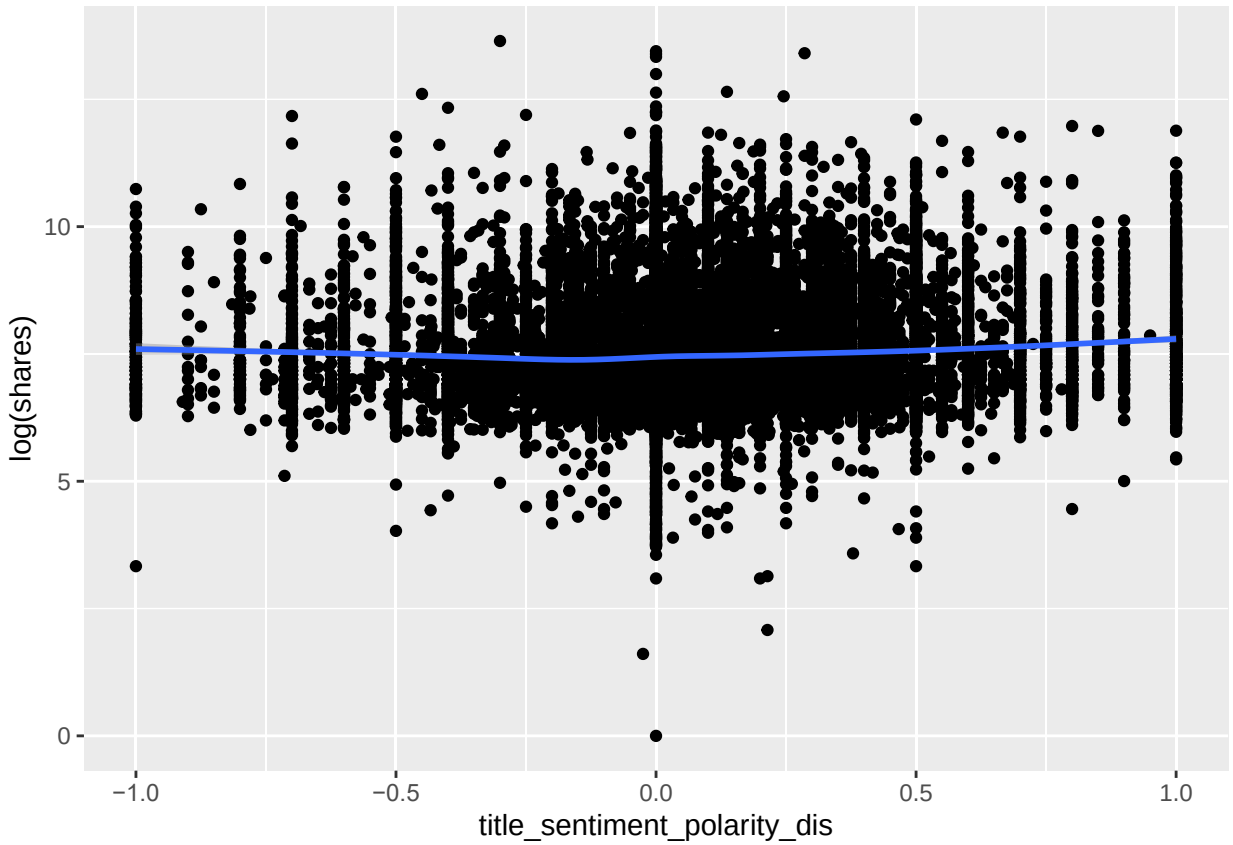
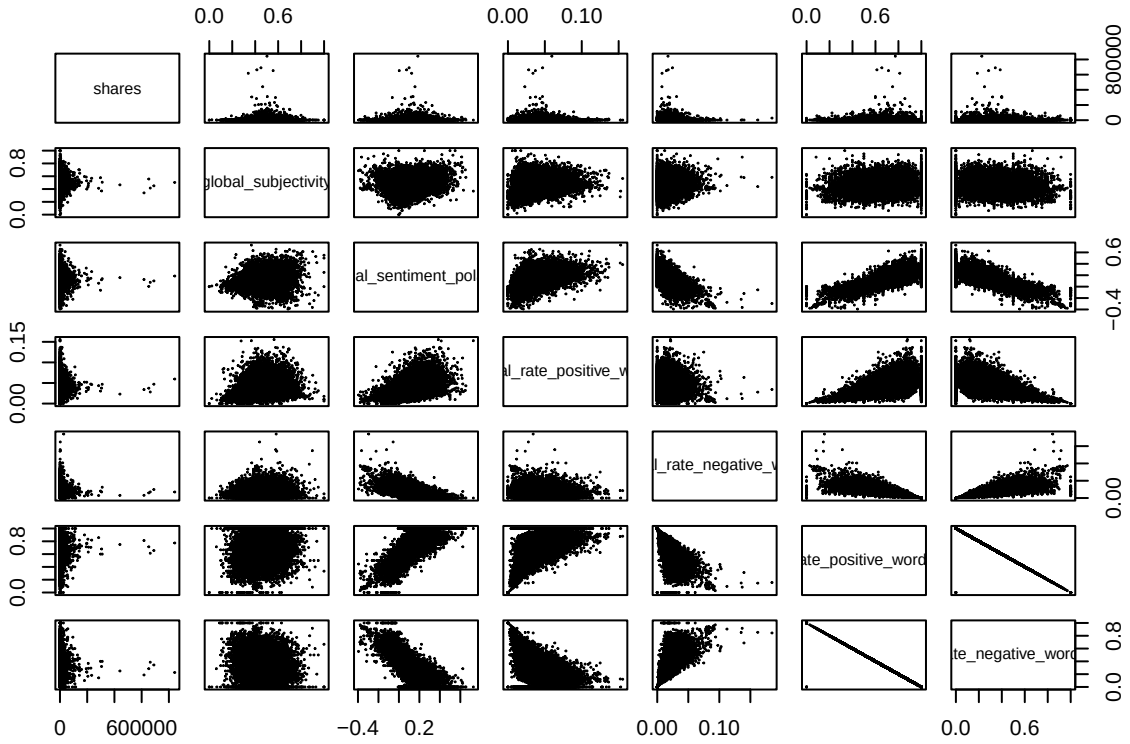



```
#"global_subjectivity"
#[41] "global_sentiment_polarity"     "global_rate_positive_words"
#[43] "global_rate_negative_words"    "rate_positive_words"
#[45] "rate_negative_words"           "avg_positive_polarity"
#[47] "min_positive_polarity"         "max_positive_polarity"
#[49] "avg_negative_polarity"         "min_negative_polarity"
#[51] "max_negative_polarity"

news_npl <- news[,c(56,40:51)]
head(news_npl)
```

```
##   shares global_subjectivity global_sentiment_polarity
## 1    593                 0.5                       0.09
## 2    711                 0.3                       0.15
## 3   1500                 0.7                       0.32
## 4   1200                 0.4                       0.10
## 5    505                 0.5                       0.28
## 6    855                 0.4                       0.07
##   global_rate_positive_words global_rate_negative_words rate_positive_words
## 1                       0.05                      0.014                 0.8
## 2                       0.04                      0.016                 0.7
## 3                       0.06                      0.009                 0.9
## 4                       0.04                      0.021                 0.7
## 5                       0.07                      0.012                 0.9
## 6                       0.03                      0.027                 0.5
##   rate_negative_words avg_positive_polarity min_positive_polarity
## 1                 0.2                   0.4                  0.10
```

```
## 2                  0.3                  0.3                  0.03
## 3                  0.1                  0.5                  0.10
## 4                  0.3                  0.4                  0.14
## 5                  0.1                  0.4                  0.03
## 6                  0.5                  0.4                  0.14
##   max_positive_polarity avg_negative_polarity min_negative_polarity
## 1                   0.7                  -0.3                  -0.6
## 2                   0.7                  -0.1                  -0.1
## 3                   1.0                  -0.5                  -0.8
## 4                   0.8                  -0.4                  -0.6
## 5                   1.0                  -0.2                  -0.5
## 6                   0.6                  -0.2                  -0.4
##   max_negative_polarity
## 1                 -0.20
## 2                 -0.10
## 3                 -0.13
## 4                 -0.17
## 5                 -0.05
## 6                 -0.10
```

```
pairs(news_npl[1:7],cex=0.1)
```




```
cor(news_npl[2:13])#  rate_negative_words and rate_positive_words are highly correlated (r= -0.9976925)
```

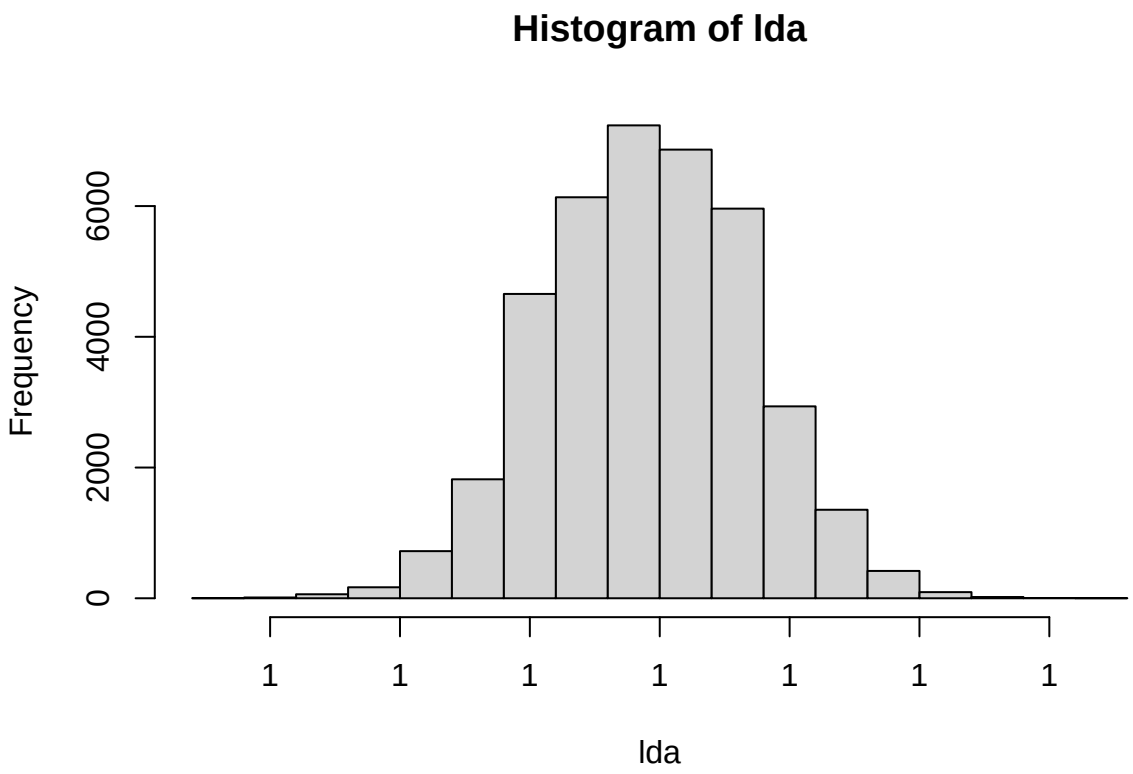
```
##                            global_subjectivity global_sentiment_polarity
## global_subjectivity                        1.0                     0.269
## global_sentiment_polarity                  0.3                     1.000
## global_rate_positive_words                 0.3                     0.541
## global_rate_negative_words                 0.1                    -0.564
## rate_positive_words                        0.1                     0.780
## rate_negative_words                       -0.1                    -0.780
```

```
## avg_positive_polarity                    0.4                    0.469
## min_positive_polarity                    0.1                    0.037
## max_positive_polarity                    0.3                    0.382
## avg_negative_polarity                   -0.3                    0.336
## min_negative_polarity                   -0.2                    0.381
## max_negative_polarity                   -0.1                   -0.002
##                            global_rate_positive_words
## global_subjectivity                            0.2947
## global_sentiment_polarity                      0.5412
## global_rate_positive_words                     1.0000
## global_rate_negative_words                    -0.0007
## rate_positive_words                            0.5301
## rate_negative_words                           -0.5288
## avg_positive_polarity                          0.1285
## min_positive_polarity                         -0.2250
## max_positive_polarity                          0.3356
## avg_negative_polarity                          0.0119
## min_negative_polarity                          0.0117
## max_negative_polarity                          0.0068
##                            global_rate_negative_words rate_positive_words
## global_subjectivity                            0.1030                0.12
## global_sentiment_polarity                     -0.5636                0.78
## global_rate_positive_words                    -0.0007                0.53
## global_rate_negative_words                     1.0000               -0.76
## rate_positive_words                           -0.7581                1.00
## rate_negative_words                            0.7599               -1.00
## avg_positive_polarity                          0.0431                0.06
## min_positive_polarity                         -0.0025               -0.13
## max_positive_polarity                          0.0479                0.17
## avg_negative_polarity                         -0.2849                0.26
## min_negative_polarity                         -0.4307                0.39
## max_negative_polarity                          0.1310               -0.10
##                            rate_negative_words avg_positive_polarity
## global_subjectivity                      -0.12                 0.39
## global_sentiment_polarity                -0.78                 0.47
## global_rate_positive_words               -0.53                 0.13
## global_rate_negative_words                0.76                 0.04
## rate_positive_words                      -1.00                 0.06
## rate_negative_words                       1.00                -0.06
## avg_positive_polarity                    -0.06                 1.00
## min_positive_polarity                     0.13                 0.41
## max_positive_polarity                    -0.17                 0.57
## avg_negative_polarity                    -0.26                -0.09
## min_negative_polarity                    -0.39                -0.06
## max_negative_polarity                     0.10                -0.03
##                            min_positive_polarity max_positive_polarity
## global_subjectivity                        0.121                 0.26
## global_sentiment_polarity                  0.037                 0.38
## global_rate_positive_words                -0.225                 0.34
## global_rate_negative_words                -0.002                 0.05
## rate_positive_words                       -0.128                 0.17
## rate_negative_words                        0.129                -0.17
## avg_positive_polarity                      0.406                 0.57
## min_positive_polarity                      1.000                -0.14
```

```
## max_positive_polarity                    -0.141                  1.00
## avg_negative_polarity                     0.022                 -0.10
## min_negative_polarity                     0.179                 -0.24
## max_negative_polarity                    -0.144                  0.12
##                            avg_negative_polarity min_negative_polarity
## global_subjectivity                       -0.30                 -0.20
## global_sentiment_polarity                  0.34                  0.38
## global_rate_positive_words                 0.01                  0.01
## global_rate_negative_words                -0.28                 -0.43
## rate_positive_words                        0.26                  0.39
## rate_negative_words                       -0.26                 -0.39
## avg_positive_polarity                     -0.09                 -0.06
## min_positive_polarity                      0.02                  0.18
## max_positive_polarity                     -0.10                 -0.24
## avg_negative_polarity                      1.00                  0.72
## min_negative_polarity                      0.72                  1.00
## max_negative_polarity                      0.56                  0.02
##                            max_negative_polarity
## global_subjectivity                       -0.125
## global_sentiment_polarity                 -0.002
## global_rate_positive_words                 0.007
## global_rate_negative_words                 0.131
## rate_positive_words                       -0.098
## rate_negative_words                        0.098
## avg_positive_polarity                     -0.029
## min_positive_polarity                     -0.144
## max_positive_polarity                      0.121
## avg_negative_polarity                      0.556
## min_negative_polarity                      0.020
## max_negative_polarity                      1.000
```

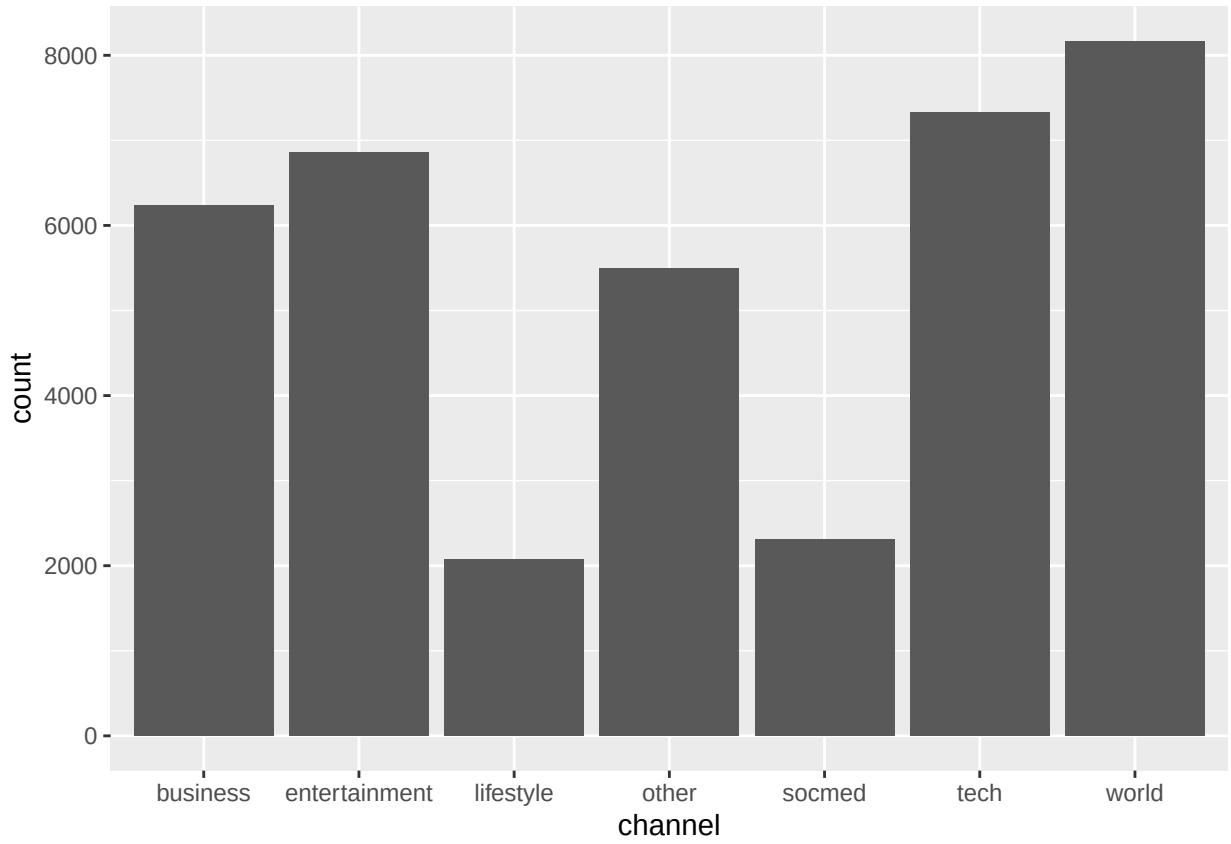
categorical data exploration

```
lda <- news$LDA_00+news$LDA_01+news$LDA_02+news$LDA_03+news$LDA_04
hist(lda)
```

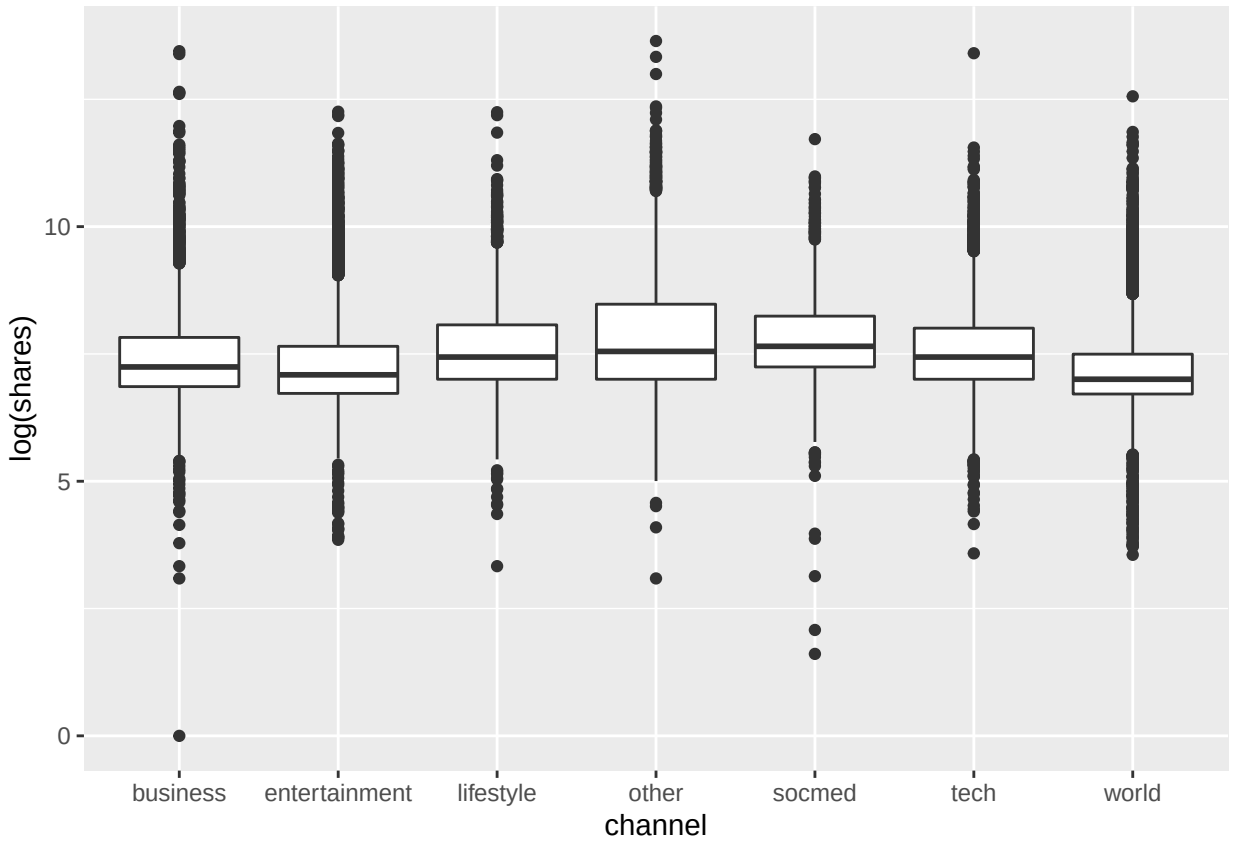



```
## the sum of this five variables =1, In this case we should drop one of these variables,we drop LDA_00
```

