109130  0.33009682   0.882            0.3785
## Paid1            0.04851222  0.04301862   1.128            0.2602
## log.comment      0.02296330  0.02436795   0.942            0.3467
## log.like         0.46281027  0.03727588  12.416 < 0.0000000000000002 ***
## log.share        0.06175024  0.04602506   1.342            0.1806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3572 on 348 degrees of freedom
## Multiple R-squared:  0.8175, Adjusted R-squared:  0.7949
## F-statistic: 36.25 on 43 and 348 DF,  p-value: < 0.00000000000000022
```

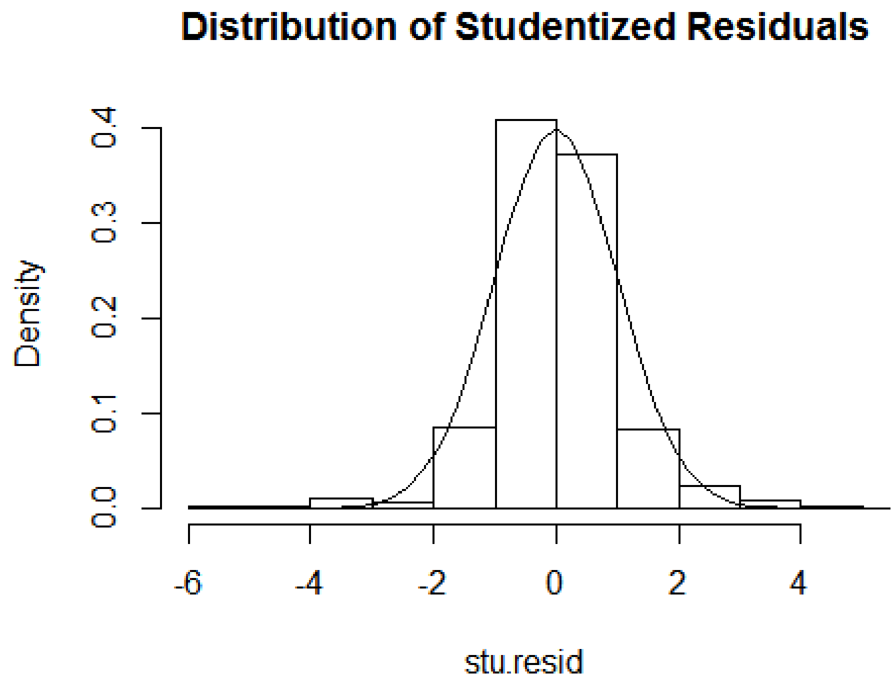
By using Transformation, we obtain R-Squared value of 0.8175. The model fits well with the data. Comparing Adjusted R-squared with the previous model value, this value is also very high

Observing the residual plots and checking for Normality

```
residuals <- rstandard(model4)
qqnorm(residuals)
qqline(residuals)
```



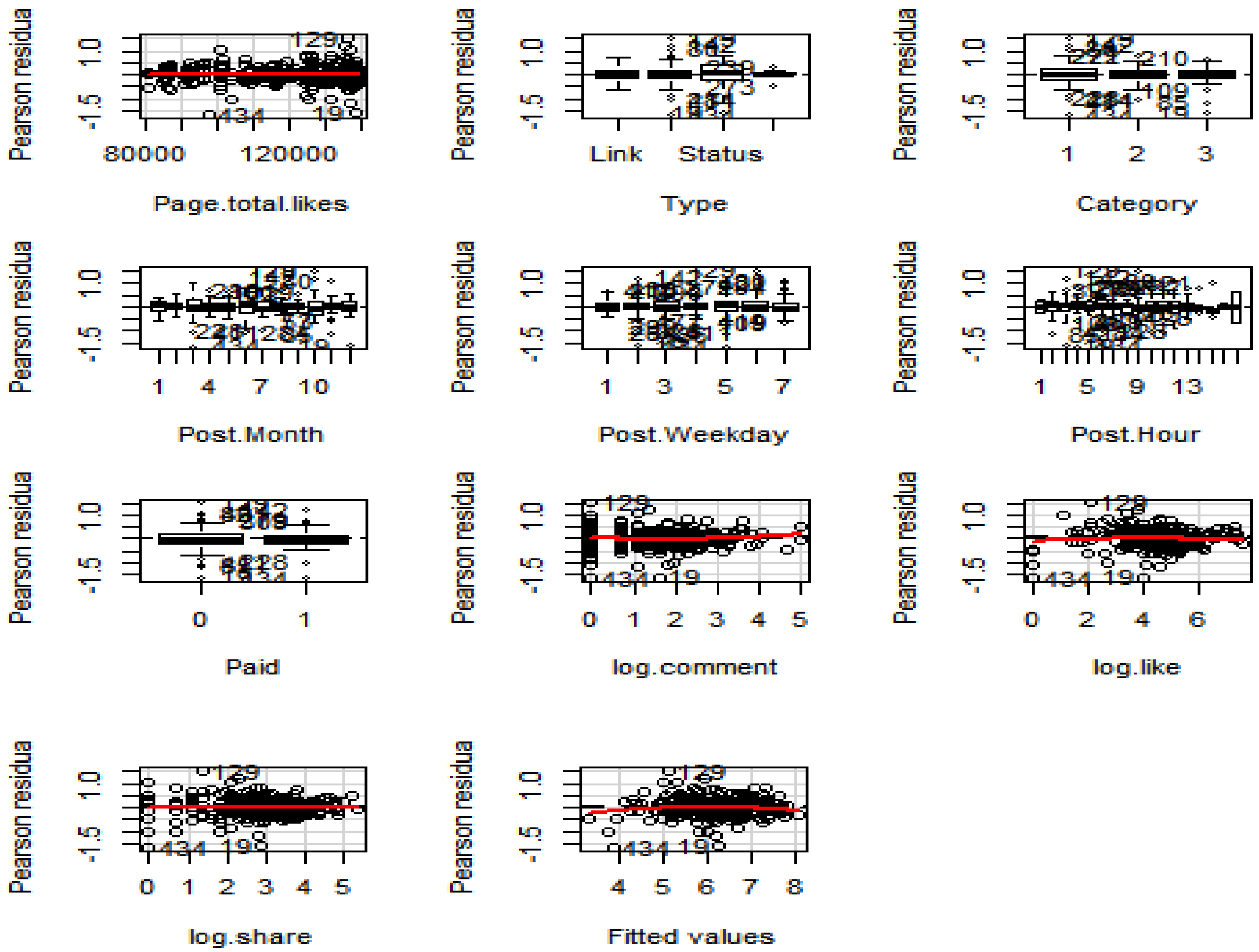

```
stu.resid <- studres(model4)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(-3.5, 7,length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



The residual plots, QQplot and Histogram, both are almost normally distributed. This means the model fits well

Residuals plot with Fitted values and other Regressors

```
residualPlots(model4,id.n=3)
```




```
##                  Test stat Pr(>|t|)
## Page.total.likes    -1.841    0.066
## Type                    NA       NA
## Category                NA       NA
## Post.Month              NA       NA
## Post.Weekday            NA       NA
## Post.Hour               NA       NA
## Paid                    NA       NA
## log.comment          1.595    0.112
## log.like            -1.026    0.305
## log.share            0.021    0.983
## Tukey test          -2.187    0.029
```

Observing the residual vs fitted plots and residuals vs regressors plot, the errors are almost randomly distributed. We see that our model fits well

Checking for Collinearity

Variance Inflation Factors

```
vif(model4)

##                    GVIF Df GVIF^(1/(2*Df))
## Page.total.likes 157.695262  1      12.557677
## Type              1.829232  3       1.105889
## Category          2.440324  2       1.249862
## Post.Month      790.590084 11       1.354332
## Post.Weekday      1.788789  6       1.049655
## Post.Hour         6.135301 16       1.058327
## Paid              1.147515  1       1.071221
## log.comment       2.043645  1       1.429561
## log.like          5.558619  1       2.357672
## log.share         5.992470  1       2.447952
```

