

# Project-3

Dongyu Wang

2023-10-25

## Project Deliverable 3 Q3

Outcome Page.total.likes, main predictor Type

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

```
library(ggplot2)  
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':  
##   method from  
##   +.gg      ggplot2
```

```
library(RNOMni)  
library(car)
```

```
## Loading required package: carData
```

```
##  
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':  
##  
##   recode
```

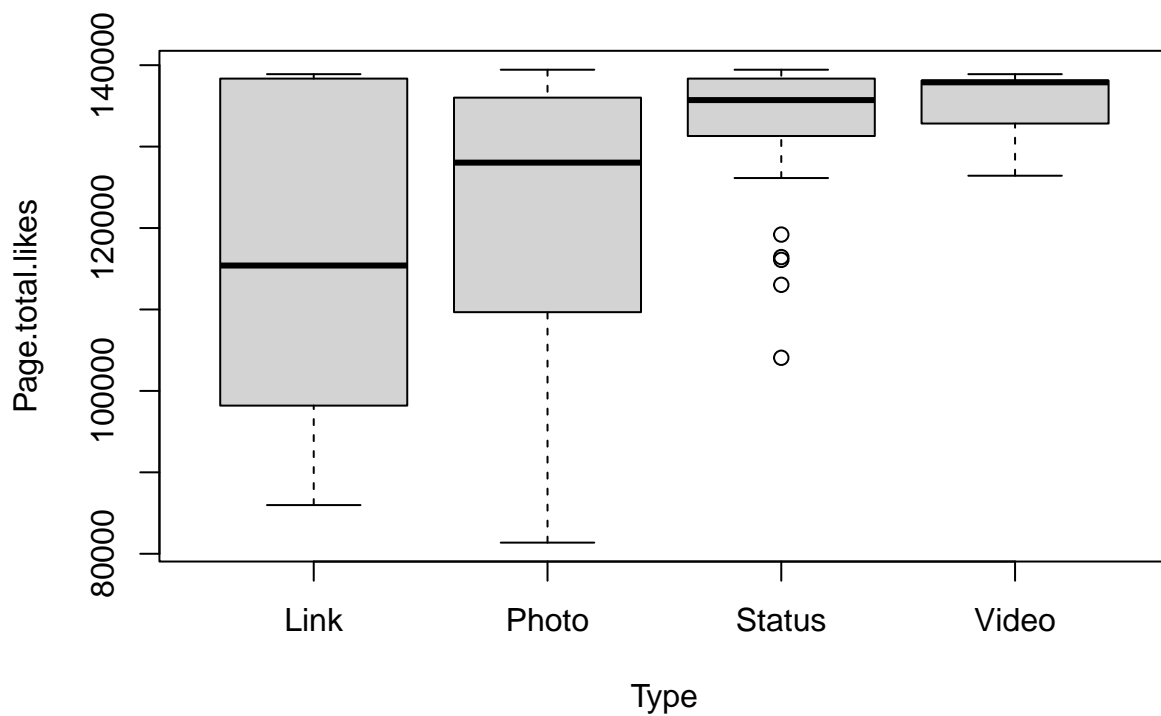
```
setwd("C:/Users/wdy24/Dropbox/BU_PHD/Coursework/BU MA series/MA575/Lab/Project/Code_dataset")
fb <- read.csv(file="dataset_Facebook(3).csv",header = T,sep = ";")
```

## Descriptive analysis

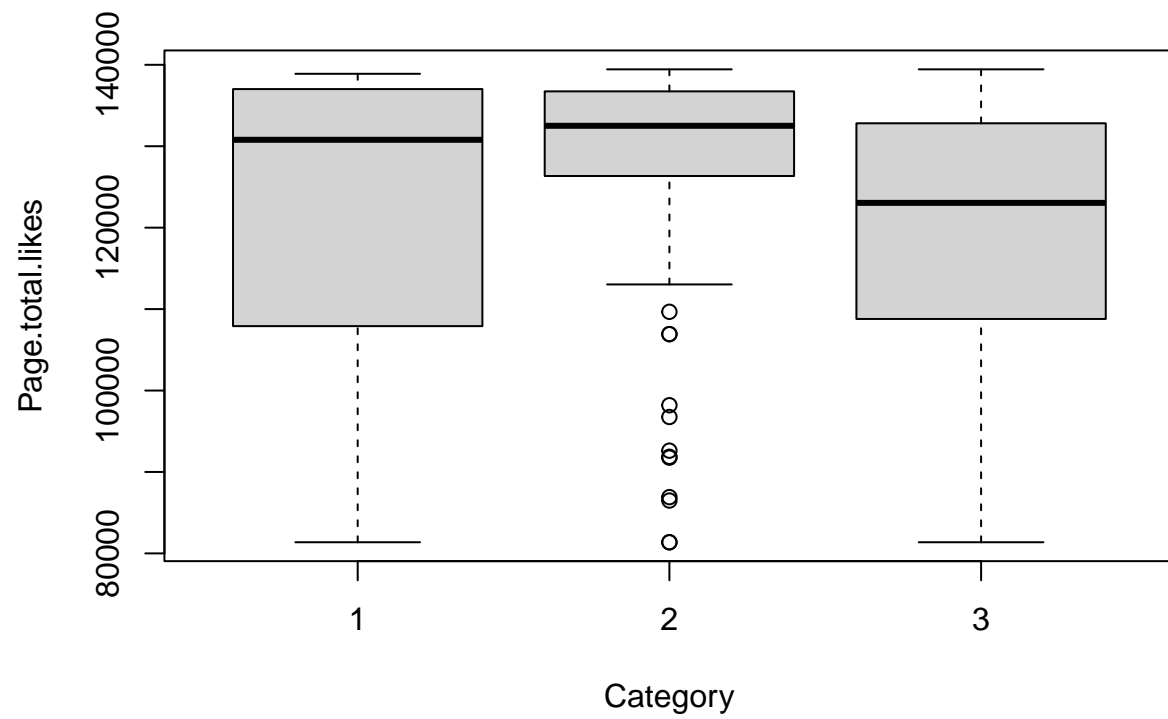
```
attach(fb)
table(fb$Type, useNA = 'always')
```

```
##
##   Link  Photo Status  Video  <NA>
##    22    426     45     7      0
```

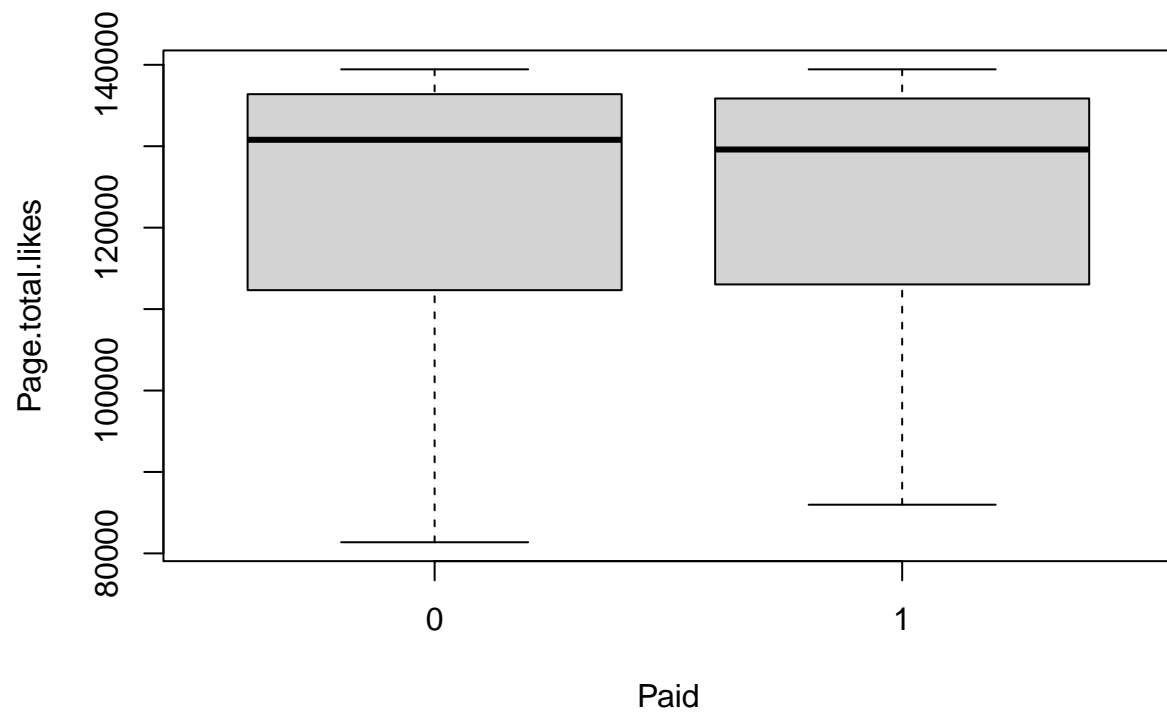
```
boxplot(Page.total.likes ~ Type)
```



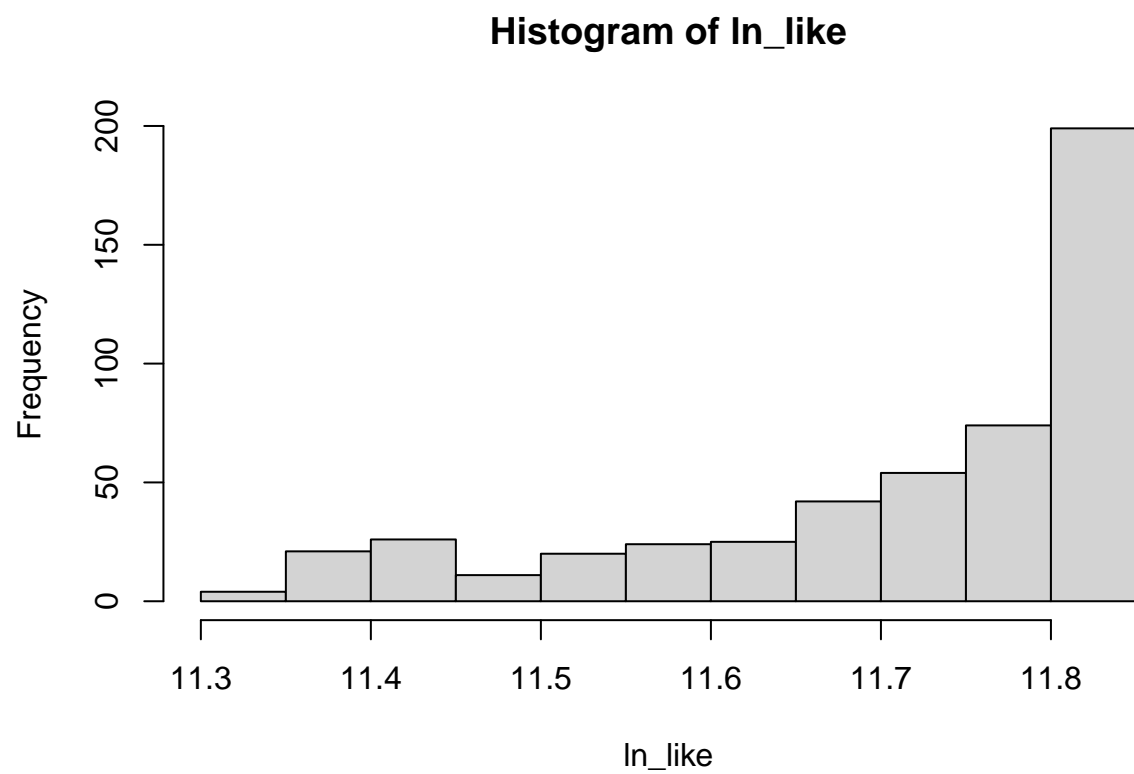
```
boxplot(Page.total.likes ~ Category)
```



```
boxplot(Page.total.likes ~ Paid)
```

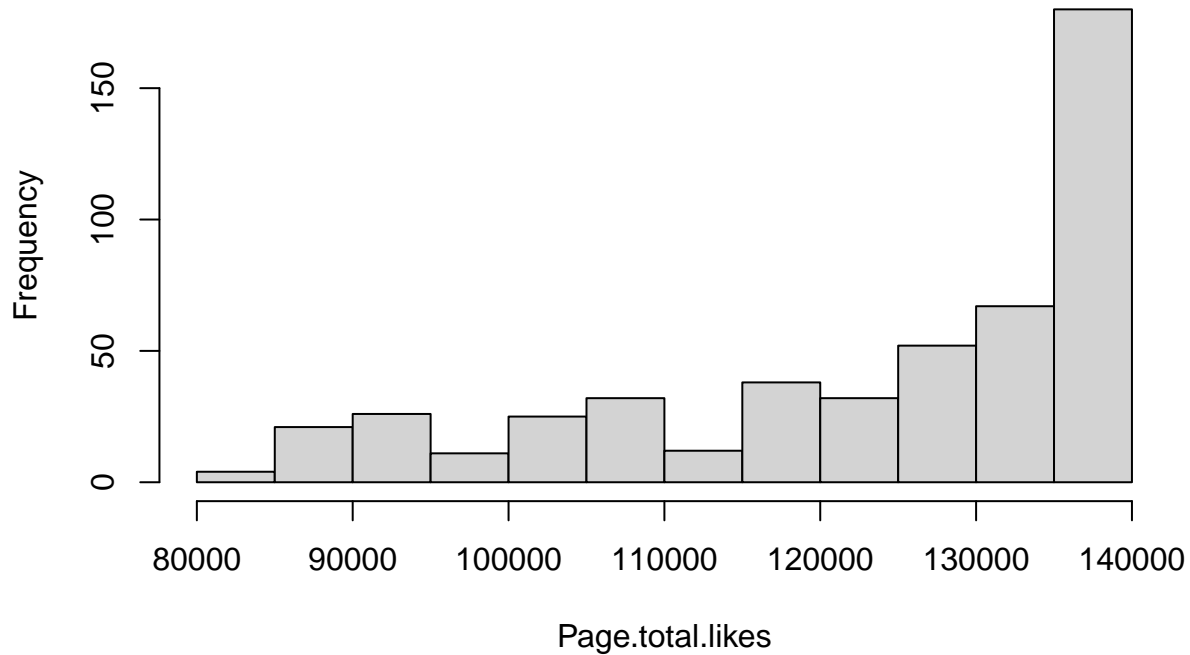


```
ln_like <- log(Page.total.likes)
hist(ln_like)
```



```
hist(Page.total.likes)
```

## Histogram of Page.total.likes



```
tapply(Page.total.likes,Type, summary)
```

```
## $Link
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  85979  98829 115396 116363 138059 138895
##
## $Photo
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  81370 109670 128032 122354 136013 139441
##
## $Status
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 104070 131300 135713 132647 138353 139441
##
## $Video
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 126424 132835 137893 135015 138111 138895
```

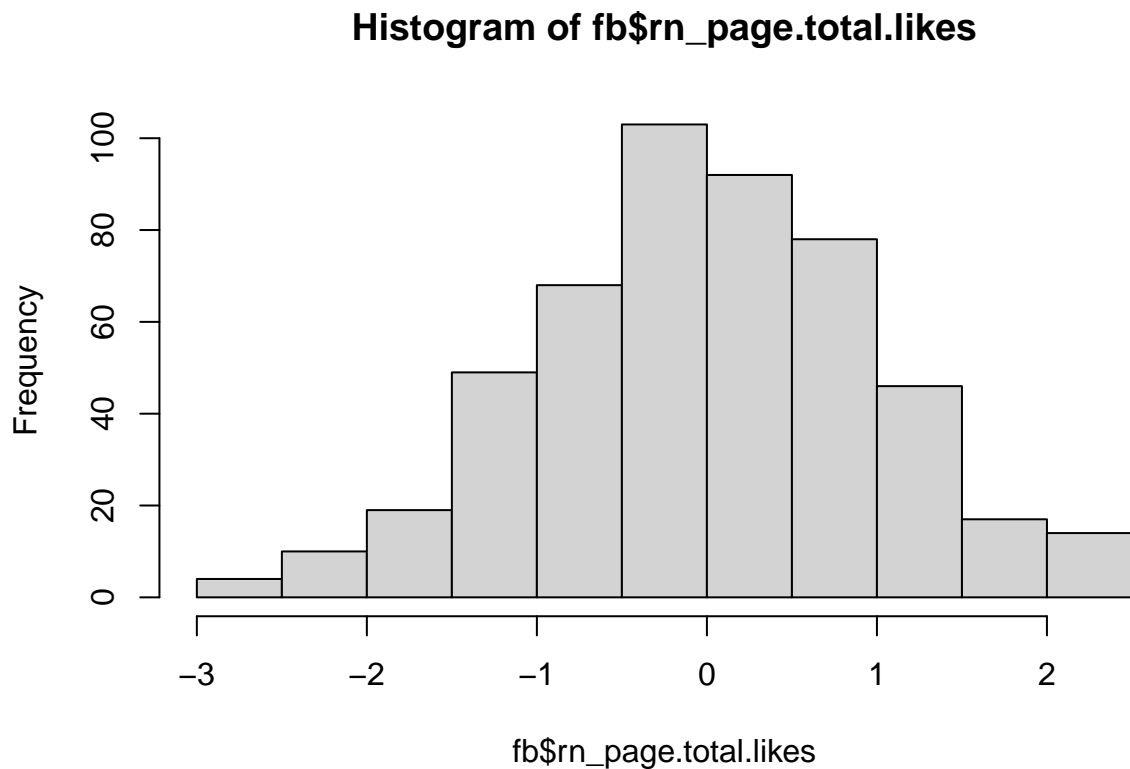
```
fb$rn_page.total.likes <- RankNorm(Page.total.likes,ties.method = "average")
summary(fb$rn_page.total.likes)
```

```
##      Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.631705 -0.664782 -0.022550 -0.002038  0.655348  2.190531
```

```
sd(fb$rn_page.total.likes)
```

```
## [1] 0.9878953
```

```
hist(fb$rn_page.total.likes)
```



```
attach(fb)
```

```
## The following objects are masked from fb (pos = 3):
```

```
##
```

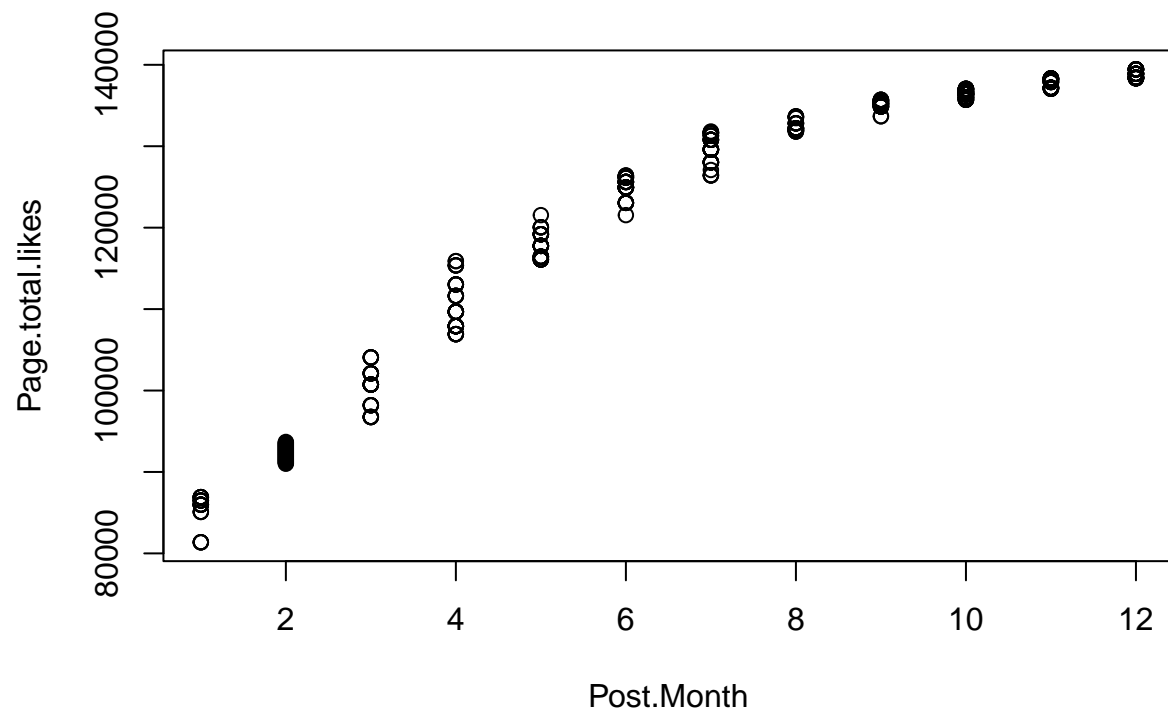
```
## Category, comment, Lifetime.Engaged.Users,  
## Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post,  
## Lifetime.Post.Consumers, Lifetime.Post.Consumptions,  
## Lifetime.Post.Impressions.by.people.who.have.liked.your.Page,  
## Lifetime.Post.reach.by.people.who.like.your.Page,  
## Lifetime.Post.Total.Impressions, Lifetime.Post.Total.Reach, like,  
## Page.total.likes, Paid, Post.Hour, Post.Month, Post.Weekday, share,  
## Total.Interactions, Type
```

```
## Page.total.likes highly skewed toward the left - need to transform but first proceed with the MV mod
```

```
## Inital variable selection from scatter plot + heatmap
```