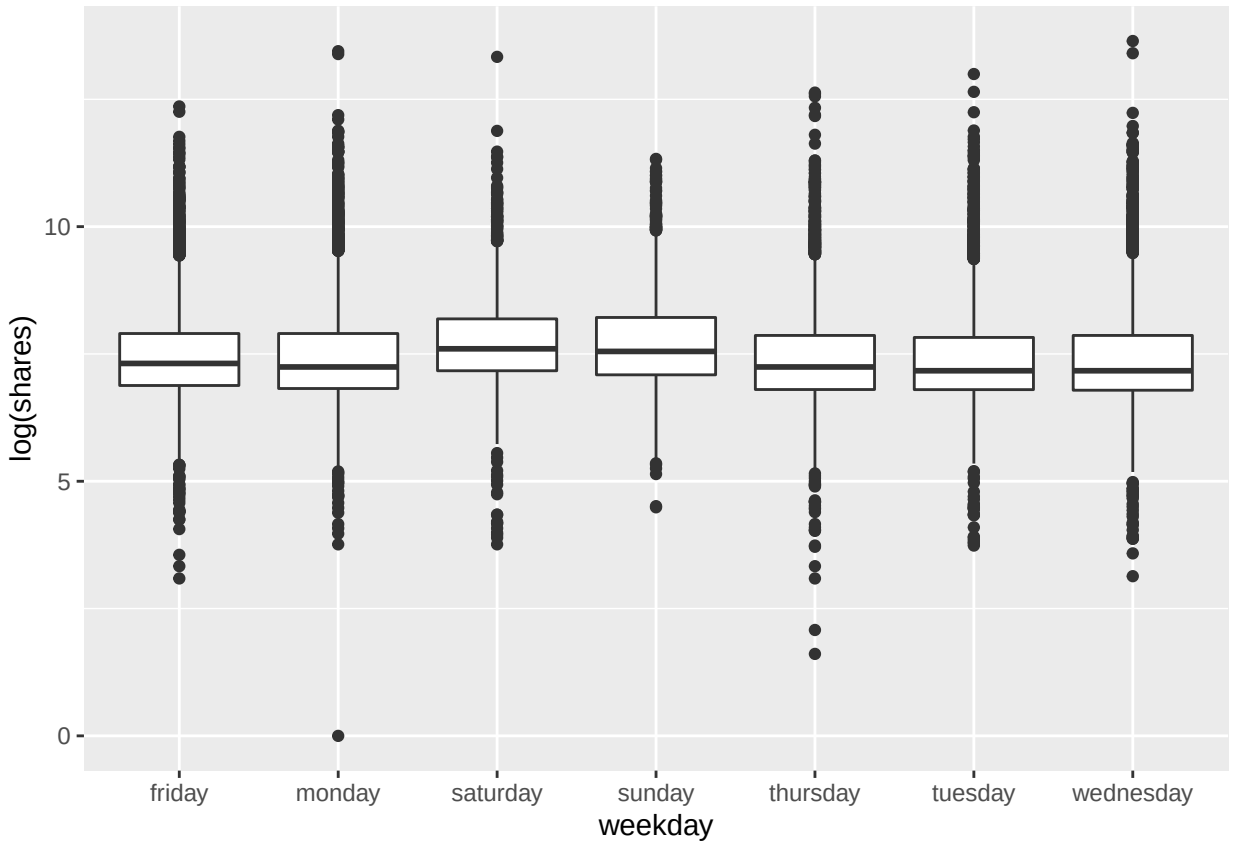
```
library(ggplot2)
# recoding the dummy variables
news$channel[news$data_channel_is_lifestyle==1] <-"lifestyle"
news$channel[news$data_channel_is_entertainment==1] <-"entertainment"
news$channel[news$data_channel_is_socmed==1] <-"socmed"
news$channel[news$data_channel_is_tech==1] <-"tech"
news$channel[news$data_channel_is_world==1] <-"world"
news$channel[news$data_channel_is_bus==1] <-"business"
news$channel[news$data_channel_is_world==0 & news$data_channel_is_lifestyle==0 & news$data_channel_is_e
news$channel <- factor(news$channel)
ggplot(news,aes(channel))+geom_bar()
```



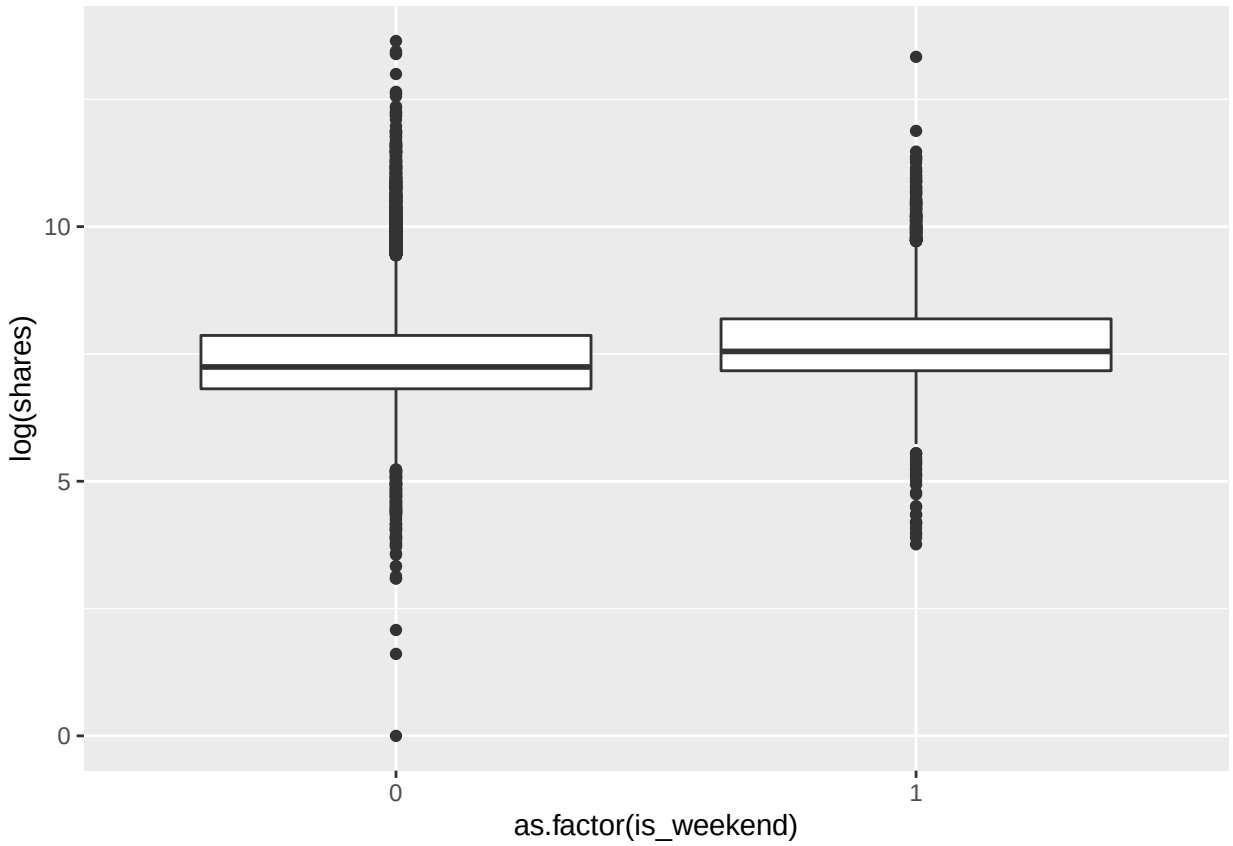

```
ggplot(news, aes(channel,log(shares)))+
  geom_boxplot()
```




```
# recoding the dummy variables
news$weekday[news$weekday_is_monday==1] <-"monday"
news$weekday[news$weekday_is_tuesday==1] <-"tuesday"
news$weekday[news$weekday_is_wednesday==1] <-"wednesday"
news$weekday[news$weekday_is_thursday==1] <-"thursday"
news$weekday[news$weekday_is_friday==1] <-"friday"
news$weekday[news$weekday_is_saturday==1] <-"saturday"
news$weekday[news$weekday_is_sunday==1] <-"sunday"
news$weekday <- factor(news$weekday)

ggplot(news, aes(weekday,log(shares)))+
  geom_boxplot()
```




```
ggplot(news, aes(as.factor(is_weekend),log(shares)))+
  geom_boxplot()
```




```
# comment: There are no differences through Monday to Friday on log(shares), median log(shares) for wee
```

drop redundant variables

```
library(tidyverse)
news1 <- news %>%
  dplyr::select(-weekday_is_monday, -weekday_is_tuesday, -weekday_is_wednesday,-weekday_is_thursday, -we

length(news1)
```

```
## [1] 36
```

```
# we have 36 variables in our final data set
names(news1)
```

```
##  [1] "n_tokens_title"              "n_tokens_content"
##  [3] "n_non_stop_unique_tokens"    "num_hrefs"
##  [5] "num_self_hrefs"              "num_imgs"
##  [7] "num_videos"                  "average_token_length"
##  [9] "num_keywords"                "kw_max_min"
## [11] "kw_min_max"                  "kw_max_max"
## [13] "kw_avg_max"                  "kw_min_avg"
## [15] "kw_avg_avg"                  "self_reference_avg_sharess"
## [17] "is_weekend"                  "LDA_01"
## [19] "LDA_02"                      "LDA_03"
## [21] "LDA_04"                      "global_subjectivity"
## [23] "global_sentiment_polarity"   "global_rate_positive_words"
## [25] "global_rate_negative_words"  "rate_positive_words"
```

```
## [27] "avg_positive_polarity"        "min_positive_polarity"
## [29] "max_positive_polarity"        "avg_negative_polarity"
## [31] "min_negative_polarity"        "max_negative_polarity"
## [33] "shares"                       "title_subjectivity_dis"
## [35] "title_sentiment_polarity_dis" "channel"
```

build the model

set up training and test data set

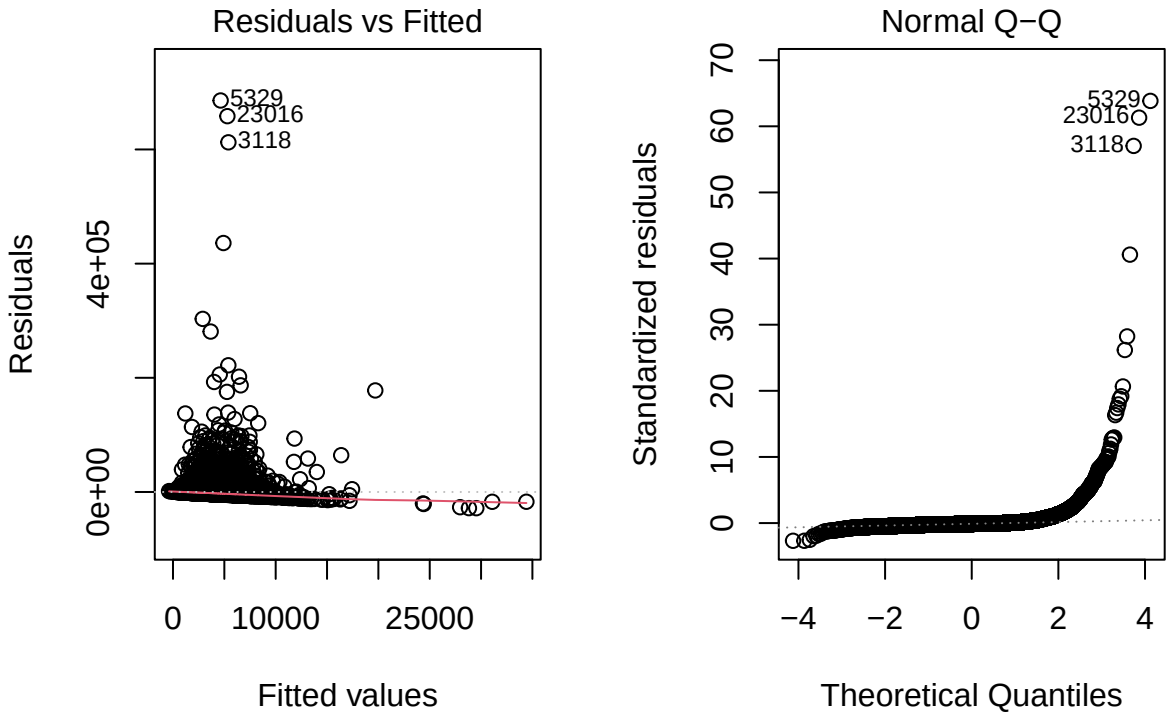
```
set.seed(123)
n<-nrow(news1)


index_train <- sample(1:n, round(0.7*n)) # we randomly choose 70% of the data as training data
newstrain <- news1[index_train,]
newstest <- news1[-index_train,]
```

```
options(scipen=0)
lmnew_full <- lm(shares~.,data=newstrain)
lmnew_full_step <- step(lmnew_full,trace=F)
summary(lmnew_full_step)
```

```
##
## Call:
## lm(formula = shares ~ n_tokens_title + n_tokens_content + num_hrefs +
##     num_self_hrefs + kw_avg_max + kw_min_avg + kw_avg_avg + self_reference_avg_sharess +
##     is_weekend + LDA_01 + LDA_02 + global_subjectivity + global_rate_positive_words +
##     min_positive_polarity + max_negative_polarity + channel,
##     data = newstrain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28338  -2143  -1279   -228 685761
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.83e+02   5.73e+02   -0.49  0.62144
## n_tokens_title               7.06e+01   3.18e+01    2.22  0.02647 *
## n_tokens_content             6.32e-01   1.69e-01    3.75  0.00018 ***
## num_hrefs                    1.84e+01   7.12e+00    2.58  0.00995 **
## num_self_hrefs              -4.79e+01   1.95e+01   -2.46  0.01398 *
## kw_avg_max                  -1.83e-03   6.10e-04   -3.01  0.00265 **
## kw_min_avg                  -1.05e-01   6.81e-02   -1.54  0.12276
## kw_avg_avg                   7.82e-01   6.70e-02   11.68  < 2e-16 ***
## self_reference_avg_sharess   1.24e-02   2.82e-03    4.42  1.0e-05 ***
## is_weekend                   4.77e+02   1.96e+02    2.44  0.01488 *
## LDA_01                      -6.25e+02   3.87e+02   -1.62  0.10612
## LDA_02                      -1.44e+03   4.39e+02   -3.27  0.00108 **
## global_subjectivity          2.38e+03   8.47e+02    2.81  0.00499 **
## global_rate_positive_words -1.25e+04   4.58e+03   -2.72  0.00650 **
## min_positive_polarity       -2.04e+03   1.04e+03   -1.95  0.05063 .
## max_negative_polarity       -2.20e+03   7.24e+02   -3.04  0.00241 **
## channelentertainment        -1.44e+02   2.72e+02   -0.53  0.59721
## channellifestyle            -1.34e+02   3.41e+02   -0.39  0.69391
```

```
## channelother                1.59e+03   2.69e+02    5.90  3.6e-09 ***
## channelsocmed               3.79e+02   3.28e+02    1.16  0.24765
## channeltech                 2.28e+02   2.34e+02    0.98  0.32870
## channelworld                1.03e+02   3.40e+02    0.30  0.76105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10700 on 26901 degrees of freedom
## Multiple R-squared:  0.0212, Adjusted R-squared:  0.0204
## F-statistic: 27.7 on 21 and 26901 DF,  p-value: <2e-16
```

```
par(mfrow=c(1,2))
plot(lmnew_full_step,1:2)
```



refit the model

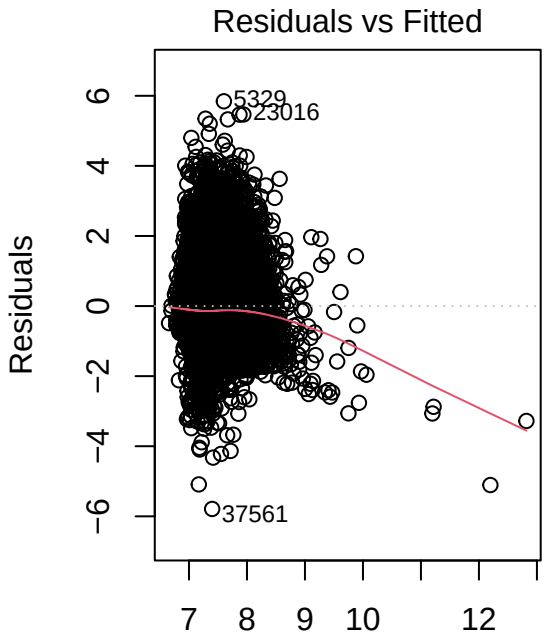
using log- transformation

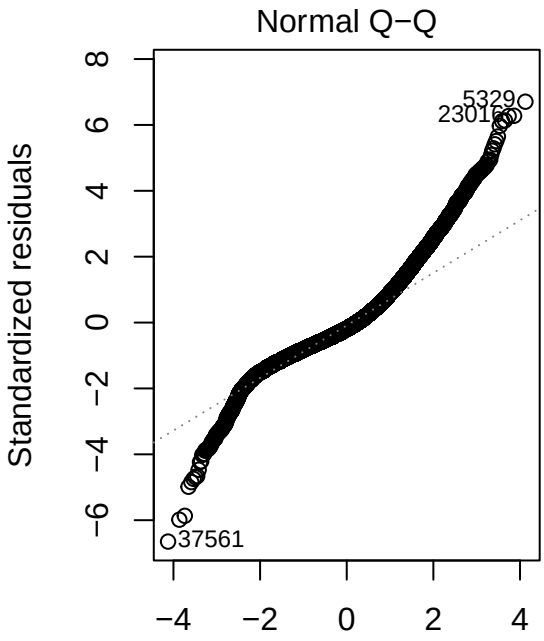
```
lmnew_full_log <- lm(log(shares)~.,data=newstrain)
summary(lmnew_full_log )
```

```
##
## Call:
## lm(formula = log(shares) ~ ., data = newstrain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.790 -0.544 -0.169  0.393  5.843
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.29e+00   1.52e-01   47.91  < 2e-16 ***
```

```
## n_tokens_title                 9.53e-04   2.61e-03    0.36  0.71523
## n_tokens_content               4.37e-05   1.80e-05    2.43  0.01493 *
## n_non_stop_unique_tokens      -1.31e-01   7.35e-02   -1.79  0.07378 .
## num_hrefs                      4.22e-03   6.15e-04    6.86  7.2e-12 ***
## num_self_hrefs                -9.66e-03   1.60e-03   -6.03  1.7e-09 ***
## num_imgs                       2.76e-03   8.23e-04    3.35  0.00080 ***
## num_videos                     3.00e-03   1.41e-03    2.12  0.03376 *
## average_token_length          -5.73e-02   2.11e-02   -2.72  0.00660 **
## num_keywords                   1.30e-02   3.39e-03    3.84  0.00012 ***
## kw_max_min                    -1.42e-05   1.59e-06   -8.89  < 2e-16 ***
## kw_min_max                    -4.64e-07   1.12e-07   -4.13  3.7e-05 ***
## kw_max_max                    -1.50e-07   3.51e-08   -4.27  1.9e-05 ***
## kw_avg_max                    -3.95e-08   7.37e-08   -0.54  0.59200
## kw_min_avg                     8.99e-06   5.99e-06    1.50  0.13328
## kw_avg_avg                     1.43e-04   6.74e-06   21.21  < 2e-16 ***
## self_reference_avg_sharess     1.92e-06   2.29e-07    8.37  < 2e-16 ***
## is_weekend                     2.70e-01   1.60e-02   16.92  < 2e-16 ***
## LDA_01                        -3.42e-01   4.86e-02   -7.04  2.0e-12 ***
## LDA_02                        -5.13e-01   4.56e-02  -11.25  < 2e-16 ***
## LDA_03                        -2.87e-01   4.63e-02   -6.21  5.4e-10 ***
## LDA_04                        -2.06e-01   4.18e-02   -4.93  8.1e-07 ***
## global_subjectivity            4.40e-01   7.69e-02    5.72  1.1e-08 ***
## global_sentiment_polarity     -1.12e-01   1.52e-01   -0.74  0.45984
## global_rate_positive_words    -5.79e-01   6.49e-01   -0.89  0.37221
## global_rate_negative_words    -1.56e+00   1.25e+00   -1.25  0.21073
## rate_positive_words            1.66e-02   1.02e-01    0.16  0.86988
## avg_positive_polarity          3.42e-02   1.23e-01    0.28  0.78162
## min_positive_polarity         -2.93e-01   1.02e-01   -2.87  0.00412 **
## max_positive_polarity         -1.29e-02   3.86e-02   -0.33  0.73809
## avg_negative_polarity         -1.65e-01   1.14e-01   -1.45  0.14769
## min_negative_polarity          2.24e-02   4.13e-02    0.54  0.58750
## max_negative_polarity          3.53e-02   9.42e-02    0.37  0.70782
## title_subjectivity_dis         5.36e-02   1.75e-02    3.06  0.00221 **
## title_sentiment_polarity_dis   3.83e-02   2.17e-02    1.76  0.07774 .
## channelentertainment           1.33e-02   3.35e-02    0.40  0.69064
## channellifestyle               9.71e-02   3.37e-02    2.88  0.00396 **
## channelother                   2.86e-01   3.49e-02    8.18  2.9e-16 ***
## channelsocmed                  3.73e-01   2.86e-02   13.04  < 2e-16 ***
## channeltech                    3.01e-01   3.03e-02    9.93  < 2e-16 ***
## channelworld                   1.33e-01   3.27e-02    4.07  4.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 26882 degrees of freedom
## Multiple R-squared:  0.118,  Adjusted R-squared:  0.116
## F-statistic: 89.7 on 40 and 26882 DF,  p-value: <2e-16
```

```
par(mfrow=c(1,2))
plot(lmnew_full_log,1:2)
```






```
table(newstrain$channel)
```

```
##
##      business entertainment     lifestyle         other        socmed
##          4332          4810          1447          3864          1606
##          tech         world
##          5147          5717
```

```
dim(newstrain)
```