Observing the Variance Inflation Factors, the values are almost less than or close to 10 (cut-off factor)

Variance Decomposition Proportion

```
colldiag(Train[,-c(2:15, 20:23)], center = TRUE)

## Condition
## Index    Variance Decomposition Proportions
##                      Page.total.likes comment like  share
## 1              1.000 0.000          0.000   0.000 0.000
## 2              1.835 0.959          0.000   0.000 0.000
## 3              2.958 0.006          0.000   0.000 0.000
## 4              3.986 0.035          0.000   0.000 0.000
## 5  16309146577994740.000 0.000          1.000   1.000 1.000
##   Total.Interactions
## 1 0.000
## 2 0.000
## 3 0.000
## 4 0.000
## 5 1.000
```

- Observing Variance Decomposition Proportion, it hints that comment, like and share are linearly correlated.
- Observing the correlation matrix also suggest high correlation between the three

We now build a model by dropping one of them, mostly the one which is least correlated with the output - Comment

---

Model 5

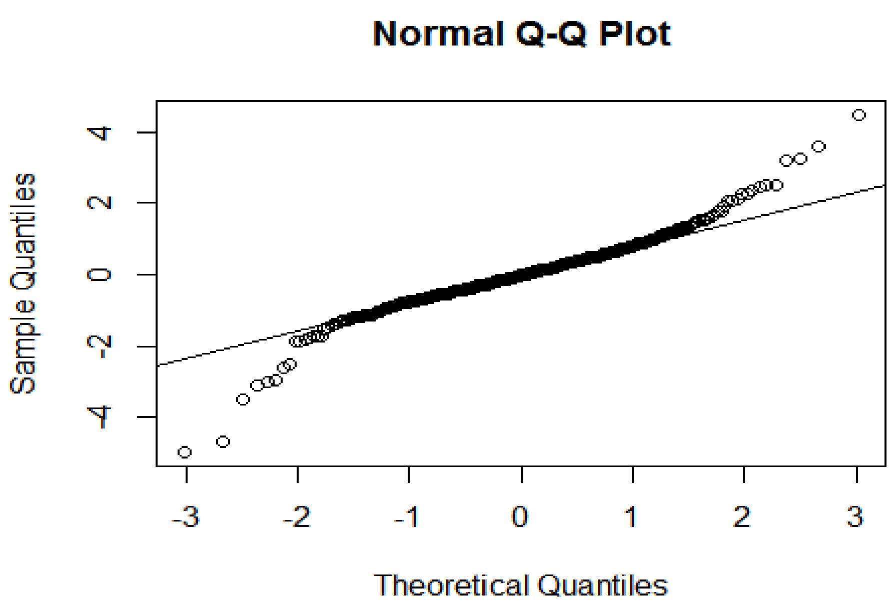
```
model5 <- lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + Post.Hour + Paid + log.like + log.sh
are,
         data = Train)
summary(model5)

##
## Call:
## lm(formula = log.Y ~ Page.total.likes + Type + Category + Post.Month +
##     Post.Weekday + Post.Hour + Paid + log.like + log.share, data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63264 -0.18679 -0.00712  0.16854  1.52826
##
## Coefficients:
##                   Estimate  Std. Error t value           Pr(>|t|)
## (Intercept)       5.62412732  1.17898624   4.770    0.00000270765100 ***
## Page.total.likes -0.00002639  0.00001361  -1.940              0.0532 .
## TypePhoto         0.71927546  0.10108739   7.115    0.00000000000638 ***
## TypeStatus        1.87741012  0.12343314  15.210 < 0.0000000000000002 ***
## TypeVideo         1.17032032  0.17346247   6.747    0.00000000006298 ***
## Category2        -0.37487409  0.06010914  -6.237    0.00000000129311 ***
## Category3        -0.46801505  0.05327246  -8.785 < 0.0000000000000002 ***
## Post.Month2       0.24284217  0.14351280   1.692              0.0915 .
## Post.Month3       0.08766509  0.22523844   0.389              0.6974
## Post.Month4       0.66012382  0.34840338   1.895              0.0590 .
## Post.Month5       0.68970732  0.44815625   1.539              0.1247
## Post.Month6       1.12294327  0.54462012   2.062              0.0400 *
## Post.Month7       1.00877858  0.60534158   1.666              0.0965 .
## Post.Month8       1.07207340  0.64963993   1.650              0.0998 .
## Post.Month9       0.98501424  0.68127660   1.446              0.1491
## Post.Month10      1.26838773  0.69888150   1.815              0.0704 .
## Post.Month11      0.49699644  0.71561358   0.695              0.4878
## Post.Month12      0.81318037  0.73406738   1.108              0.2687
## Post.Weekday2    -0.03454393  0.07208045  -0.479              0.6321
## Post.Weekday3     0.04235939  0.07559341   0.560              0.5756
## Post.Weekday4    -0.16666498  0.07341560  -2.270              0.0238 *
## Post.Weekday5    -0.12956936  0.07321411  -1.770              0.0776 .
## Post.Weekday6     0.01591256  0.07011160   0.227              0.8206
## Post.Weekday7     0.12935323  0.06953312   1.860              0.0637 .
## Post.Hour2        0.06142372  0.19850395   0.309              0.7572
## Post.Hour3       -0.04580579  0.18897213  -0.242              0.8086
## Post.Hour4        0.06950675  0.19906434   0.349              0.7272
## Post.Hour5        0.07562980  0.21666515   0.349              0.7273
## Post.Hour6       -0.21733362  0.21146749  -1.028              0.3048
## Post.Hour7        0.05395116  0.21922021   0.246              0.8057
## Post.Hour8       -0.17533458  0.22439271  -0.781              0.4351
## Post.Hour9        0.01727549  0.19996256   0.086              0.9312
## Post.Hour10      -0.00154621  0.18979571  -0.008              0.9935
## Post.Hour11      -0.07747443  0.19385753  -0.400              0.6897
## Post.Hour12       0.20873570  0.19954568   1.046              0.2963
## Post.Hour13       0.06607627  0.19283176   0.343              0.7321
## Post.Hour14       0.15479544  0.21758250   0.711              0.4773
```

```
## Post.Hour15     0.52905116  0.26152466   2.023            0.0438 *
## Post.Hour17     0.25485358  0.29427196   0.866            0.3871
## Post.Hour18     0.29119491  0.33004384   0.882            0.3782
## Paid1          0.05030013  0.04296987   1.171            0.2426
## log.like       0.47349941  0.03550244  13.337 < 0.0000000000000002 ***
## log.share       0.06725331  0.04564576   1.473            0.1416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3572 on 349 degrees of freedom
## Multiple R-squared:  0.817,  Adjusted R-squared:  0.795
## F-statistic:  37.1 on 42 and 349 DF,  p-value: < 0.00000000000000022
```

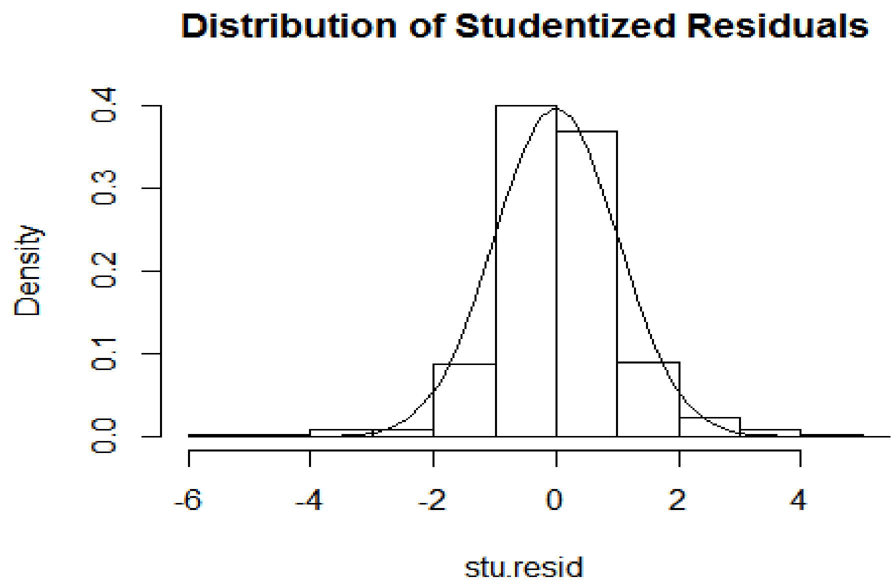
The R-Squared value increases to 0.817

Observing the residual plots and checking for Normality

```
residuals <- rstandard(model5)

qqnorm(residuals)
qqline(residuals)
```