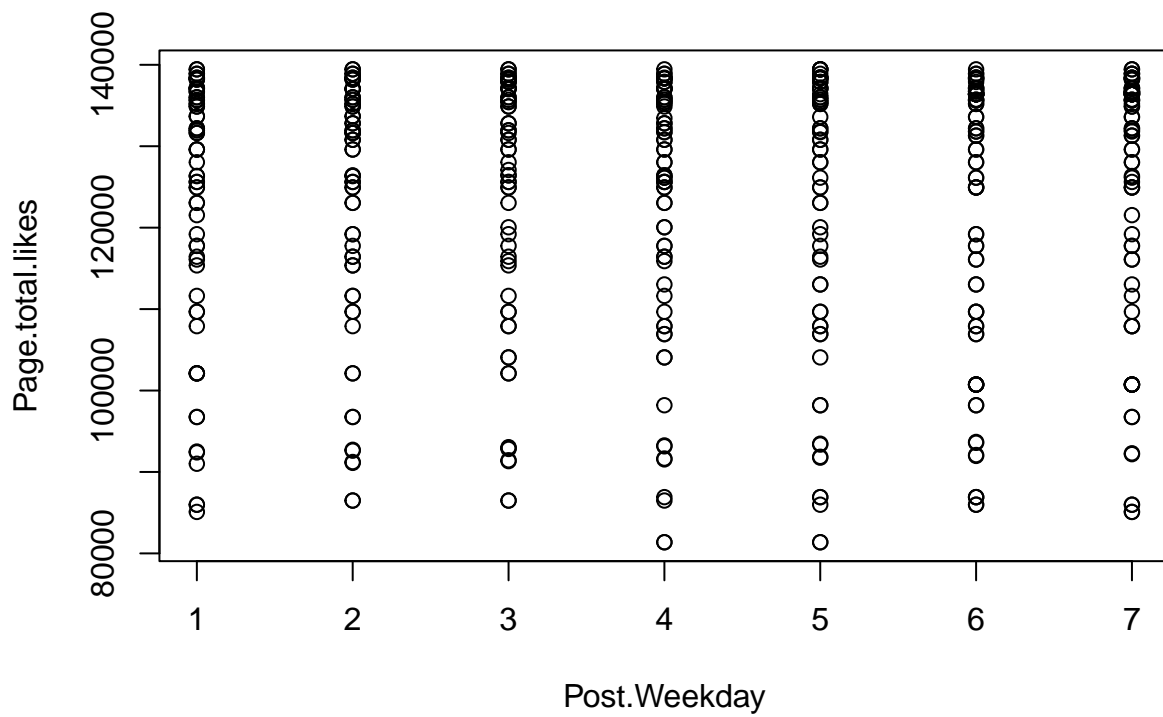
```
## Type Post.Month Post.Weekday Post.Hour Total.Interactions Lifetime.Post.Consumers
```

```
plot(Page.total.likes ~ Post.Month)
```

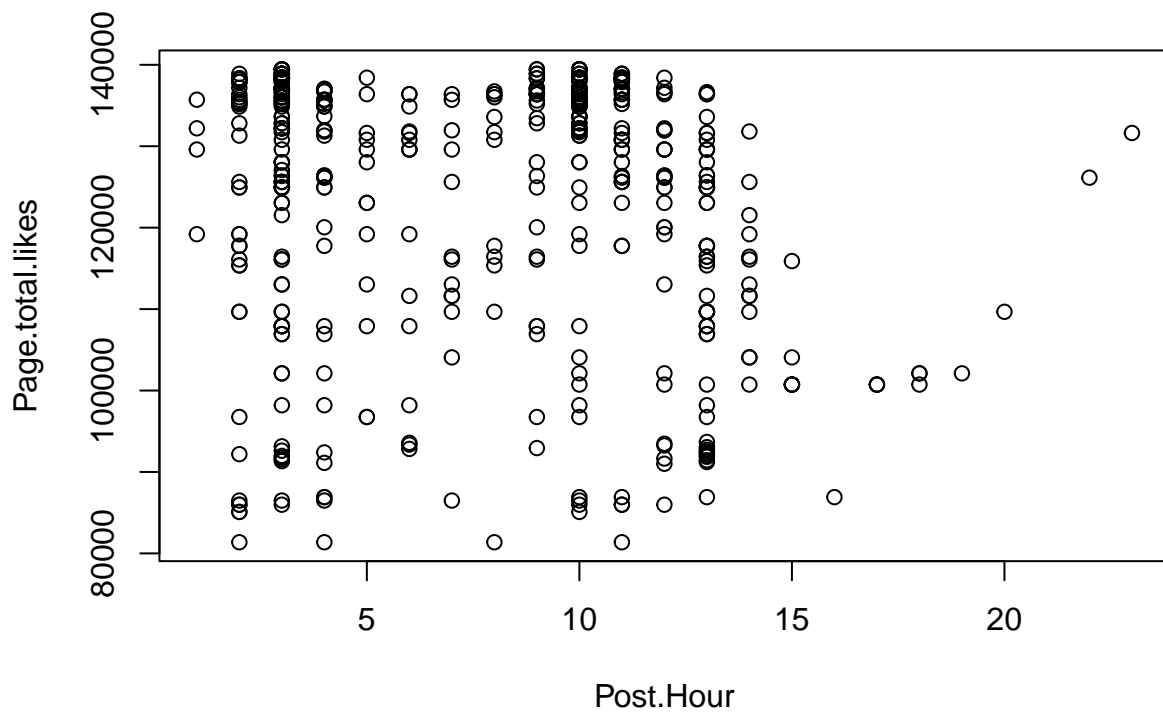


```
plot(Page.total.likes ~ Post.Weekday)
```

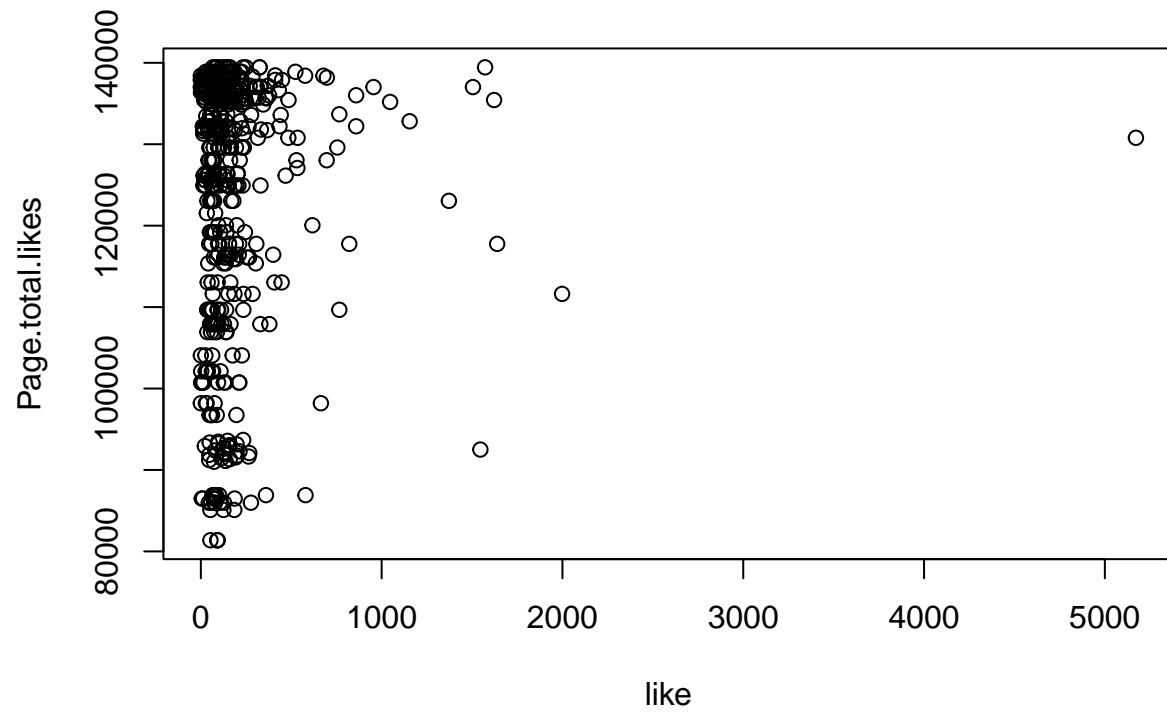




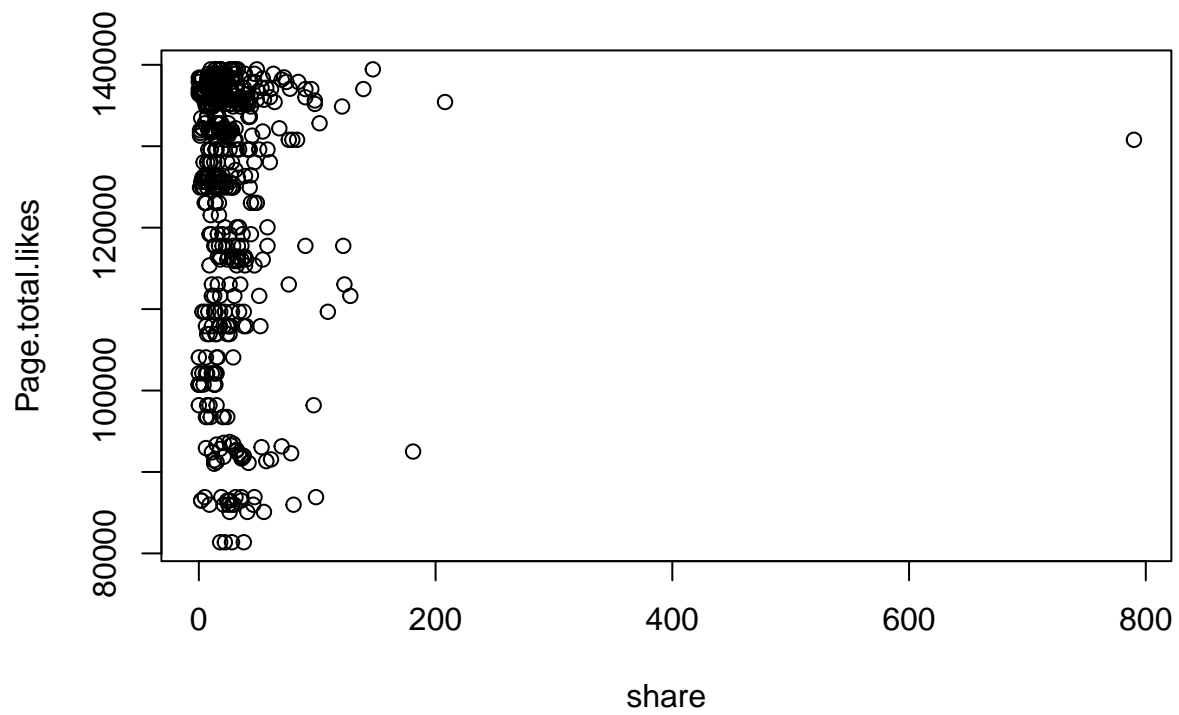
```
plot(Page.total.likes ~ Post.Hour)
```



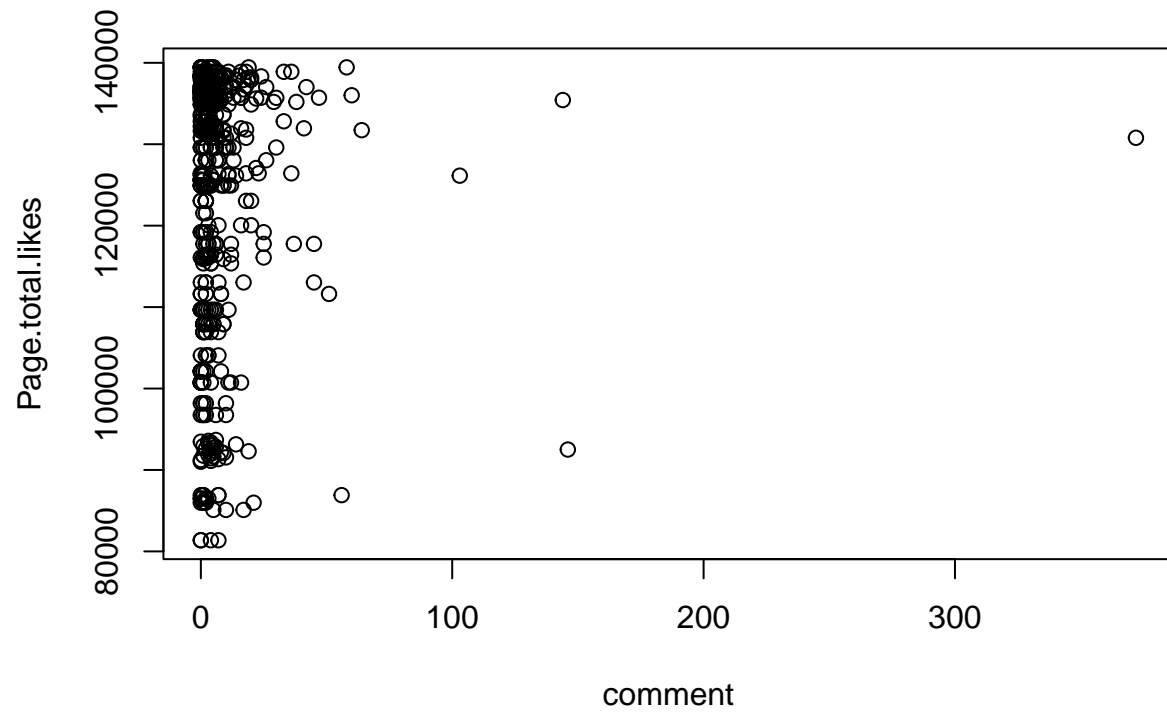
```
plot(Page.total.likes ~ like)
```



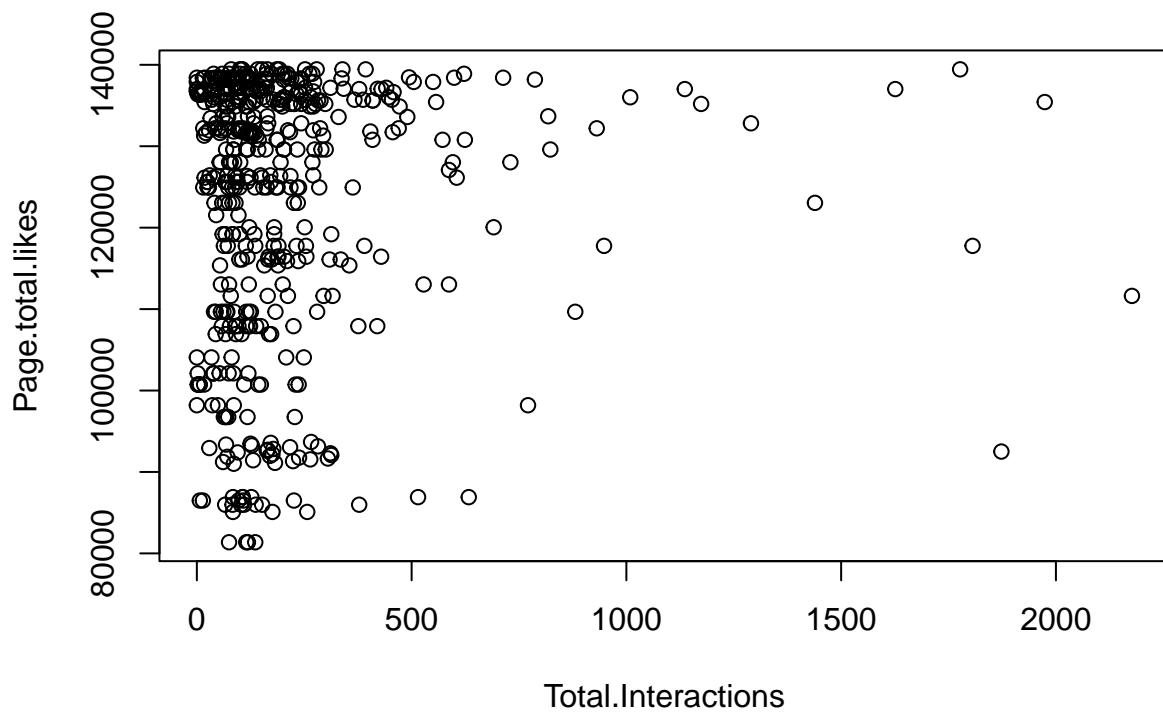
```
plot(Page.total.likes ~ share)
```



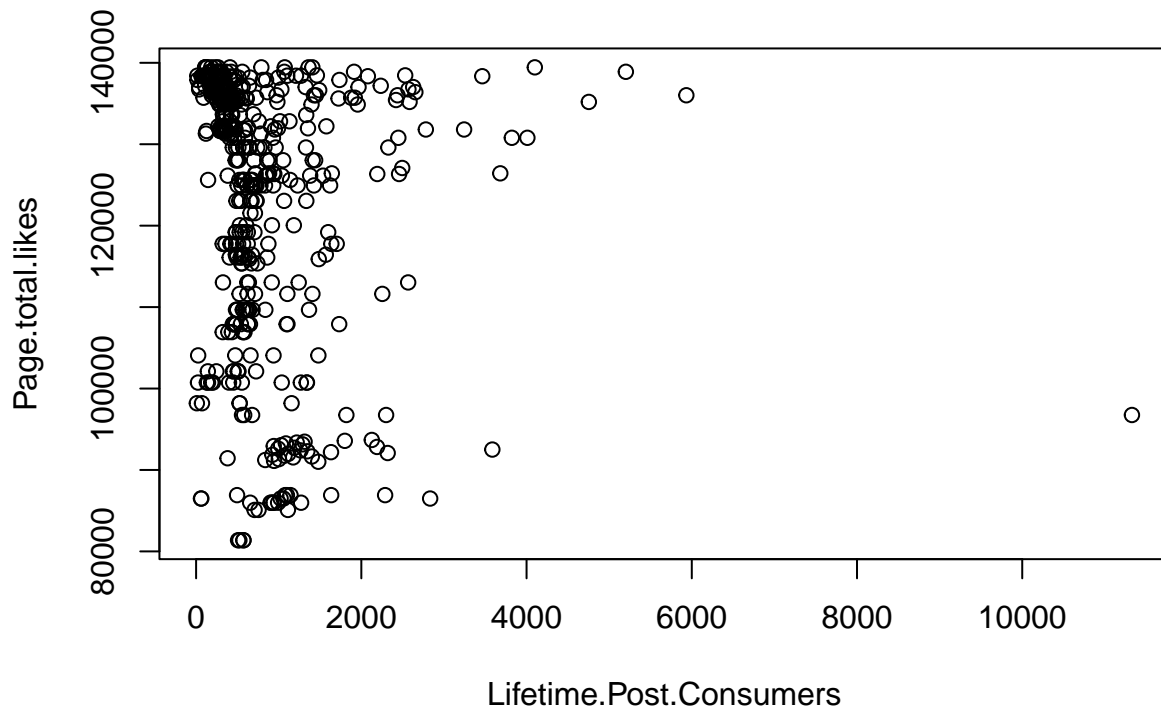
```
plot(Page.total.likes ~ comment)
```



```
plot(Page.total.likes ~ Total.Interactions, subset = Total.Interactions < 6000)
```



```
plot(Page.total.likes ~ Lifetime.Post.Consumers)
```



```
# temp_data <- fb %>% select("Page.total.likes",starts_with("Lifetime"))
# pair1<- ggpairs(temp_data,
#               lower=list(continuous=wrap("cor", alpha=0.5), combo="box"),
#               upper=list(continuous=wrap("points", alpha=0.3, size=0.1)), labeller = label_wrap_gen(5, mult
# pair1<- pair1 + theme(axis.text.x = element_text(angle = 40, hjust = 1, size = 7), axis.text.y = elem
# pair1
```