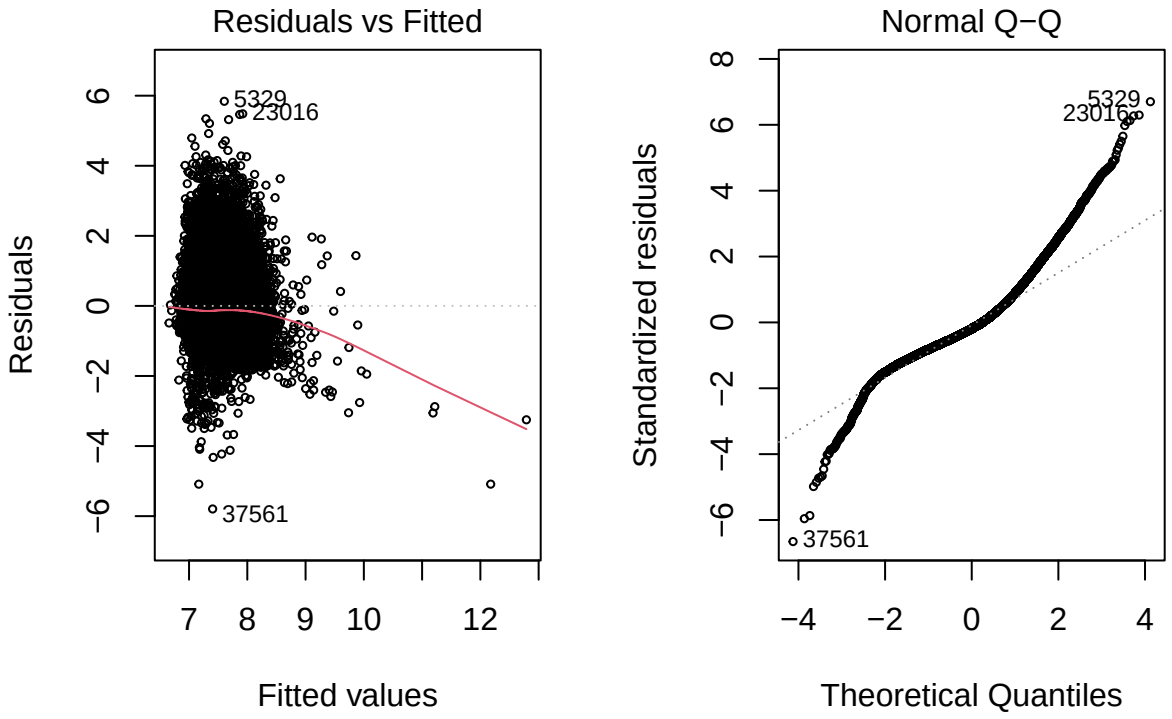
```
## [1] 26923    36
```

```
lmnew_full_logs <- step(lmnew_full_log,trace=F)
summary(lmnew_full_logs)
```

```
##
## Call:
## lm(formula = log(shares) ~ n_tokens_content + n_non_stop_unique_tokens +
##     num_hrefs + num_self_hrefs + num_imgs + num_videos + average_token_length +
##     num_keywords + kw_max_min + kw_min_max + kw_max_max + kw_min_avg +
##     kw_avg_avg + self_reference_avg_sharess + is_weekend + LDA_01 +
##     LDA_02 + LDA_03 + LDA_04 + global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + min_positive_polarity + avg_negative_polarity +
##     title_subjectivity_dis + title_sentiment_polarity_dis + channel,
##     data = newstrain)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.795 -0.543 -0.169  0.393  5.840
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.30e+00   1.28e-01   57.16  < 2e-16 ***
```

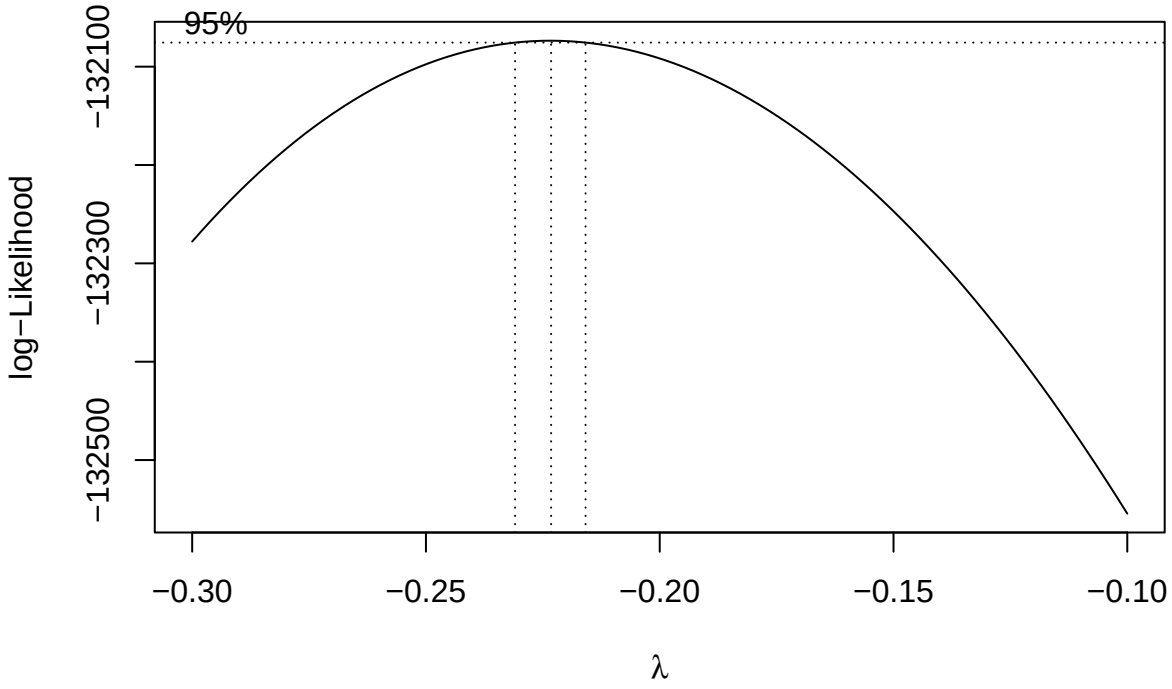
```
## n_tokens_content              3.98e-05   1.62e-05    2.45  0.01417 *
## n_non_stop_unique_tokens     -1.30e-01   7.30e-02   -1.78  0.07545 .
## num_hrefs                     4.18e-03   6.10e-04    6.85  7.4e-12 ***
## num_self_hrefs               -9.63e-03   1.60e-03   -6.01  1.8e-09 ***
## num_imgs                      2.77e-03   8.20e-04    3.38  0.00074 ***
## num_videos                    2.92e-03   1.39e-03    2.10  0.03593 *
## average_token_length         -5.68e-02   2.09e-02   -2.71  0.00663 **
## num_keywords                  1.36e-02   3.17e-03    4.29  1.8e-05 ***
## kw_max_min                   -1.41e-05   1.58e-06   -8.90  < 2e-16 ***
## kw_min_max                   -4.86e-07   1.05e-07   -4.63  3.6e-06 ***
## kw_max_max                   -1.61e-07   2.63e-08   -6.13  8.7e-10 ***
## kw_min_avg                    8.69e-06   5.96e-06    1.46  0.14525
## kw_avg_avg                    1.42e-04   6.60e-06   21.54  < 2e-16 ***
## self_reference_avg_sharess    1.92e-06   2.29e-07    8.36  < 2e-16 ***
## is_weekend                    2.70e-01   1.59e-02   16.96  < 2e-16 ***
## LDA_01                       -3.41e-01   4.85e-02   -7.02  2.2e-12 ***
## LDA_02                       -5.13e-01   4.55e-02  -11.27  < 2e-16 ***
## LDA_03                       -2.90e-01   4.61e-02   -6.28  3.4e-10 ***
## LDA_04                       -2.04e-01   4.17e-02   -4.90  9.5e-07 ***
## global_subjectivity           4.21e-01   7.19e-02    5.86  4.8e-09 ***
## global_rate_positive_words   -8.49e-01   3.85e-01   -2.21  0.02736 *
## global_rate_negative_words   -1.26e+00   5.51e-01   -2.28  0.02251 *
## min_positive_polarity        -2.93e-01   8.56e-02   -3.43  0.00061 ***
## avg_negative_polarity        -1.34e-01   4.89e-02   -2.73  0.00634 **
## title_subjectivity_dis        5.51e-02   1.74e-02    3.16  0.00158 **
## title_sentiment_polarity_dis  3.52e-02   2.15e-02    1.64  0.10106
## channelentertainment          1.68e-02   3.29e-02    0.51  0.61014
## channellifestyle              9.86e-02   3.31e-02    2.97  0.00294 **
## channelother                  2.86e-01   3.49e-02    8.20  2.6e-16 ***
## channelsocmed                 3.76e-01   2.77e-02   13.59  < 2e-16 ***
## channeltech                   3.03e-01   3.01e-02   10.09  < 2e-16 ***
## channelworld                  1.36e-01   3.21e-02    4.22  2.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 26890 degrees of freedom
## Multiple R-squared:  0.118,  Adjusted R-squared:  0.117
## F-statistic:  112 on 32 and 26890 DF,  p-value: <2e-16
```

```
par(mfrow=c(1,2))
plot(lmnew_full_logs,1:2,cex=0.5)
```



using boxcox

```
options(scipen=0,digits = 3)
library(MASS)
library(car)
boxcox(lmnew_full, plotit = TRUE, lambda = seq(-0.1, -0.3, by = -0.05))
```




```
summary(powerTransform(lmnew_full))
```

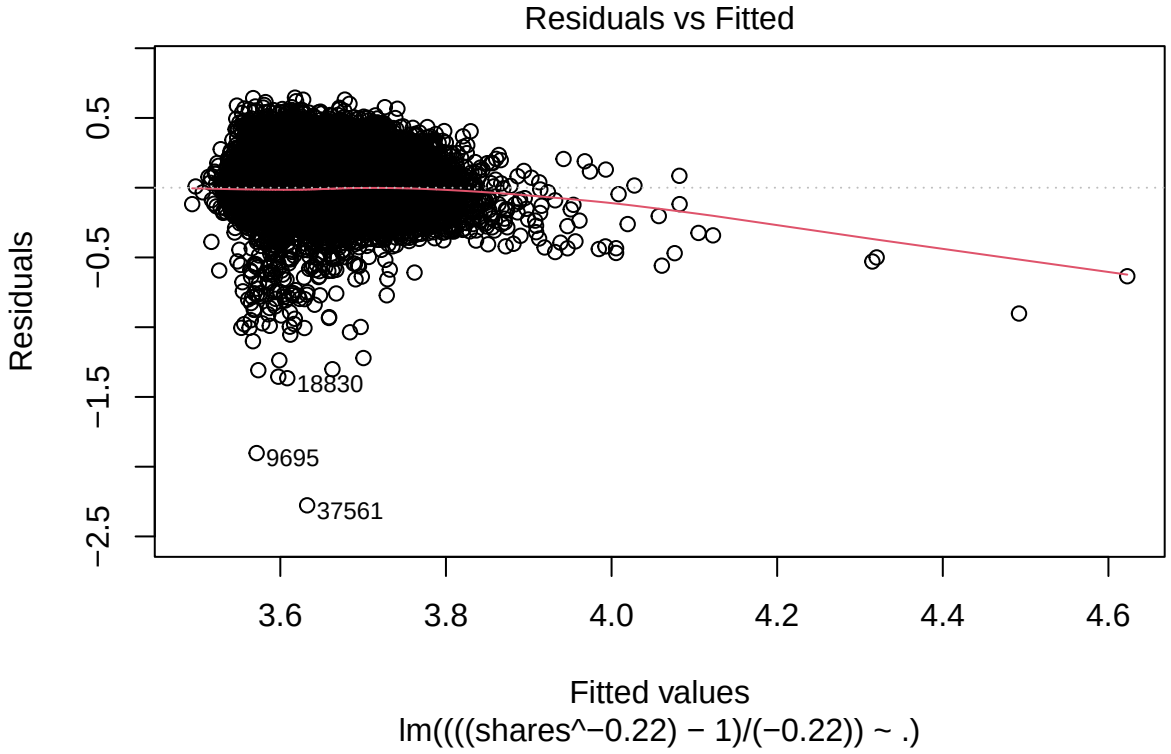
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    -0.223       -0.22       -0.231       -0.216
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                          LRT df   pval
## LR test, lambda = (0) 3062  1 <2e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df   pval
## LR test, lambda = (1) 108118  1 <2e-16
```

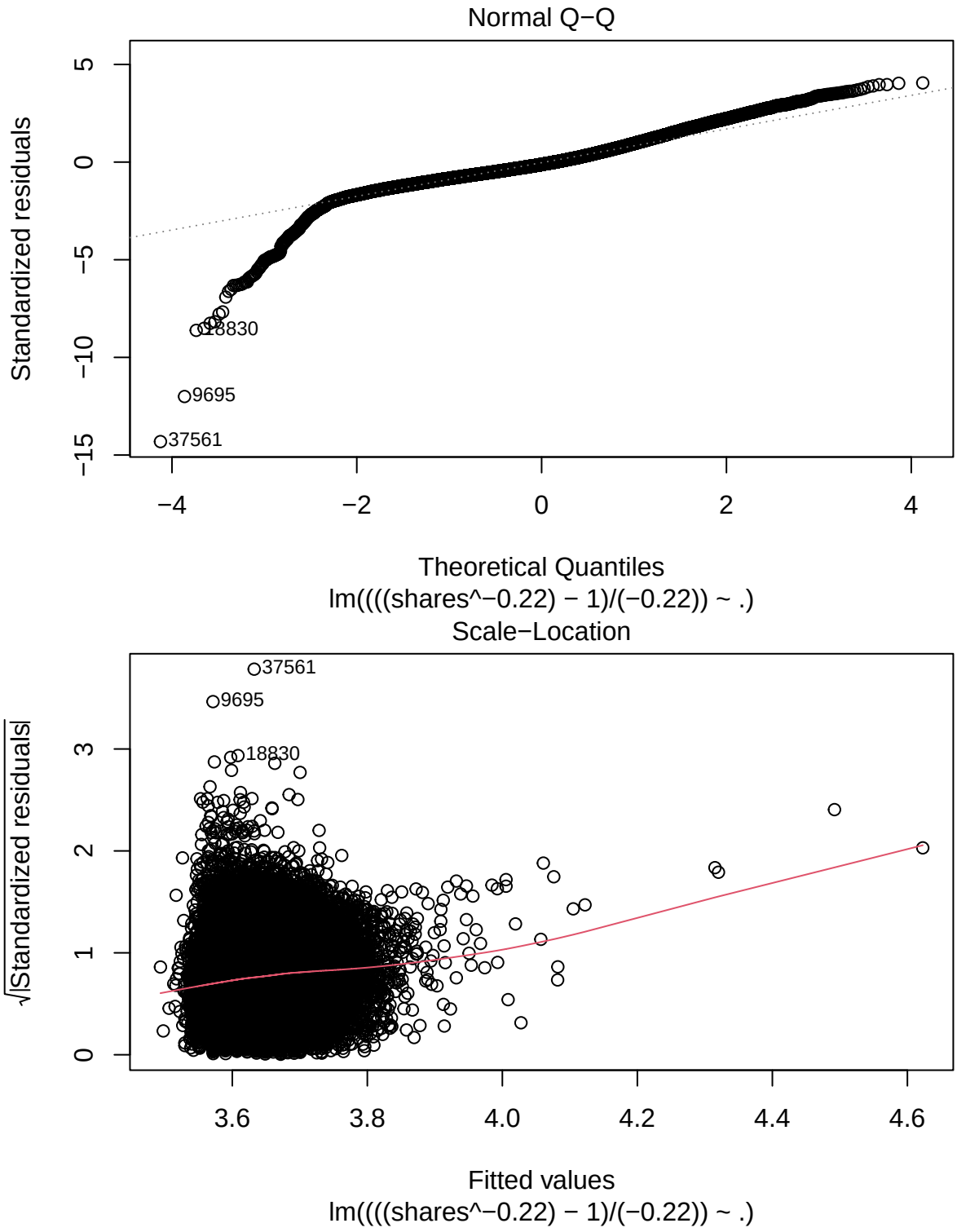
```
lmnew_full_t <- lm((((shares ^ -0.22) - 1) / (-0.22))~.,data=newstrain)
summary(lmnew_full_t )
```

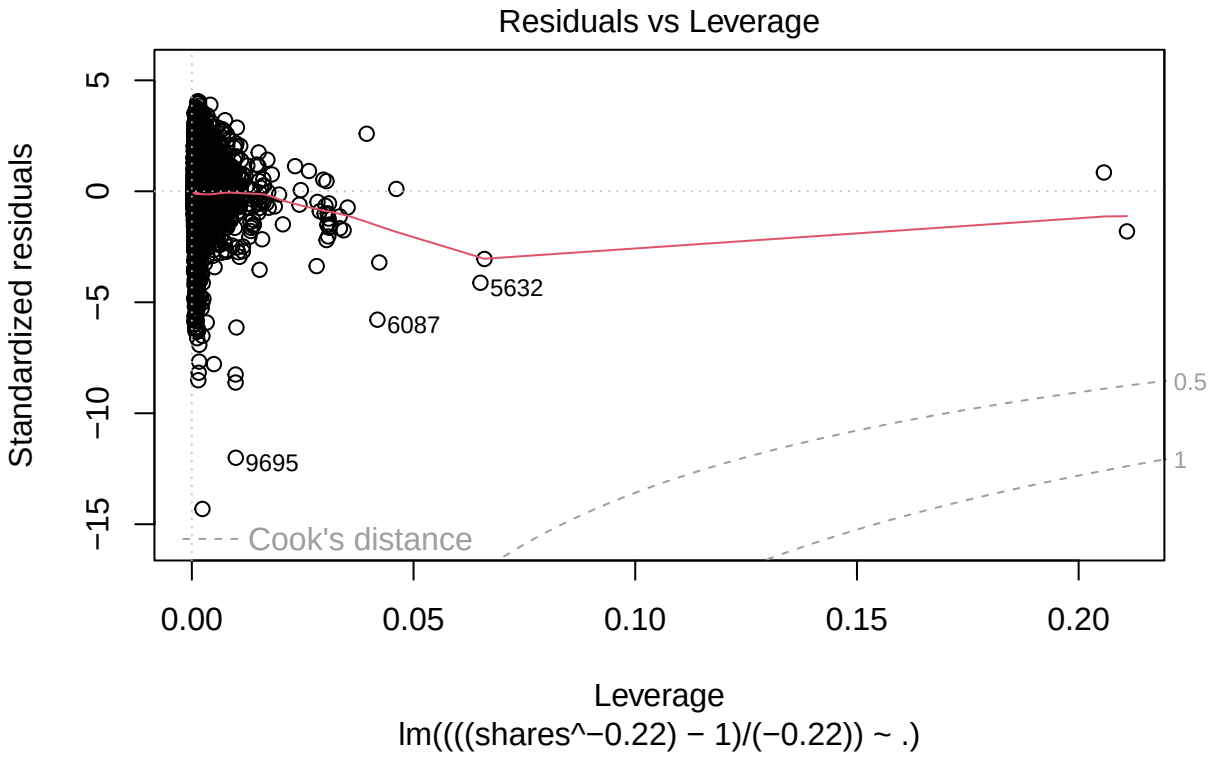
```
##
## Call:
## lm(formula = (((shares^-0.22) - 1)/(-0.22)) ~ ., data = newstrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2770 -0.0972 -0.0191  0.0878  0.6455
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.63e+00   2.78e-02  130.53  < 2e-16 ***
## n_tokens_title                2.25e-04   4.77e-04    0.47   0.6373
## n_tokens_content              6.49e-06   3.28e-06    1.98   0.0480 *
## n_non_stop_unique_tokens     -3.68e-02   1.34e-02   -2.74   0.0062 **
## num_hrefs                     7.64e-04   1.12e-04    6.80  1.1e-11 ***
## num_self_hrefs               -1.49e-03   2.93e-04   -5.10  3.4e-07 ***
## num_imgs                      4.77e-04   1.50e-04    3.17   0.0015 **
## num_videos                    5.94e-04   2.58e-04    2.30   0.0214 *
## average_token_length         -1.10e-02   3.85e-03   -2.84   0.0044 **
## num_keywords                  2.60e-03   6.20e-04    4.19  2.7e-05 ***
## kw_max_min                   -2.63e-06   2.91e-07   -9.04  < 2e-16 ***
## kw_min_max                   -1.17e-07   2.05e-08   -5.70  1.2e-08 ***
## kw_max_max                   -2.92e-08   6.41e-09   -4.55  5.4e-06 ***
## kw_avg_max                   -4.03e-09   1.35e-08   -0.30   0.7646
## kw_min_avg                    2.86e-06   1.09e-06    2.62   0.0089 **
## kw_avg_avg                    2.55e-05   1.23e-06   20.74  < 2e-16 ***
## self_reference_avg_sharess    3.39e-07   4.19e-08    8.11  5.4e-16 ***
## is_weekend                    5.43e-02   2.92e-03   18.61  < 2e-16 ***
## LDA_01                       -7.07e-02   8.87e-03   -7.97  1.7e-15 ***
## LDA_02                       -9.95e-02   8.33e-03  -11.94  < 2e-16 ***
## LDA_03                       -6.03e-02   8.46e-03   -7.13  1.0e-12 ***
## LDA_04                       -3.98e-02   7.63e-03   -5.21  1.9e-07 ***
## global_subjectivity           7.57e-02   1.41e-02    5.39  7.1e-08 ***
## global_sentiment_polarity    -2.60e-02   2.77e-02   -0.94   0.3490
## global_rate_positive_words   -1.02e-01   1.19e-01   -0.86   0.3903
## global_rate_negative_words   -2.29e-01   2.28e-01   -1.00   0.3163
## rate_positive_words           1.25e-02   1.86e-02    0.67   0.5021
## avg_positive_polarity         8.12e-03   2.25e-02    0.36   0.7182
## min_positive_polarity        -5.84e-02   1.87e-02   -3.12   0.0018 **
```

