

Final Project

Ethan Lee

2024-02-21

```
tiktok_data <- read.csv("tiktok4.csv")

tiktok_observations <- nrow(tiktok_data)
tiktok_variables <- ncol(tiktok_data)

print(paste("number of observations:", tiktok_observations))

## [1] "number of observations: 950"
print(paste("number of variables:", tiktok_variables))

## [1] "number of variables: 8"
print(names(tiktok_data))

## [1] "Rank"          "Tiktoker.name"  "Tiktok.name"   "followers"
## [5] "views.avg."    "likes.avg.."    "comments.avg.." "shares.avg.."

# Function to convert values with 'M' and 'K' suffixes to numeric
convert_followers <- function(x) {
  # Remove 'M' and 'K' suffixes and convert to numeric
  num <- as.numeric(sub("[^0-9.]", "", x))
  # Multiply by the necessary value to get the number
  if (endsWith(x, "M")) {
    num <- num * 1e6
  } else if (endsWith(x, "K")) {
    num <- num * 1e3
  }
  num
}

tiktok_data$followers_numeric <- sapply(tiktok_data$followers, convert_followers)
tiktok_data$views_numeric <- sapply(tiktok_data$views.avg., convert_followers)
tiktok_data$likes_numeric <- sapply(tiktok_data$likes.avg., convert_followers)
tiktok_data$comments_numeric <- sapply(tiktok_data$comments.avg., convert_followers)
tiktok_data$shares_numeric <- sapply(tiktok_data$shares.avg., convert_followers)

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
```

```
sorted_tiktok <- arrange(tiktok_data, desc(views_numeric))
```

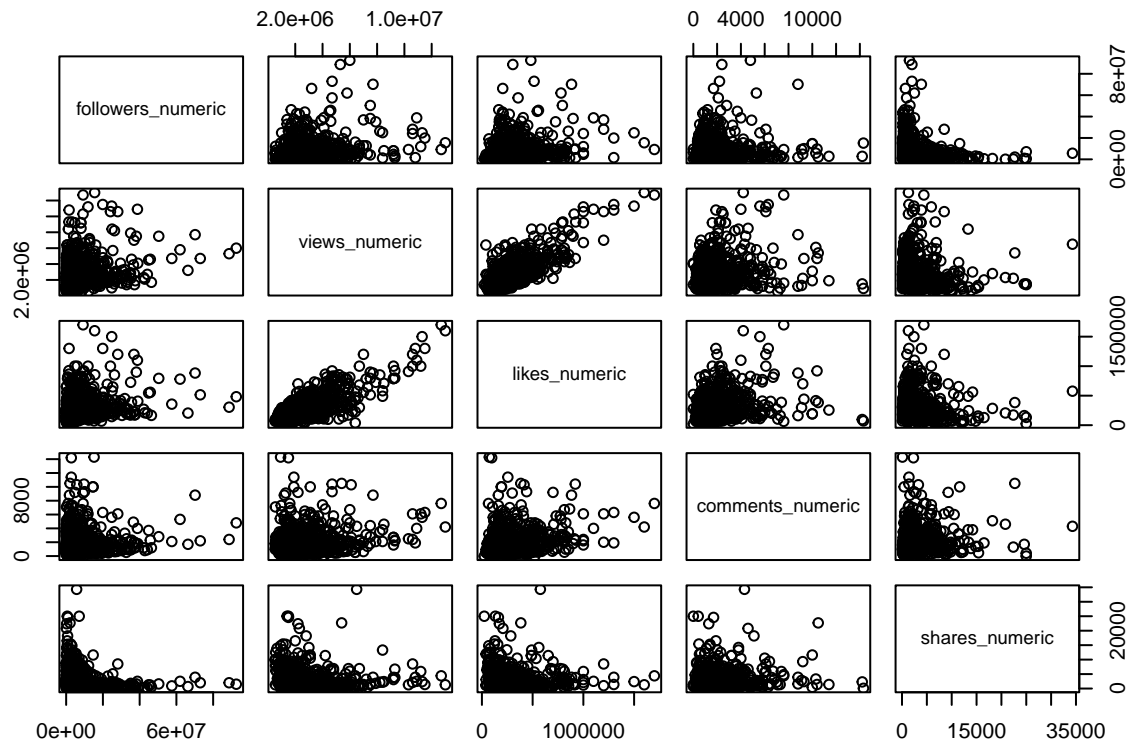
```
head(sorted_tiktok, 10)
```

```
##      Rank      Tiktok.name      Tiktok.name followers views.avg. likes.avg..
## 1      27 nicollefigueroaa Nicolle Figueroa    15.5M      13M      1.6M
## 2      18      fadiljaidi      Fadil Jaidi     9.2M      12.7M     1.7M
## 3      26      surthycooks      Surthycooks    19.9M      11.5M     1.3M
## 4      24 lexibrookerivera      Lexi Rivera    24.8M      11.3M     1.5M
## 5      25      hotspanishmx      HotSpanish    11.9M      11.2M      1M
## 6      34 bayashi.tiktok      Bayashi      38.8M      10.9M     1.1M
## 7      33 slaterkodish      slaterkodish    1.6M      10.8M     1.3M
## 8      19      rosalia      La Rosalia    28.1M      10.6M     1.2M
## 9      49      mmmjoemele      Joe Mele     24.1M      10.6M      1M
## 10     62 bigchungus.tik      BigChungus    9.2M      10.3M    909.6K
##      comments.avg.. shares.avg.. followers_numeric views_numeric likes_numeric
## 1      4.2K      1.3K      15500000      13000000      1600000
## 2      7.6K      4.4K      9200000      12700000      1700000
## 3      6.3K      1.4K      19900000      11500000      1300000
## 4      5.6K      2.4K      24800000      11300000      1500000
## 5      6.1K      5.9K      11900000      11200000      1000000
## 6      4K      2.5K      38800000      10900000      1100000
## 7      1.9K      3.9K      1600000      10800000      1300000
## 8      6.1K      8.5K      28100000      10600000      1200000
## 9      2.4K      2.1K      24100000      10600000      1000000
## 10     2.5K      1.3K      9200000      10300000      909600
##      comments_numeric shares_numeric
## 1      4200      1300
## 2      7600      4400
## 3      6300      1400
## 4      5600      2400
## 5      6100      5900
## 6      4000      2500
## 7      1900      3900
## 8      6100      8500
## 9      2400      2100
## 10     2500      1300
```

```
wanted_columns <- c("followers_numeric", "views_numeric", "likes_numeric", "comments_numeric", "shares_
```

```
tiktok_subset <- tiktok_data[wanted_columns]
```

```
pairs(tiktok_subset)
```