## Fit the MV model

```
## dummy variable for type
fb$type.photo <- ifelse(Type == c("Photo"),1,0)
fb$type.status <- ifelse(Type == c("Status"),1,0)
fb$type.video <- ifelse(Type == c("Video"),1,0)

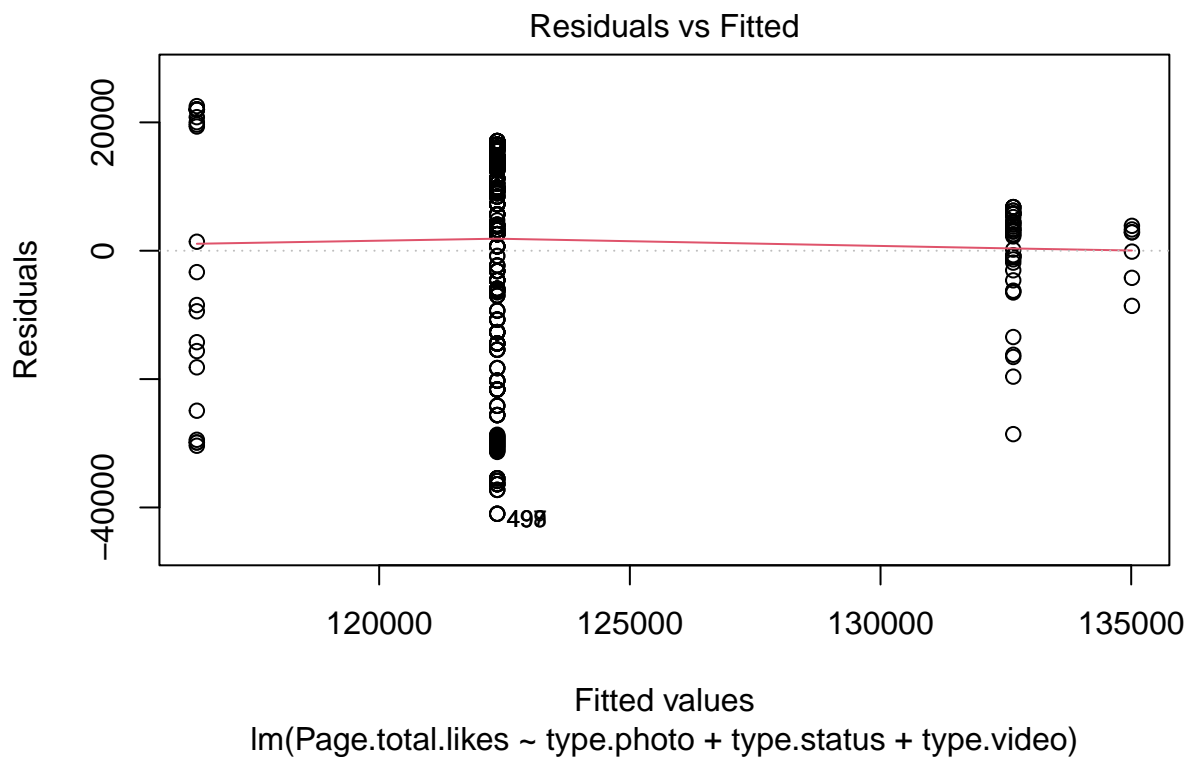
## simple linear model

m00 <- lm(Page.total.likes~type.photo+type.status+type.video, data=fb)
summary(m00)

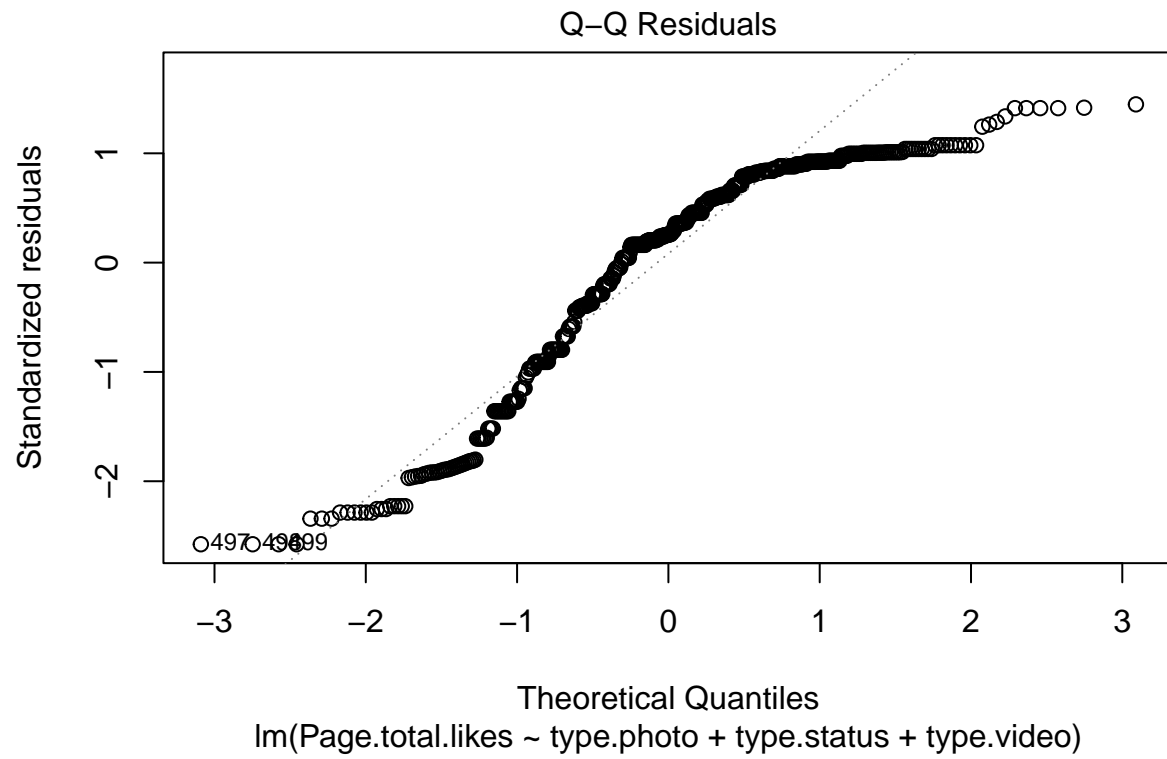
##
## Call:
## lm(formula = Page.total.likes ~ type.photo + type.status + type.video,
##     data = fb)
```

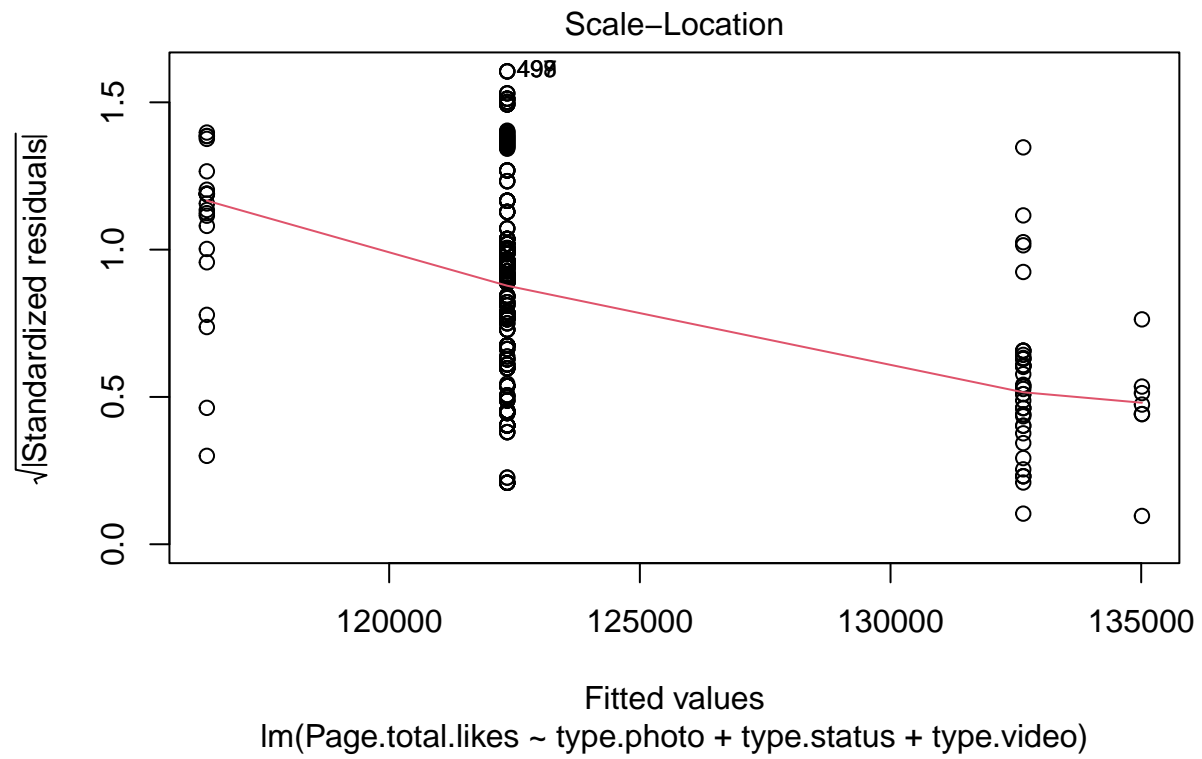
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -40984 -10734   4070  13359  22532
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   116363     3396   34.270 < 2e-16 ***
## type.photo      5991      3482    1.721  0.0860 .
## type.status    16284     4143    3.930 9.69e-05 ***
## type.video     18652     6911    2.699  0.0072 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15930 on 496 degrees of freedom
## Multiple R-squared:  0.04788,    Adjusted R-squared:  0.04212
## F-statistic: 8.314 on 3 and 496 DF,  p-value: 2.102e-05
```

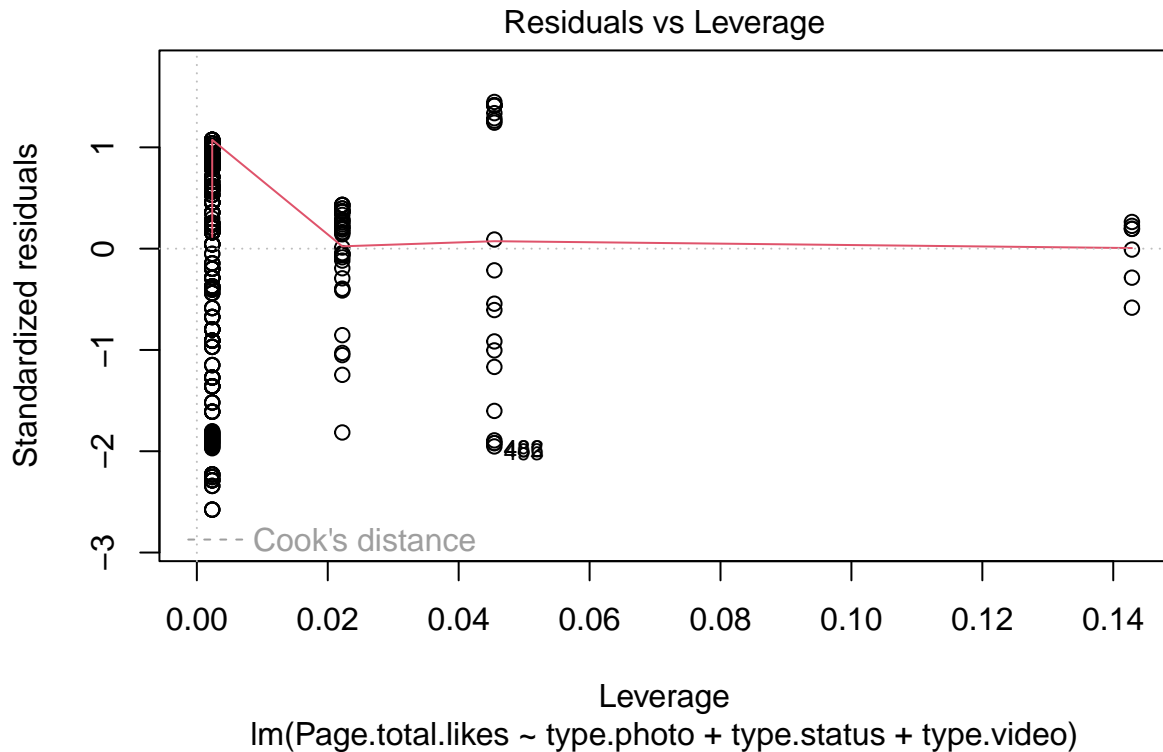
```
plot(m00)
```







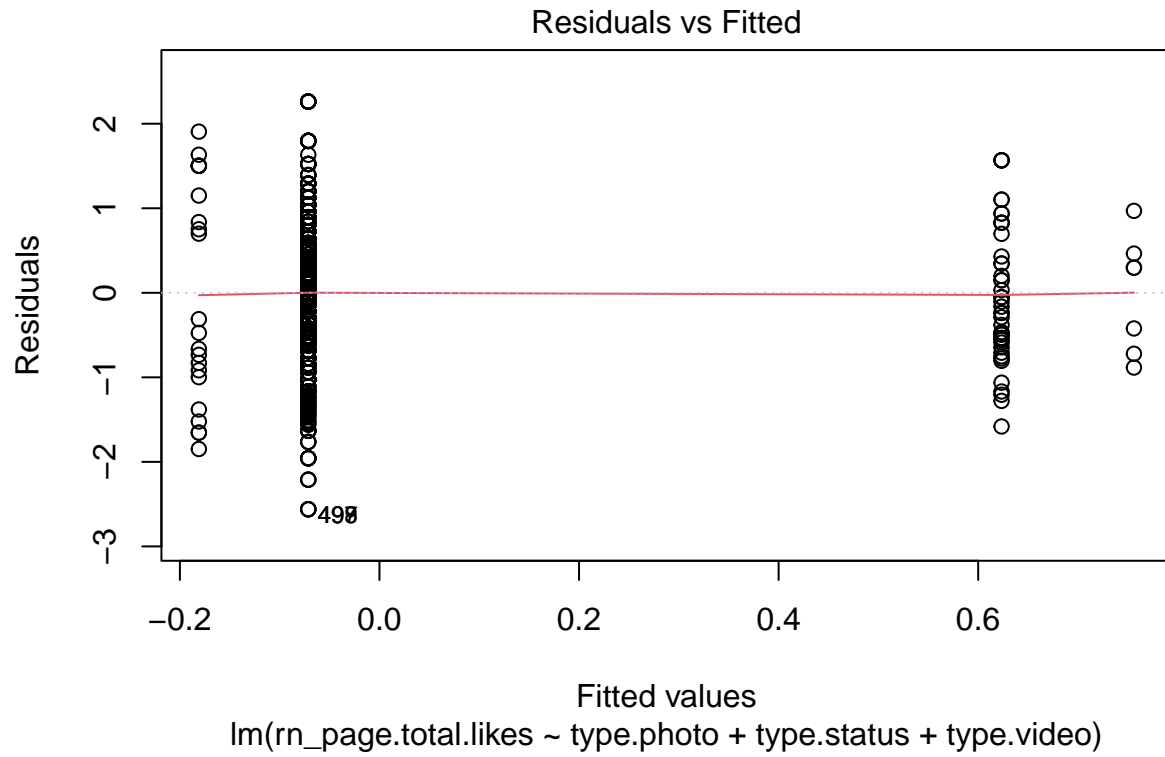


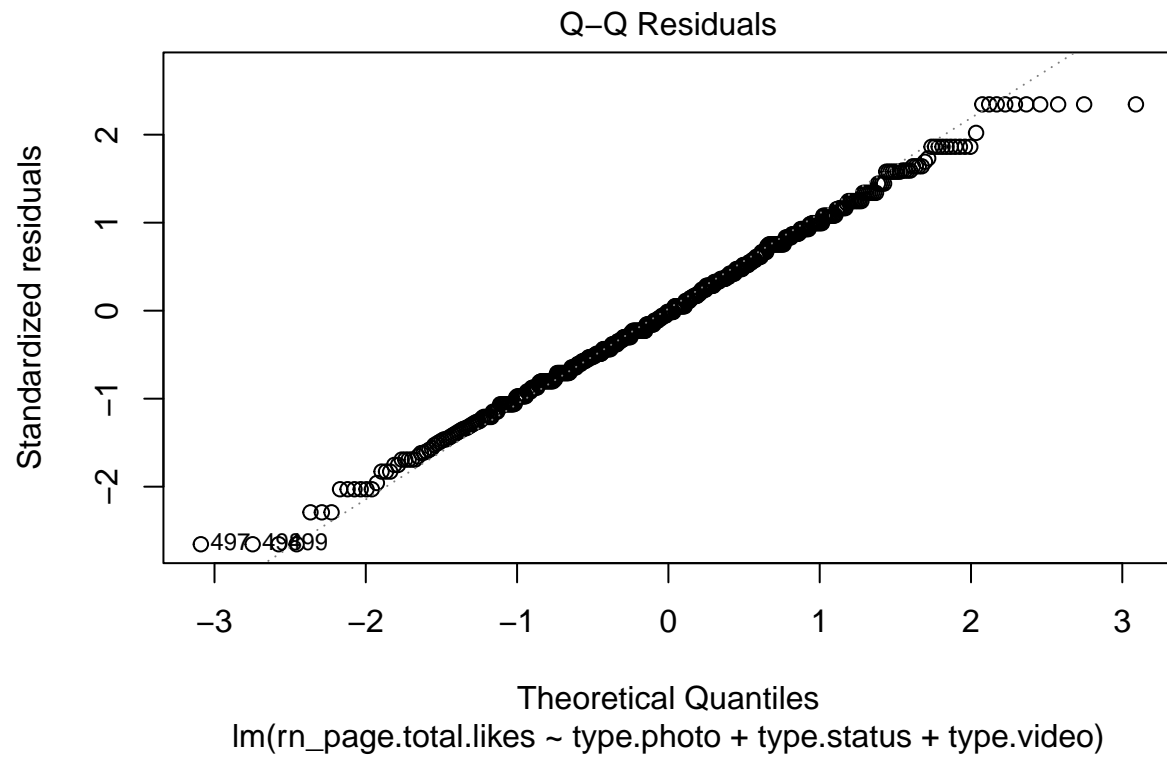


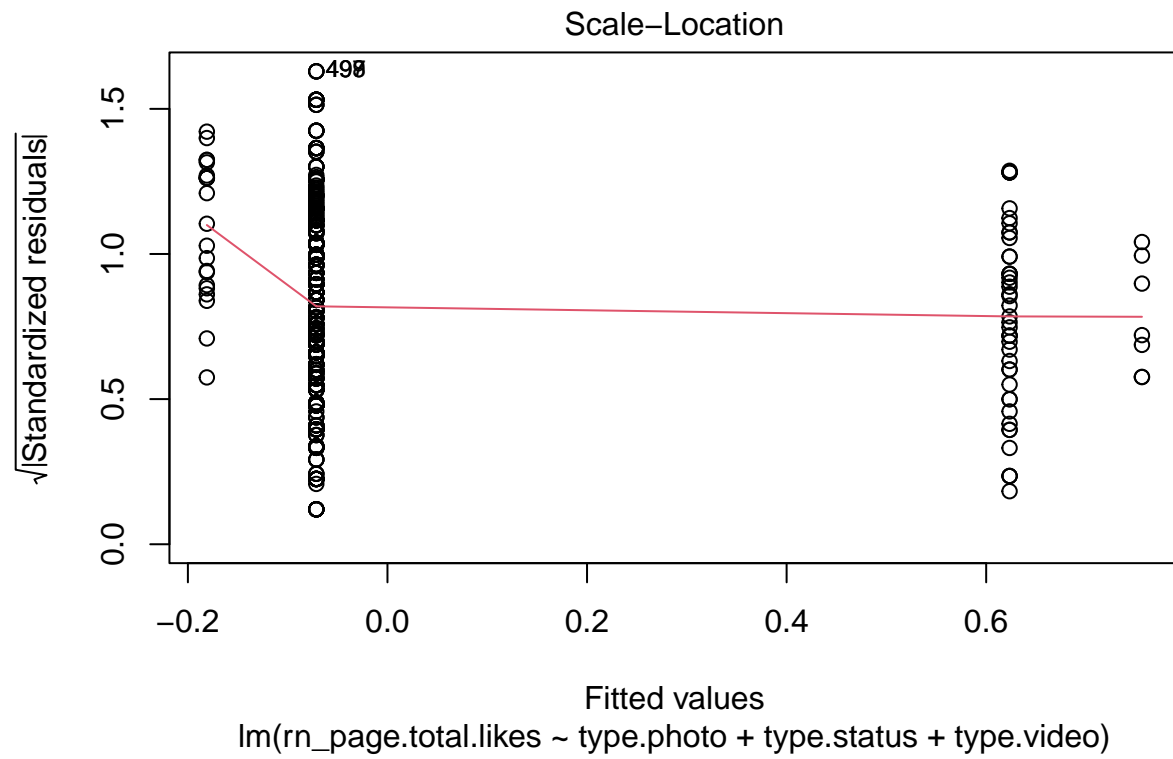
```
## simple linear model - rank normalized
m0 <- lm(rn_page.total.likes~type.photo+type.status+type.video, data=fb)
summary(m0)
```

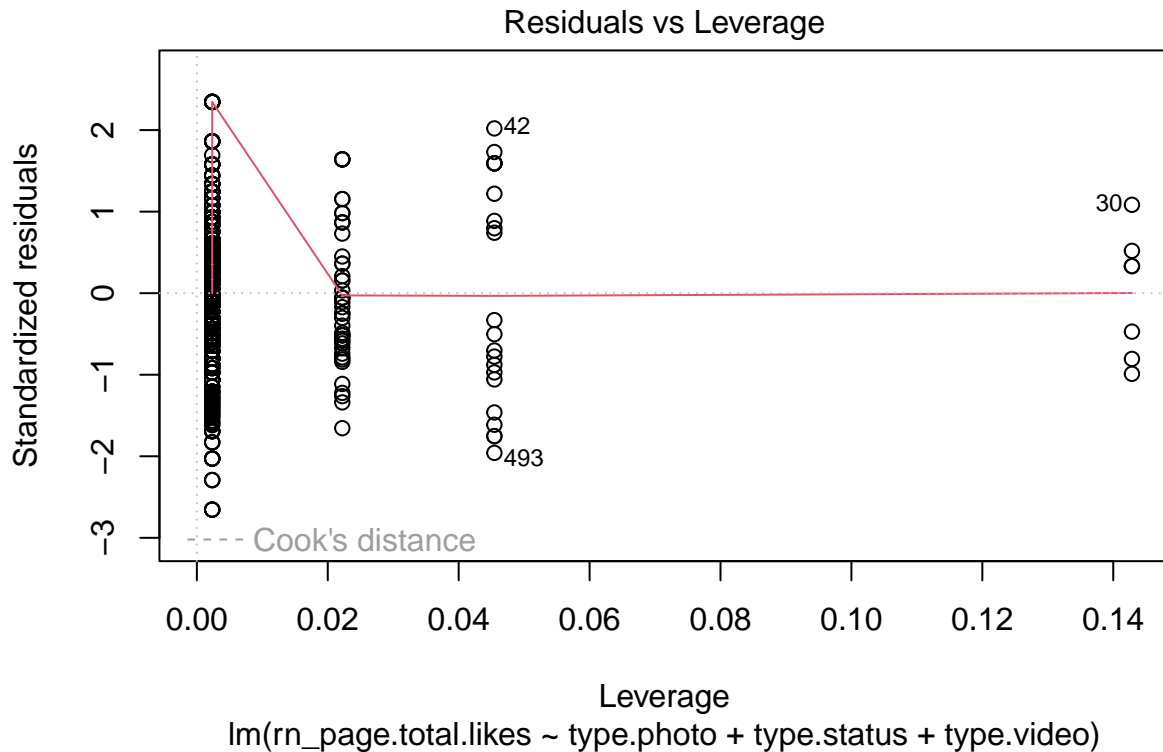
```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video, data = fb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56038 -0.68363 -0.01396  0.72667  2.26186
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1810     0.2059  -0.879  0.37977
## type.photo     0.1097     0.2111   0.519  0.60368
## type.status    0.8045     0.2512   3.202  0.00145 **
## type.video     0.9371     0.4191   2.236  0.02579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9658 on 496 degrees of freedom
## Multiple R-squared:  0.05006,    Adjusted R-squared:  0.04432
## F-statistic: 8.713 on 3 and 496 DF,  p-value: 1.214e-05
```