```
## max_positive_polarity        -3.71e-03   7.05e-03   -0.53   0.5985
## avg_negative_polarity        -2.82e-02   2.08e-02   -1.36   0.1750
## min_negative_polarity         3.81e-03   7.55e-03    0.51   0.6134
## max_negative_polarity         8.02e-03   1.72e-02    0.47   0.6413
## title_subjectivity_dis        9.91e-03   3.20e-03    3.09   0.0020 **
## title_sentiment_polarity_dis  6.97e-03   3.96e-03    1.76   0.0785 .
## channelentertainment          2.34e-03   6.12e-03    0.38   0.7027
## channellifestyle              1.77e-02   6.16e-03    2.87   0.0041 **
## channelother                  5.06e-02   6.38e-03    7.93  2.3e-15 ***
## channelsocmed                 7.06e-02   5.23e-03   13.51  < 2e-16 ***
## channeltech                   5.78e-02   5.54e-03   10.43  < 2e-16 ***
## channelworld                  2.38e-02   5.97e-03    3.98  6.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.159 on 26882 degrees of freedom
## Multiple R-squared:  0.122,  Adjusted R-squared:  0.12
## F-statistic: 93.1 on 40 and 26882 DF,  p-value: <2e-16
```

```
plot(lmnew_full_t)
```




Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
18830
9695
37561
−4
−2
0
2
4
Theoretical Quantiles
lm((((shares^−0.22) − 1)/(−0.22)) ~ .)
Scale−Location
√|Standardized residuals|
37561
9695
18830
3
2
1
0
3.6
3.8
4.0
4.2
4.4
4.6
Fitted values
lm((((shares^−0.22) − 1)/(−0.22)) ~ .)



improve the model

```
dim(newstrain)
```

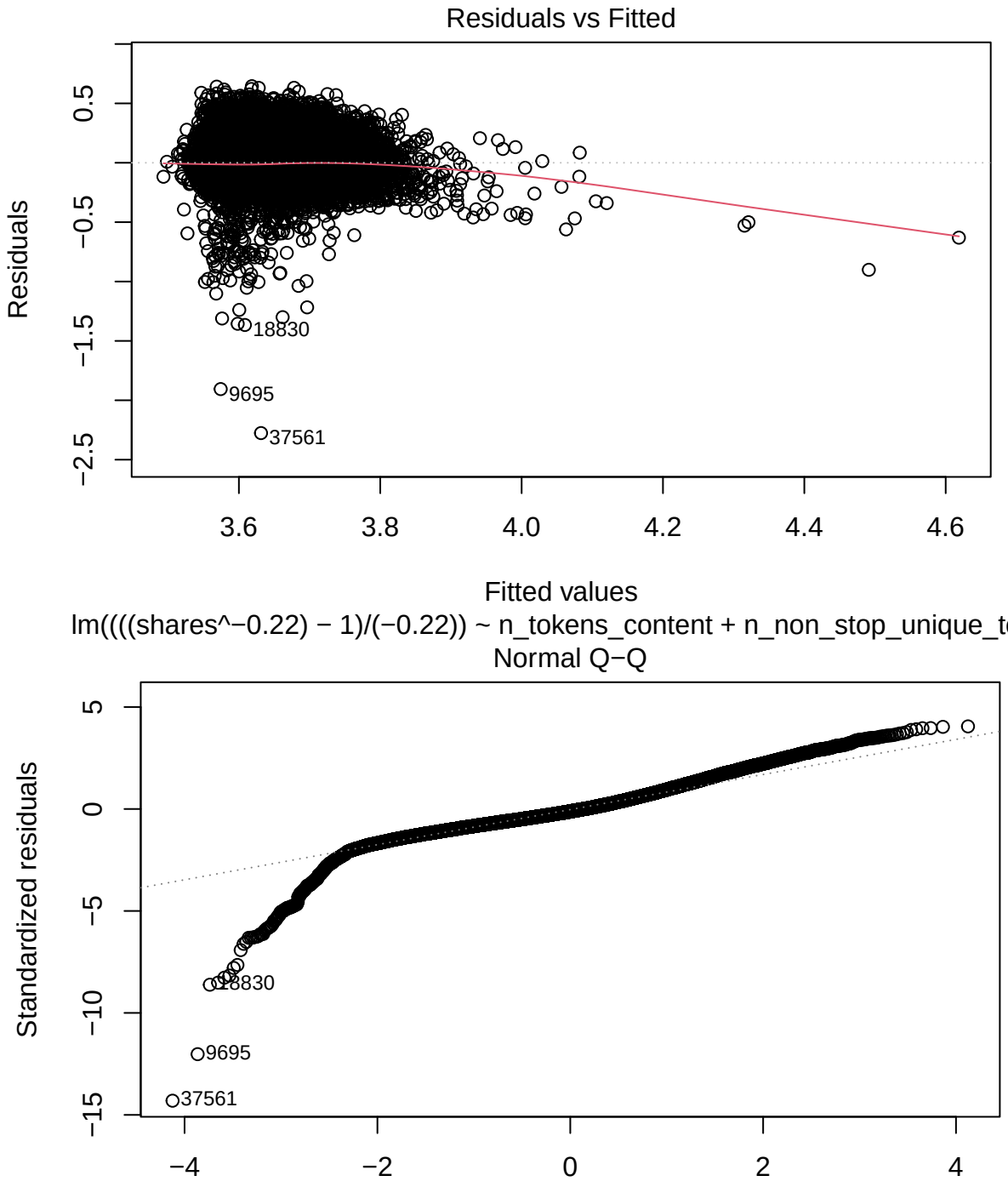
```
## [1] 26923    36
```

```
lmnew_full_ts <- step(lmnew_full_t,trace=F)
summary(lmnew_full_ts)
```

```
##
## Call:
## lm(formula = (((shares^-0.22) - 1)/(-0.22)) ~ n_tokens_content +
##     n_non_stop_unique_tokens + num_hrefs + num_self_hrefs + num_imgs +
##     num_videos + average_token_length + num_keywords + kw_max_min +
##     kw_min_max + kw_max_max + kw_min_avg + kw_avg_avg + self_reference_avg_sharess +
##     is_weekend + LDA_01 + LDA_02 + LDA_03 + LDA_04 + global_subjectivity +
##     global_sentiment_polarity + global_rate_negative_words +
##     min_positive_polarity + avg_negative_polarity + title_subjectivity_dis +
##     title_sentiment_polarity_dis + channel, data = newstrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2759 -0.0971 -0.0190  0.0877  0.6448
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.64e+00   2.33e-02  155.93  < 2e-16 ***
## n_tokens_content              5.66e-06   2.96e-06    1.91  0.05625 .
## n_non_stop_unique_tokens     -3.75e-02   1.33e-02   -2.83  0.00468 **
## num_hrefs                     7.65e-04   1.11e-04    6.89  5.9e-12 ***
```

```
## num_self_hrefs               -1.50e-03   2.92e-04   -5.13  2.9e-07 ***
## num_imgs                      4.89e-04   1.50e-04    3.27  0.00109 **
## num_videos                    6.02e-04   2.55e-04    2.36  0.01804 *
## average_token_length         -1.10e-02   3.83e-03   -2.87  0.00416 **
## num_keywords                  2.66e-03   5.80e-04    4.58  4.6e-06 ***
## kw_max_min                   -2.62e-06   2.89e-07   -9.08  < 2e-16 ***
## kw_min_max                   -1.19e-07   1.92e-08   -6.23  4.8e-10 ***
## kw_max_max                   -3.01e-08   4.80e-09   -6.27  3.8e-10 ***
## kw_min_avg                    2.82e-06   1.09e-06    2.59  0.00968 **
## kw_avg_avg                    2.55e-05   1.21e-06   21.11  < 2e-16 ***
## self_reference_avg_sharess    3.41e-07   4.18e-08    8.14  4.0e-16 ***
## is_weekend                    5.42e-02   2.91e-03   18.62  < 2e-16 ***
## LDA_01                       -7.04e-02   8.86e-03   -7.94  2.0e-15 ***
## LDA_02                       -9.93e-02   8.31e-03  -11.94  < 2e-16 ***
## LDA_03                       -6.01e-02   8.42e-03   -7.14  9.7e-13 ***
## LDA_04                       -3.96e-02   7.62e-03   -5.19  2.1e-07 ***
## global_subjectivity           7.69e-02   1.38e-02    5.56  2.8e-08 ***
## global_sentiment_polarity    -3.06e-02   1.51e-02   -2.02  0.04368 *
## global_rate_negative_words   -3.85e-01   1.23e-01   -3.12  0.00179 **
## min_positive_polarity        -5.18e-02   1.50e-02   -3.45  0.00057 ***
## avg_negative_polarity        -1.69e-02   9.62e-03   -1.75  0.07952 .
## title_subjectivity_dis        9.88e-03   3.18e-03    3.11  0.00187 **
## title_sentiment_polarity_dis  6.77e-03   3.95e-03    1.71  0.08710 .
## channelentertainment          2.79e-03   6.01e-03    0.46  0.64233
## channellifestyle              1.79e-02   6.06e-03    2.95  0.00316 **
## channelother                  5.04e-02   6.37e-03    7.91  2.6e-15 ***
## channelsocmed                 7.08e-02   5.06e-03   13.99  < 2e-16 ***
## channeltech                   5.80e-02   5.49e-03   10.56  < 2e-16 ***
## channelworld                  2.40e-02   5.87e-03    4.10  4.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.159 on 26890 degrees of freedom
## Multiple R-squared:  0.122,  Adjusted R-squared:  0.121
## F-statistic:  116 on 32 and 26890 DF,  p-value: <2e-16
```

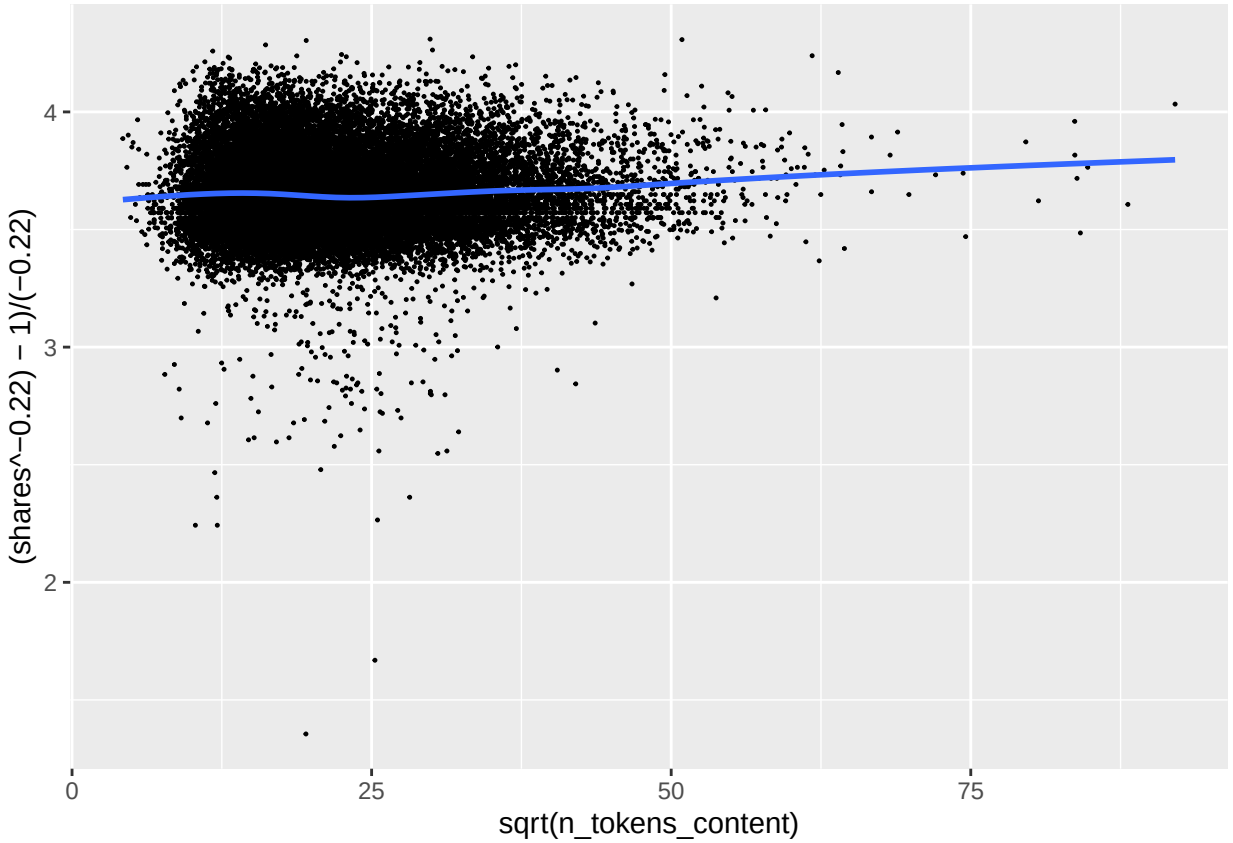
```
#par(mfrow=c(1,2))
plot(lmnew_full_ts,1:2)
```




```
plotpoint <- function(z,x,y) {
  ggplot(z,aes(x,y))+
  geom_point(cex=0.2)+
  geom_smooth(se=F)
}
```

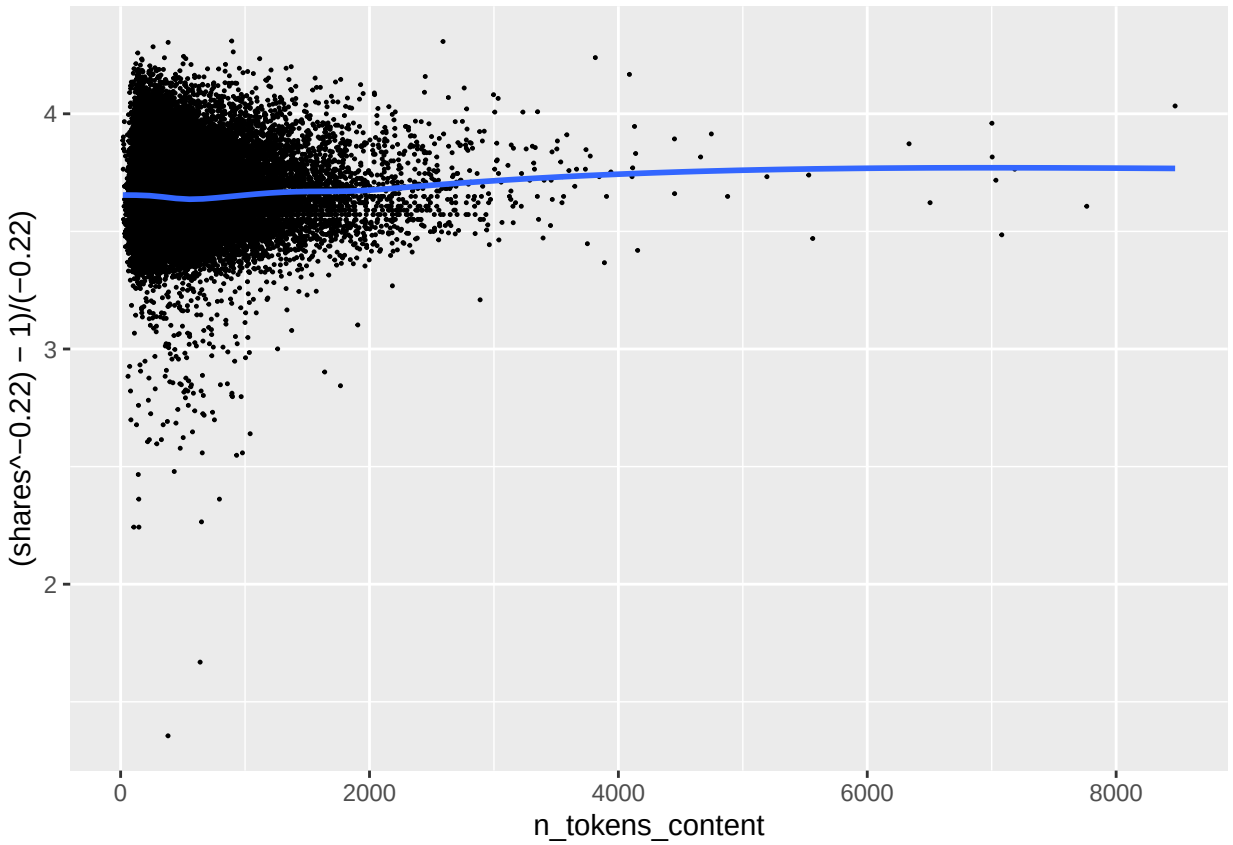
```
par(mfrow=c(1,2))
plotpoint(newstrain,sqrt(newstrain$n_tokens_content),((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("sq
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



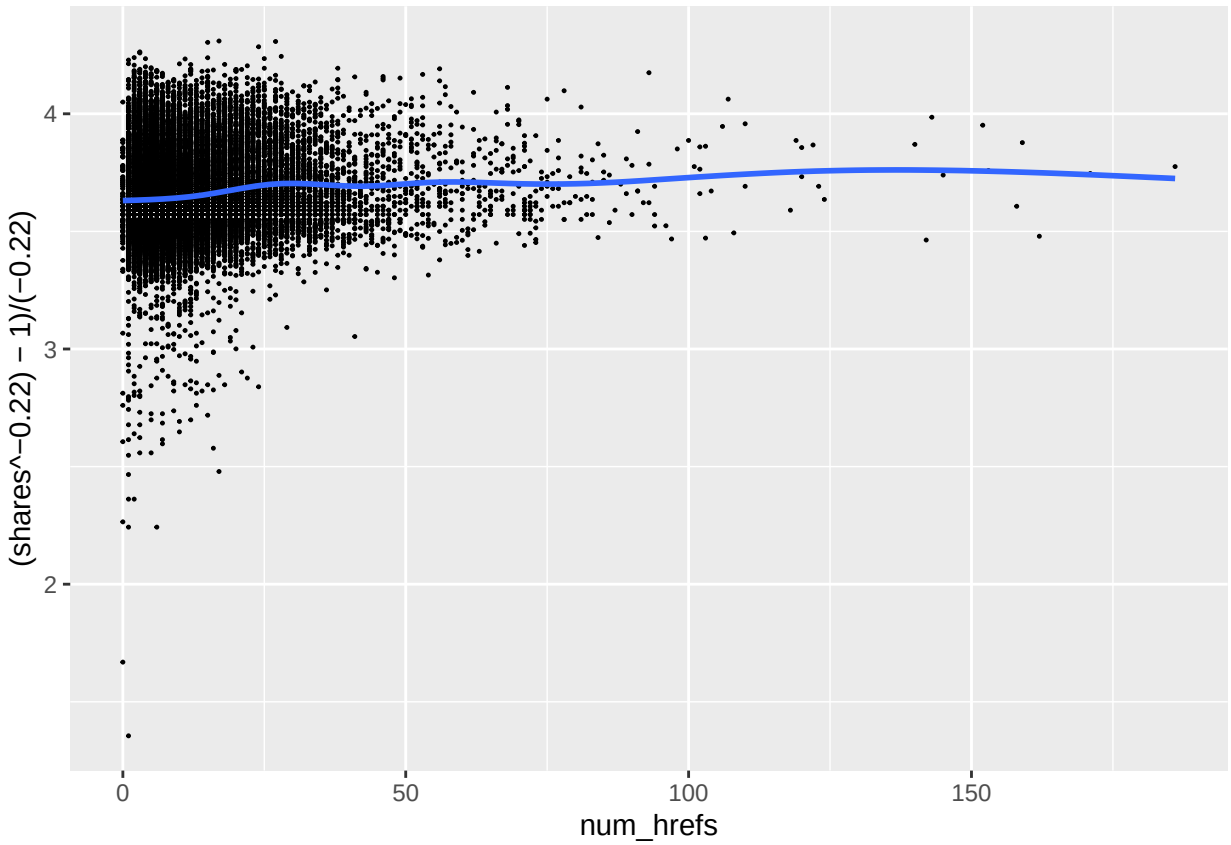

```
plotpoint(newstrain,newstrain$n_tokens_content,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("n_tokens
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



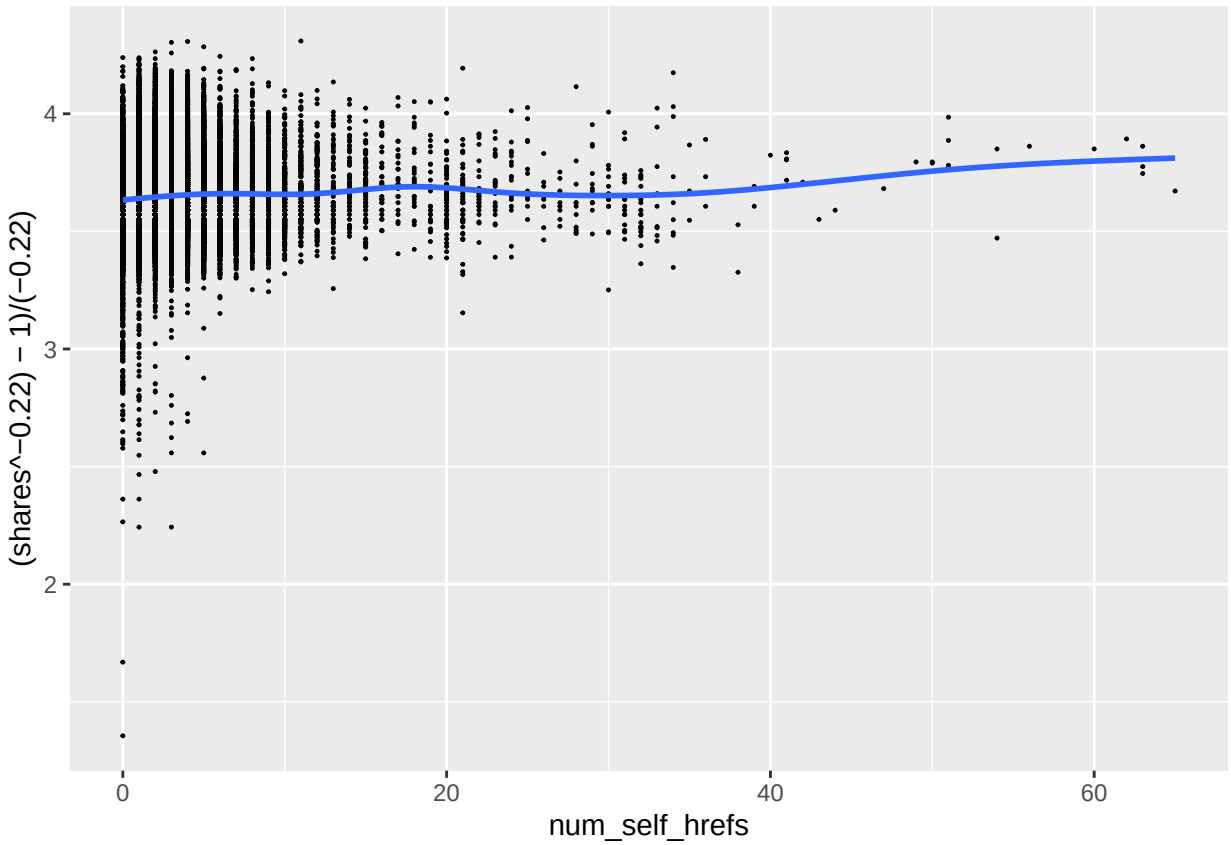