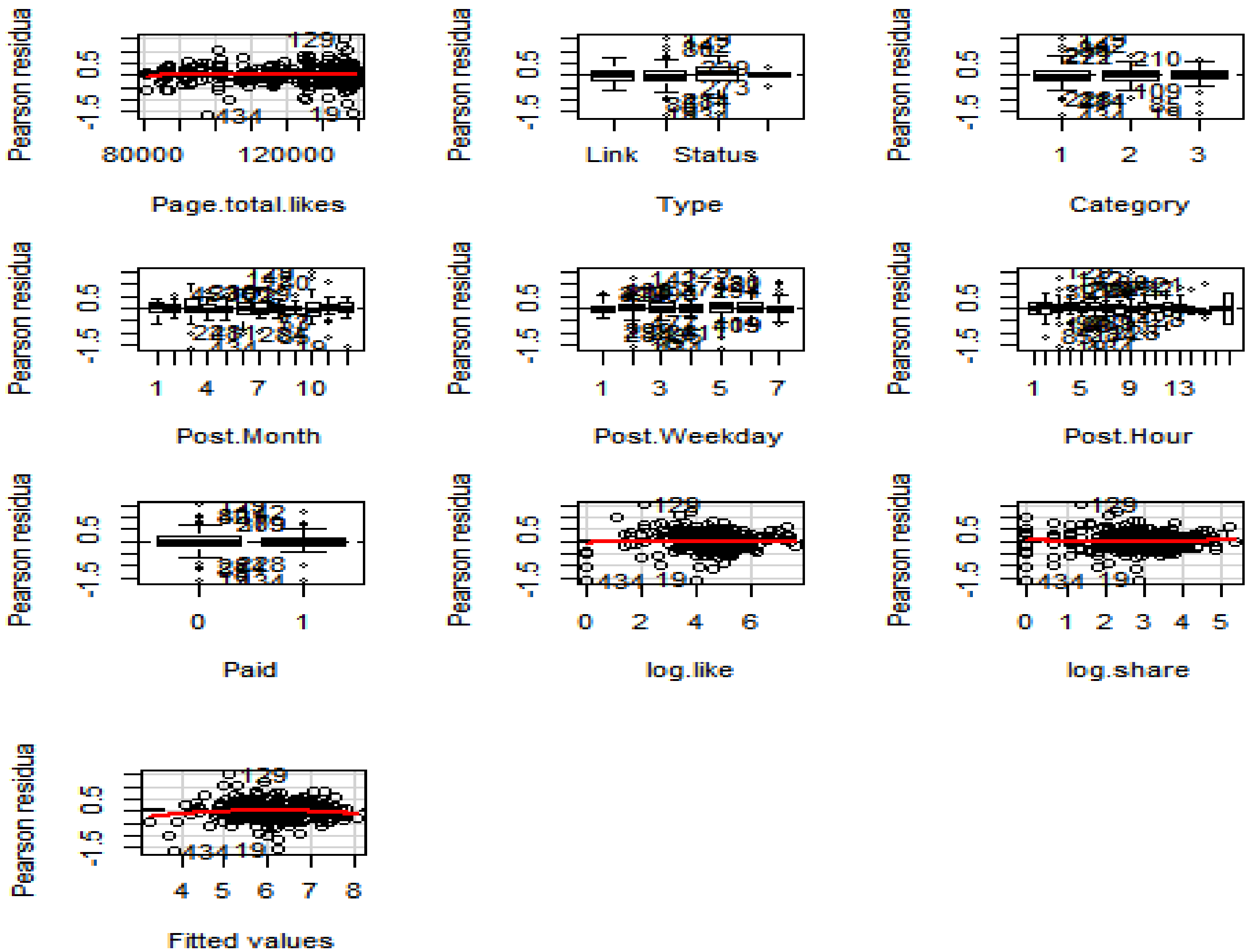



```
stu.resid <- studres(model5)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(-3.5, 7,length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Residuals plot with Fitted values and other Regressors

```
residualPlots(model5,id.n=3)
```




```
##                  Test stat Pr(>|t|)
## Page.total.likes    -1.857    0.064
## Type                    NA       NA
## Category                NA       NA
## Post.Month              NA       NA
## Post.Weekday            NA       NA
## Post.Hour               NA       NA
## Paid                    NA       NA
## log.like            -0.691    0.490
## log.share            0.366    0.715
## Tukey test          -1.980    0.048
```

We have built a model with R-Squared equal close to 0.82. The model fits the data well which can be even confirmed from the residual plots

Next we try to see if interations can improve the performance of the model.

Model 6

We try to see how interactions between categorical variables can improve the performance of the model. Interactions will help to identify how a particular post of particular kind when uploaded at a particular hour/month and if paid or not is able to attract maximum engagement from the user

On checking different permutations and combinations, we observed that interactions between Type, Post.Weekday and Post.Hour improves the performance of the model significantly. This interaction will help us determine which type of post when uploaded at what particular weekday and hour attracts maximum engagement from the user

```
model6 = lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + Post.Hour + Paid + log.like + log.sha
re +
         Type*Post.Weekday*Post.Hour , data = Train)
summary(model6)
```

(Complete Output not shown)

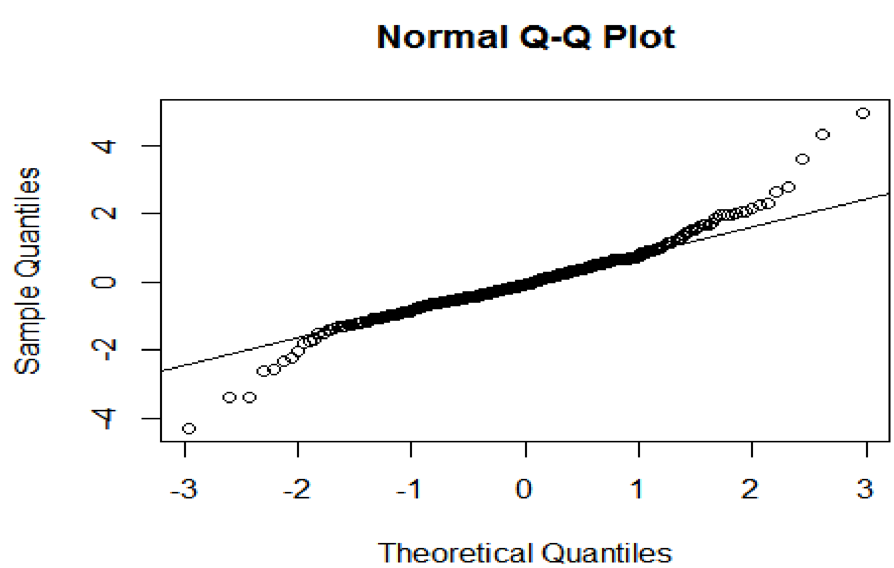
```
##
## Call:
## lm(formula = log.Y ~ Page.total.likes + Type + Category + Post.Month +
##     Post.Weekday + Post.Hour + Paid + log.like + log.share +
##     Type * Post.Weekday * Post.Hour, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9463 -0.1237  0.0000  0.1059  1.3538
##
## Coefficients: (339 not defined because of singularities)
##                                  Estimate  Std. Error t value
## (Intercept)                    6.94382533  1.85391396   3.745
## Page.total.likes              -0.00002743  0.00001547  -1.773
## TypePhoto                      1.39667315  0.95007800   1.470
## TypeStatus                     3.45614804  1.06117365   3.257
## TypeVideo                      1.14536920  0.89456069   1.280
## Category2                     -0.41929123  0.06405323  -6.546
## Category3                     -0.40538588  0.05570633  -7.277
## Post.Month2                    0.12936171  0.14939841   0.866
## Post.Month3                   -0.02802130  0.24936702  -0.112
## Post.Month4                    0.48147535  0.38043885   1.266
## Post.Month5                    0.52069277  0.50402292   1.033
## Post.Month6                    0.95504180  0.60866303   1.569
## Post.Month7                    0.89160163  0.67749465   1.316
## Post.Month8                    0.99677340  0.72804153   1.369
## Post.Month9                    0.91882522  0.76402528   1.203
## Post.Month10                   1.13908002  0.78847890   1.445
## Post.Month11                   0.44710979  0.80328434   0.557
## Post.Month12                   0.73959693  0.82640201   0.895
## Post.Weekday2                 -2.41209511  0.88351524  -2.730
## Post.Weekday3                 -2.16065670  0.91236374  -2.368
## Post.Weekday4                 -2.57533648  1.42397950  -1.809
## Post.Weekday5                 -1.90686467  1.18995522  -1.602
## Post.Weekday6                 -1.66315573  1.06425744  -1.563
## Post.Weekday7                 -0.80500325  1.13751048  -0.708
## Post.Hour2                    -0.54449018  1.30254588  -0.418
```