```
plot(m0)
```









```
## first model - initial try
```

```
m1 <- lm(Page.total.likes~type.photo+type.status+type.video+Post.Month+Post.Weekday+Post.Hour+Total.Int
summary(m1)
```

```
##
```

```
## Call:
```

```
## lm(formula = Page.total.likes ~ type.photo + type.status + type.video +
##      Post.Month + Post.Weekday + Post.Hour + Total.Interactions +
##      Lifetime.Post.Consumers, data = fb)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -14529.3 -4098.5    80.7   4538.7   9056.2
```

```
##
```

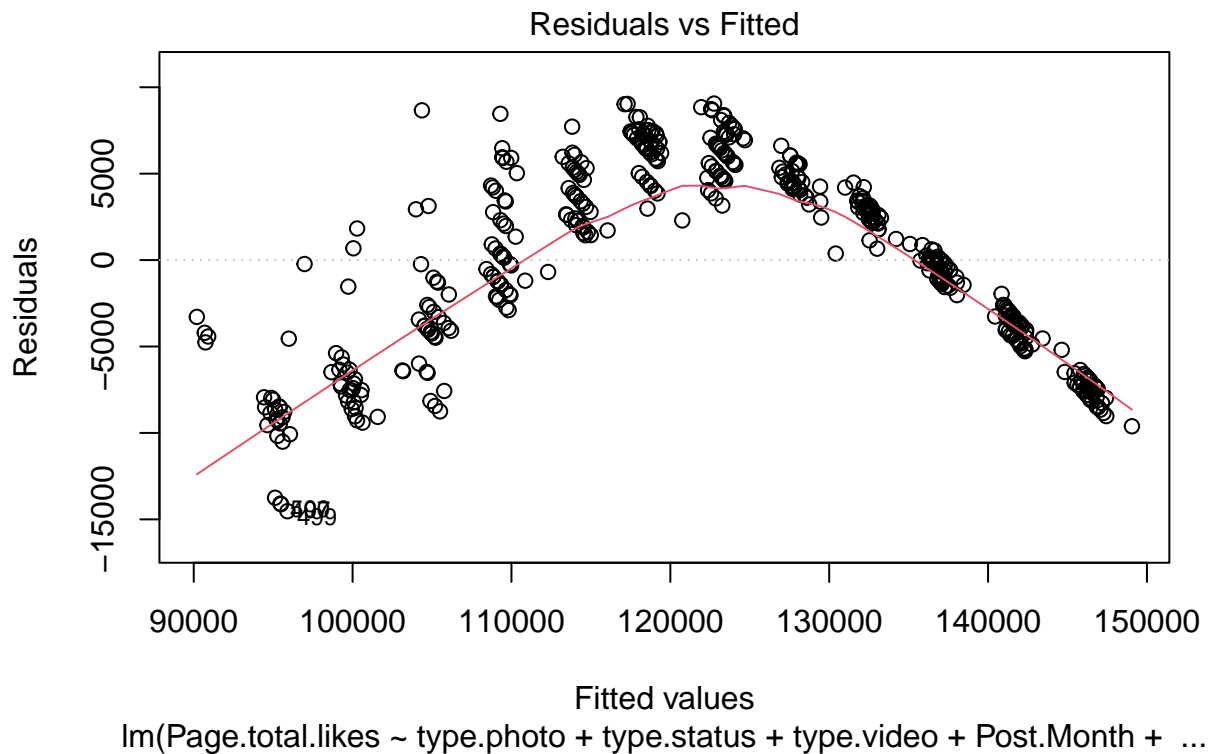
```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    86547.0321   1428.6112   60.581 < 2e-16 ***
## type.photo      4933.9233   1193.8861    4.133 4.22e-05 ***
## type.status     5933.8613   1542.7534    3.846 0.000136 ***
## type.video      5323.9892   2408.4439    2.211 0.027527 *
## Post.Month      4586.8315    79.5925   57.629 < 2e-16 ***
## Post.Weekday    -164.8681   119.7304   -1.377 0.169142
## Post.Hour        64.6117    56.6151    1.141 0.254324
## Total.Interactions  1.5845    0.6965    2.275 0.023330 *
## Lifetime.Post.Consumers -0.7583  0.3454   -2.195 0.028615 *
```

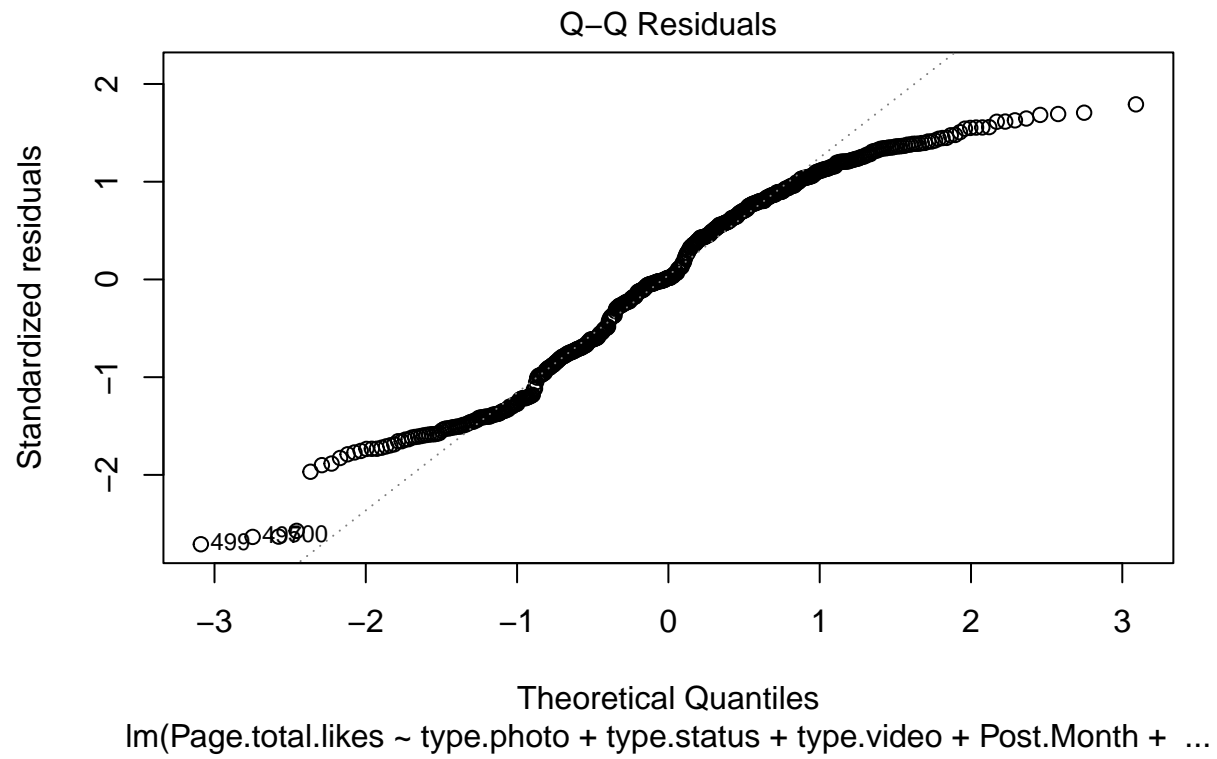
```
## ---
```

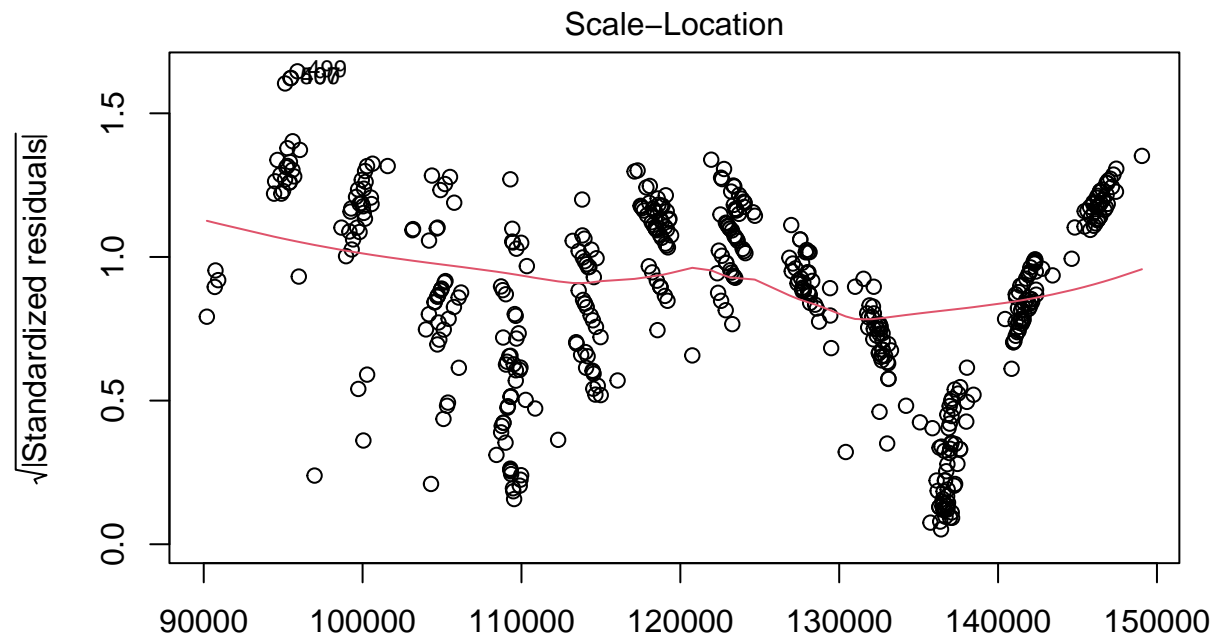
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 5389 on 491 degrees of freedom  
## Multiple R-squared:  0.8921, Adjusted R-squared:  0.8903  
## F-statistic: 507.3 on 8 and 491 DF,  p-value: < 2.2e-16
```

```
plot(m1)
```

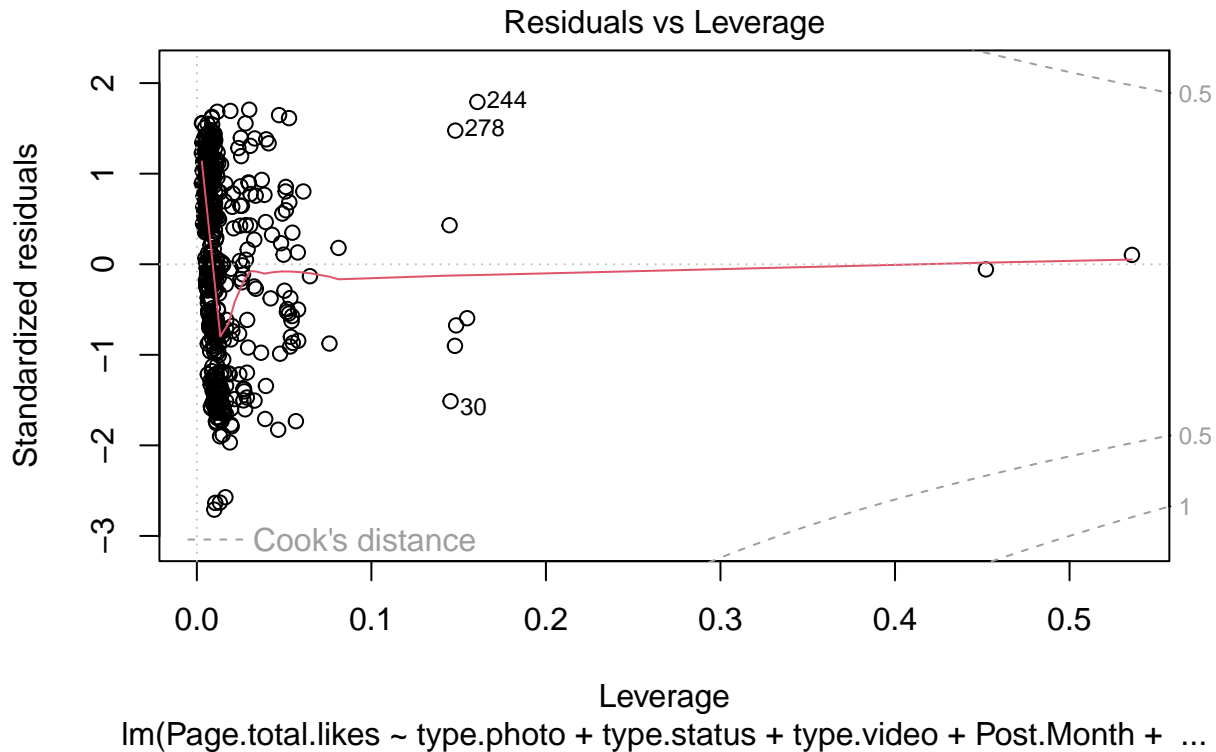








$\text{lm}(\text{Page.total.likes} \sim \text{type.photo} + \text{type.status} + \text{type.video} + \text{Post.Month} + \dots)$



```
## model diagnostics are way off... will need to: 1) transform the outcome
##                                                  2) add in a squared term
##                                                  3) check collinearity
##                                                  4) create one time variable instead of 3
```