```
plotpoint(newstrain,newstrain$num_hrefs,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("num_hrefs")+ylab
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



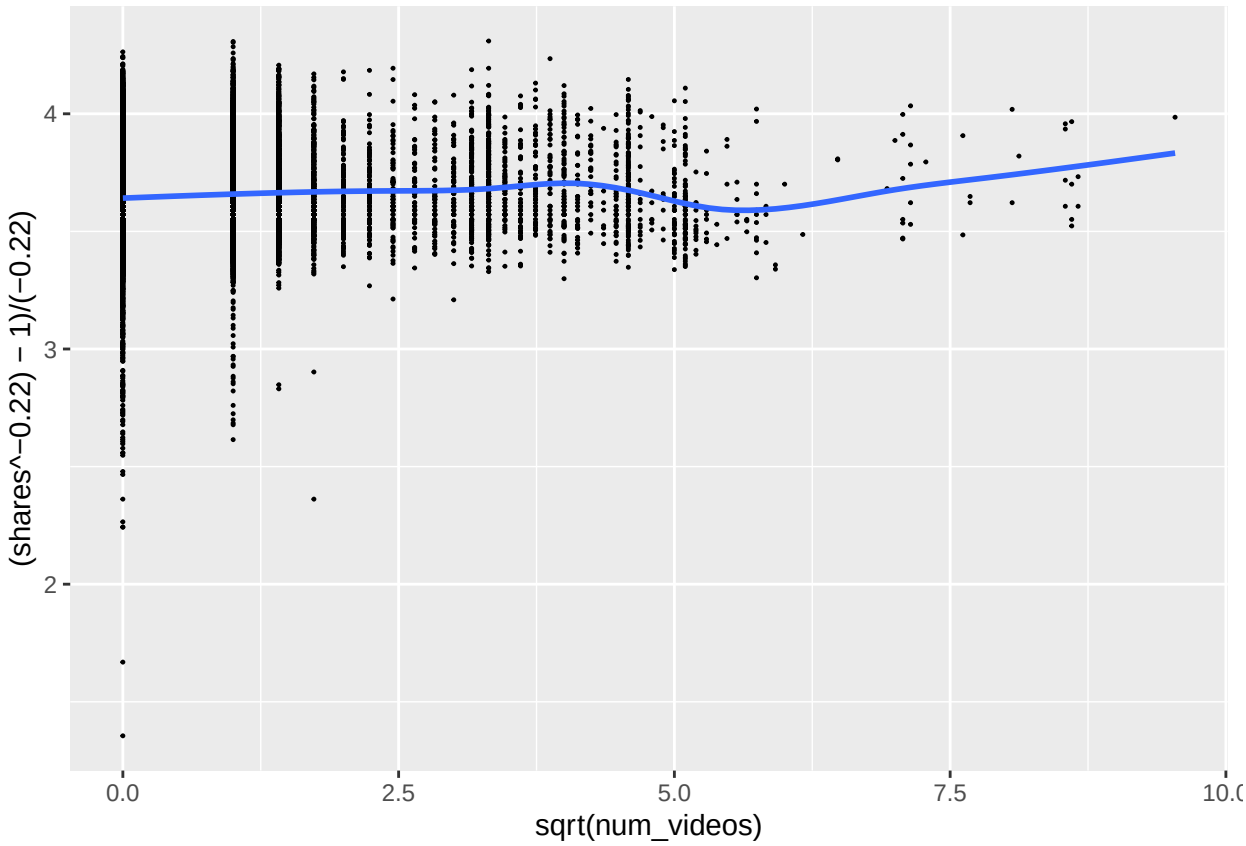

```
plotpoint(newstrain,newstrain$num_self_hrefs,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("num_self_hr
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



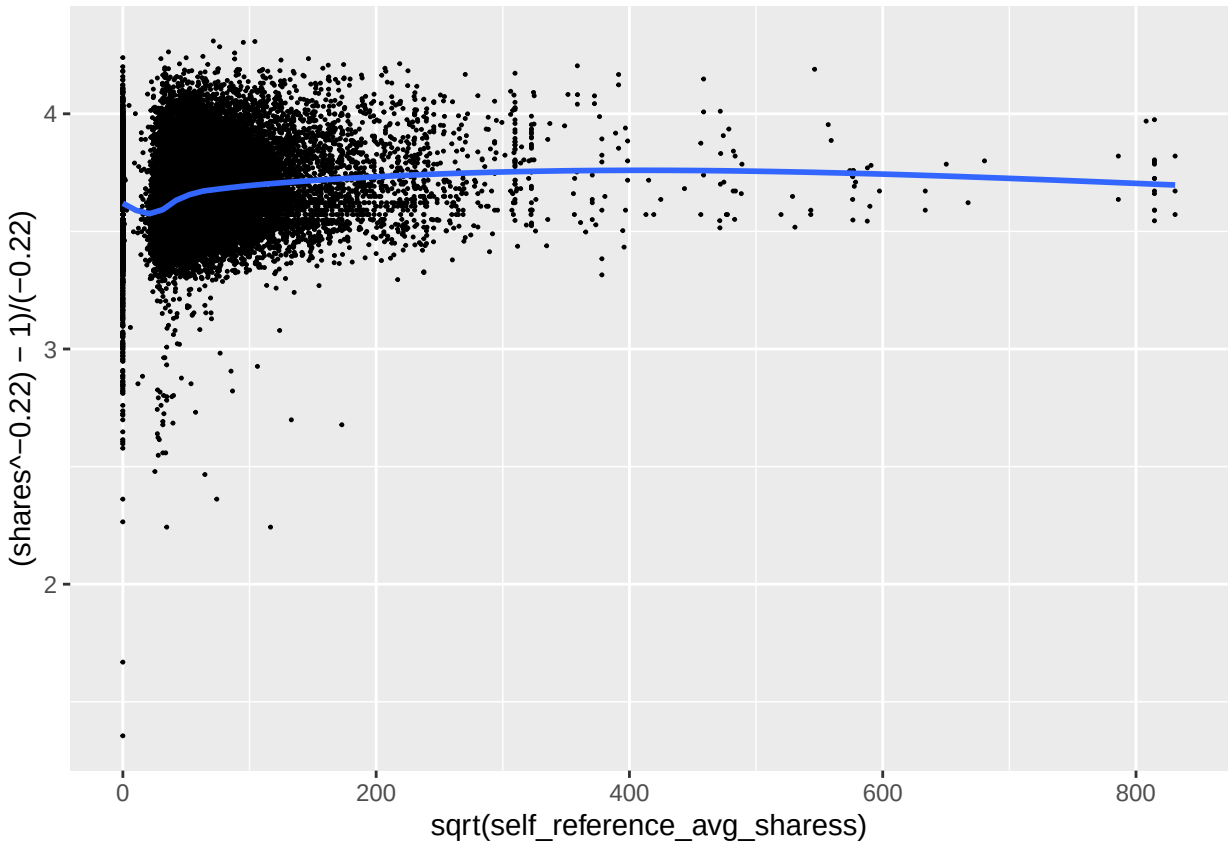

```
plotpoint(newstrain,sqrt(newstrain$num_videos),((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("sqrt(num
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



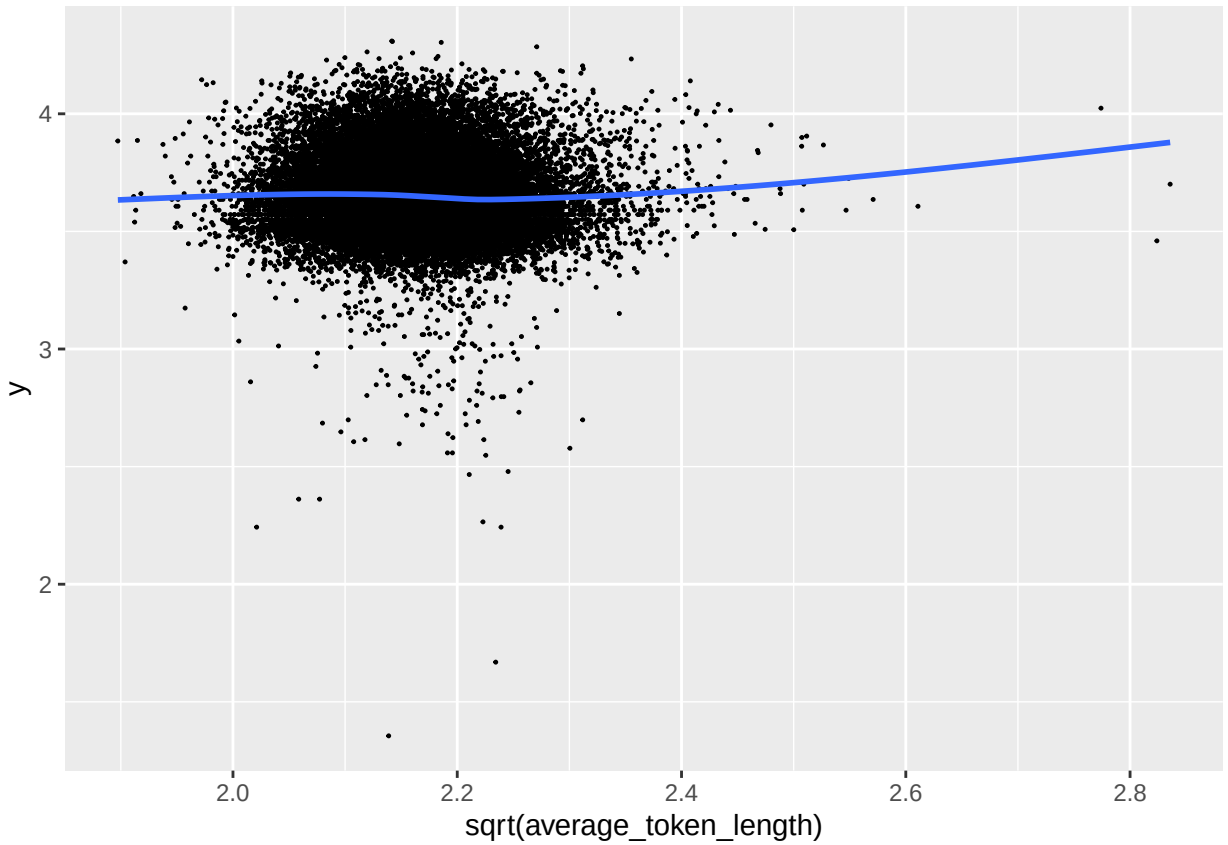

```
plotpoint(newstrain,sqrt(newstrain$self_reference_avg_sharess),((newstrain$shares^-0.22) - 1)/(-0.22)) +
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
plotpoint(newstrain,sqrt(newstrain$average_token_length),((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab
```

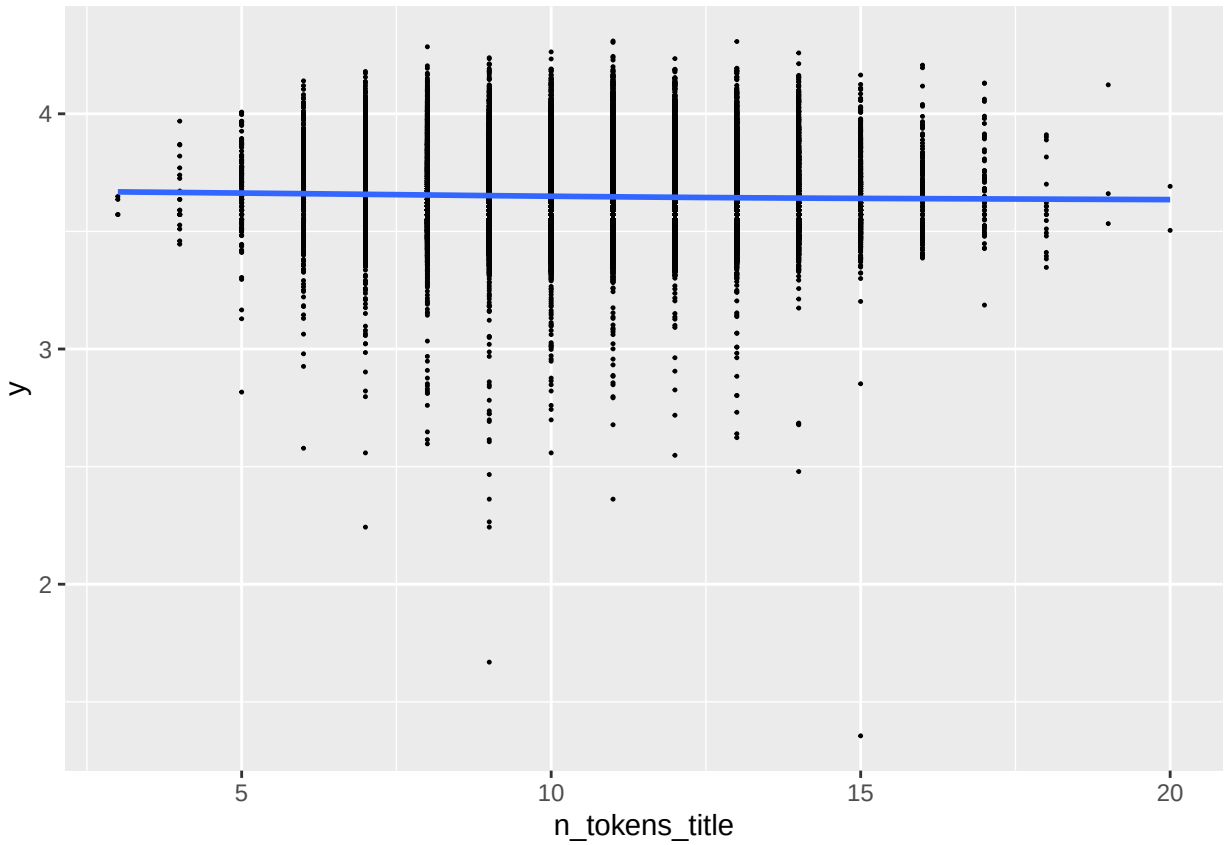
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



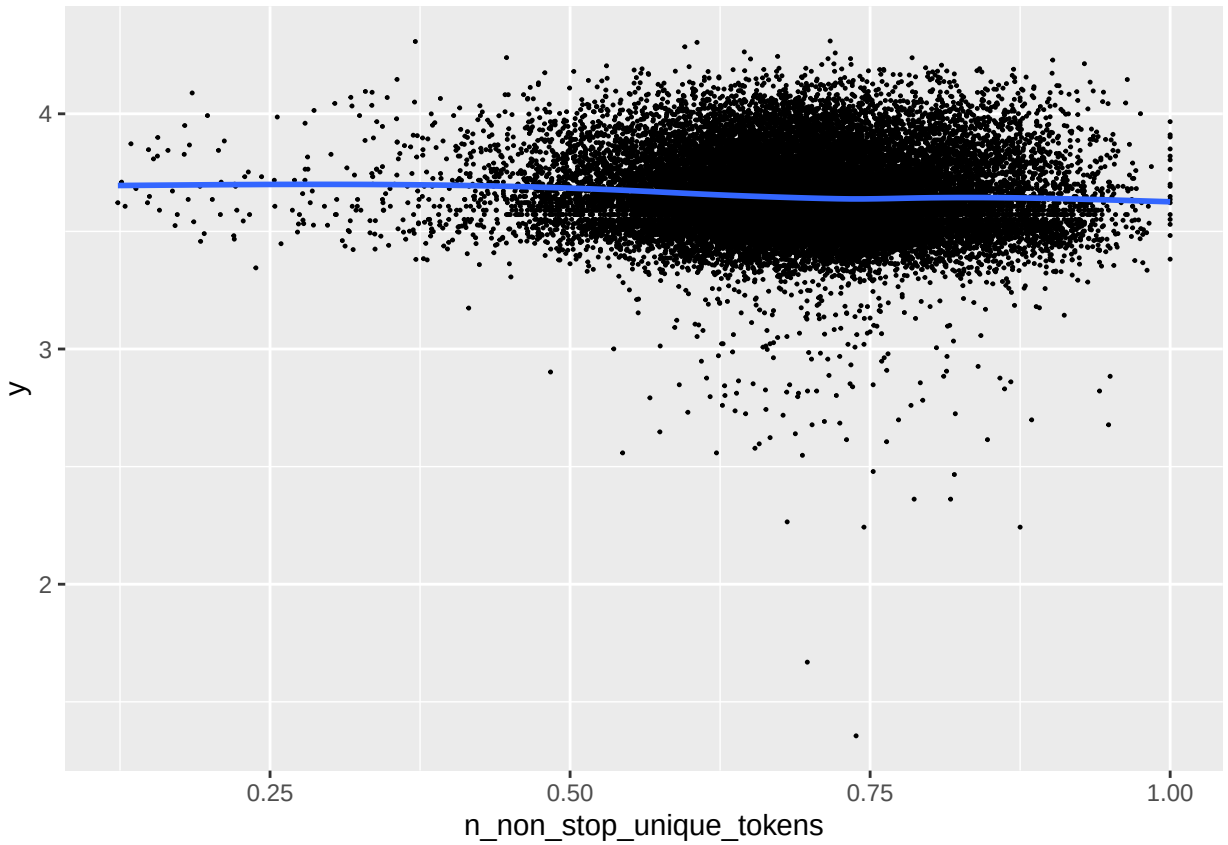

```
## it seems there is no relationship between n_tokens_title and the transformer response variable, the
```

```
plotpoint(newstrain,newstrain$n_tokens_title,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("n_tokens_t
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

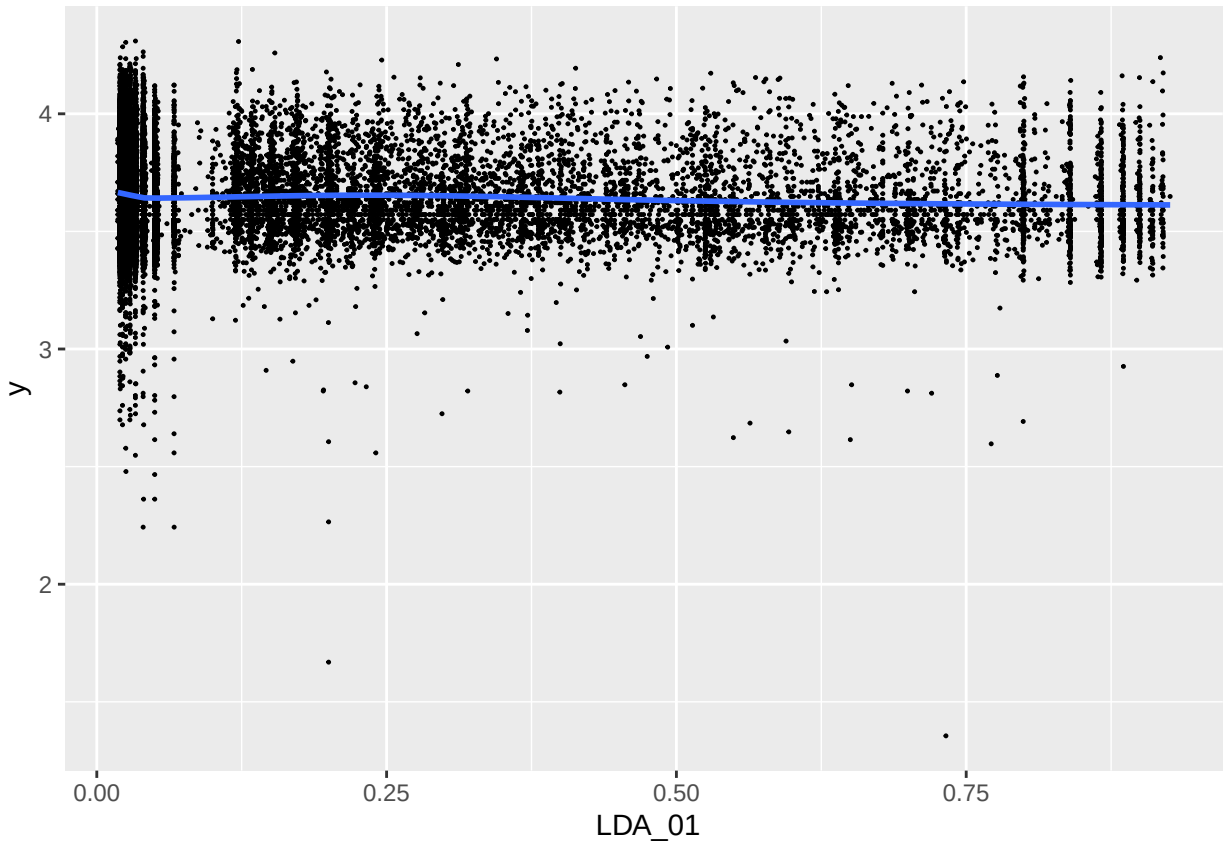