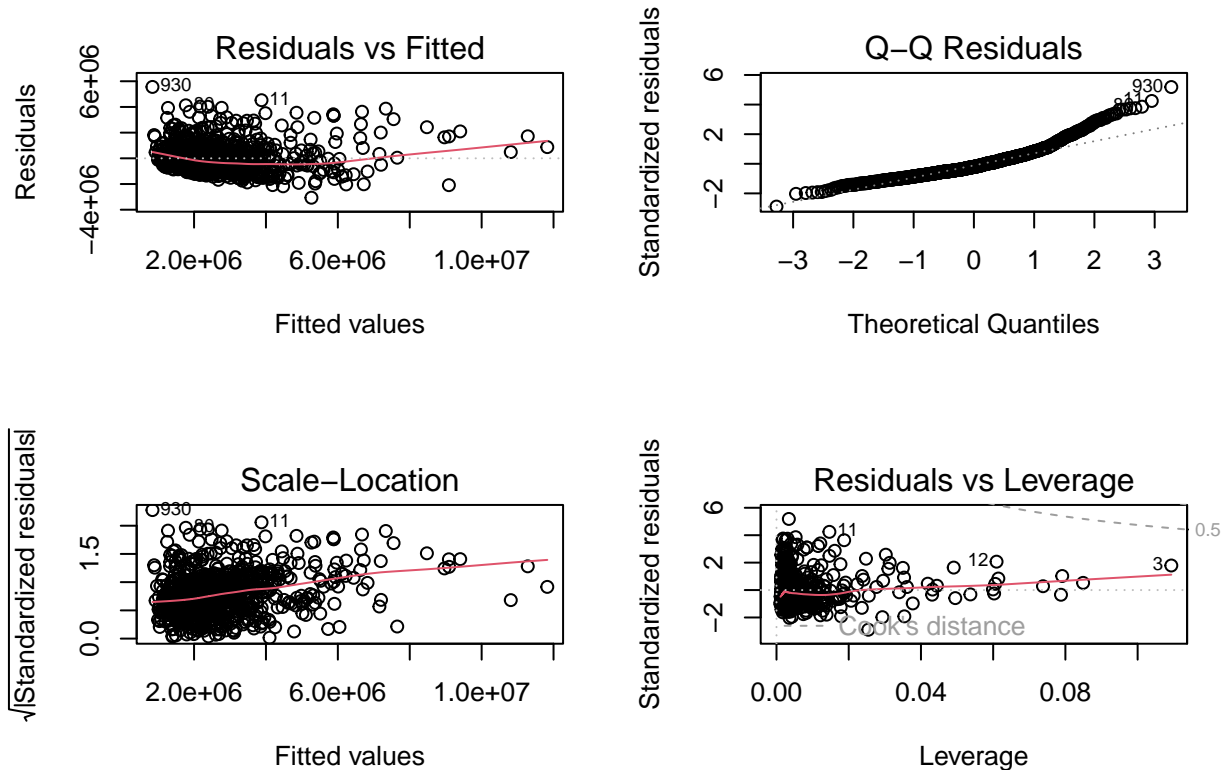
```
m1 <- lm(views_numeric ~ followers_numeric + likes_numeric + comments_numeric + shares_numeric, data = tiktok_subset)
summary(m1)
```

```
##
## Call:
## lm(formula = views_numeric ~ followers_numeric + likes_numeric +
##     comments_numeric + shares_numeric, data = tiktok_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3065842  -710505  -221989   469605   5557204
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.706e+05  7.376e+04   7.736 2.63e-14 ***
## followers_numeric  1.895e-02  3.711e-03   5.107 3.96e-07 ***
## likes_numeric     6.516e+00  1.904e-01  34.231 < 2e-16 ***
## comments_numeric  -4.001e+00  2.234e+01  -0.179  0.858
## shares_numeric    7.876e+00  1.119e+01   0.704  0.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1074000 on 945 degrees of freedom
## Multiple R-squared:  0.6042, Adjusted R-squared:  0.6025
## F-statistic: 360.6 on 4 and 945 DF, p-value: < 2.2e-16
anova(m1)

## Analysis of Variance Table
##
## Response: views_numeric
```

```
##              Df      Sum Sq   Mean Sq   F value Pr(>F)
## followers_numeric  1 2.2028e+14 2.2028e+14  191.0690 <2e-16 ***
## likes_numeric      1 1.4422e+15 1.4422e+15 1250.9967 <2e-16 ***
## comments_numeric   1 1.0229e+09 1.0229e+09   0.0009 0.9762
## shares_numeric     1 5.7154e+11 5.7154e+11   0.4958 0.4815
## Residuals         945 1.0895e+15 1.1529e+12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2,2))
plot(m1)
```



```
library(car)
```