```
## Post.Hour3                    -0.29340856  1.48150862  -0.198
## Post.Hour4                    -0.83500582  1.54677795  -0.540
## Post.Hour5                    -1.92327977  0.76433066  -2.516
......................

......................

## TypePhoto:Post.Weekday5:Post.Hour18              NA
## TypeStatus:Post.Weekday5:Post.Hour18             NA
## TypeVideo:Post.Weekday5:Post.Hour18              NA
## TypePhoto:Post.Weekday6:Post.Hour18              NA
## TypeStatus:Post.Weekday6:Post.Hour18             NA
## TypeVideo:Post.Weekday6:Post.Hour18              NA
## TypePhoto:Post.Weekday7:Post.Hour18              NA
## TypeStatus:Post.Weekday7:Post.Hour18             NA
## TypeVideo:Post.Weekday7:Post.Hour18              NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.316 on 238 degrees of freedom
## Multiple R-squared:  0.9024, Adjusted R-squared:  0.8396
## F-statistic: 14.38 on 153 and 238 DF,  p-value: < 0.00000000000000022
```

The R-Squared of the model has increased to 0.9024. The Adjusted R-Squared has also improved

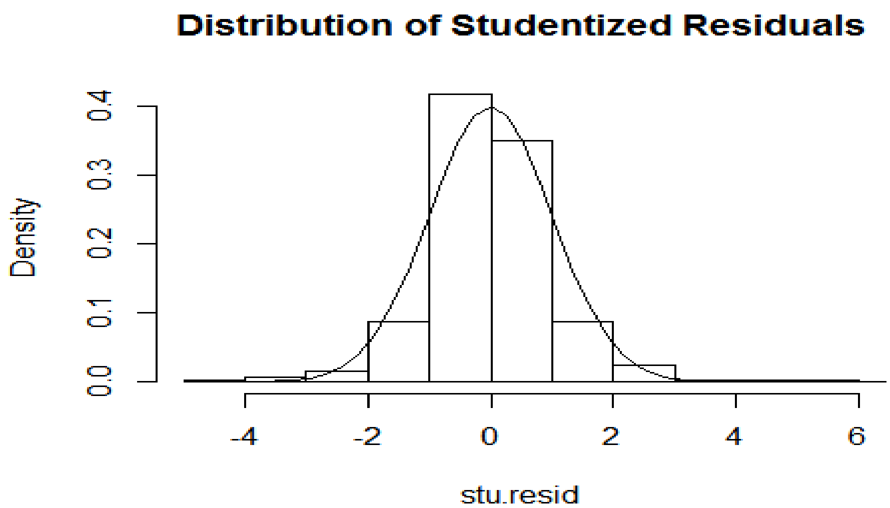
Observing the residual plots and checking for Normality

```
residuals <- rstandard(model6)
qqnorm(residuals)
qqline(residuals)
```



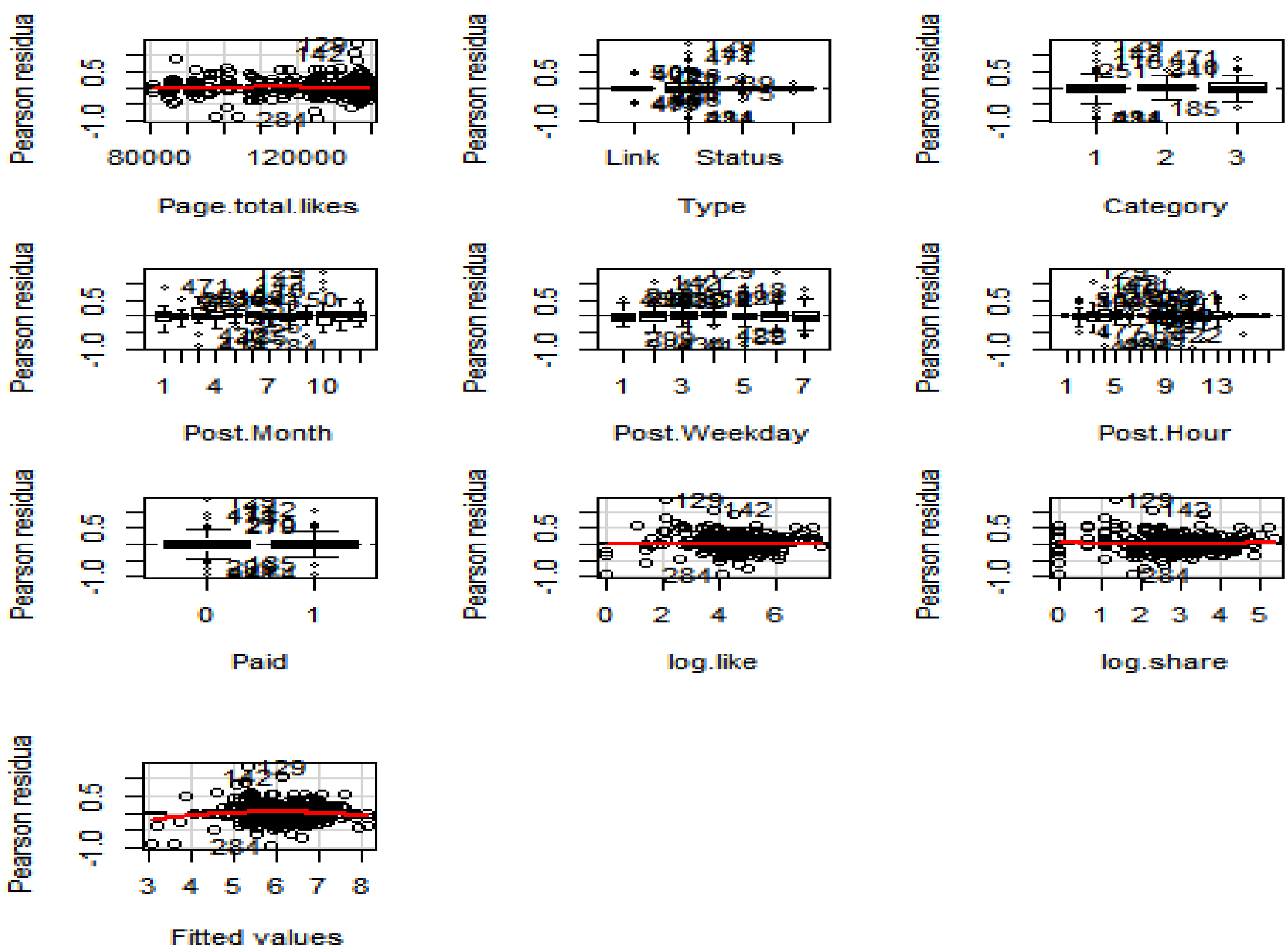