```
# these following variables have linear relationship with our transformed response variable.

plotpoint(newstrain,newstrain$n_non_stop_unique_tokens,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



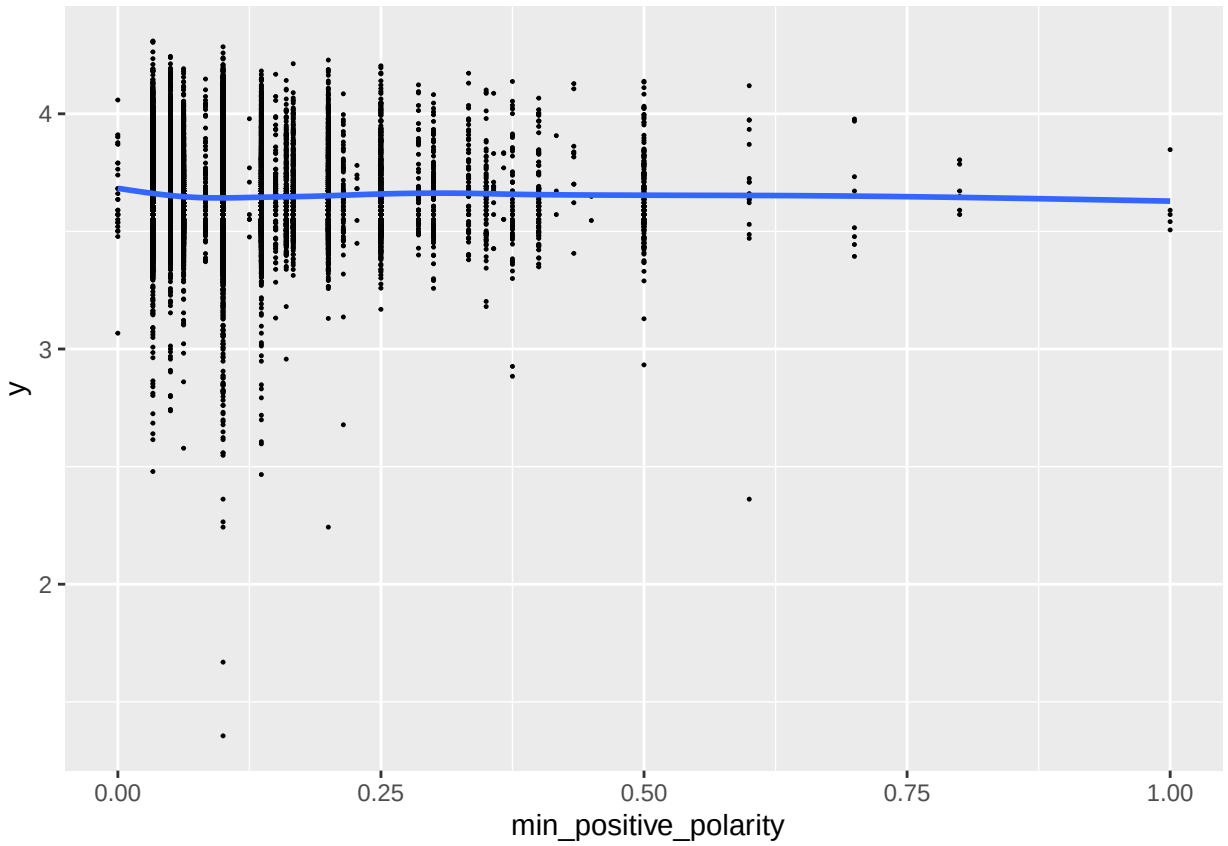

```
plotpoint(newstrain,newstrain$LDA_01,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab(" LDA_01")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



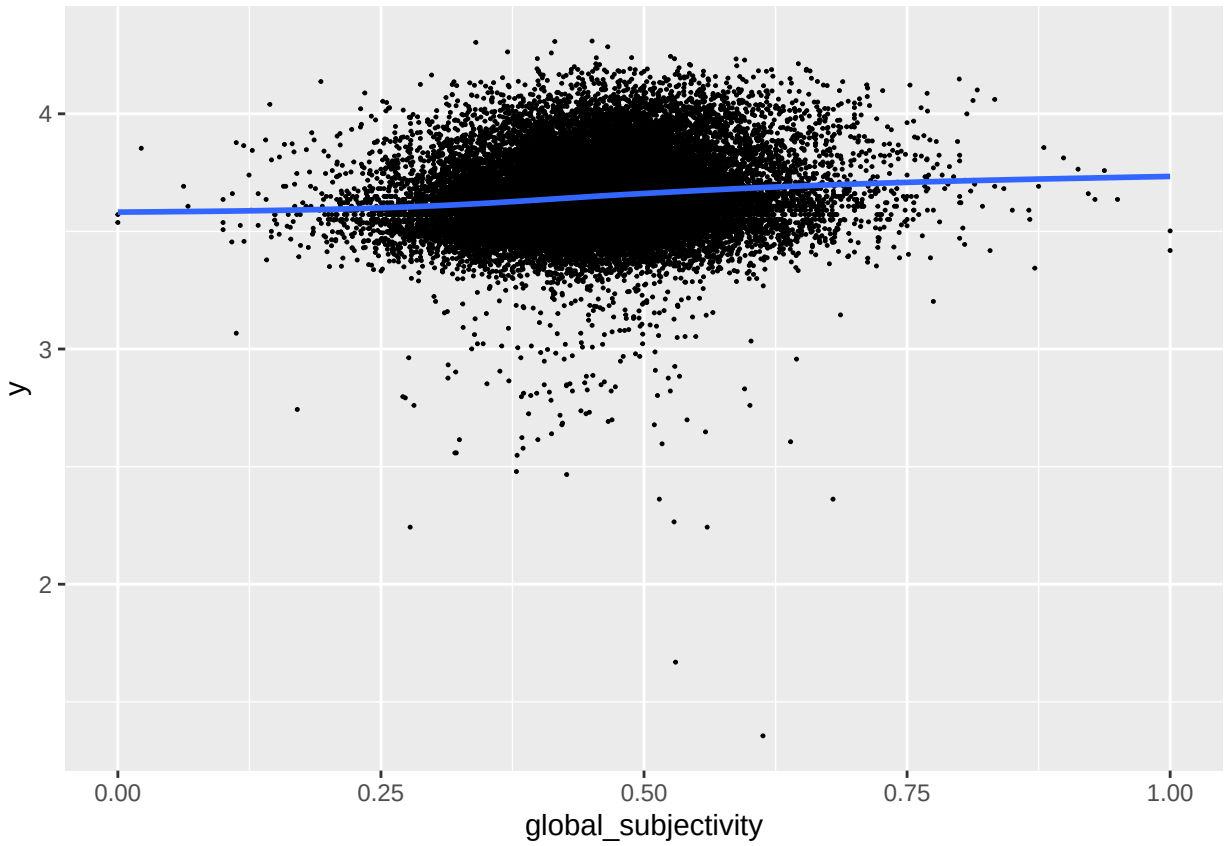

```
plotpoint(newstrain,newstrain$ min_positive_polarity,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab(" m
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



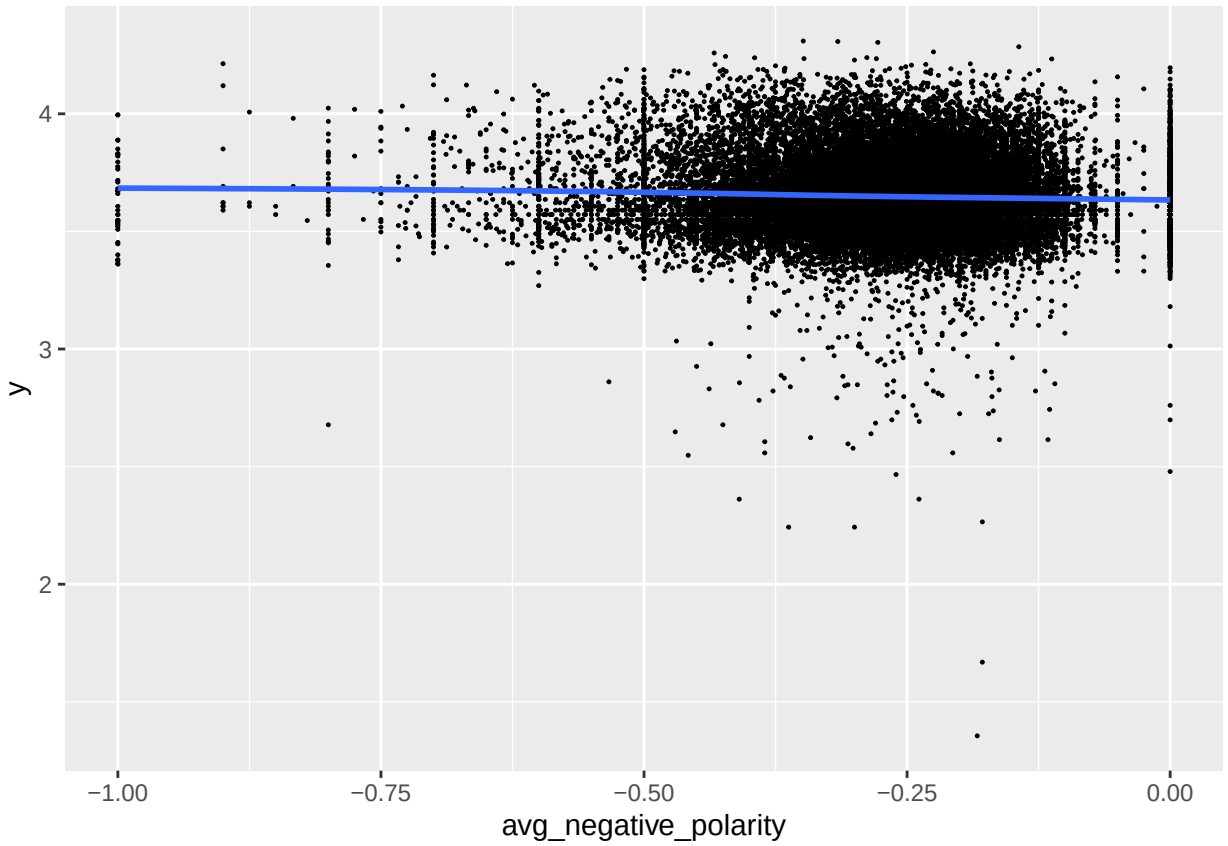

```
plotpoint(newstrain,newstrain$global_subjectivity,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("globa
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



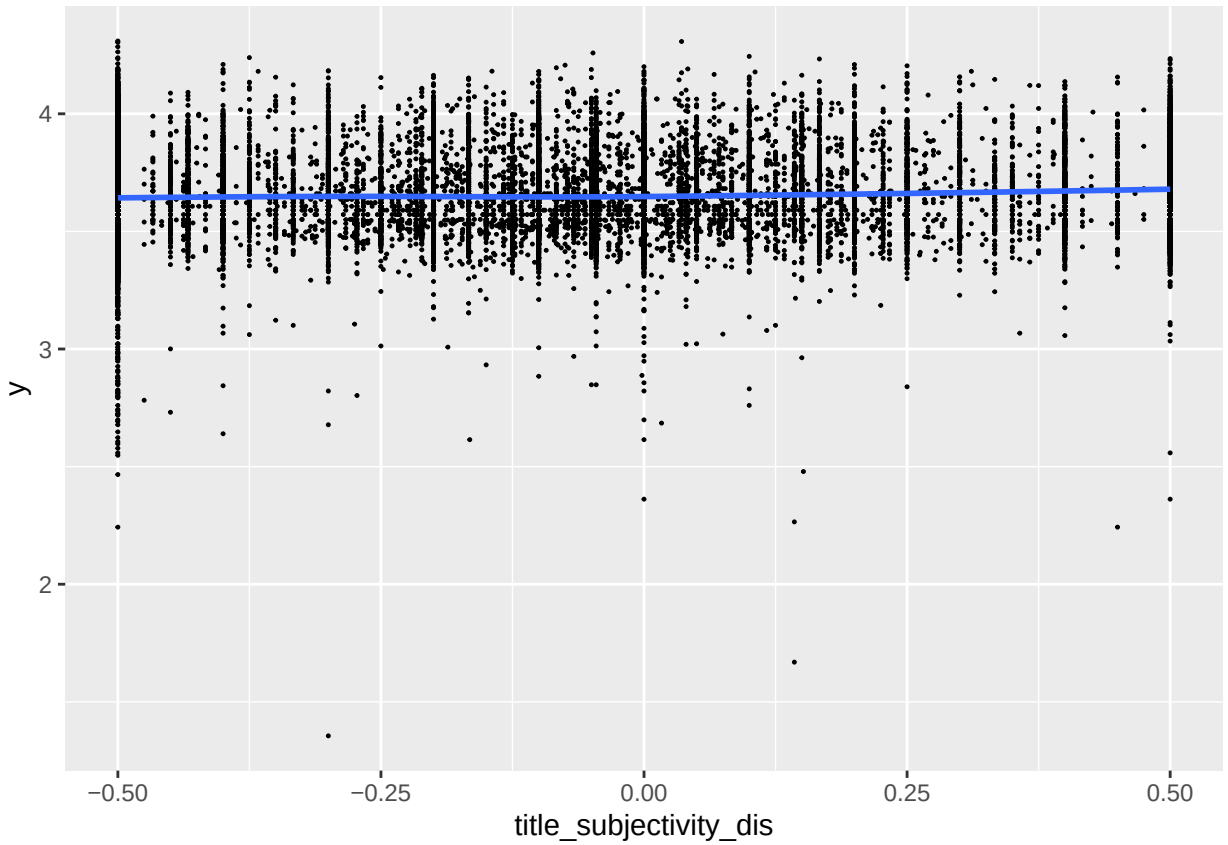

```
plotpoint(newstrain,newstrain$avg_negative_polarity,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("avg
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



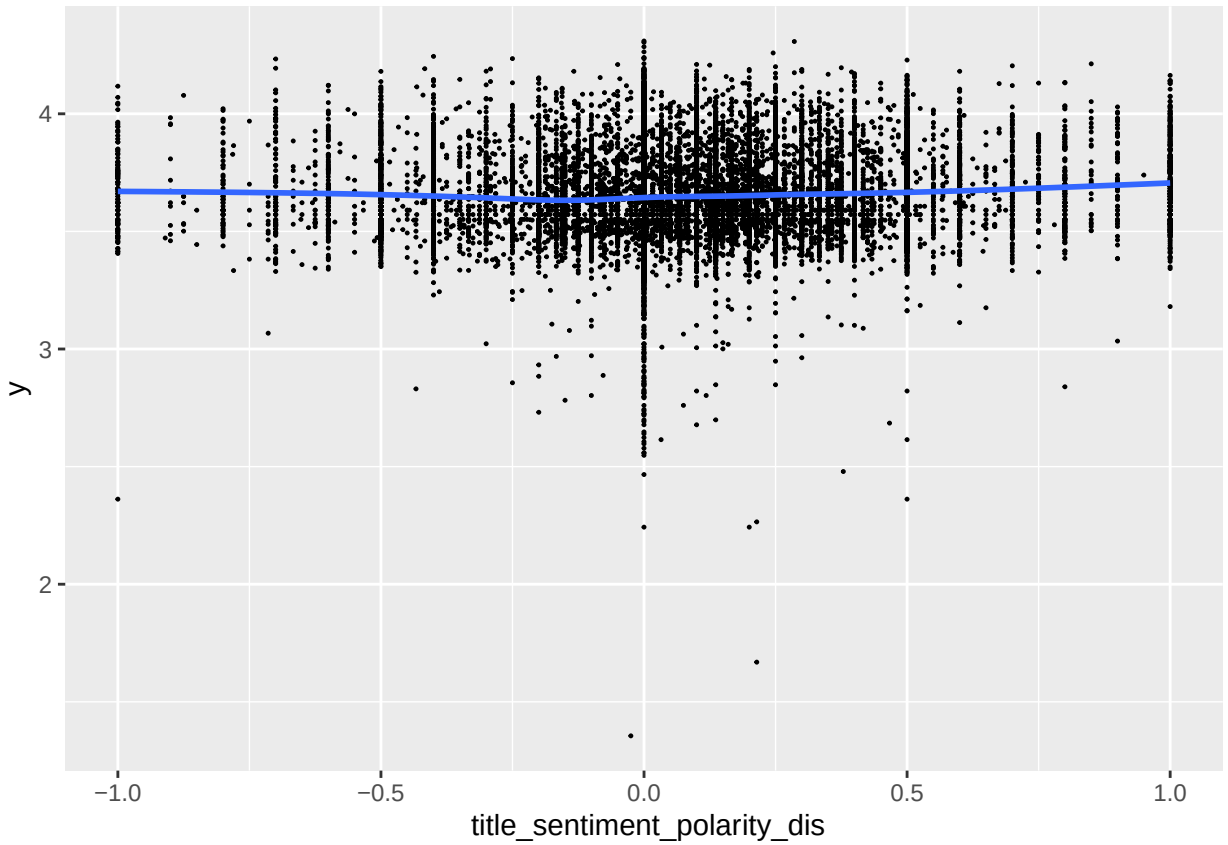

```
plotpoint(newstrain,newstrain$title_subjectivity_dis,((newstrain$shares^-0.22) - 1)/(-0.22)) + xlab("ti
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
plotpoint(newstrain,newstrain$ title_sentiment_polarity_dis,((newstrain$shares^-0.22) - 1)/(-0.22)) + xl
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



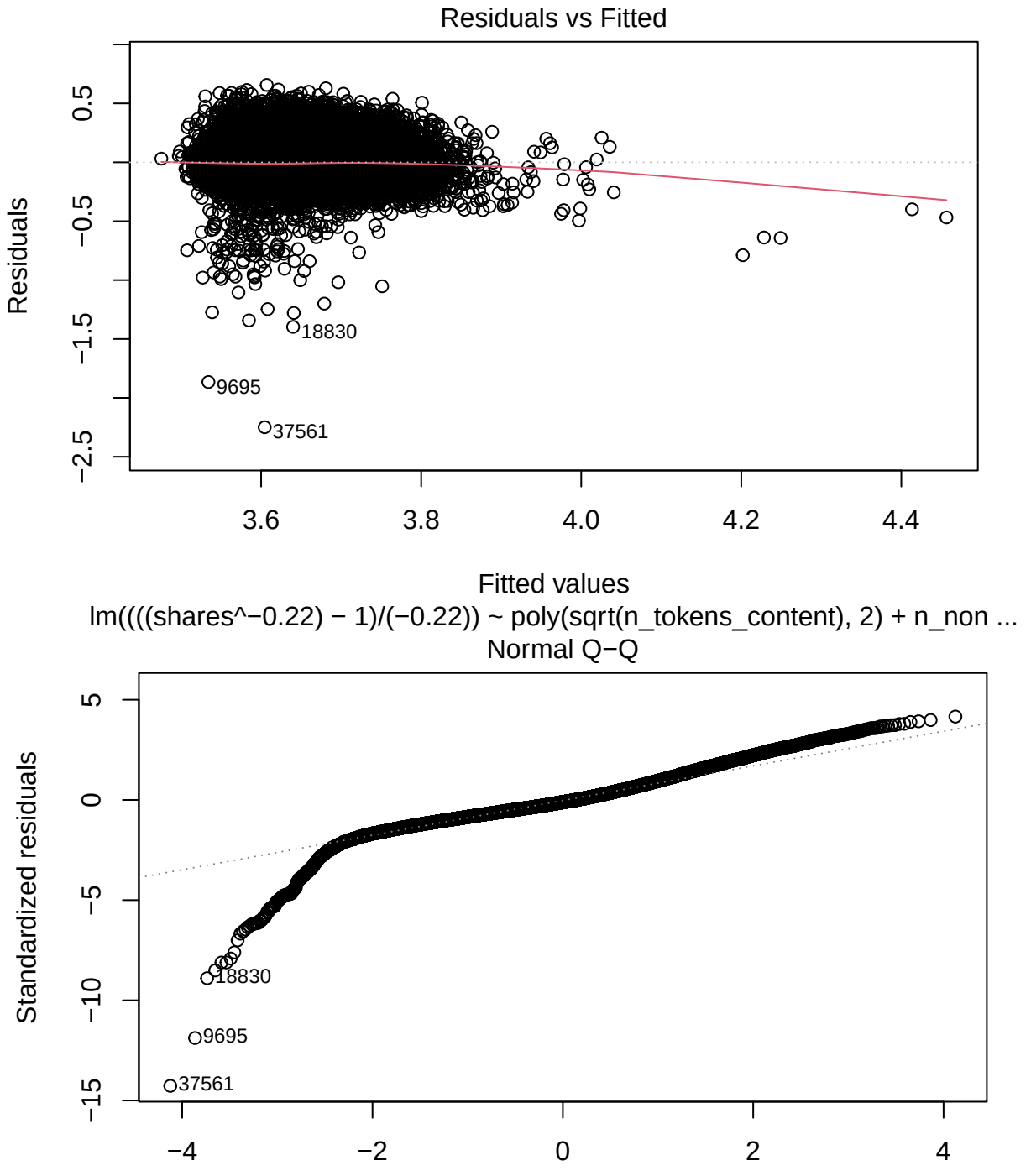