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

```
##
```

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

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

```
##
```

```
##      recode
```

```
tiktok_subset$comments_numeric[tiktok_subset$comments_numeric == 0] <- 1
```

```
summary(tranxy <- powerTransform(cbind(views_numeric, followers_numeric, likes_numeric,
                                         comments_numeric, shares_numeric) ~ 1, data = tiktok_subset))
```

```
## bcPower Transformations to Multinormality
```

```
##              Est Power Rounded Pwr Wald Lwr Bnd Wald Up Bnd
## views_numeric   -0.0887      0.00   -0.1782      0.0008
## followers_numeric 0.1855      0.19    0.1493      0.2218
## likes_numeric    0.1452      0.15    0.0689      0.2215
```

```

## comments_numeric      0.1925      0.19      0.1534      0.2316
## shares_numeric        -0.0760     -0.08     -0.1277     -0.0243
##
## Likelihood ratio test that transformation parameters are equal to 0
## (all log transformations)
##
## LRT df      pval
## LR test, lambda = (0 0 0 0 0) 269.7237 5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##
## LRT df      pval
## LR test, lambda = (1 1 1 1 1) 4817.547 5 < 2.22e-16

tiktok_subset$transformed_followers <- (tiktok_subset$followers_numeric)^(0.19)
tiktok_subset$transformed_likes <- (tiktok_subset$likes_numeric)^(0.15)
tiktok_subset$transformed_comments <- (tiktok_subset$comments_numeric)^(0.19)
tiktok_subset$transformed_shares <- (tiktok_subset$shares_numeric)^(-0.08)
tiktok_subset$transformed_views <- log(tiktok_subset$views_numeric)

m2 <- lm(transformed_views ~ transformed_followers +
          transformed_likes + transformed_shares + transformed_comments, data = tiktok_subset)
summary(m2)

##
## Call:
## lm(formula = transformed_views ~ transformed_followers + transformed_likes +
##     transformed_shares + transformed_comments, data = tiktok_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88112 -0.29365 -0.02945  0.26185  2.08864
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.095899   0.233258  43.282 < 2e-16 ***
## transformed_followers  0.014737   0.003250   4.535 6.51e-06 ***
## transformed_likes     0.611160   0.024904  24.541 < 2e-16 ***
## transformed_shares    0.614070   0.299008   2.054  0.0403 *
## transformed_comments -0.008977   0.023663  -0.379  0.7045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3993 on 945 degrees of freedom
## Multiple R-squared:  0.4629, Adjusted R-squared:  0.4606
## F-statistic: 203.6 on 4 and 945 DF, p-value: < 2.2e-16

anova(m2)

## Analysis of Variance Table
##
## Response: transformed_views
##
## Df Sum Sq Mean Sq F value Pr(>F)
## transformed_followers  1  27.951  27.951 175.2730 < 2e-16 ***
## transformed_likes      1 101.079 101.079 633.8397 < 2e-16 ***
## transformed_shares     1   0.804   0.804   5.0412 0.02498 *
## transformed_comments   1   0.023   0.023   0.1439 0.70452
## Residuals            945 150.699   0.159