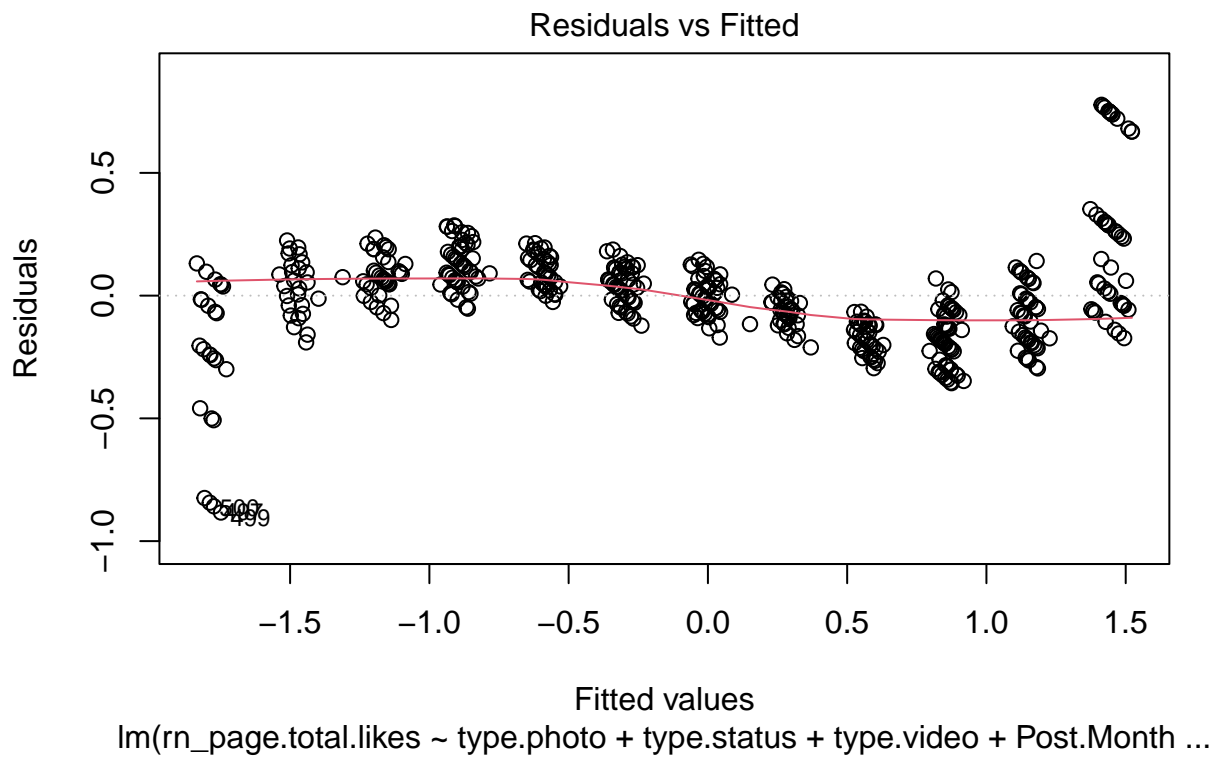
```
## M2 Rank normalize the outcome
```

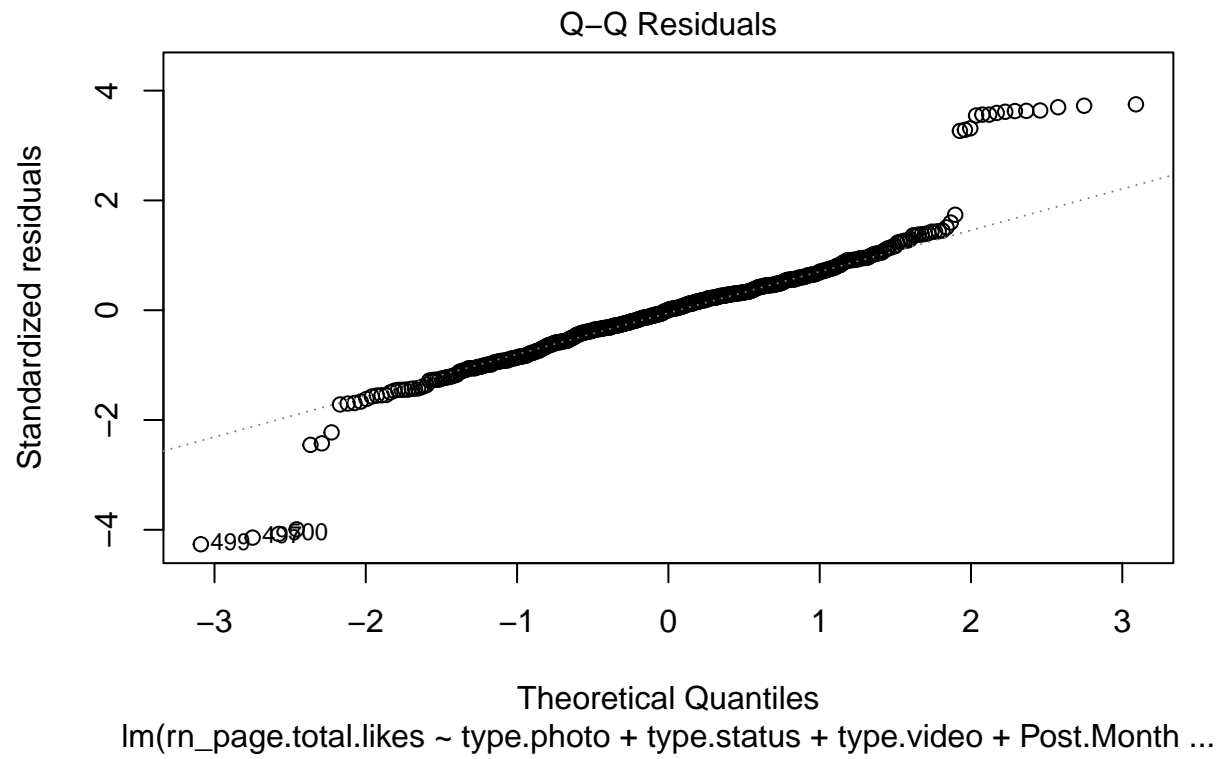
```
m2 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Post.Month+Post.Weekday+Post.Hour+Total.
summary(m2)
```

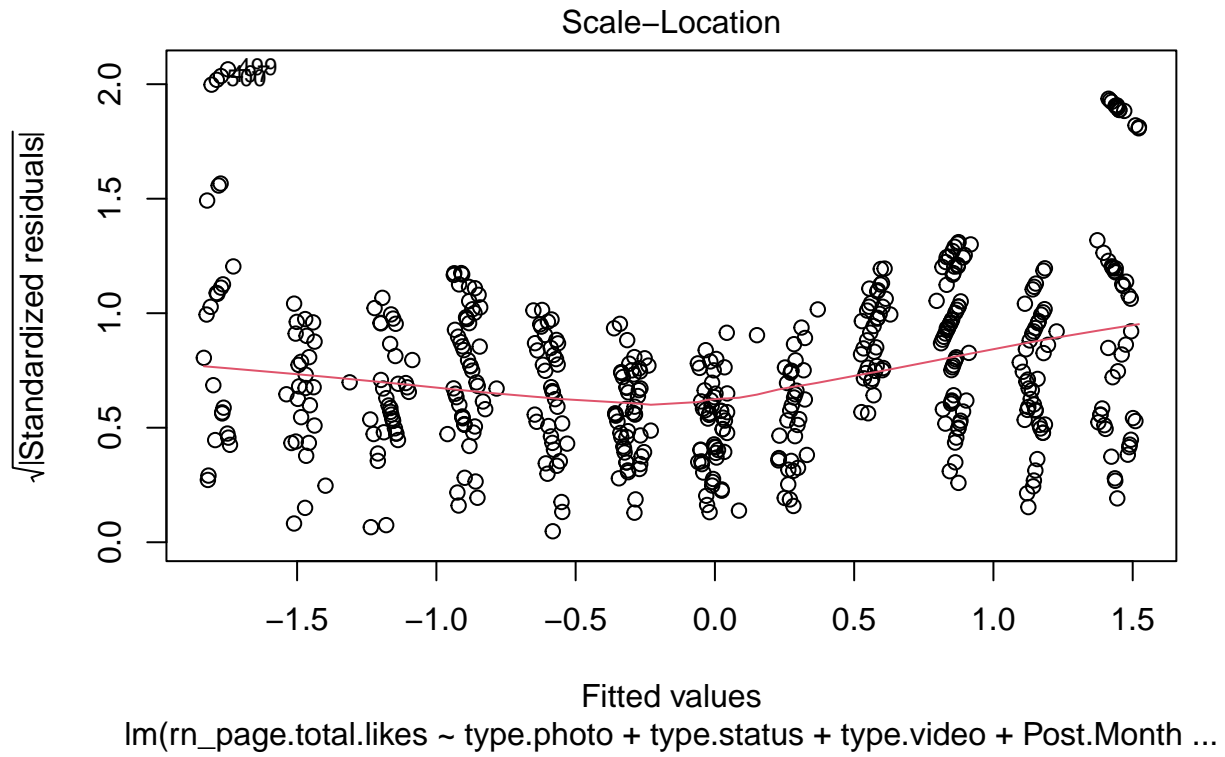
```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + Post.Month + Post.Weekday + Post.Hour + Total.Interactions +
##     Lifetime.Post.Consumers, data = fb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88342 -0.11577  0.00034  0.09434  0.77723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.108e+00  5.521e-02 -38.181  < 2e-16 ***
## type.photo      3.303e-02  4.614e-02   0.716  0.47443
## type.status     8.938e-02  5.962e-02   1.499  0.13448
## type.video      5.294e-02  9.307e-02   0.569  0.56971
```

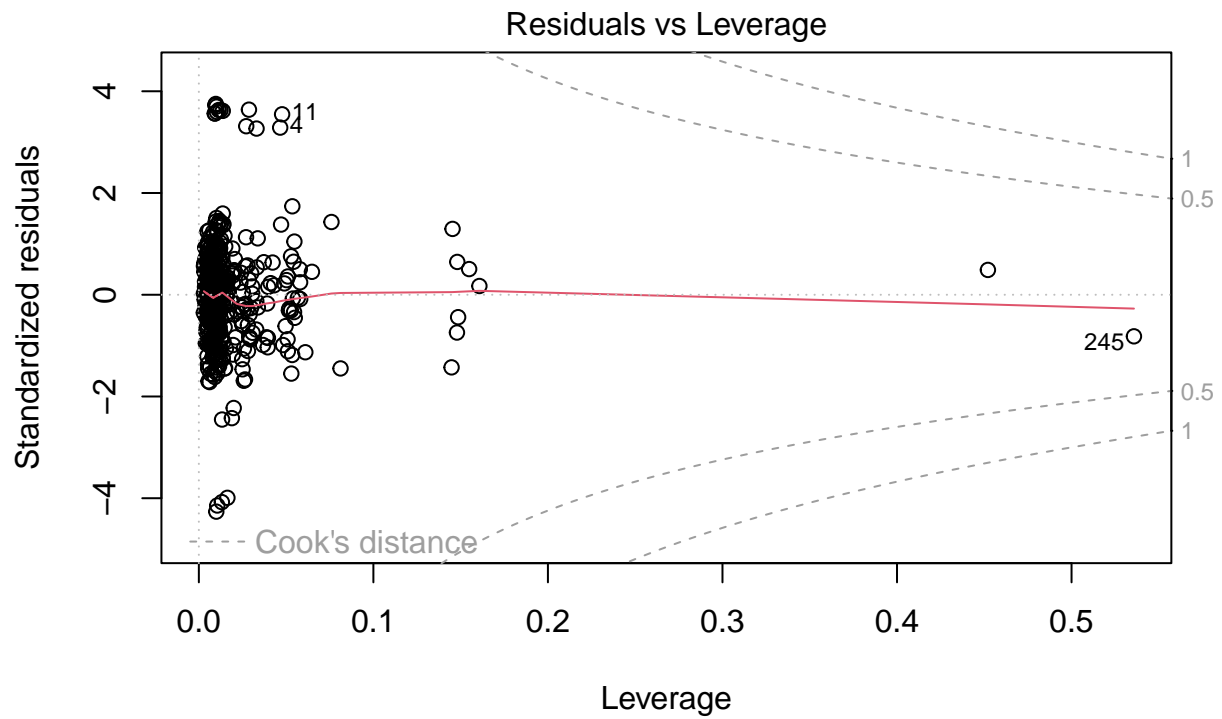
```
## Post.Month          2.917e-01  3.076e-03  94.839 < 2e-16 ***
## Post.Weekday        -6.655e-03  4.627e-03  -1.438  0.15098
## Post.Hour           5.730e-03  2.188e-03   2.619  0.00909 **
## Total.Interactions   3.462e-05  2.691e-05   1.286  0.19888
## Lifetime.Post.Consumers -1.098e-05  1.335e-05  -0.822  0.41133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2083 on 491 degrees of freedom
## Multiple R-squared:  0.9563, Adjusted R-squared:  0.9556
## F-statistic: 1342 on 8 and 491 DF, p-value: < 2.2e-16
```

```
plot(m2)
```



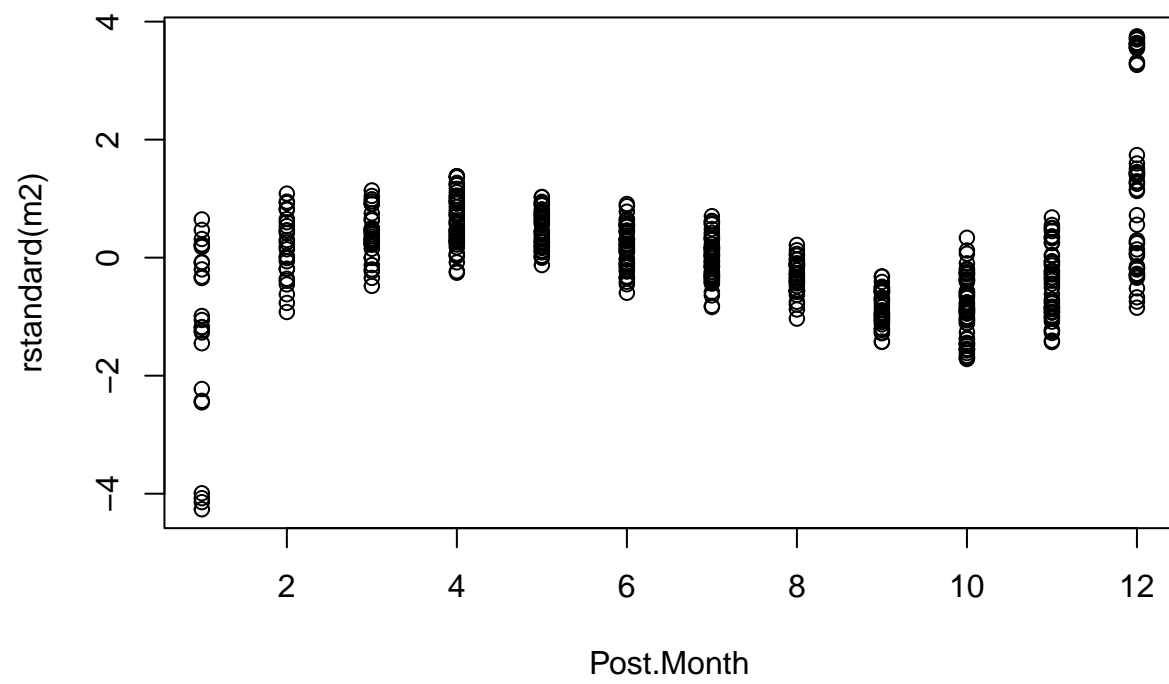






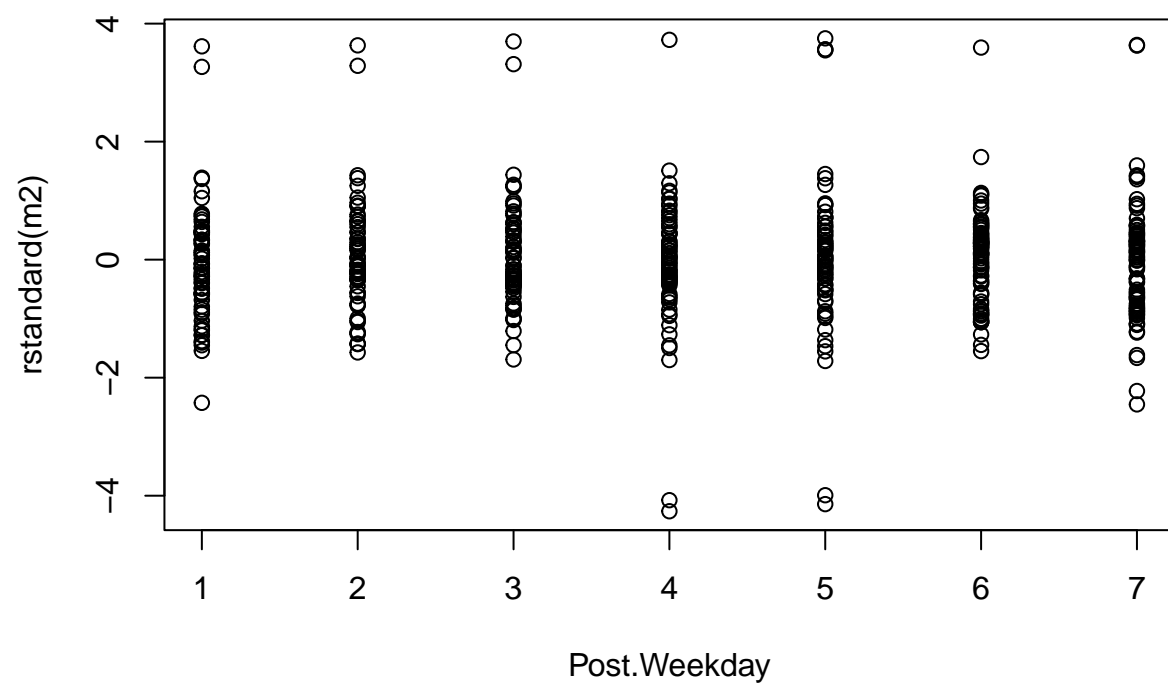
lm(rn\_page.total.likes ~ type.photo + type.status + type.video + Post.Month ...)

```
plot(rstandard(m2) ~ Post.Month)
```

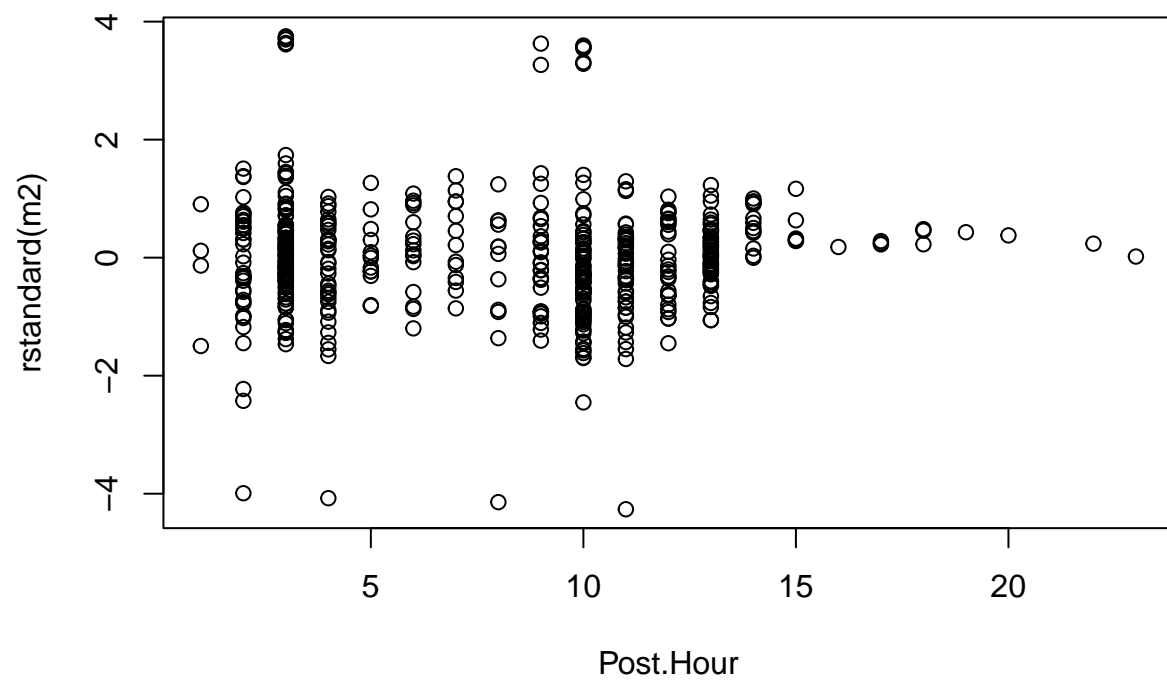


```
plot(rstandard(m2)~Post.Weekday)
```