```
# we will refit our model using the findings above,add the polynomial effect and drop the none signific

lmnew_full_ts1 <- lm(formula = (((shares^-0.22) - 1)/(-0.22)) ~
    poly(sqrt(n_tokens_content),2)  + n_non_stop_unique_tokens +
    sqrt(num_hrefs) + poly(sqrt(num_self_hrefs),2) + poly(sqrt(num_imgs),2) + poly(sqrt(num_videos),3) +
    num_keywords +channel +sqrt(kw_max_min)  +  +
    kw_min_max + kw_avg_max  + kw_avg_avg  + poly(sqrt(self_reference_avg_sharess),3) + is_weekend + LDA
    global_sentiment_polarity + global_rate_positive_words +
    rate_positive_words + min_positive_polarity + avg_negative_polarity +
    title_subjectivity_dis + title_sentiment_polarity_dis ,
    data = newstrain)

summary(lmnew_full_ts1)
```

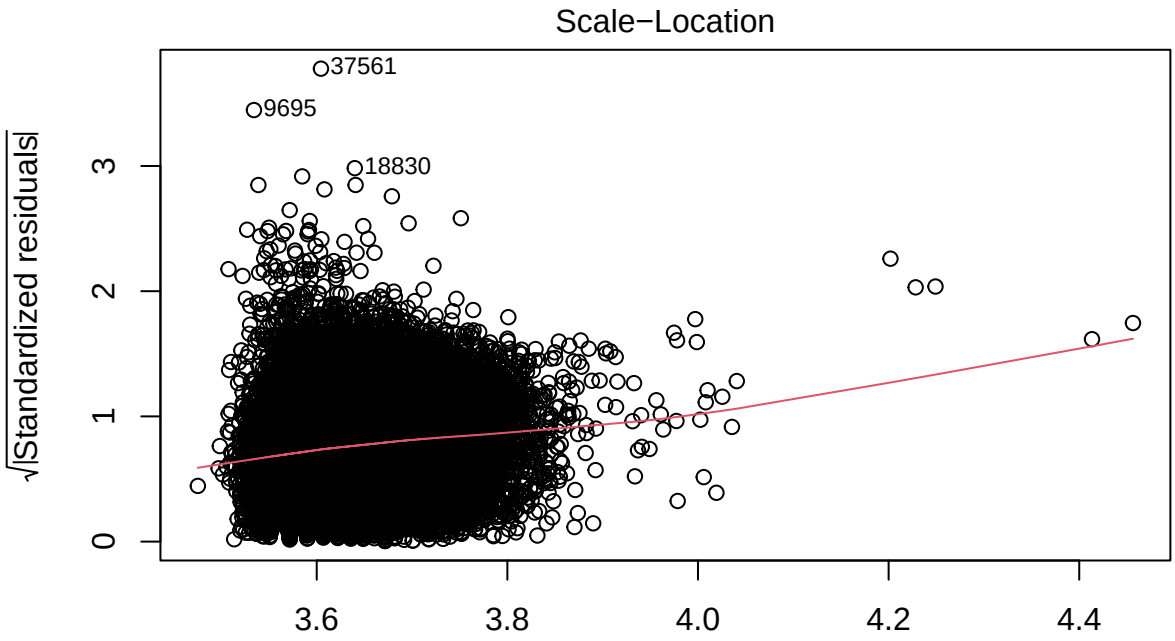
```
##
## Call:
## lm(formula = (((shares^-0.22) - 1)/(-0.22)) ~ poly(sqrt(n_tokens_content),
##     2) + n_non_stop_unique_tokens + sqrt(num_hrefs) + poly(sqrt(num_self_hrefs),
##     2) + poly(sqrt(num_imgs), 2) + poly(sqrt(num_videos), 3) +
##     average_token_length + num_keywords + channel + sqrt(kw_max_min) +
##     +kw_min_max + kw_avg_max + kw_avg_avg + poly(sqrt(self_reference_avg_sharess),
##     3) + is_weekend + LDA_01 + LDA_02 + LDA_03 + LDA_04 + global_subjectivity +
##     global_sentiment_polarity + global_rate_positive_words +
##     rate_positive_words + min_positive_polarity + avg_negative_polarity +
##     title_subjectivity_dis + title_sentiment_polarity_dis, data = newstrain)
##
## Residuals:
```

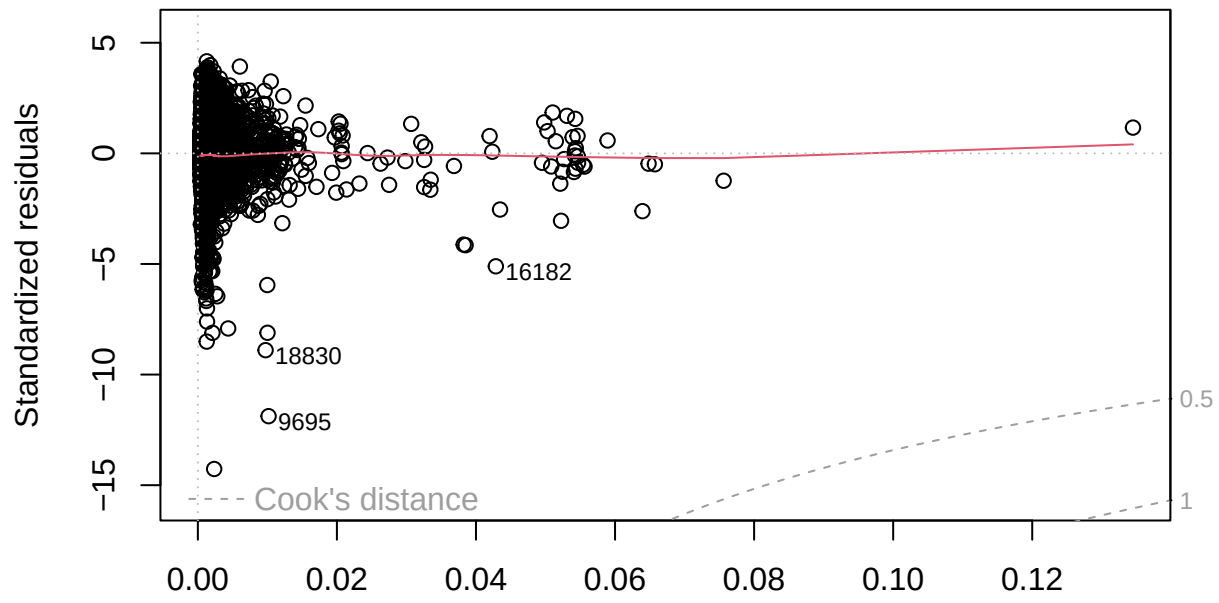
```
##     Min      1Q  Median      3Q     Max
## -2.2491 -0.0964 -0.0177  0.0873  0.6562
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              3.63e+00   2.41e-02  150.55
## poly(sqrt(n_tokens_content), 2)1          7.07e-02   2.64e-01    0.27
## poly(sqrt(n_tokens_content), 2)2          9.15e-01   1.78e-01    5.13
## n_non_stop_unique_tokens                 -5.21e-02   1.49e-02   -3.49
## sqrt(num_hrefs)                           8.32e-03   9.71e-04    8.56
## poly(sqrt(num_self_hrefs), 2)1           -2.09e+00   2.01e-01  -10.41
## poly(sqrt(num_self_hrefs), 2)2           -1.65e-01   1.81e-01   -0.91
## poly(sqrt(num_imgs), 2)1                  4.17e-01   2.18e-01    1.91
## poly(sqrt(num_imgs), 2)2                 -1.95e-01   1.71e-01   -1.14
## poly(sqrt(num_videos), 3)1                1.09e+00   1.87e-01    5.86
## poly(sqrt(num_videos), 3)2               -1.04e+00   1.62e-01   -6.41
## poly(sqrt(num_videos), 3)3                6.43e-01   1.60e-01    4.01
## average_token_length                     -6.09e-03   3.85e-03   -1.58
## num_keywords                              6.96e-04   5.85e-04    1.19
## channelentertainment                     -8.64e-03   6.03e-03   -1.43
## channellifestyle                          9.80e-03   6.10e-03    1.61
## channelother                              4.28e-02   6.34e-03    6.75
## channelsocmed                             5.77e-02   5.15e-03   11.21
## channeltech                               5.00e-02   5.48e-03    9.14
## channelworld                              1.29e-02   5.89e-03    2.20
## sqrt(kw_max_min)                         -2.31e-04   6.33e-05   -3.65
## kw_min_max                               -6.27e-08   1.95e-08   -3.21
## kw_avg_max                               -7.06e-08   1.03e-08   -6.86
## kw_avg_avg                                2.09e-05   1.09e-06   19.15
## poly(sqrt(self_reference_avg_sharess), 3)1  3.64e+00   1.79e-01   20.29
## poly(sqrt(self_reference_avg_sharess), 3)2 -2.24e+00   1.73e-01  -12.98
## poly(sqrt(self_reference_avg_sharess), 3)3  9.13e-01   1.75e-01    5.23
## is_weekend                                5.62e-02   2.89e-03   19.48
## LDA_01                                   -6.89e-02   8.80e-03   -7.82
## LDA_02                                   -9.44e-02   8.25e-03  -11.44
## LDA_03                                   -6.08e-02   8.40e-03   -7.24
## LDA_04                                   -4.01e-02   7.56e-03   -5.30
## global_subjectivity                       6.84e-02   1.36e-02    5.04
## global_sentiment_polarity                -2.06e-02   1.93e-02   -1.06
## global_rate_positive_words               -4.39e-02   8.23e-02   -0.53
## rate_positive_words                       1.92e-02   1.12e-02    1.71
## min_positive_polarity                    -6.38e-02   1.66e-02   -3.84
## avg_negative_polarity                    -1.60e-02   9.66e-03   -1.66
## title_subjectivity_dis                    8.59e-03   3.16e-03    2.72
## title_sentiment_polarity_dis              8.44e-03   3.92e-03    2.15
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## poly(sqrt(n_tokens_content), 2)1          0.78884
## poly(sqrt(n_tokens_content), 2)2          2.9e-07 ***
## n_non_stop_unique_tokens                  0.00049 ***
## sqrt(num_hrefs)                           < 2e-16 ***
## poly(sqrt(num_self_hrefs), 2)1            < 2e-16 ***
## poly(sqrt(num_self_hrefs), 2)2            0.36128
## poly(sqrt(num_imgs), 2)1                  0.05562 .
```

```
## poly(sqrt(num_imgs), 2)2                    0.25416
## poly(sqrt(num_videos), 3)1                  4.6e-09 ***
## poly(sqrt(num_videos), 3)2                  1.5e-10 ***
## poly(sqrt(num_videos), 3)3                  6.2e-05 ***
## average_token_length                        0.11347
## num_keywords                                0.23414
## channelentertainment                        0.15165
## channellifestyle                            0.10797
## channelother                                1.5e-11 ***
## channelsocmed                               < 2e-16 ***
## channeltech                                 < 2e-16 ***
## channelworld                                0.02792 *
## sqrt(kw_max_min)                            0.00026 ***
## kw_min_max                                  0.00132 **
## kw_avg_max                                  7.2e-12 ***
## kw_avg_avg                                  < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)1  < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)2  < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)3  1.7e-07 ***
## is_weekend                                  < 2e-16 ***
## LDA_01                                      5.4e-15 ***
## LDA_02                                      < 2e-16 ***
## LDA_03                                      4.7e-13 ***
## LDA_04                                      1.2e-07 ***
## global_subjectivity                         4.6e-07 ***
## global_sentiment_polarity                   0.28728
## global_rate_positive_words                  0.59424
## rate_positive_words                         0.08754 .
## min_positive_polarity                       0.00012 ***
## avg_negative_polarity                       0.09776 .
## title_subjectivity_dis                      0.00654 **
## title_sentiment_polarity_dis                0.03122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.158 on 26883 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.137
## F-statistic:  110 on 39 and 26883 DF,  p-value: <2e-16
```

```
plot(lmnew_full_ts1)
```


Residuals vs Fitted
Residuals
0.5
−0.5
−1.5
−2.5
18830
9695
37561
3.6
3.8
4.0
4.2
4.4
Fitted values
lm((((shares^−0.22) − 1)/(−0.22)) ~ poly(sqrt(n_tokens_content), 2) + n_non ...
Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
18830
9695
37561
−4
−2
0
2
4
Theoretical Quantiles
lm((((shares^−0.22) − 1)/(−0.22)) ~ poly(sqrt(n_tokens_content), 2) + n_non ...





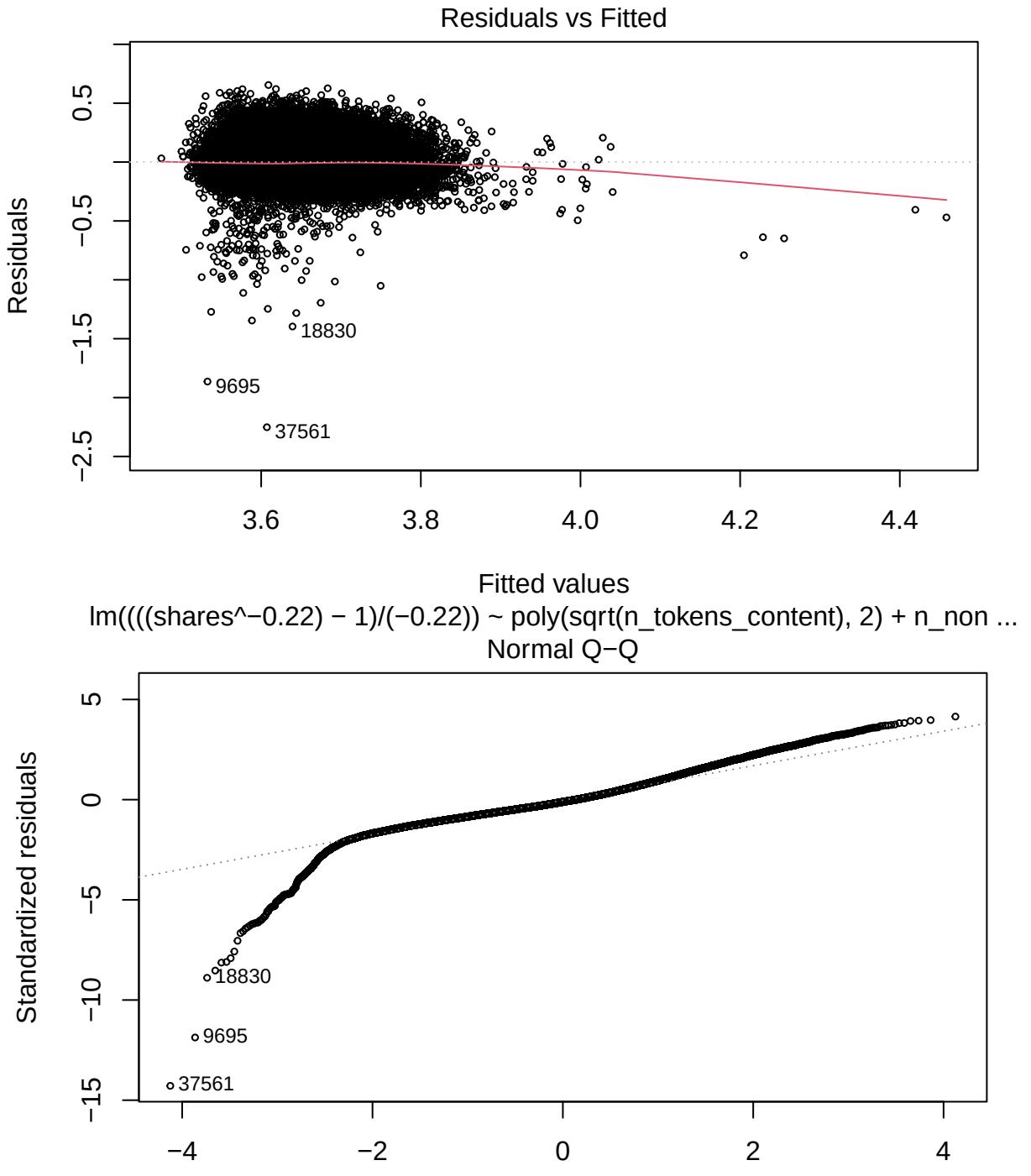

```
# in our final model, we can see multiple R-squared:0.1362, Adjusted R-squared:  0.1351 both of these t
options(scipen=0)
lmnew_final <- step(lmnew_full_ts1,trace=F)
summary(lmnew_final)
```

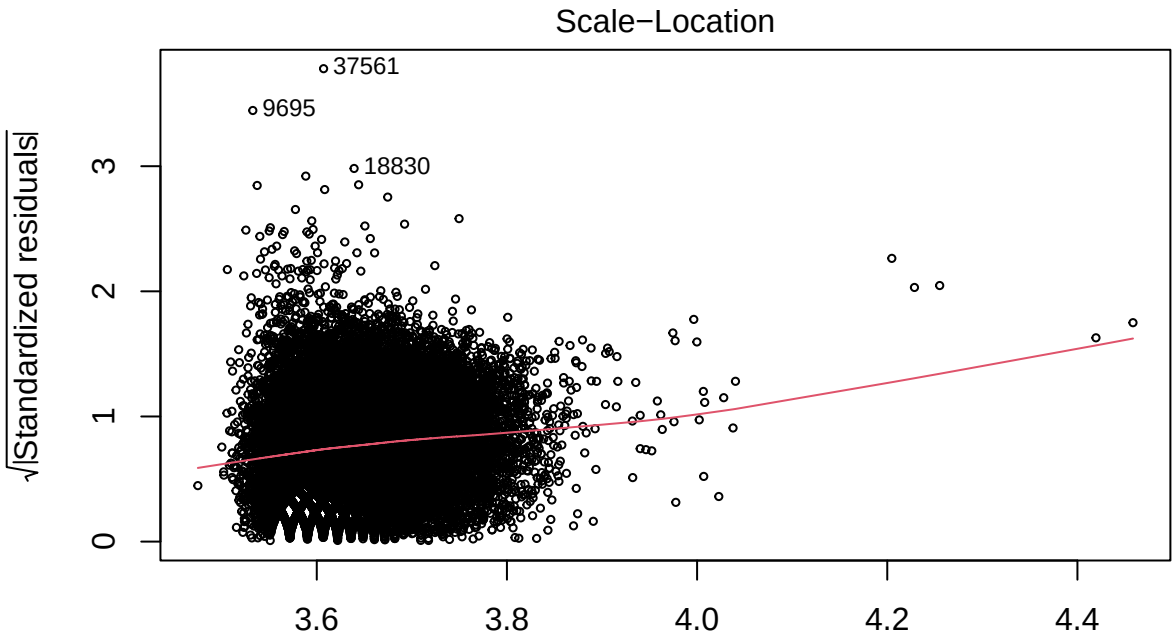
```
##
```

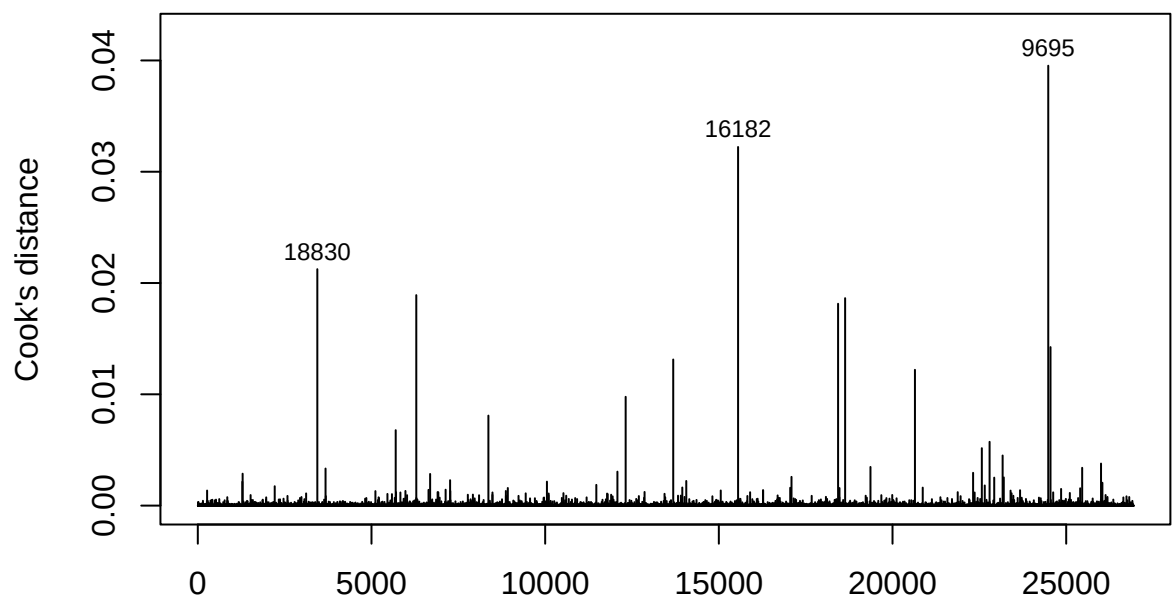
```
## Call:
## lm(formula = (((shares^-0.22) - 1)/(-0.22)) ~ poly(sqrt(n_tokens_content),
##     2) + n_non_stop_unique_tokens + sqrt(num_hrefs) + poly(sqrt(num_self_hrefs),
##     2) + poly(sqrt(num_imgs), 2) + poly(sqrt(num_videos), 3) +
##     average_token_length + channel + sqrt(kw_max_min) + kw_min_max +
##     kw_avg_max + kw_avg_avg + poly(sqrt(self_reference_avg_sharess),
##     3) + is_weekend + LDA_01 + LDA_02 + LDA_03 + LDA_04 + global_subjectivity +
##     min_positive_polarity + avg_negative_polarity + title_subjectivity_dis +
##     title_sentiment_polarity_dis, data = newstrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2518 -0.0963 -0.0178  0.0871  0.6541
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              3.65e+00   2.27e-02  160.64
## poly(sqrt(n_tokens_content), 2)1         1.61e-02   2.61e-01    0.06
## poly(sqrt(n_tokens_content), 2)2         9.30e-01   1.78e-01    5.24
## n_non_stop_unique_tokens                -5.42e-02   1.49e-02   -3.65
## sqrt(num_hrefs)                          8.42e-03   9.60e-04    8.77
## poly(sqrt(num_self_hrefs), 2)1          -2.08e+00   2.01e-01  -10.35
## poly(sqrt(num_self_hrefs), 2)2          -1.69e-01   1.81e-01   -0.94
## poly(sqrt(num_imgs), 2)1                 4.07e-01   2.18e-01    1.87
## poly(sqrt(num_imgs), 2)2                -1.94e-01   1.71e-01   -1.13
## poly(sqrt(num_videos), 3)1               1.09e+00   1.86e-01    5.87
## poly(sqrt(num_videos), 3)2              -1.07e+00   1.62e-01   -6.61
## poly(sqrt(num_videos), 3)3               6.62e-01   1.60e-01    4.13
## average_token_length                    -5.71e-03   3.84e-03   -1.49
## channelentertainment                    -8.68e-03   6.03e-03   -1.44
## channellifestyle                         9.95e-03   6.08e-03    1.64
## channelother                             4.29e-02   6.32e-03    6.79
## channelsocmed                            5.72e-02   5.14e-03   11.15
## channeltech                              5.04e-02   5.47e-03    9.21
## channelworld                             1.32e-02   5.87e-03    2.25
## sqrt(kw_max_min)                        -2.20e-04   6.29e-05   -3.51
## kw_min_max                              -6.61e-08   1.93e-08   -3.42
## kw_avg_max                              -7.33e-08   9.96e-09   -7.36
## kw_avg_avg                               2.09e-05   1.09e-06   19.19
## poly(sqrt(self_reference_avg_sharess), 3)1  3.65e+00   1.79e-01   20.36
## poly(sqrt(self_reference_avg_sharess), 3)2 -2.24e+00   1.73e-01  -12.98
## poly(sqrt(self_reference_avg_sharess), 3)3  9.13e-01   1.74e-01    5.24
## is_weekend                               5.62e-02   2.88e-03   19.52
## LDA_01                                  -6.94e-02   8.80e-03   -7.89
## LDA_02                                  -9.48e-02   8.23e-03  -11.53
## LDA_03                                  -6.11e-02   8.38e-03   -7.30
## LDA_04                                  -3.95e-02   7.54e-03   -5.24
## global_subjectivity                      6.56e-02   1.25e-02    5.24
## min_positive_polarity                   -6.85e-02   1.54e-02   -4.43
## avg_negative_polarity                   -1.69e-02   8.65e-03   -1.95
## title_subjectivity_dis                   8.29e-03   3.13e-03    2.65
## title_sentiment_polarity_dis             8.60e-03   3.82e-03    2.25
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
```