```
stu.resid <- studres(model6)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(-3.5, 7,length=40)
```

```
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Residuals plot with Fitted values and other Regressors

```
residualPlots(model6,id.n=3)
```



```
##                  Test stat Pr(>|t|)
## Page.total.likes    -0.756    0.450
## Type                    NA       NA
## Category                NA       NA
## Post.Month              NA       NA
## Post.Weekday            NA       NA
## Post.Hour               NA       NA
## Paid                    NA       NA
## log.like            -0.234    0.815
## log.share            0.797    0.426
## Tukey test          -2.651    0.008
```

Observing the residuals plots, the model fits well.

Next we select the best subset model using stepwise regression

---

Best Subset selection

We will use the AIC criterion for obtaining the best subset.

```r
step <- stepAIC(model6, direction="both")
step$anova # display results
```

```
## Start:  AIC=-790.88
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + 
##     Post.Hour + Paid + log.like + log.share + Type * Post.Weekday * 
##     Post.Hour
## 
##                              Df Sum of Sq    RSS     AIC
## - Type:Post.Weekday:Post.Hour  6    0.1815 23.941 -799.90
## - log.share                    1    0.0901 23.850 -791.40
## <none>                                     23.760 -790.88
## - Paid                         1    0.1411 23.901 -790.56
## - Page.total.likes             1    0.3138 24.074 -787.74
## - Category                     2    6.2428 30.003 -703.43
## - Post.Month                  11   10.6932 34.453 -667.21
## - log.like                     1   15.5567 39.317 -595.45
## 
## Step:  AIC=-799.9
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + 
##     Post.Hour + Paid + log.like + log.share + Type:Post.Weekday + 
##     Type:Post.Hour + Post.Weekday:Post.Hour
## 
##                              Df Sum of Sq    RSS     AIC
## - Post.Weekday:Post.Hour      71   10.0982 34.040 -803.95
## - log.share                    1    0.0652 24.007 -800.83
## - Paid                         1    0.0986 24.040 -800.29
## <none>                                     23.941 -799.90
## - Page.total.likes             1    0.3345 24.276 -796.46
## + Type:Post.Weekday:Post.Hour  6    0.1815 23.760 -790.88
```

```
## - Type:Post.Hour           16    2.7328 26.674 -789.53
## - Type:Post.Weekday        12    4.1031 28.044 -761.89
## - Category                  2    6.3790 30.320 -711.30
## - Post.Month               11   10.7900 34.731 -676.06
## - log.like                  1   16.5228 40.464 -596.17
##
## Step:  AIC=-803.95
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##     Post.Hour + Paid + log.like + log.share + Type:Post.Weekday +
##     Type:Post.Hour
##
##                        Df Sum of Sq    RSS     AIC
## - log.share             1    0.0773 34.117 -805.06
## <none>                              34.040 -803.95
## - Paid                  1    0.1758 34.215 -803.93
## - Page.total.likes      1    0.2635 34.303 -802.92
## + Post.Weekday:Post.Hour 71   10.0982 23.941 -799.90
## - Type:Post.Hour       18    4.7812 38.821 -788.42
## - Type:Post.Weekday    12    3.8198 37.859 -786.25
## - Category              2    9.1732 43.213 -714.41
## - Post.Month           11   16.6826 50.722 -669.60
## - log.like              1   21.3224 55.362 -615.29
##
## Step:  AIC=-805.06
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##     Post.Hour + Paid + log.like + Type:Post.Weekday + Type:Post.Hour
##
##                        Df Sum of Sq     RSS     AIC
## - Paid                  1     0.155  34.272 -805.28
## <none>                               34.117 -805.06
## - Page.total.likes      1     0.241  34.358 -804.30
## + log.share             1     0.077  34.040 -803.95
## + Post.Weekday:Post.Hour 71    10.110  24.007 -800.83
## - Type:Post.Hour       18     4.804  38.921 -789.42
## - Type:Post.Weekday    12     3.877  37.994 -786.86
## - Category              2     9.647  43.764 -711.44
## - Post.Month           11    16.607  50.723 -671.59
## - log.like              1    86.827 120.943 -310.97
##
## Step:  AIC=-805.28
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##     Post.Hour + log.like + Type:Post.Weekday + Type:Post.Hour
##
##                        Df Sum of Sq     RSS     AIC
## <none>                               34.272 -805.28
## + Paid                  1     0.155  34.117 -805.06
## - Page.total.likes      1     0.223  34.494 -804.74
## + log.share             1     0.056  34.215 -803.93
## + Post.Weekday:Post.Hour 71    10.182  24.089 -801.48
## - Type:Post.Hour       18     4.833  39.105 -789.57
## - Type:Post.Weekday    12     3.918  38.189 -786.85
## - Category              2     9.726  43.997 -711.36
## - Post.Month           11    16.645  50.916 -672.10
```

```
## - log.like              1    87.357 121.628 -310.75
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##     Post.Hour + Paid + log.like + log.share + Type * Post.Weekday *
##     Post.Hour
##
## Final Model:
## log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
##     Post.Hour + log.like + Type:Post.Weekday + Type:Post.Hour
##
##
##                        Step Df    Deviance Resid. Df Resid. Dev
## 1                                              238   23.75987
## 2 - Type:Post.Weekday:Post.Hour  6  0.18145316       244   23.94132
## 3      - Post.Weekday:Post.Hour 71 10.09819607       315   34.03952
## 4                  - log.share  1  0.07725916       316   34.11678
## 5                       - Paid  1  0.15475534       317   34.27153
##        AIC
## 1 -790.8795
## 2 -799.8971
## 3 -803.9460
## 4 -805.0573
## 5 -805.2832
```

It is observed that many regressors have dropped. The best model is

lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday + Post.Hour + log.like + Type:Post.Weekday + Type:Post.Hour, data = Train)

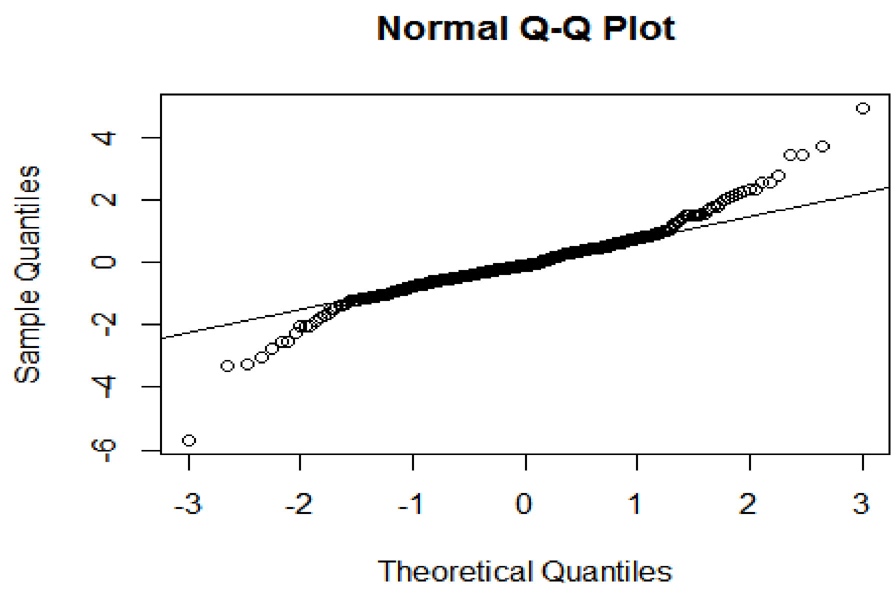
```r
BestModel <- lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
             Post.Hour + log.like + Type:Post.Weekday + Type:Post.Hour, data = Train)
summary(BestModel)
```

```
##
## Call:
## lm(formula = log.Y ~ Page.total.likes + Type + Category + Post.Month +
##     Post.Weekday + Post.Hour + log.like + Type:Post.Weekday +
##     Type:Post.Hour, data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71563 -0.14492 -0.02016  0.13705  1.54899
##
## Coefficients: (32 not defined because of singularities)
##                        Estimate  Std. Error t value
## (Intercept)             4.28971731  1.41044749   3.041
## Page.total.likes       -0.00001861  0.00001297  -1.435
## TypePhoto               1.48363424  0.77329938   1.919
## TypeStatus              3.39249746  0.80399652   4.220
## TypeVideo               0.83143779  0.80490701   1.033
```