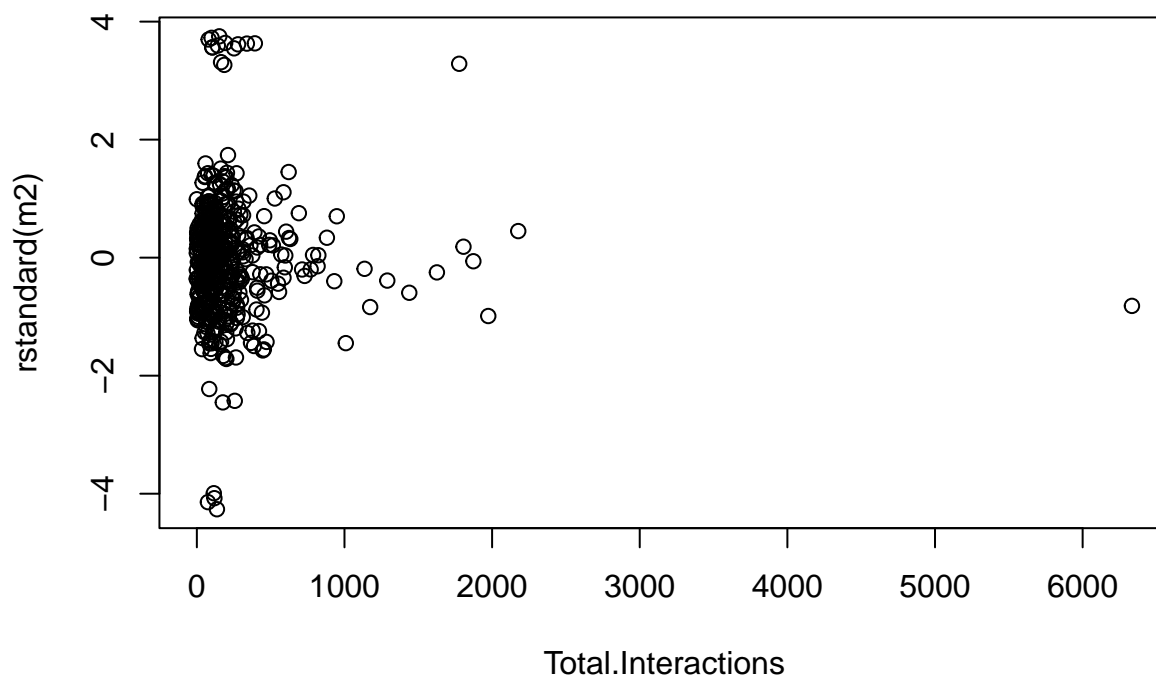




```
plot(rstandard(m2)~Post.Hour)
```



```
plot(rstandard(m2)~Total.Interactions)
```

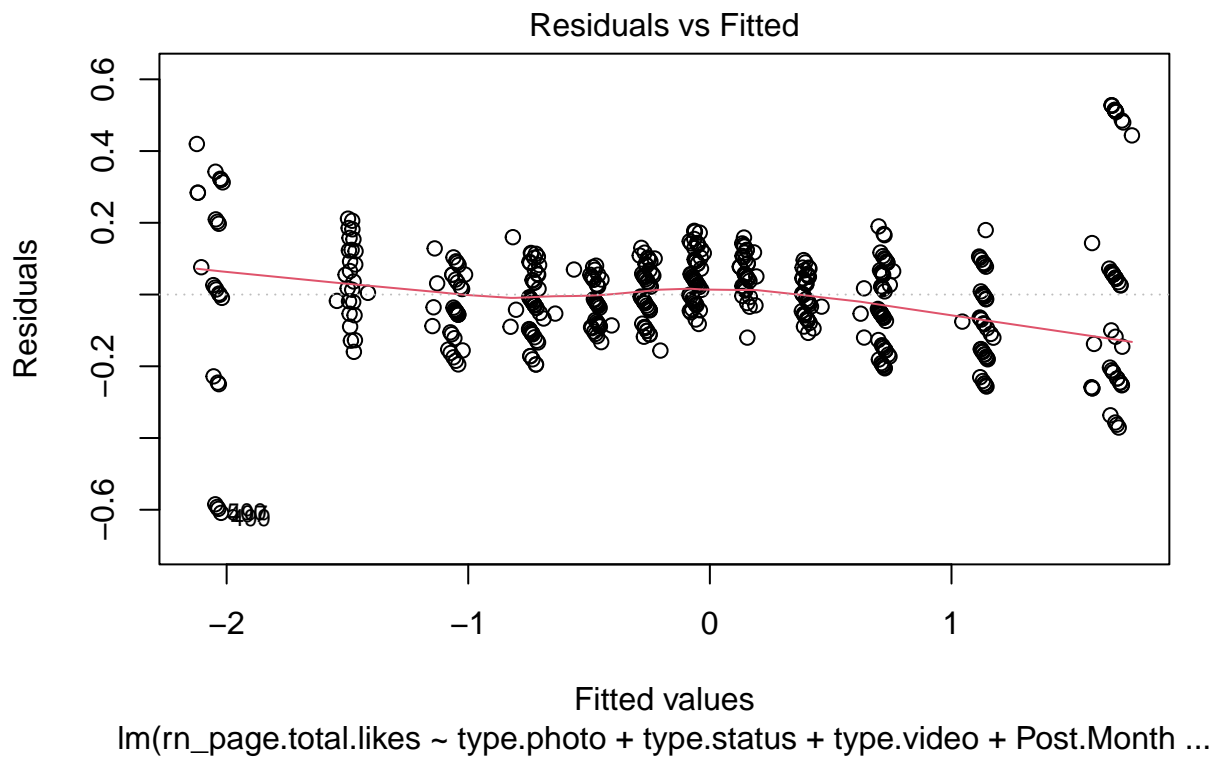


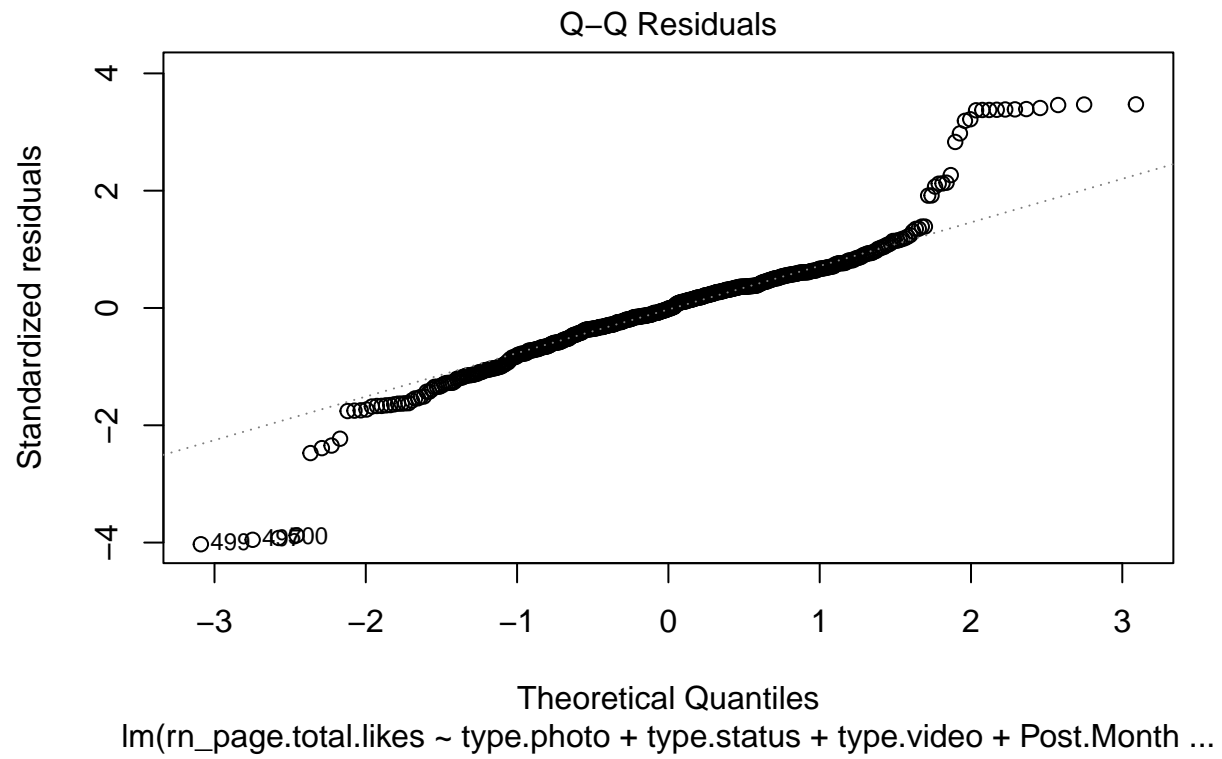
```
## better but still not good - quadratic trend with post.month- try adding a cubic term and a squared t
m3 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Post.Month+I(Post.Month^2)+I(Post.Month^3) +
summary(m3)
```

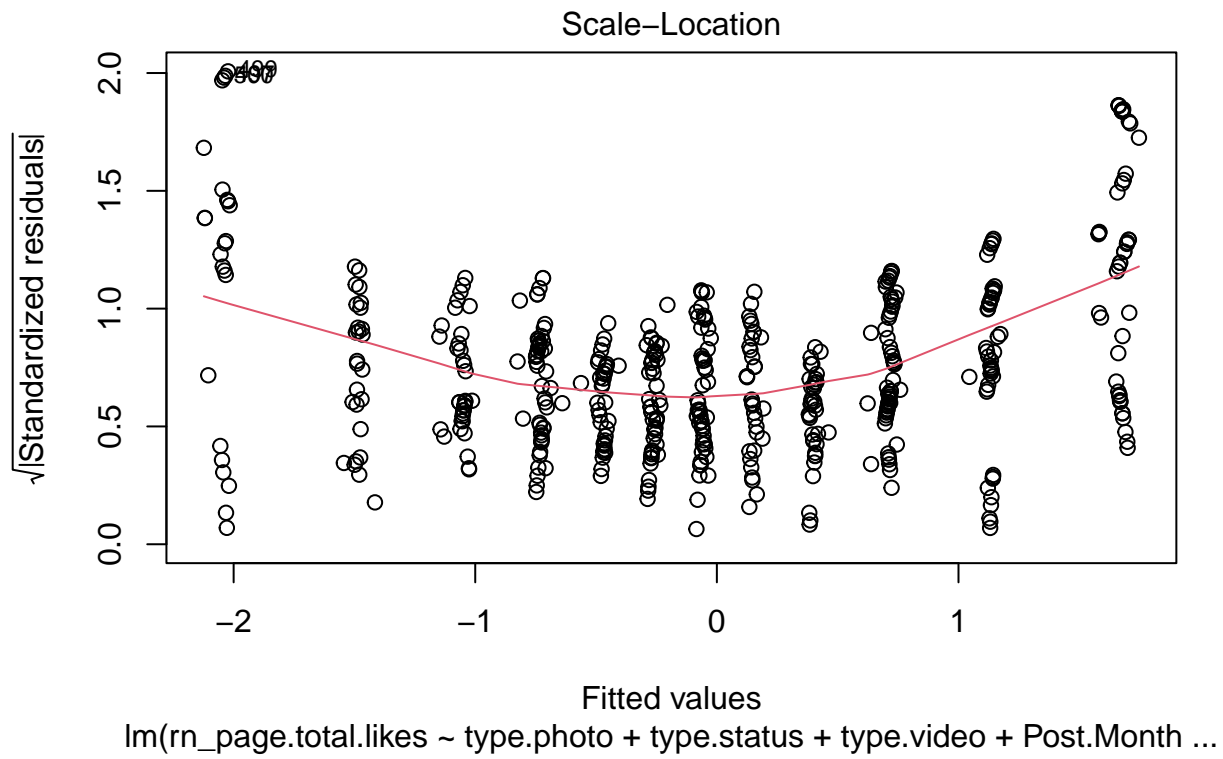
```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + Post.Month + I(Post.Month^2) + I(Post.Month^3) +
##     Post.Weekday + Post.Hour + Total.Interactions + Lifetime.Post.Consumers,
##     data = fb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60865 -0.08043 -0.00148  0.07222  0.52785
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.823e+00  5.851e-02 -48.255  <2e-16 ***
## type.photo       7.899e-02  3.468e-02   2.277   0.0232 *
## type.status      1.105e-01  4.451e-02   2.483   0.0134 *
## type.video       1.011e-01  6.891e-02   1.467   0.1430
## Post.Month       7.906e-01  2.750e-02  28.748  <2e-16 ***
## I(Post.Month^2)  -9.099e-02  4.624e-03 -19.677  <2e-16 ***
## I(Post.Month^3)   4.642e-03  2.294e-04  20.233  <2e-16 ***
## Post.Weekday    -2.288e-03  3.422e-03  -0.669   0.5041
## Post.Hour        2.303e-03  1.624e-03   1.418   0.1569
```

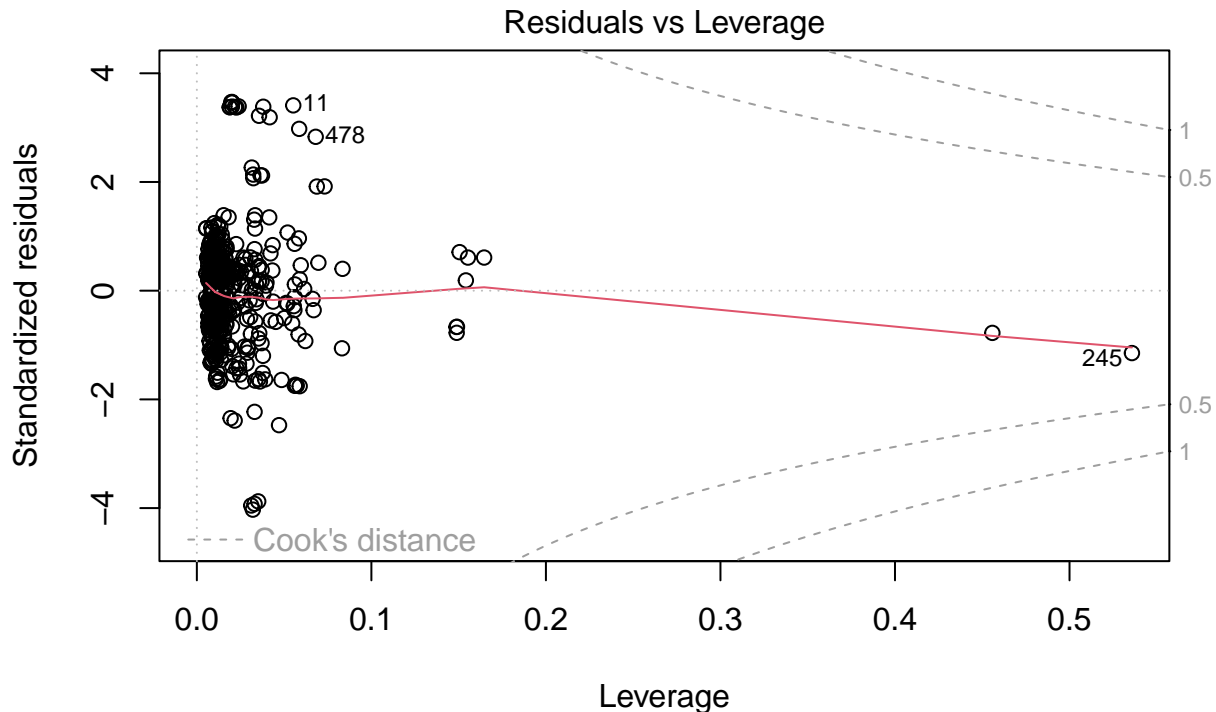
```
## Total.Interactions      4.104e-05  1.990e-05   2.062   0.0397 *
## Lifetime.Post.Consumers -8.286e-06  9.868e-06  -0.840   0.4015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1536 on 489 degrees of freedom
## Multiple R-squared:  0.9763, Adjusted R-squared:  0.9758
## F-statistic: 2016 on 10 and 489 DF, p-value: < 2.2e-16
```

```
plot(m3)
```









lm(rn\_page.total.likes ~ type.photo + type.status + type.video + Post.Month ...

```
## not good - remove non-sig variables
```

```
m4 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Post.Month+I(Post.Month^2)+I(Post.Month^3)+
summary(m4)
```

```
##
```

```
## Call:
```

```
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + Post.Month + I(Post.Month^2) + I(Post.Month^3) +
##     Total.Interactions + Lifetime.Post.Consumers, data = fb)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.60059 -0.07755 -0.00001  0.06588  0.51893
```

```
##
```

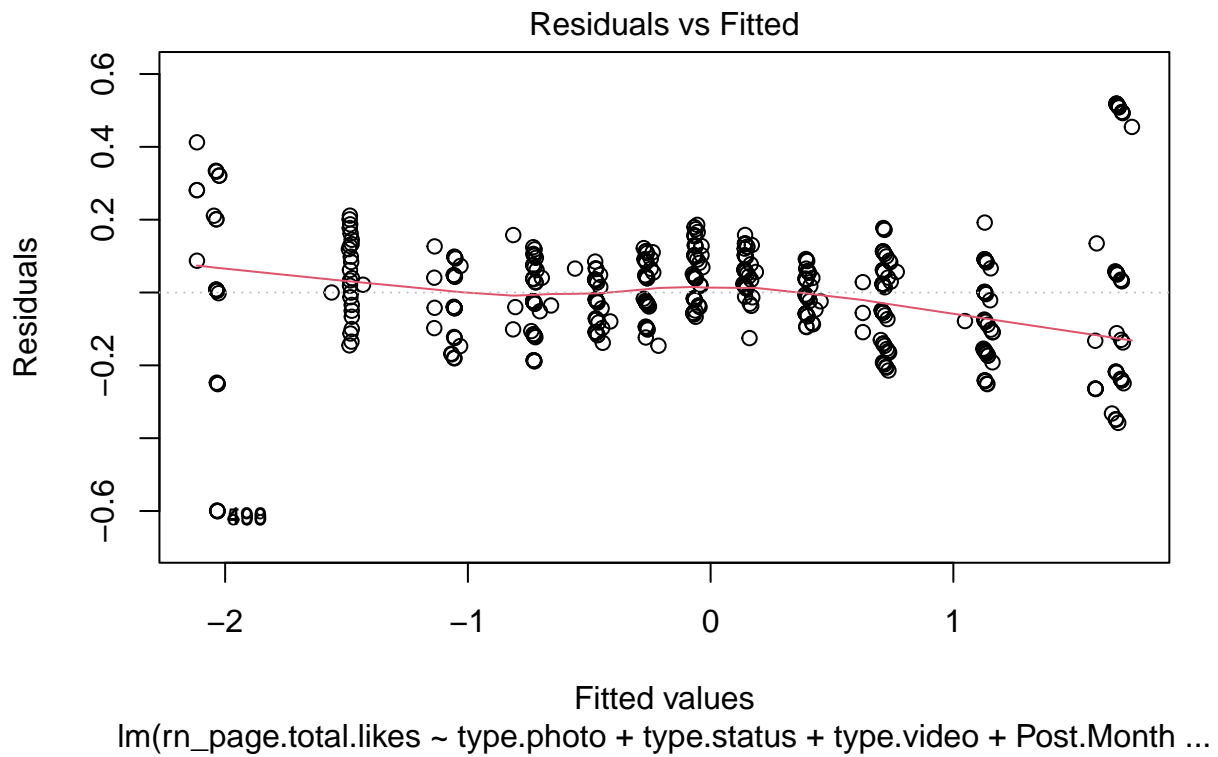
```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.825e+00  5.542e-02 -50.975  <2e-16 ***
## type.photo       8.441e-02  3.449e-02   2.447   0.0147 *
## type.status      1.143e-01  4.435e-02   2.577   0.0103 *
## type.video       1.110e-01  6.862e-02   1.618   0.1064
## Post.Month       7.957e-01  2.730e-02  29.143  <2e-16 ***
## I(Post.Month^2)  -9.191e-02  4.586e-03 -20.042  <2e-16 ***
## I(Post.Month^3)   4.685e-03  2.277e-04  20.576  <2e-16 ***
## Total.Interactions  4.112e-05  1.987e-05   2.070   0.0390 *
## Lifetime.Post.Consumers -8.275e-06  9.866e-06  -0.839   0.4020
```