```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

tt_model_empty <- lm(transformed_views ~ 1, data = tiktok_subset)
tt_forward_model <- step(tt_model_empty, scope = transformed_views ~ transformed_followers + transformed_views + transformed_comments + transformed_shares, data = tiktok_subset, direction = "forward")

## Start:  AIC=-1156.71
## transformed_views ~ 1
##
##              Df Sum of Sq  RSS    AIC
## + transformed_likes      1   124.547 156.01 -1712.2
## + transformed_followers  1    27.951 252.60 -1254.4
## + transformed_comments   1     5.464 275.09 -1173.4
## + transformed_shares     1     4.578 275.98 -1170.3
## <none>                      280.56 -1156.7
##
## Step:  AIC=-1712.22
## transformed_views ~ transformed_likes
##
##              Df Sum of Sq  RSS    AIC
## + transformed_followers  1     4.4824 151.53 -1737.9
## + transformed_shares     1     2.0286 153.98 -1722.7
## <none>                      156.01 -1712.2
## + transformed_comments   1     0.1752 155.83 -1711.3
##
## Step:  AIC=-1737.92
## transformed_views ~ transformed_likes + transformed_followers
##
##              Df Sum of Sq  RSS    AIC
## + transformed_shares     1    0.80393 150.72 -1741.0
## <none>                      151.53 -1737.9
## + transformed_comments   1    0.15429 151.37 -1736.9
##
## Step:  AIC=-1740.97
## transformed_views ~ transformed_likes + transformed_followers +
##   transformed_shares
##
##              Df Sum of Sq  RSS    AIC
## <none>                      150.72 -1741.0
## + transformed_comments   1  0.022948 150.70 -1739.1

m3 <- lm(transformed_views ~ transformed_followers +
          transformed_likes + transformed_shares, data = tiktok_subset)
summary(m3)