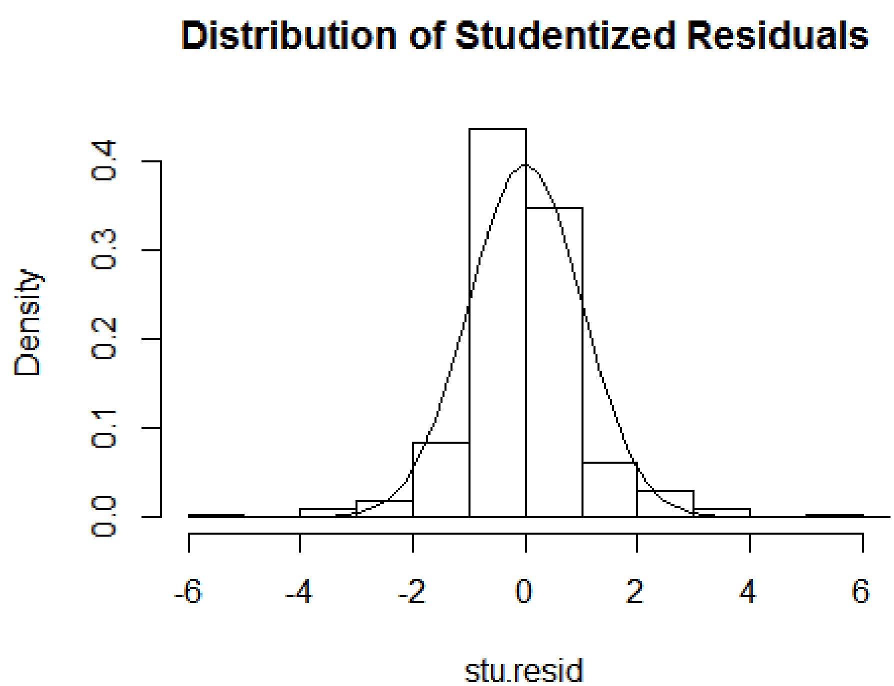
```
## Category2            -0.39966068  0.05688039  -7.026
## Category3            -0.43802327  0.04856116  -9.020
## Post.Month2            0.12649177  0.13805673   0.916
## Post.Month3           -0.10738353  0.21637312  -0.496
## Post.Month4            0.42831639  0.33177593   1.291
## Post.Month5            0.38399626  0.42738775   0.898
## Post.Month6            0.72600669  0.51745093   1.403
## Post.Month7            0.58286255  0.57814267   1.008
## Post.Month8            0.62914500  0.61771340   1.019
## Post.Month9            0.54763016  0.64930574   0.843
## Post.Month10           0.80164131  0.66674052   1.202
## Post.Month11           0.05235743  0.68272223   0.077
## Post.Month12           0.30656392  0.69879593   0.439
## Post.Weekday2         -0.68599074  0.41121287  -1.668
……..

………

## TypePhoto:Post.Hour13             0.01747 *
## TypeStatus:Post.Hour13                 NA
## TypeVideo:Post.Hour13                  NA
## TypePhoto:Post.Hour14             0.03665 *
## TypeStatus:Post.Hour14                 NA
## TypeVideo:Post.Hour14                  NA
## TypePhoto:Post.Hour15                  NA
## TypeStatus:Post.Hour15                 NA
## TypeVideo:Post.Hour15                  NA
## TypePhoto:Post.Hour17                  NA
## TypeStatus:Post.Hour17                 NA
## TypeVideo:Post.Hour17                  NA
## TypePhoto:Post.Hour18                  NA
## TypeStatus:Post.Hour18                 NA
## TypeVideo:Post.Hour18                  NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3288 on 317 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8263
## F-statistic: 26.13 on 74 and 317 DF,  p-value: < 0.00000000000000022
```

Observing the residual plots and checking for Normality

```
residuals <- rstandard(BestModel)
qqnorm(residuals)
qqline(residuals)
```

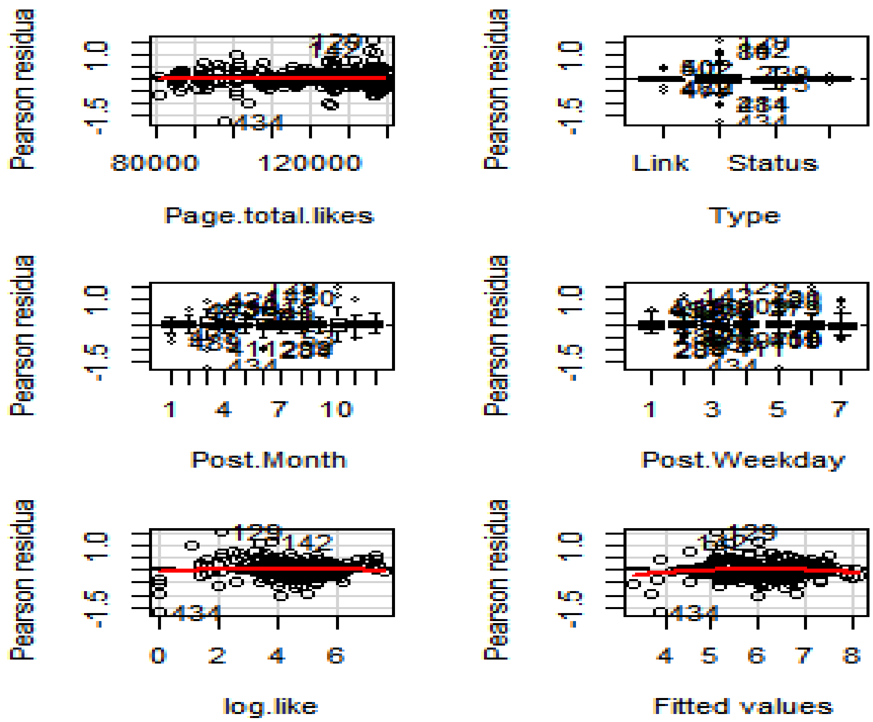



```
stu.resid <- studres(BestModel)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(-3.5, 7,length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Residuals plot with Fitted values and other Regressors

```
residualPlots(BestModel,id.n=3)
```




```
##                Test stat Pr(>|t|)
## Page.total.likes    -2.078    0.039
## Type                    NA       NA
## Category                NA       NA
## Post.Month              NA       NA
## Post.Weekday            NA       NA
## Post.Hour               NA       NA
## log.like            -0.392    0.695
## Tukey test          -2.576    0.010
```

We have successfully built a model which explains almost 86% variability in the data with most significant regressors

---

Validation

We test our model using the test data set and use the model BestModel for predictions