```
## poly(sqrt(n_tokens_content), 2)1           0.95086
## poly(sqrt(n_tokens_content), 2)2           1.6e-07 ***
## n_non_stop_unique_tokens                   0.00026 ***
## sqrt(num_hrefs)                            < 2e-16 ***
## poly(sqrt(num_self_hrefs), 2)1             < 2e-16 ***
## poly(sqrt(num_self_hrefs), 2)2             0.34872
## poly(sqrt(num_imgs), 2)1                   0.06131 .
## poly(sqrt(num_imgs), 2)2                   0.25703
## poly(sqrt(num_videos), 3)1                 4.4e-09 ***
## poly(sqrt(num_videos), 3)2                 4.0e-11 ***
## poly(sqrt(num_videos), 3)3                 3.6e-05 ***
## average_token_length                       0.13687
## channelentertainment                       0.14967
## channellifestyle                           0.10198
## channelother                               1.1e-11 ***
## channelsocmed                              < 2e-16 ***
## channeltech                                < 2e-16 ***
## channelworld                               0.02422 *
## sqrt(kw_max_min)                           0.00045 ***
## kw_min_max                                 0.00062 ***
## kw_avg_max                                 2.0e-13 ***
## kw_avg_avg                                 < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)1  < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)2  < 2e-16 ***
## poly(sqrt(self_reference_avg_sharess), 3)3  1.7e-07 ***
## is_weekend                                 < 2e-16 ***
## LDA_01                                     3.1e-15 ***
## LDA_02                                     < 2e-16 ***
## LDA_03                                     3.0e-13 ***
## LDA_04                                     1.6e-07 ***
## global_subjectivity                        1.6e-07 ***
## min_positive_polarity                      9.4e-06 ***
## avg_negative_polarity                      0.05061 .
## title_subjectivity_dis                     0.00808 **
## title_sentiment_polarity_dis               0.02439 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.158 on 26887 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.137
## F-statistic:  123 on 35 and 26887 DF,  p-value: <2e-16
```

```
plot(lmnew_final,1:4,cex=0.5)
```


Residuals vs Fitted
Residuals
0.5
−0.5
−1.5
−2.5
18830
9695
37561
3.6
3.8
4.0
4.2
4.4
Fitted values
lm((((shares^−0.22) − 1)/(−0.22)) ~ poly(sqrt(n_tokens_content), 2) + n_non ...
Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
18830
9695
37561
−4
−2
0
2
4
Theoretical Quantiles
lm((((shares^−0.22) − 1)/(−0.22)) ~ poly(sqrt(n_tokens_content), 2) + n_non ...






```
newstrain[c(18830,17019,9695),]
```

```
##       n_tokens_title n_tokens_content n_non_stop_unique_tokens num_hrefs
## 17903              8             1334                    0.617        13
## 15739             15              720                    0.705         1
## 3606              14              548                    0.682        11
```

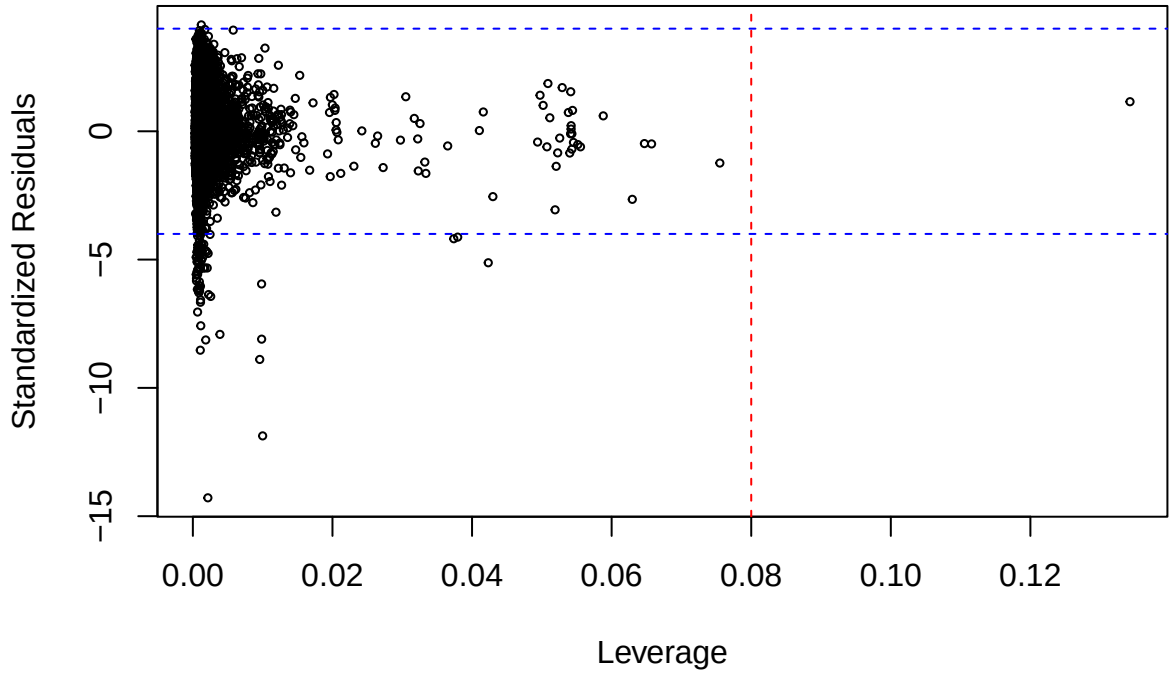
```
##       num_self_hrefs num_imgs num_videos average_token_length num_keywords
## 17903              2       12          0                 4.45            5
## 15739              0        1          0                 4.34            5
## 3606               4        1          1                 4.60            4
##       kw_max_min kw_min_max kw_max_max kw_avg_max kw_min_avg kw_avg_avg
## 17903       4600       4600     843300     264860       2950       3618
## 15739        178        752     843300     246070        465       2644
## 3606         968       3800     617900     182800       2200       2590
##       self_reference_avg_sharess is_weekend LDA_01 LDA_02 LDA_03 LDA_04
## 17903                        983          0 0.5094 0.0400 0.3706  0.040
## 15739                          0          0 0.0403 0.2385 0.2403  0.441
## 3606                        5350          0 0.7992 0.0501 0.0507  0.050
##       global_subjectivity global_sentiment_polarity global_rate_positive_words
## 17903               0.507                    0.0252                    0.0352
## 15739               0.538                    0.1484                    0.0514
## 3606                0.395                    0.1244                    0.0547
##       global_rate_negative_words rate_positive_words avg_positive_polarity
## 17903                     0.0300               0.540                 0.332
## 15739                     0.0222               0.698                 0.437
## 3606                      0.0146               0.789                 0.288
##       min_positive_polarity max_positive_polarity avg_negative_polarity
## 17903                  0.05                   0.7                -0.311
## 15739                  0.10                   1.0                -0.333
## 3606                   0.10                   0.9                -0.268
##       min_negative_polarity max_negative_polarity shares title_subjectivity_dis
## 17903                  -1.0                -0.050  10900                  -0.5
## 15739                  -0.8                -0.125   1000                  -0.1
## 3606                   -0.5                -0.125    838                  -0.5
##       title_sentiment_polarity_dis       channel
## 17903                          0.0 entertainment
## 15739                          0.4          tech
## 3606                           0.0 entertainment
```

```
options(scipen = 999,digits=3)
#library(faraway)
## check collinearity vif
## there are no serious collinearity problems in our model
car::vif(lmnew_final)
```

```
##                                         GVIF Df GVIF^(1/(2*Df))
## poly(sqrt(n_tokens_content), 2)         3.29  2           1.35
## n_non_stop_unique_tokens                2.42  1           1.56
## sqrt(num_hrefs)                         1.80  1           1.34
## poly(sqrt(num_self_hrefs), 2)           2.09  2           1.20
## poly(sqrt(num_imgs), 2)                 2.23  2           1.22
## poly(sqrt(num_videos), 3)               1.50  3           1.07
## average_token_length                    1.29  1           1.13
## channel                                82.00  6           1.44
## sqrt(kw_max_min)                        1.41  1           1.19
## kw_min_max                              1.24  1           1.11
## kw_avg_max                              1.86  1           1.36
## kw_avg_avg                              2.11  1           1.45
## poly(sqrt(self_reference_avg_sharess), 3)  1.75  3           1.10
## is_weekend                              1.02  1           1.01
## LDA_01                                  4.12  1           2.03
```

```
## LDA_02                                5.83  1           2.41
## LDA_03                                6.29  1           2.51
## LDA_04                                5.22  1           2.28
## global_subjectivity                   1.32  1           1.15
## min_positive_polarity                 1.28  1           1.13
## avg_negative_polarity                 1.19  1           1.09
## title_subjectivity_dis                1.11  1           1.05
## title_sentiment_polarity_dis          1.08  1           1.04
```

```
### check outliers
p<-35
n<-nrow(newstrain)
plot(hatvalues(lmnew_final), rstandard(lmnew_final),cex=0.5,
xlab='Leverage', ylab='Standardized Residuals')

abline(v=0.08,col="red",lty=2)
abline(h=c(-4,4),col="blue",lty=2)
```




```
ind <- which(hatvalues(lmnew_final)>0.08)
newstrain[ind,]
```

```
##       n_tokens_title n_tokens_content n_non_stop_unique_tokens num_hrefs
## 17978              9             1322                    0.595       143
##       num_self_hrefs num_imgs num_videos average_token_length num_keywords
## 17978             51        0         91                 5.06            7
##       kw_max_min kw_min_max kw_max_max kw_avg_max kw_min_avg kw_avg_avg
## 17978       2100       4800     843300     591286       3450       5421
##       self_reference_avg_sharess is_weekend LDA_01 LDA_02 LDA_03 LDA_04
## 17978                      83223          0 0.0286 0.0286  0.886 0.0286
##       global_subjectivity global_sentiment_polarity global_rate_positive_words
## 17978               0.503                    0.0128                     0.0257
##       global_rate_negative_words rate_positive_words avg_positive_polarity
## 17978                     0.0212               0.548                 0.367
```

```
##       min_positive_polarity max_positive_polarity avg_negative_polarity
## 17978                   0.1                   0.8               -0.393
##       min_negative_polarity max_negative_polarity shares title_subjectivity_dis
## 17978                  -0.8                  -0.1  13600                      0
##       title_sentiment_polarity_dis channel
## 17978                          0.5   other
```

```
sum(abs(rstandard(lmnew_final))>4) # absolute value that greater than 4
```

```
## [1] 75
```

```
sum(abs(rstandard(lmnew_final))>4)/nrow(newstrain)*100
```

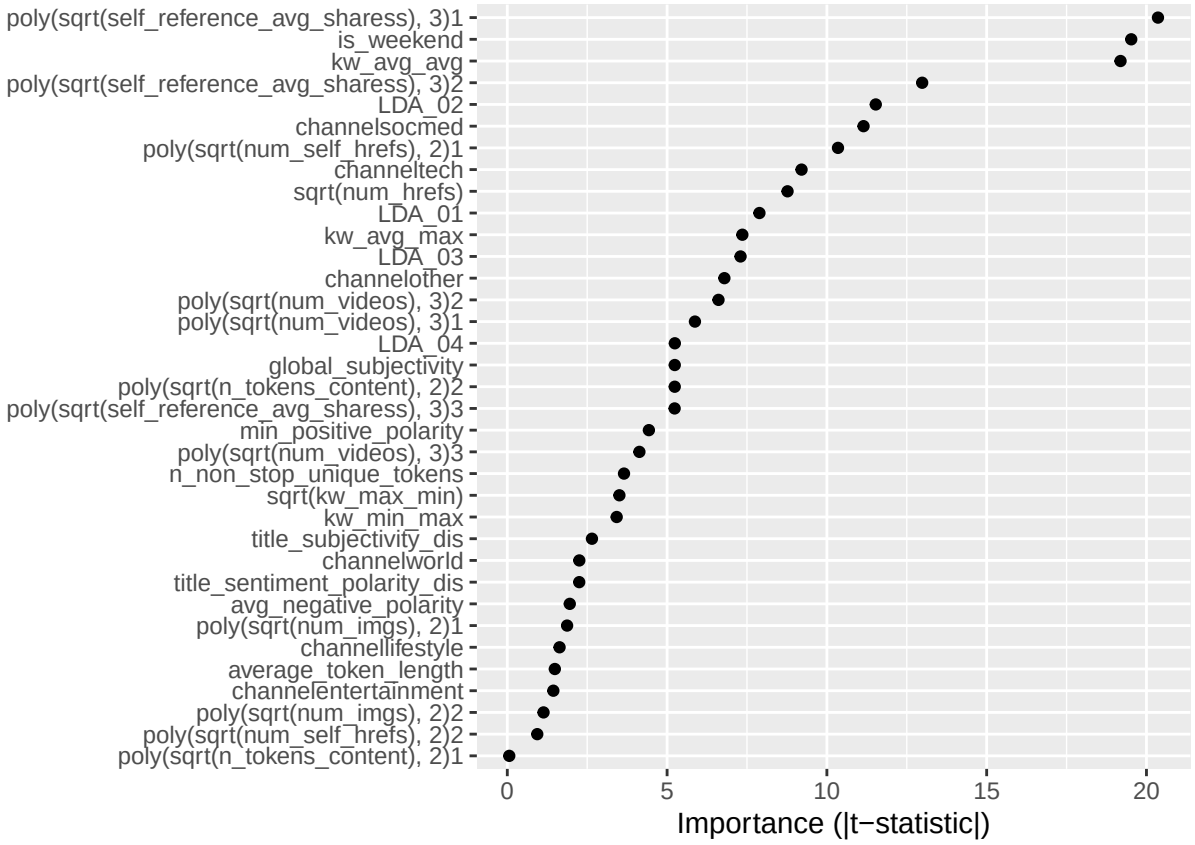
```
## [1] 0.279
```

```
# variable importance
library(vip)
```

```
##
## Attaching package: 'vip'
```

```
## The following object is masked from 'package:utils':
##
##     vi
```

```
 vip(lmnew_final, num_features = 35, geom = "point", include_type = TRUE)
```



Evaluate model proformance

generate predictions for the testing dataset

```
lm_pred_final <- predict(lmnew_final,newstest) # using the final model
lm_pred_ts <- predict(lmnew_full_ts,newstest) # using the boxcox transformed model with no polynomial e
lm_pred_log <- predict(lmnew_full_logs,newstest) # using the log transformed model
lm_pred1 <- predict(lmnew_full_step,newstest)
```

```
library(tidyverse)
 library(caret) # for cross-validation methods
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Make predictions and compute the R2, RMSE and MAE
predictions <- lmnew_final %>% predict(newstest)
data.frame( R2 = R2(predictions, ((newstest$shares^-0.22) - 1)/(-0.22)),
            RMSE = RMSE(predictions, ((newstest$shares^-0.22) - 1)/(-0.22)),
            MAE = MAE(predictions, ((newstest$shares^-0.22) - 1)/(-0.22)))
```

```
##      R2  RMSE   MAE
## 1 0.127 0.159 0.118
```

evaluate the model using adjusted Rsquare.

```
# calculate r_square and adjusted r_square for the final model
y<-((newstest$shares^-0.22) - 1)/(-0.22)
SST <-sum((y-mean(y))^2)
Res4<- y-lm_pred_final
SSR4 <-sum(Res4^2)
Rsquared_final <- 1-(SSR4/SST)
Rsquared_final
```

```
## [1] 0.127
```

```
n <- nrow(newstest)
# p_final <-35
adjust_Rsquared_final <- 1-(SSR4/SST)*((n-1)/(n-35-1))
adjust_Rsquared_final
```

```
## [1] 0.124
```

```
 # calculate r_square and adjusted r_square for the boxcox transformed model

y<-((newstest$shares^-0.22) - 1)/(-0.22)
SST <-sum((y-mean(y))^2)
Res3<- y-lm_pred_ts
SSR3 <-sum(Res3^2)
```

```
Rsquared_ts <- 1-(SSR3/SST)
Rsquared_ts
```

```
## [1] 0.108
```

```
adjust_Rsquared_ts <- 1-(SSR3/SST)*((n-1)/(n-31-1))
adjust_Rsquared_ts
```

```
## [1] 0.105
```

```
# calculate r_square and adjusted r_square for the log transformed model
y_log <- log(newstest$shares)
SST_log <-sum((y_log-mean(y_log))^2)
Res2<- y_log-lm_pred_log
SSR2 <-sum(Res2^2)
Rsquerd_log <- 1-(SSR2/SST_log)
Rsquerd_log
```

```
## [1] 0.106
```

```
adjust_Rsquared_log <- 1-(SSR2/SST_log)*((n-1)/(n-30-1))
adjust_Rsquared_log
```

```
## [1] 0.103
```

```
 # calculate r_square and adjusted r_square for the first model

y1<-newstest$shares
SST1 <-sum((y1-mean(y1))^2)
Res1<- y1-lm_pred1
SSR1 <-sum(Res1^2)
Rsquared1 <- 1-(SSR1/SST1)
Rsquared1
```

```
## [1] 0.0156
```

```
adjust_Rsquared1 <- 1-(SSR1/SST1)*((n-1)/(n-21-1))
adjust_Rsquared1
```

```
## [1] 0.0138
```

```
adjust_Rsquared_ts
```

```
## [1] 0.105
```

```
adjust_Rsquared_log
```

```
## [1] 0.103
```

```
adjust_Rsquared_final
```

```
## [1] 0.124
```

```
Rsquared_final
```

```
## [1] 0.127
```

```
Rsquared_ts
```

```
## [1] 0.108
```

```
adjust_Rsquared_log
```

```
## [1] 0.103
```

```
adjust_Rsquared <- c(adjust_Rsquared1,adjust_Rsquared_log,adjust_Rsquared_ts,adjust_Rsquared_final)
Rsquared <-c(Rsquared1,Rsquerd_log,Rsquared_ts,Rsquared_final)
Model <- c("No_Trans Model","Log_Trans Model", "Box_Cox_Trans Model","Final Model")
result <- data.frame(Model,Rsquared,adjust_Rsquared)
result
```

```
##                 Model Rsquared adjust_Rsquared
## 1      No_Trans Model   0.0156          0.0138
## 2     Log_Trans Model   0.1056          0.1033
## 3 Box_Cox_Trans Model   0.1076          0.1052
## 4         Final Model   0.1266          0.1240
```

```
cor(y_log,lm_pred_log)
```

```
## [1] 0.326
```

```
cor(y,lm_pred_ts)
```

```
## [1] 0.329
```

```
cor(y,lm_pred_final)
```

```
## [1] 0.356
```