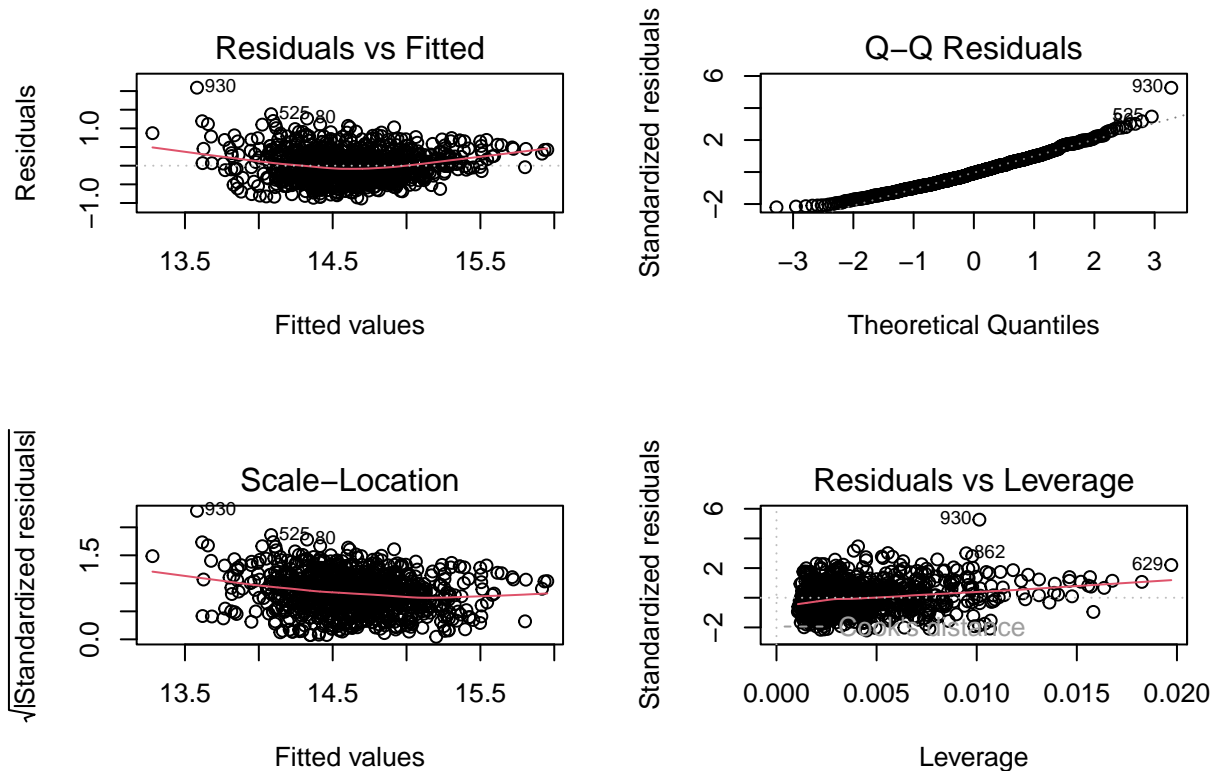
##
## Call:
## lm(formula = transformed_views ~ transformed_followers + transformed_likes +
##   transformed_shares, data = tiktok_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87867 -0.29251 -0.03013  0.26080  2.09035
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.05889   0.211779  47.497 < 2e-16 ***
## transformed_followers 0.014659  0.003242   4.522 6.91e-06 ***
## transformed_likes    0.608857  0.024141  25.221 < 2e-16 ***
## transformed_shares    0.645339  0.287292   2.246  0.0249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3992 on 946 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4611
## F-statistic: 271.6 on 3 and 946 DF, p-value: < 2.2e-16
```

```
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: transformed_views
##               Df Sum Sq Mean Sq F value Pr(>F)
## transformed_followers 1  27.951  27.951 175.4318 < 2e-16 ***
## transformed_likes     1 101.079 101.079 634.4139 < 2e-16 ***
## transformed_shares     1   0.804   0.804   5.0458 0.02492 *
## Residuals           946 150.722   0.159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2,2))
plot(m3)
```

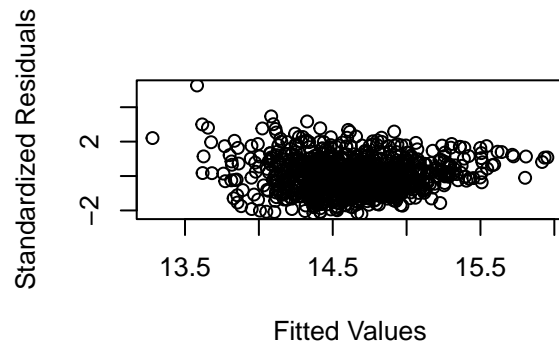
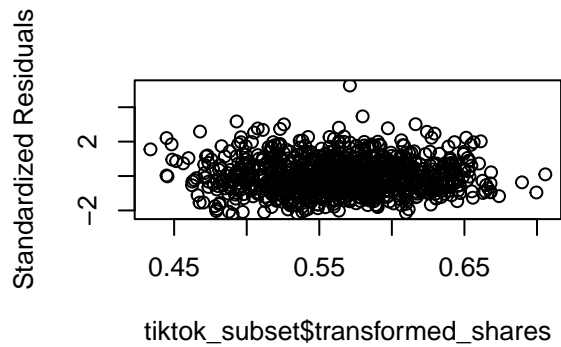
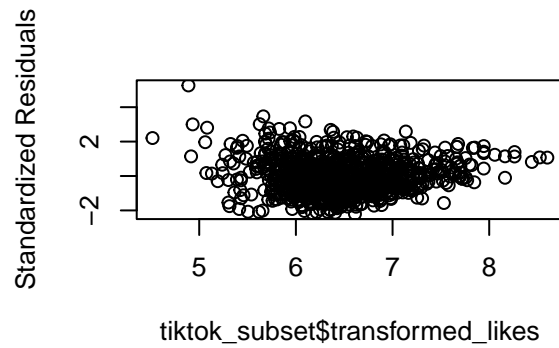
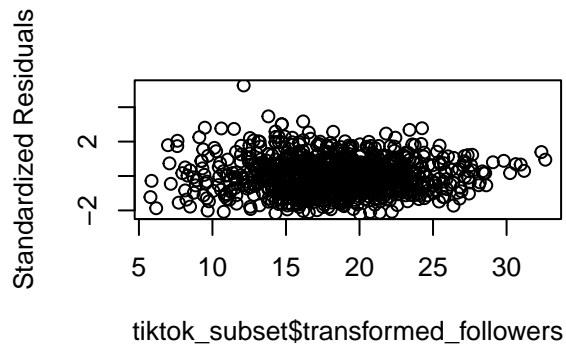