```
Test$log.Y <- log(Test$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)
Test$log.like <- log(Test$like+1)

y_hat <- predict.lm(BestModel, newdata = Test, se.fit=TRUE)$fit
y_hat <- as.vector(y_hat)
dev <- Test$log.Y - (y_hat)
num <- sum(dev^2)
dev1 <- Test$log.Y - mean(log(Test$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post))
den <- sum(dev1^2)
Predicted.Rsq <- 1 - (num/den)
Predicted.Rsq

## [1] 0.6230459
```

We obtain an R-Squared value of 0.623. Overall the R-Squared value is well

PRESS Statistics

```
press <- PRESS(BestModel)
press$P.square

sum(press$residuals^2)
sum(BestModel$residuals^2)

## .........10.........20.........30.........40.........50
## .........60.........70.........80.........90.........100
## .........110.........120.........130.........140.........150
## .........160.........170.........180.........190.........200
## .........210.........220.........230.........240.........250
## .........260.........270.........280.........290.........300
## .........310.........320.........330.........340.........350
## .........360.........370.........380.........390..
## [1] 0.6817248
## [1] 77.44487
## [1] 34.27153
```

- A low value of PRESS statistics is a good indicator that the model is good for predictions
- This can be further confirmed by comparing the sum of PRESS residuals and sum of Best Model residuals, since the two residuals are close, the model can be used for predictions

---

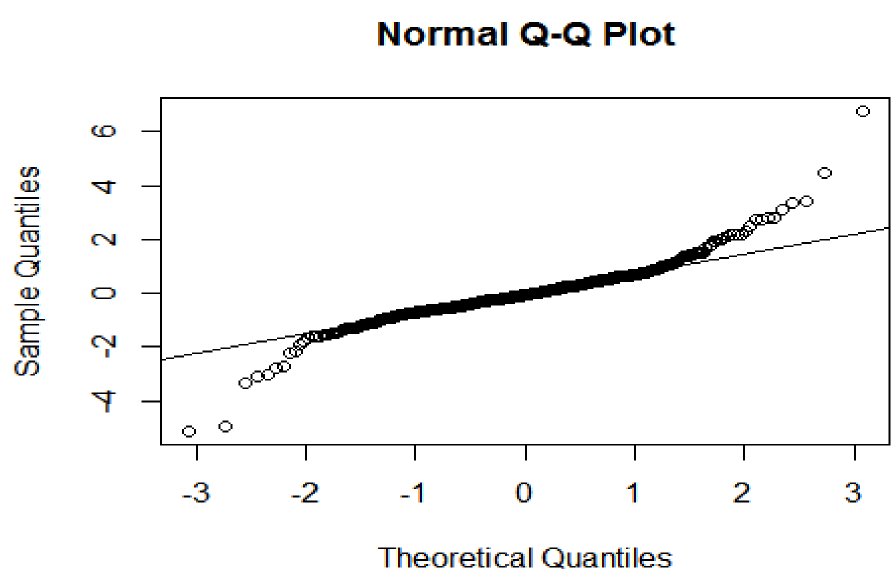
Running the model on our original data. [Using the entire data(n = 490)]

```
fb.raw$log.Y <- log(fb.raw$Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post)
fb.raw$log.like <- log(fb.raw$like+1)

FbModel <- lm(log.Y ~ Page.total.likes + Type + Category + Post.Month + Post.Weekday +
            Post.Hour + log.like + Type:Post.Weekday + Type:Post.Hour, data = fb.raw)
summary(FbModel)

##
## Call:
## lm(formula = log.Y ~ Page.total.likes + Type + Category + Post.Month +
##     Post.Weekday + Post.Hour + log.like + Type:Post.Weekday +
##     Type:Post.Hour, data = fb.raw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72856 -0.18196 -0.01136  0.15224  2.44822
##
## Coefficients: (30 not defined because of singularities)
##                       Estimate  Std. Error t value
## (Intercept)           3.01855779  1.55880658   1.936
## Page.total.likes     -0.00002122  0.00001306  -1.625
## TypePhoto             3.00473332  1.01670620   2.955
```

```
## TypeStatus            3.95929498  0.93241062   4.246
## TypeVideo             1.46853081  0.56927235   2.580
## Category2            -0.40395934  0.05479461  -7.372
## Category3            -0.44257688  0.04806866  -9.207
....

.....

## TypePhoto:Post.Hour14             0.010637 *
## TypeStatus:Post.Hour14                  NA
## TypeVideo:Post.Hour14                   NA
## TypePhoto:Post.Hour15                   NA
## TypeStatus:Post.Hour15                  NA
## TypeVideo:Post.Hour15                   NA
## TypePhoto:Post.Hour17                   NA
## TypeStatus:Post.Hour17                  NA
## TypeVideo:Post.Hour17                   NA
## TypePhoto:Post.Hour18                   NA
## TypeStatus:Post.Hour18                  NA
## TypeVideo:Post.Hour18                   NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3738 on 413 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.7948
## F-statistic: 25.92 on 76 and 413 DF,  p-value: < 0.0000000000000022
```

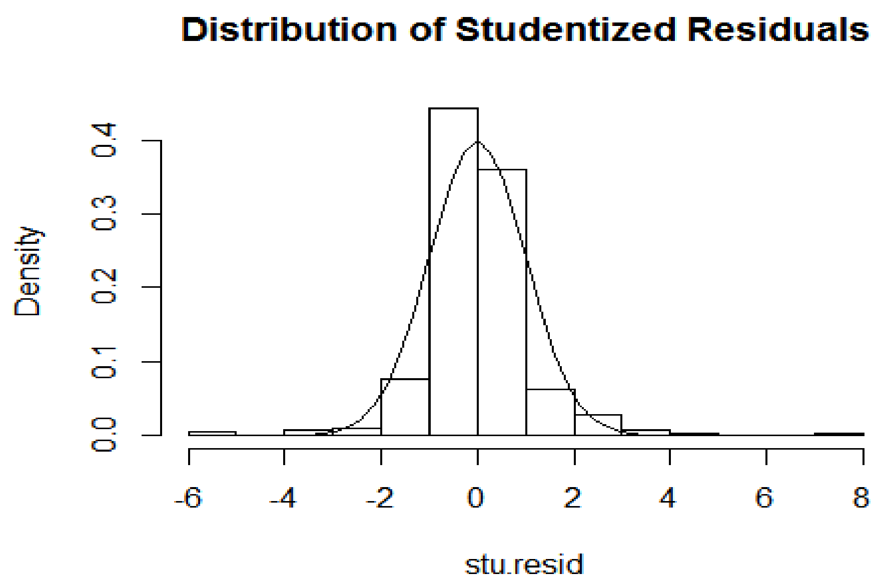
Observing the residual plots and checking for Normality