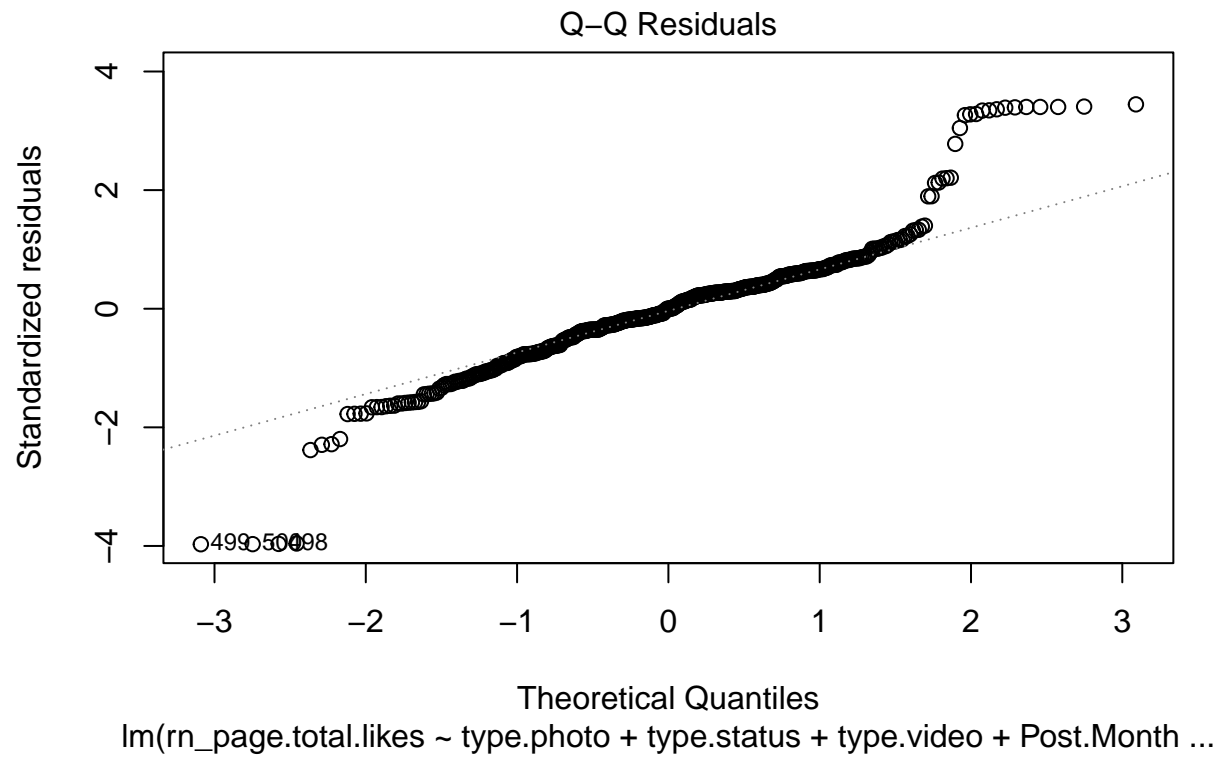
```
## ---
```

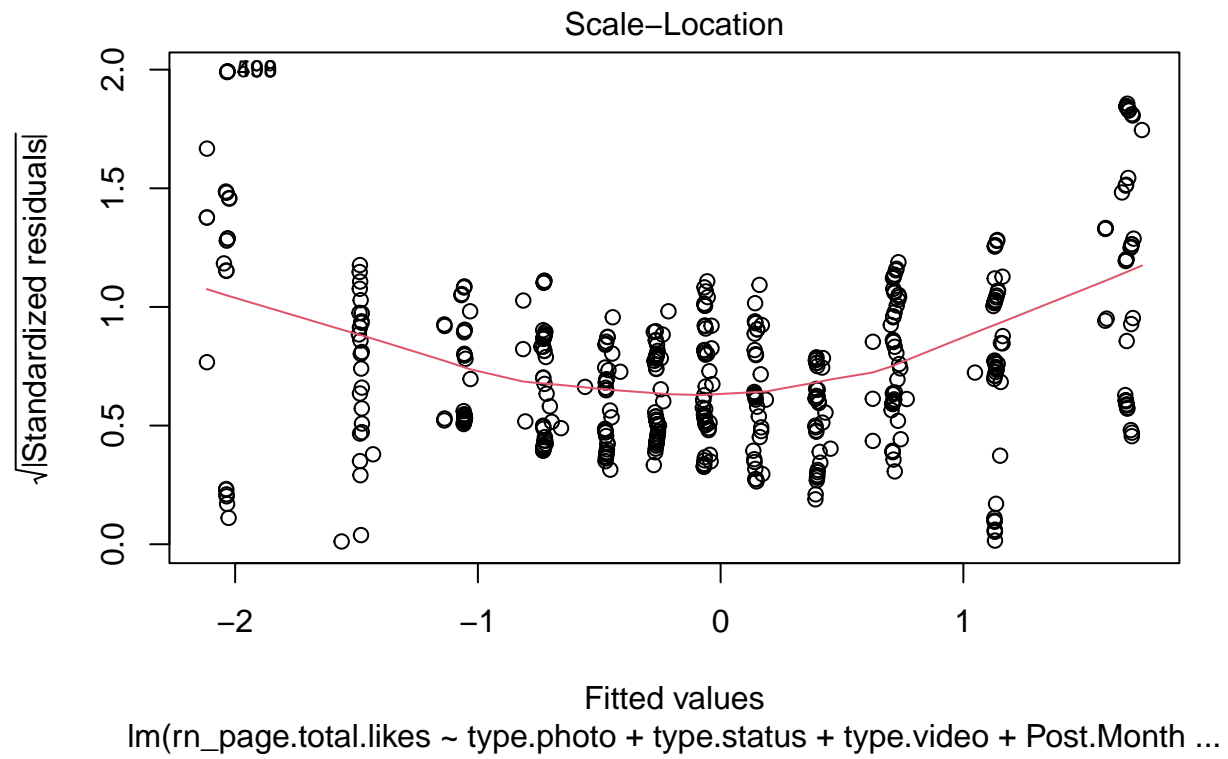
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1536 on 491 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9758
## F-statistic: 2518 on 8 and 491 DF,  p-value: < 2.2e-16
```

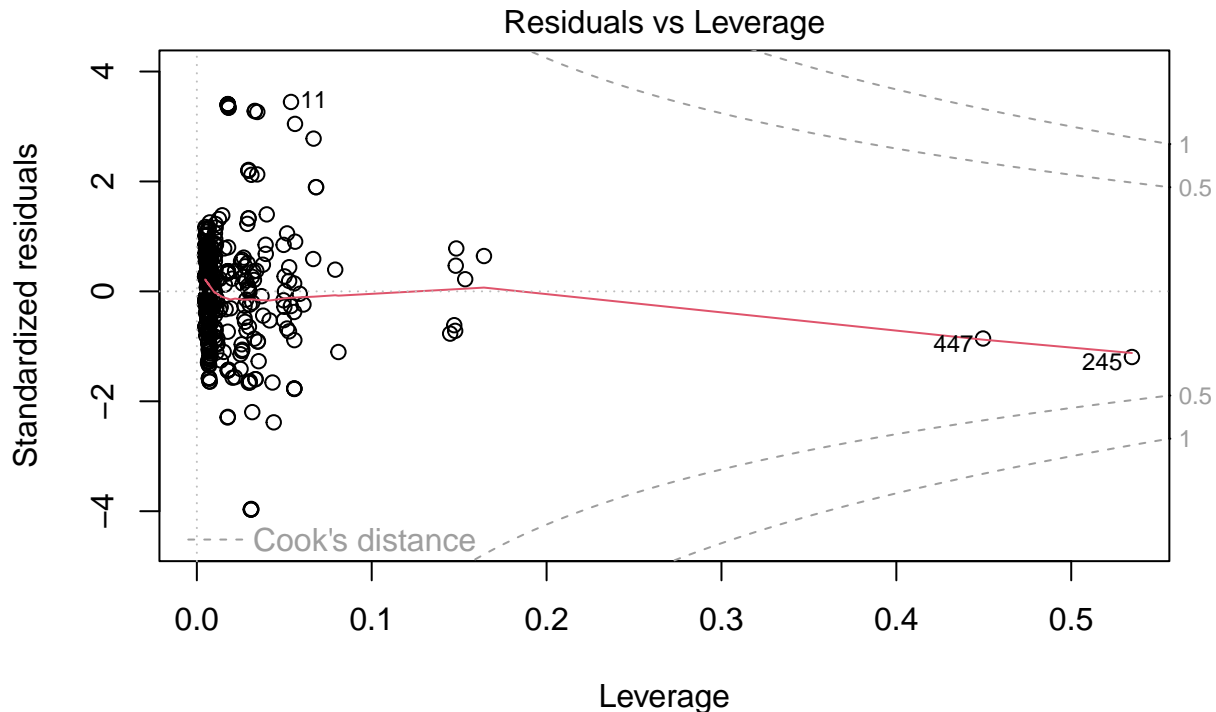
```
plot(m4)
```











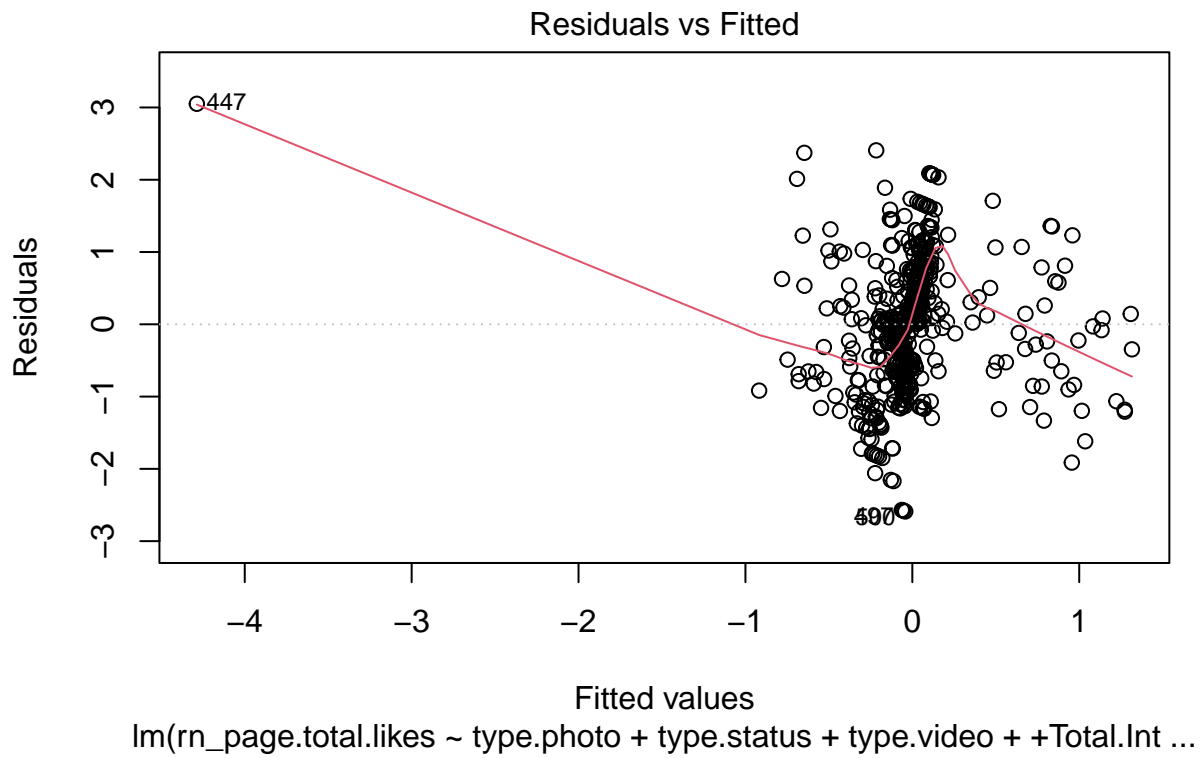
lm(rn\_page.total.likes ~ type.photo + type.status + type.video + Post.Month ...

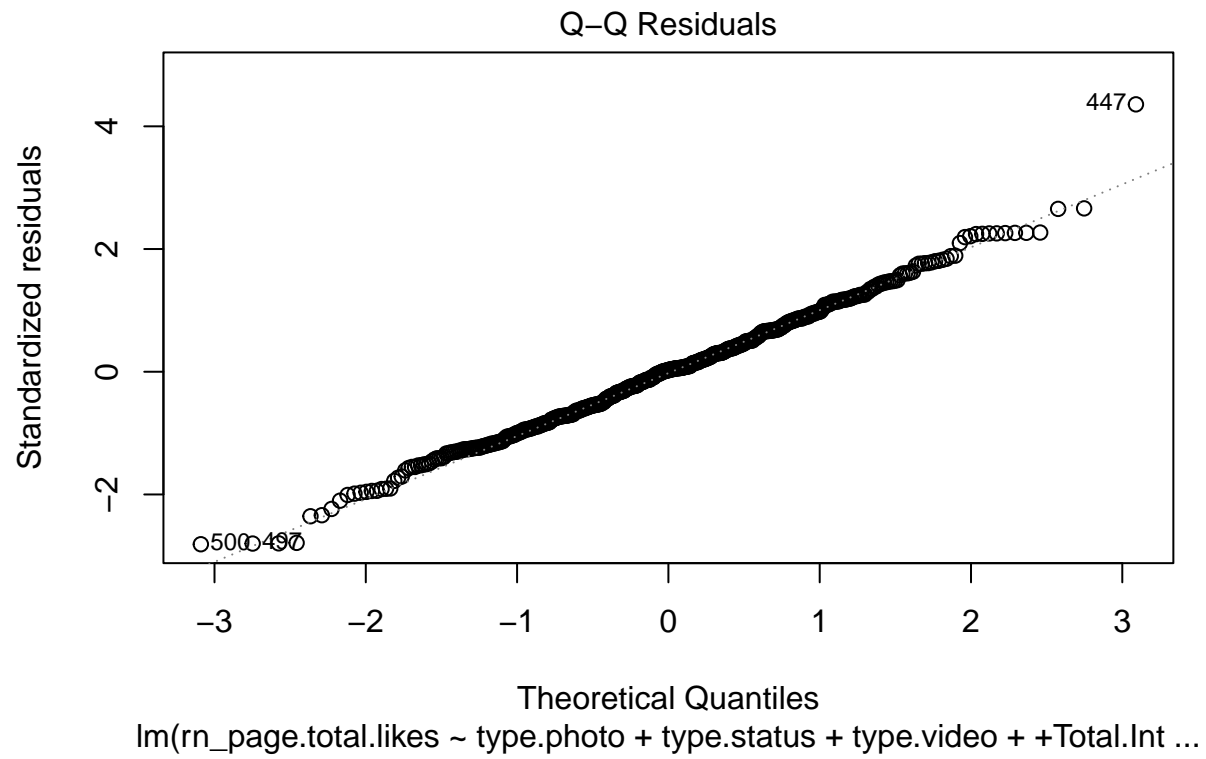
```
## still not meeting all assumptions - remove post.hour since it's not a linear association with page t
m5 <- lm(rn_page.total.likes~type.photo+type.status+type.video++Total.Interactions+Lifetime.Post.Consum
summary(m5)
```

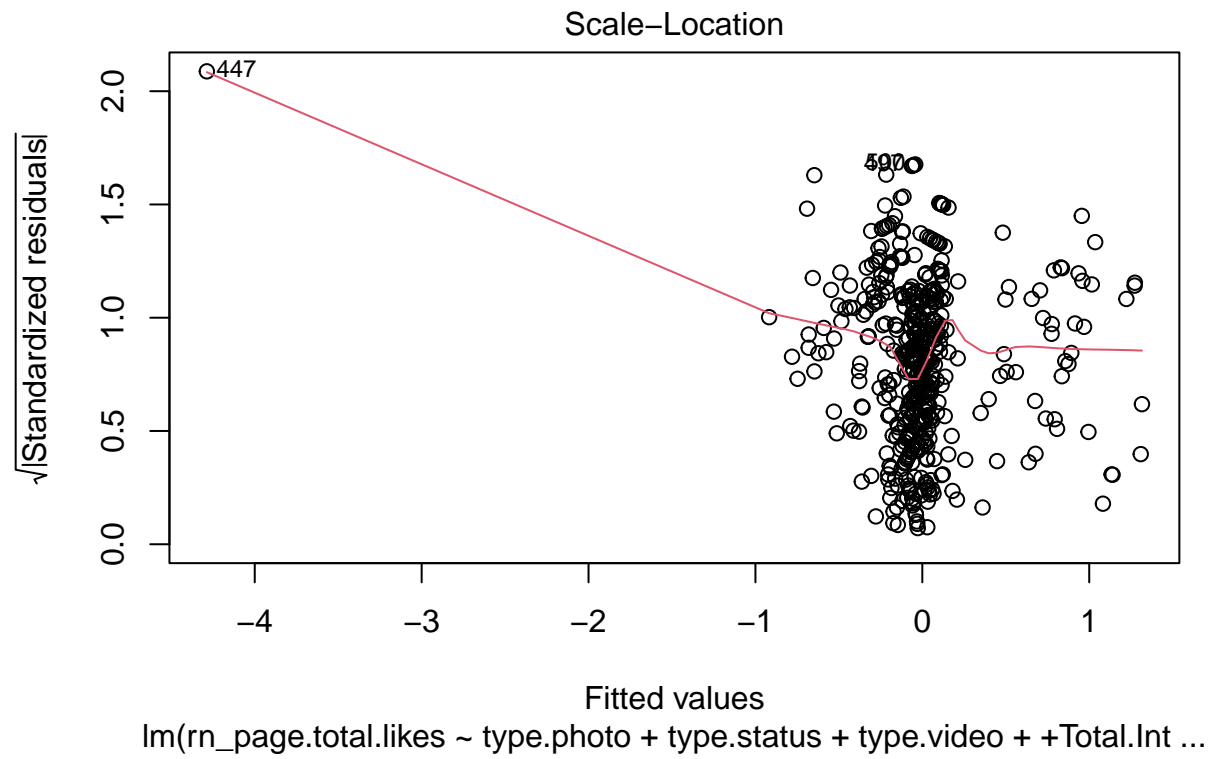
```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + +Total.Interactions + Lifetime.Post.Consumers,
##     data = fb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58932 -0.65049  0.02413  0.61369  3.05087
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.873e-02  1.973e-01  -0.501  0.616914
## type.photo     2.190e-01  2.028e-01   1.080  0.280689
## type.status    1.413e+00  2.558e-01   5.525  5.34e-08 ***
## type.video     1.371e+00  4.058e-01   3.379  0.000786 ***
## Total.Interactions  3.800e-04  1.179e-04   3.224  0.001346 **
## Lifetime.Post.Consumers -3.967e-04  5.633e-05  -7.043  6.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