```
par(mfrow=c(2,2))
StanRes2 <- rstandard(m3)
```

```

plot(tiktok_subset$transformed_followers,StanRes2,ylab="Standardized Residuals")
plot(tiktok_subset$transformed_likes,StanRes2,ylab="Standardized Residuals")
plot(tiktok_subset$transformed_shares,StanRes2,ylab="Standardized Residuals")
plot(m3$fitted.values,StanRes2,ylab="Standardized Residuals",xlab="Fitted Values")

```

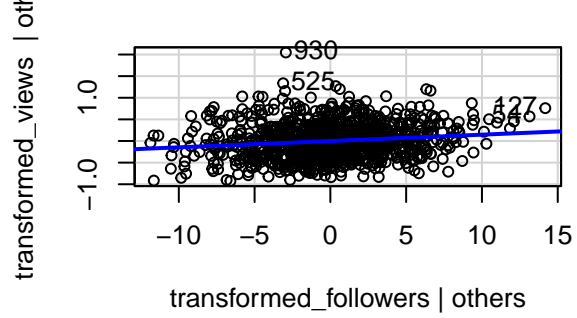


```

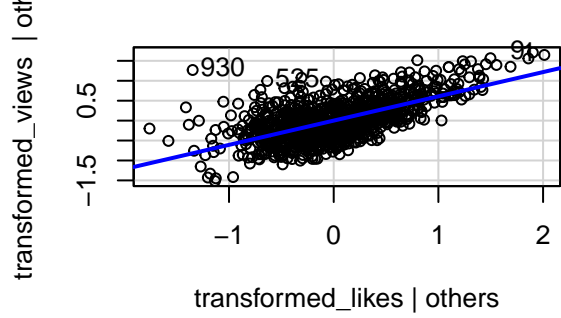
library(car)
par(mfrow=c(2,2))
avPlot(m3, variable = "transformed_followers", ask=FALSE)
avPlot(m3, variable = "transformed_likes", ask=FALSE)
avPlot(m3, variable = "transformed_shares", ask=FALSE)

```

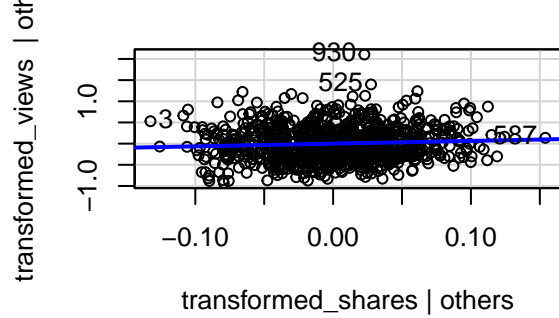

Added-Variable Plot: transformed_follow



Added-Variable Plot: transformed_like



Added-Variable Plot: transformed_shar



```
vif(m3)
```

```
## transformed_followers    transformed_likes    transformed_shares
##                1.175534                1.093710                1.079565
```