```r
residuals <- rstandard(FbModel)
qqnorm(residuals)
qqline(residuals)
```



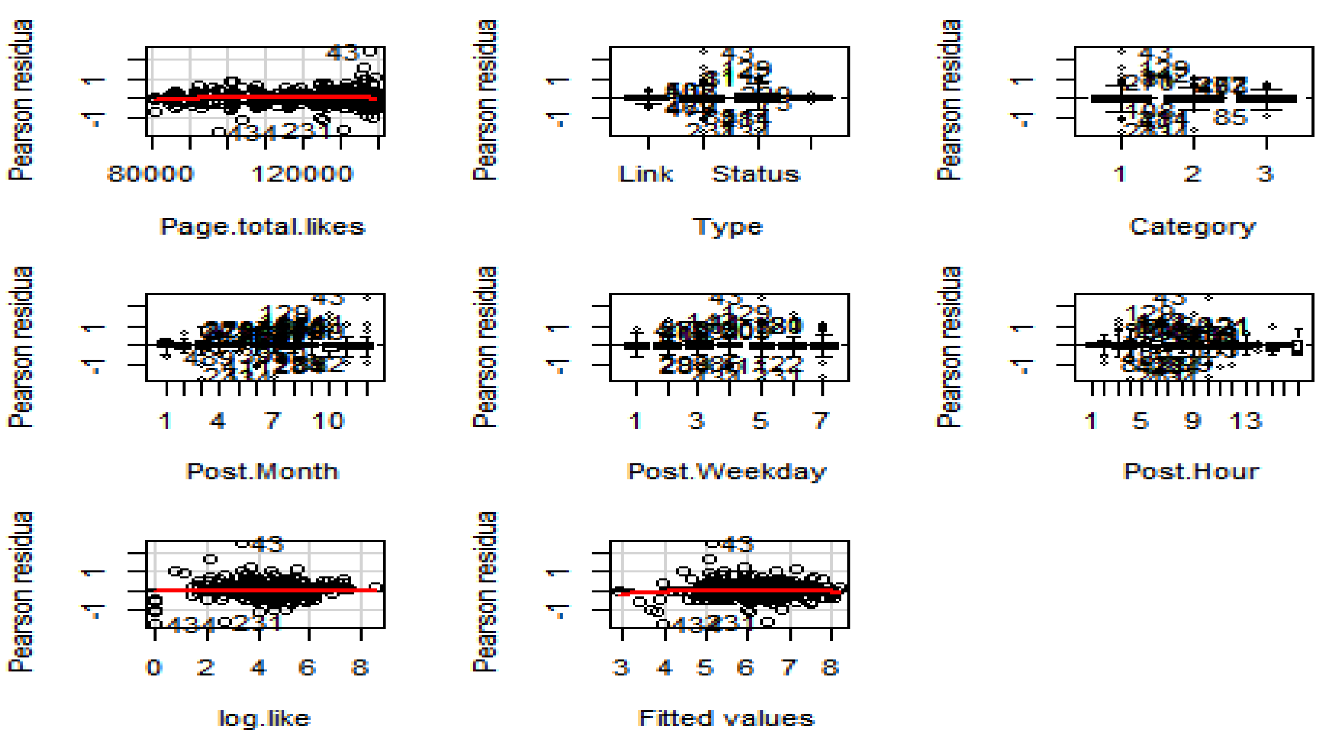

```r
stu.resid <- studres(FbModel)
hist(stu.resid, freq=FALSE, main="Distribution of Studentized Residuals")
```

```
xfit<-seq(-3.5, 7,length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Residuals plot with Fitted values and other Regressors

```
residualPlots(FbModel,id.n=3)
```



```
##              Test stat Pr(>|t|)
## Page.total.likes    -3.090    0.002
## Type                  NA       NA
## Category              NA       NA
## Post.Month            NA       NA
## Post.Weekday          NA       NA
## Post.Hour             NA       NA
## log.like          -0.303    0.762
## Tukey test        -2.004    0.045
```

We are able to build a model to predict the performance of the page in terms of Lifetime people who have liked your page and engaged with your post which explains close to 83% variability.

INTERPRETATION

```
FbModel$coefficients1 <- FbModel$coefficients[!is.na(FbModel$coefficients)]
```

Positive coefficients

```
sort(FbModel$coefficients1[FbModel$coefficients1 >0], decreasing = T)

##              TypeStatus            Post.Hour10             Post.Hour3
##              3.95929498             3.18429267             3.11646277
##             (Intercept)              TypePhoto             Post.Hour6
##              3.01855779             3.00473332             2.95083450
##              Post.Hour7            Post.Hour13             Post.Hour2
##              2.91073530             2.70287984             2.64174197
##              Post.Hour4            Post.Hour11            Post.Hour14
##              2.61455364             2.53720947             1.59799659
##               TypeVideo            Post.Hour12           Post.Month10
##              1.46853081             1.07374862             0.87650400
##              Post.Hour9            Post.Month6            Post.Month8
##              0.86395972             0.83892138             0.83027192
##  TypePhoto:Post.Weekday5            Post.Month9            Post.Month7
##              0.73459941             0.70456693             0.67487153
##  TypePhoto:Post.Weekday7  TypePhoto:Post.Weekday6  TypePhoto:Post.Weekday2
##              0.66618948             0.65189695             0.62118930
##            Post.Month12               log.like            Post.Hour15
##              0.50680170             0.50098850             0.47601360
##  TypePhoto:Post.Weekday4            Post.Month4            Post.Month5
##              0.47166813             0.46720562             0.46629500
## TypeStatus:Post.Weekday7  TypePhoto:Post.Weekday3 TypeStatus:Post.Weekday6
##              0.40130336             0.37067232             0.36966180
## TypeStatus:Post.Weekday5 TypeStatus:Post.Weekday3           Post.Month11
##              0.31108257             0.29449208             0.22309295
##             Post.Month2            Post.Hour17             Post.Hour5
##              0.17569164             0.13248821             0.08618151
##  TypeVideo:Post.Weekday3            Post.Month3
##              0.04815655             0.01913903
```

Negative coefficients

```
sort(FbModel$coefficients1[FbModel$coefficients1 < 0], decreasing = F)

##    TypePhoto:Post.Hour10     TypePhoto:Post.Hour3     TypePhoto:Post.Hour6
##         -3.24070070727         -3.15783374832         -3.10368460543
##    TypeStatus:Post.Hour7    TypePhoto:Post.Hour13     TypePhoto:Post.Hour7
##         -2.71658644033         -2.66742679692         -2.66207828620
##    TypePhoto:Post.Hour11     TypePhoto:Post.Hour2     TypePhoto:Post.Hour4
##         -2.66004649892         -2.60333841326         -2.59846367974
##    TypeStatus:Post.Hour3   TypeStatus:Post.Hour10    TypeStatus:Post.Hour2
##         -2.48376804938         -2.40794440983         -2.34367765213
##    TypeStatus:Post.Hour4    TypeStatus:Post.Hour6   TypeStatus:Post.Hour11
##         -2.30808527027         -2.29009822395         -1.74019030647
##    TypePhoto:Post.Hour14   TypeStatus:Post.Hour13 TypeStatus:Post.Weekday2
##         -1.57321576896         -1.40122912970         -1.15742990170
##     TypePhoto:Post.Hour9    TypePhoto:Post.Hour12           Post.Weekday5
##         -0.90062676439         -0.87047188938         -0.78396486560
##            Post.Weekday6            Post.Weekday4            Post.Weekday2
##         -0.62160594811         -0.56582244666         -0.56314248263
##  TypeVideo:Post.Weekday2            Post.Weekday7               Category3
##         -0.53449179501         -0.52893335848         -0.44257688432
##                Category2  TypeVideo:Post.Weekday5            Post.Weekday3
##         -0.40395933760         -0.39212148398         -0.36303331988
##  TypeVideo:Post.Weekday4    TypeVideo:Post.Hour10               Post.Hour8
##         -0.18205492840         -0.11869049104         -0.08050849230
## TypeStatus:Post.Weekday4              Post.Hour18         Page.total.likes
##         -0.03127278968         -0.02002780958         -0.00002121669
```

- A page can get maximum engagement from people based on what type of content is uploaded, during what time, the category of the page, how many likes the post has received and number of people who have liked the page
- The base model with just the intercept (Type: Link, Category: 1, Post.Month: 1, Post.Weekday: 1, Post.Hour: 1) suggest that on average, close to 20 people (exp(3.01855778735)) who have liked the page will also engage with the post
- One percent increase in the number of likes increases the engagement level by 0.5%
- Engagement level increases when the post is

    1. Type Photo is uploaded on Weekday3, Weekday5, Weekday7
    2. Type Photo is uploaded at hour 7,13,11,2,4
    3. Type Status is uploaded on Weekday3, Weekday5, Weekday6, Weekday7
    4. Type Status is uploaded at hour 2,4,11
    5. Significantly, Type Photo, Status is uploaded at hour 3,6,10
    6. Type Video is uploaded on Weekday3

- Engagement level increases very little when the post is
    1. Type Video is uploaded on Weekday4, Weekday5
    2. Type Video is uploaded on Hour10