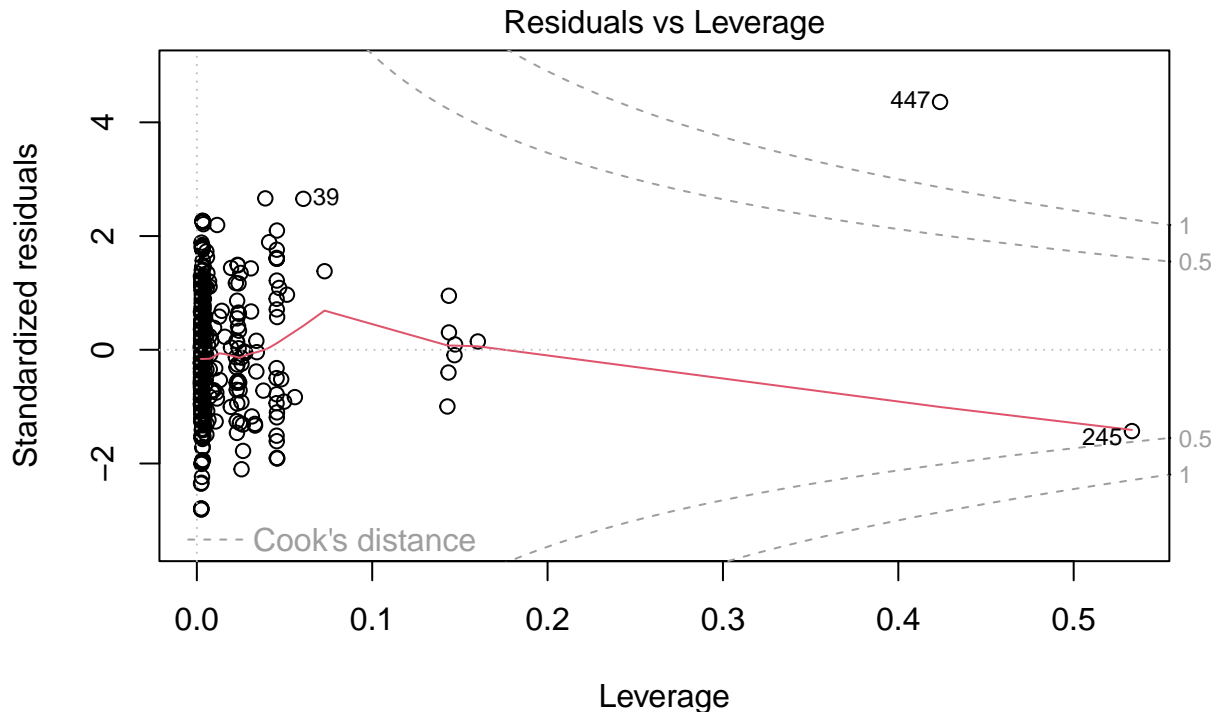
```
## Residual standard error: 0.9222 on 494 degrees of freedom
## Multiple R-squared:  0.1373, Adjusted R-squared:  0.1285
## F-statistic: 15.72 on 5 and 494 DF,  p-value: 2.264e-14
```

```
plot(m5)
```









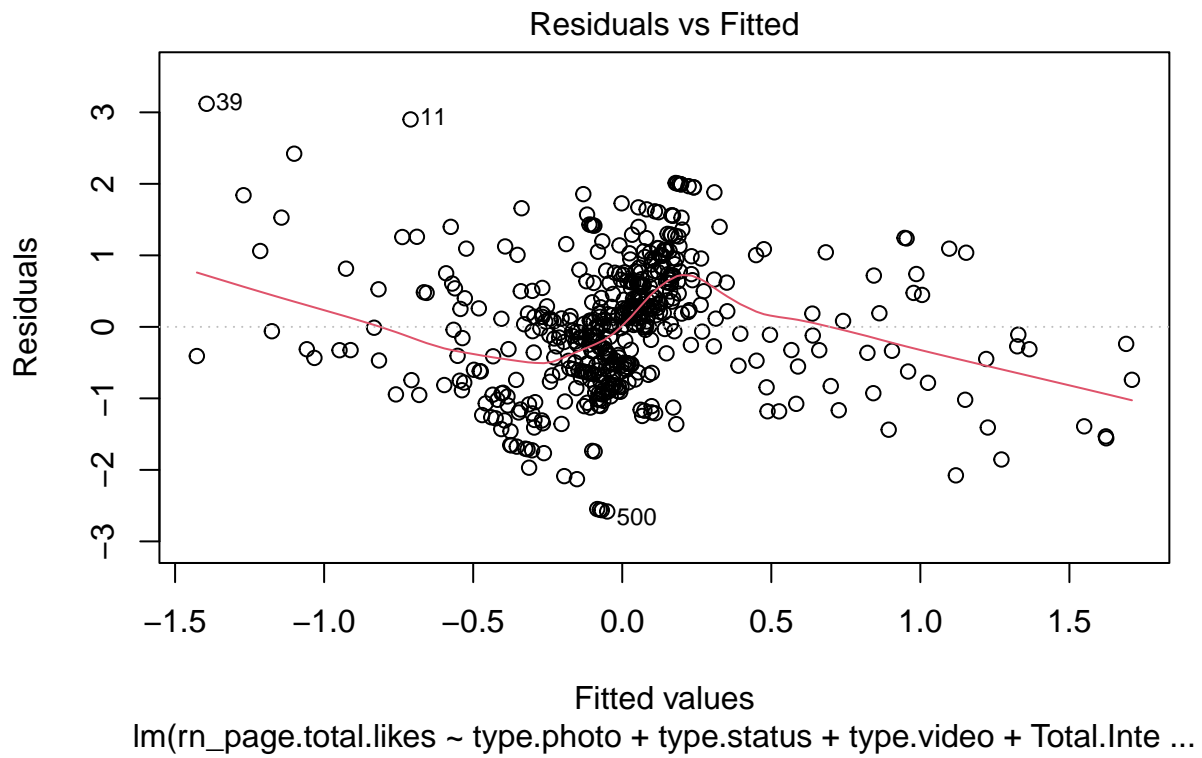
lm(rn\_page.total.likes ~ type.photo + type.status + type.video + +Total.Int ...)

```
## significant better but spotted a bad outlier #447 - remove and fit again
fb2 <- fb[-c(447,245),]
m6 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Total.Interactions+Lifetime.Post.Consumers,
summary(m6)
```

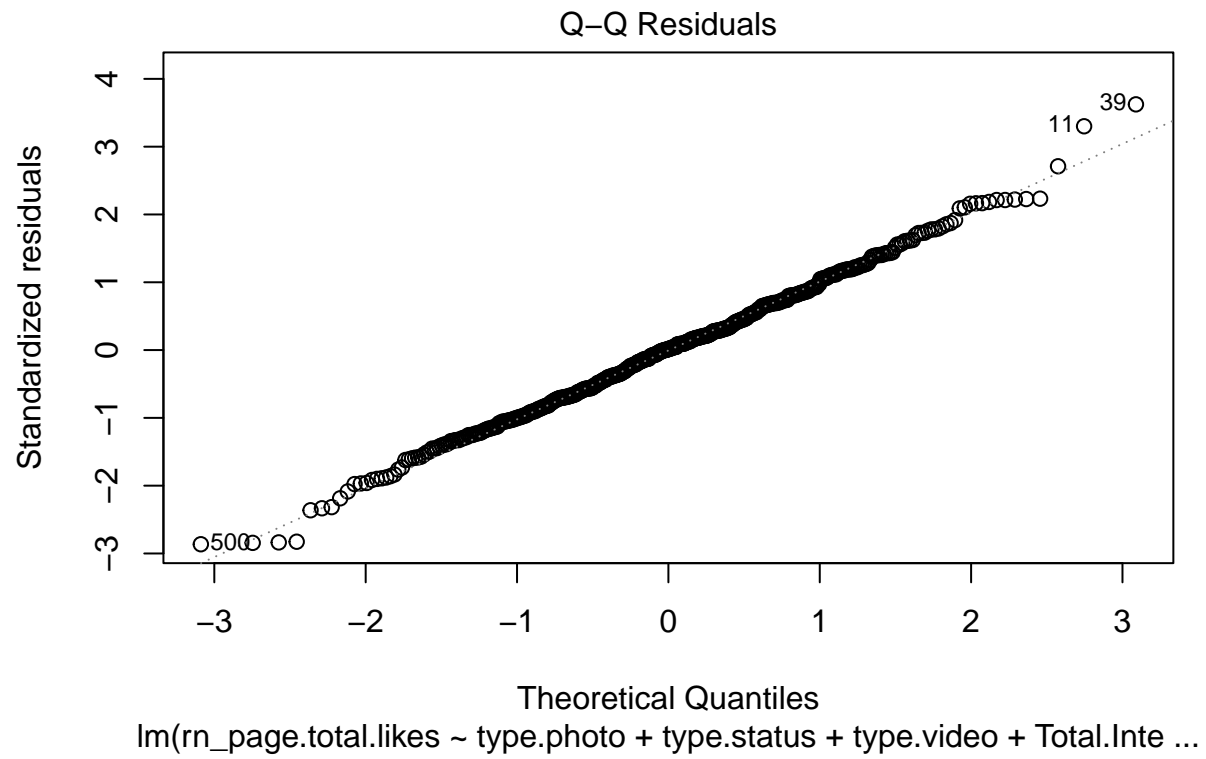
```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + Total.Interactions + Lifetime.Post.Consumers,
##     data = fb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58144 -0.62227  0.01378  0.60747  3.11957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.638e-02  1.935e-01  -0.343   0.732
## type.photo     2.534e-01  1.989e-01   1.274   0.203
## type.status    1.752e+00  2.609e-01   6.714 5.23e-11 ***
## type.video     1.591e+00  4.008e-01   3.969 8.28e-05 ***
## Total.Interactions  7.934e-04  1.735e-04   4.574 6.07e-06 ***
## Lifetime.Post.Consumers -6.330e-04  7.375e-05  -8.584 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

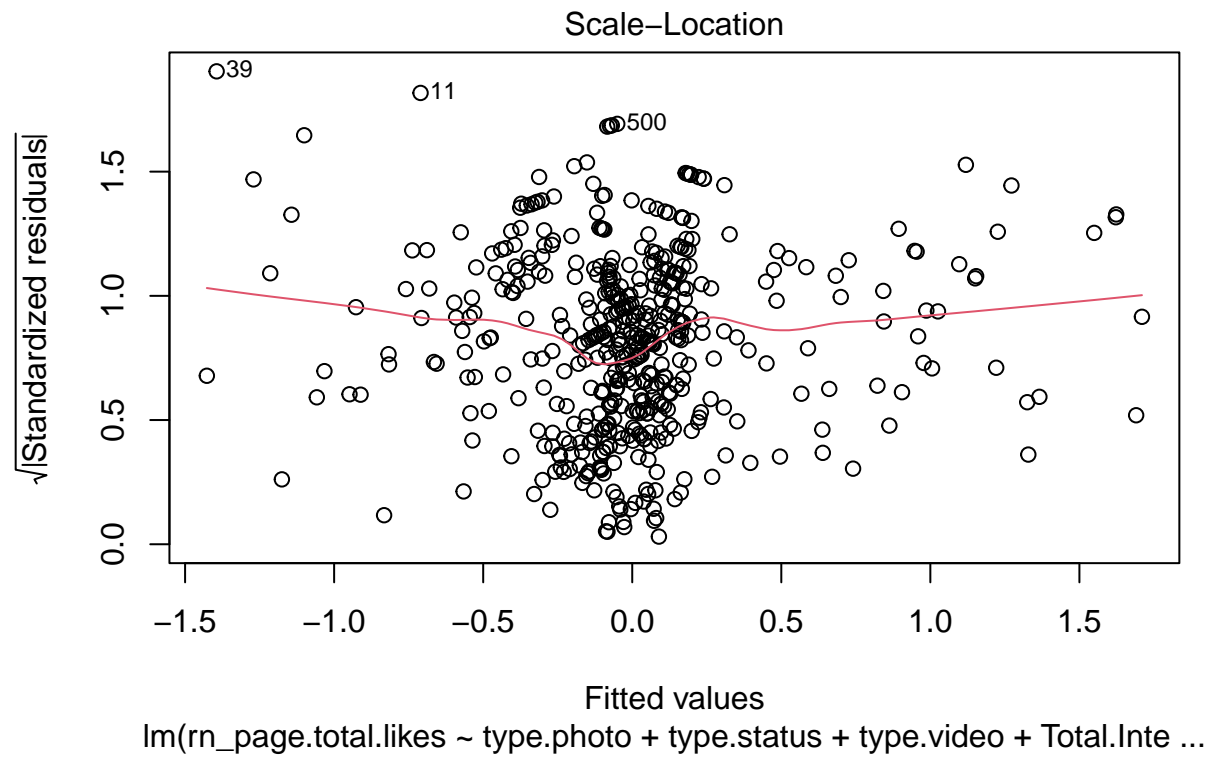
```
## Residual standard error: 0.9026 on 492 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.166
## F-statistic: 20.79 on 5 and 492 DF,  p-value: < 2.2e-16
```

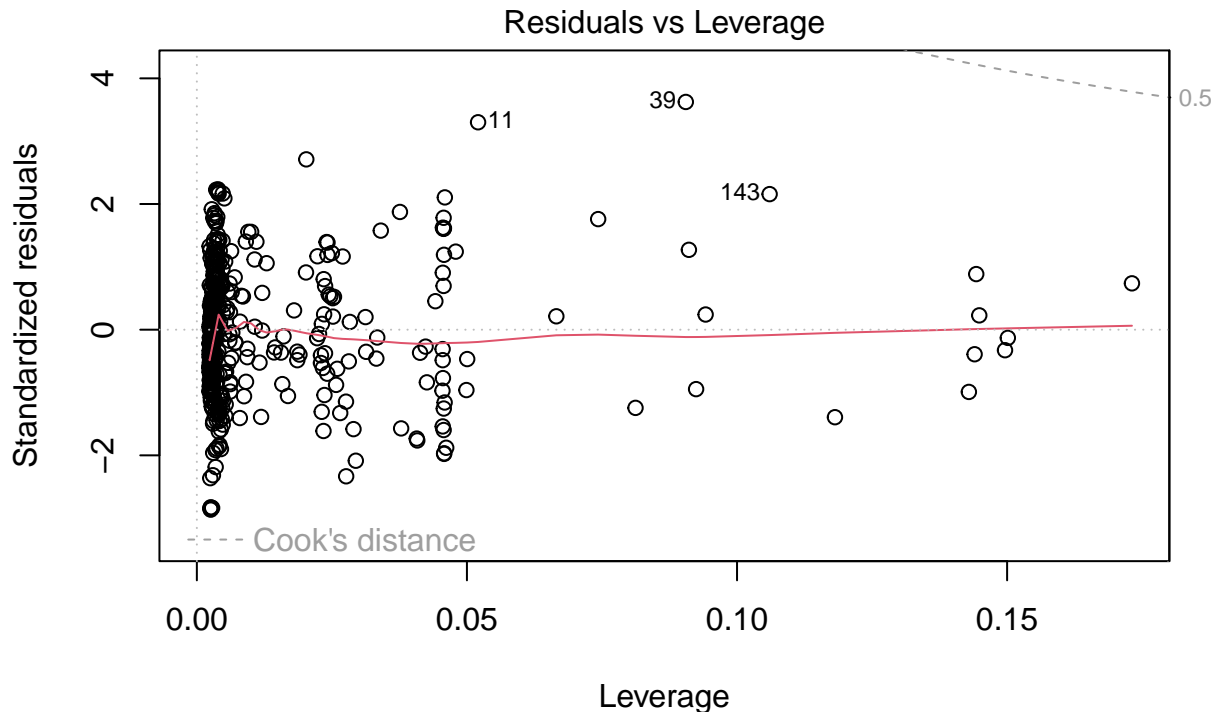
```
plot(m6)
```











lm(rn\_page.total.likes ~ type.photo + type.status + type.video + Total.Inte ...

```
## model comparison using F test
anova(m0,m5)
```

```
## Analysis of Variance Table
##
## Model 1: rn_page.total.likes ~ type.photo + type.status + type.video
## Model 2: rn_page.total.likes ~ type.photo + type.status + type.video +
##   +Total.Interactions + Lifetime.Post.Consumers
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      496 462.61
## 2      494 420.15  2    42.461 24.962 4.706e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## check collinearity
vif(m6)
```

```
##           type.photo           type.status           type.video
##           3.060808           3.420518           1.360975
## Total.Interactions Lifetime.Post.Consumers
##           1.277676           1.783045
```

```
## final model
```

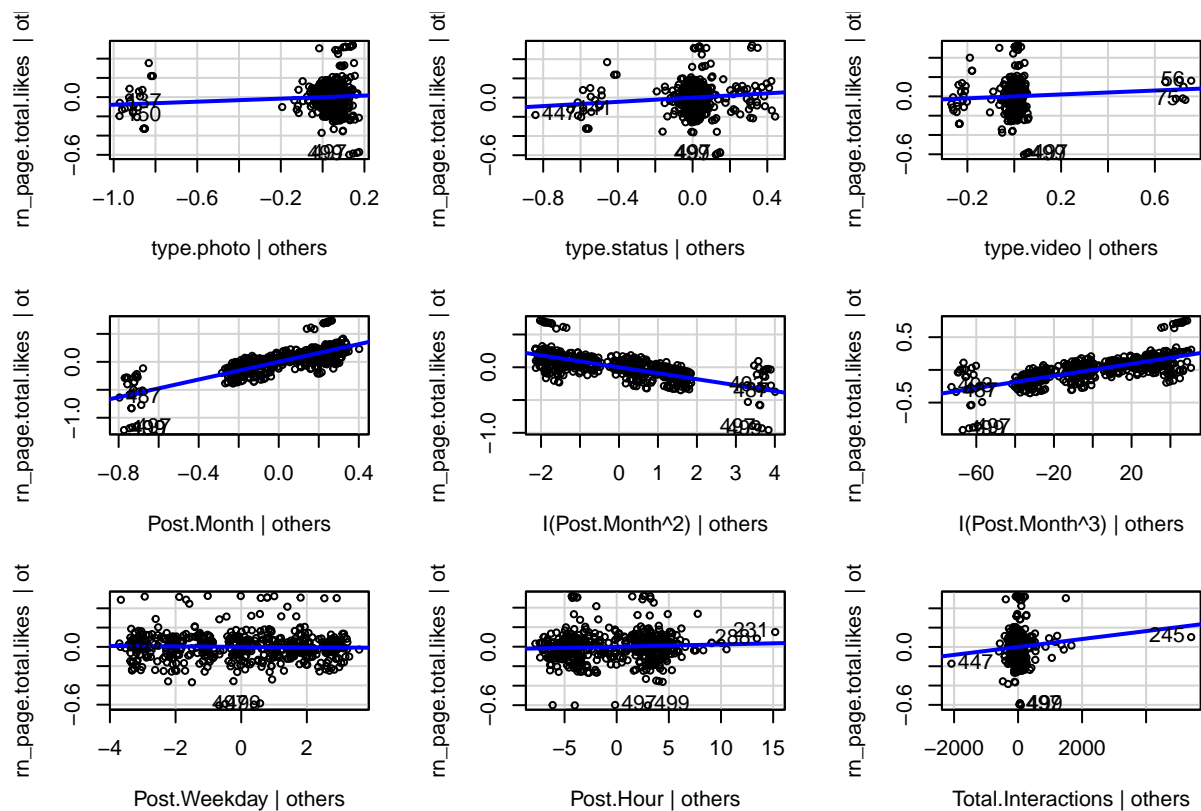
```
summary(m6)
```

```
##
## Call:
## lm(formula = rn_page.total.likes ~ type.photo + type.status +
##     type.video + Total.Interactions + Lifetime.Post.Consumers,
##     data = fb2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58144 -0.62227  0.01378  0.60747  3.11957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.638e-02  1.935e-01  -0.343   0.732
## type.photo     2.534e-01  1.989e-01   1.274   0.203
## type.status    1.752e+00  2.609e-01   6.714 5.23e-11 ***
## type.video     1.591e+00  4.008e-01   3.969 8.28e-05 ***
## Total.Interactions  7.934e-04  1.735e-04   4.574 6.07e-06 ***
## Lifetime.Post.Consumers -6.330e-04  7.375e-05  -8.584 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9026 on 492 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.166
## F-statistic: 20.79 on 5 and 492 DF,  p-value: < 2.2e-16
```

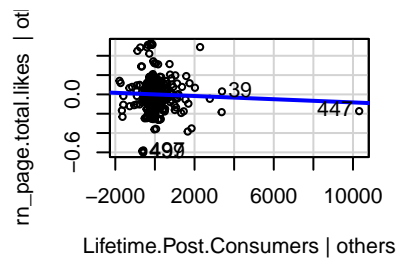
## Visualization

```
## m3
```

```
avPlots(m3)
```



## Added-Variable Plots



```
## m6  
avPlots(m6)
```

## Added-Variable Plots

