HW week 12

w203: Statistics for Data Science

Shan He

OLS Inference

```
library(car)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

The file videos.txt contains data scraped from Youtube.com.

1. Fit a linear model predicting the number of views (views), from the length of a video (length) and its average user rating (rate).

```
setwd("/Users/shanhe/Desktop/W203/Homework/Week 12")
df <- read.table("videos.txt", header = TRUE, sep = "\t")
head(df)
```

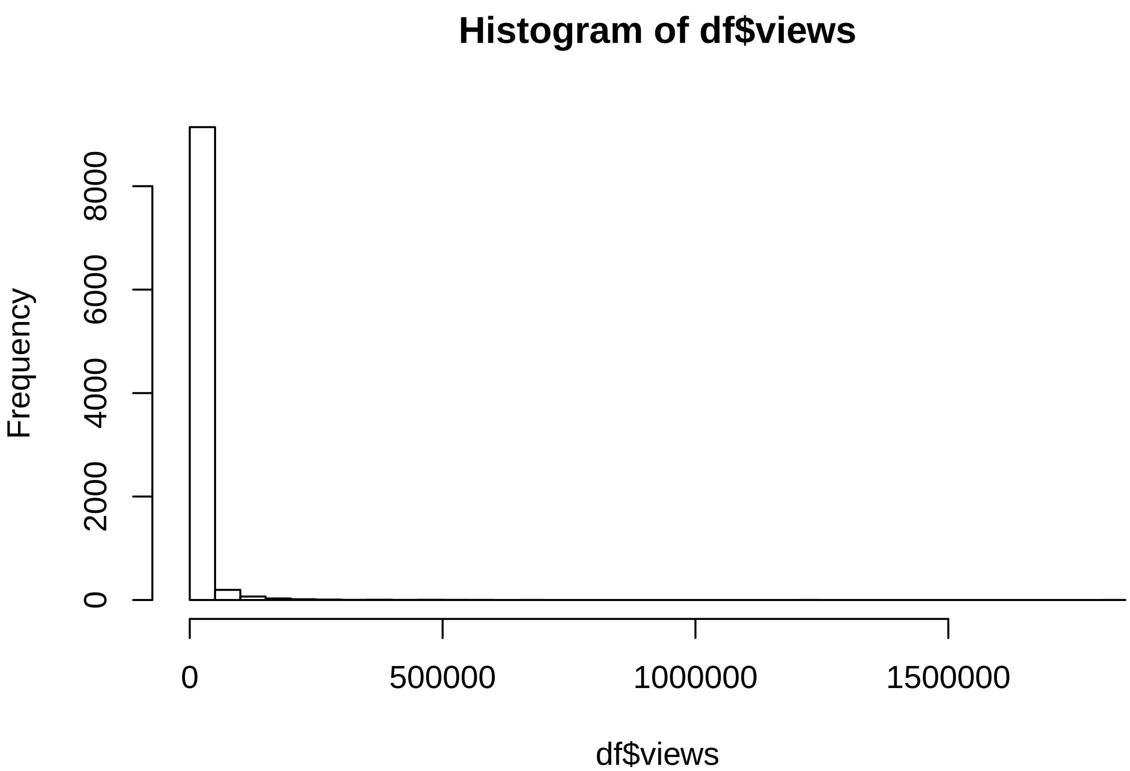
```
##      video_id            uploader  age      category length views rate
## 1 9QR1tni70fo              BHJJYP 1131        Comedy    126   204 3.00
## 2 l1DCSqAJ740          musicalrox 1236         Music    243  1652 3.91
## 3 ZES_o3XYGjM        tessaceleste 1243 Entertainment    105   898 4.48
## 4 4I8b40cViDE booloveswondergirls 1237 Entertainment    278   928 5.00
## 5 Elp6Bf0HJIM  Fizz101Productionz 1252        Comedy     26   392 1.50
## 6 VPuKu7aU9GY         slytherin66 1236 Entertainment    252   318 5.00
##   ratings comments
## 1       2        1
## 2      11        4
## 3      81       36
## 4      24       13
## 5       8       17
## 6       2        3
```

It might make sense for us to take the logrithms of the views and length as the variables in our linear model. We can check to see whether they look reasonable

```r
summary(df$views)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        3      348     1454     9374     6207  1807640        9
```

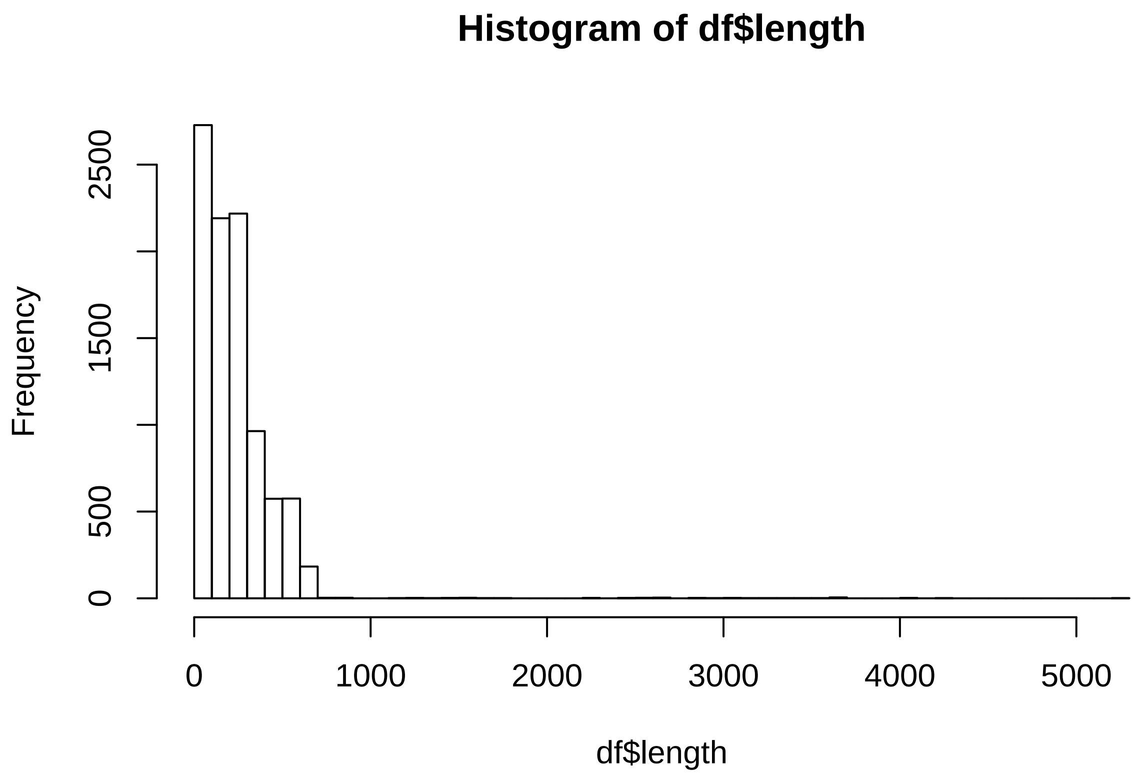
```r
hist(df$views, breaks = 50)
```




```r
summary(df$length)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##      1.0     83.0    193.0    226.7    298.2   5289.0        9
```

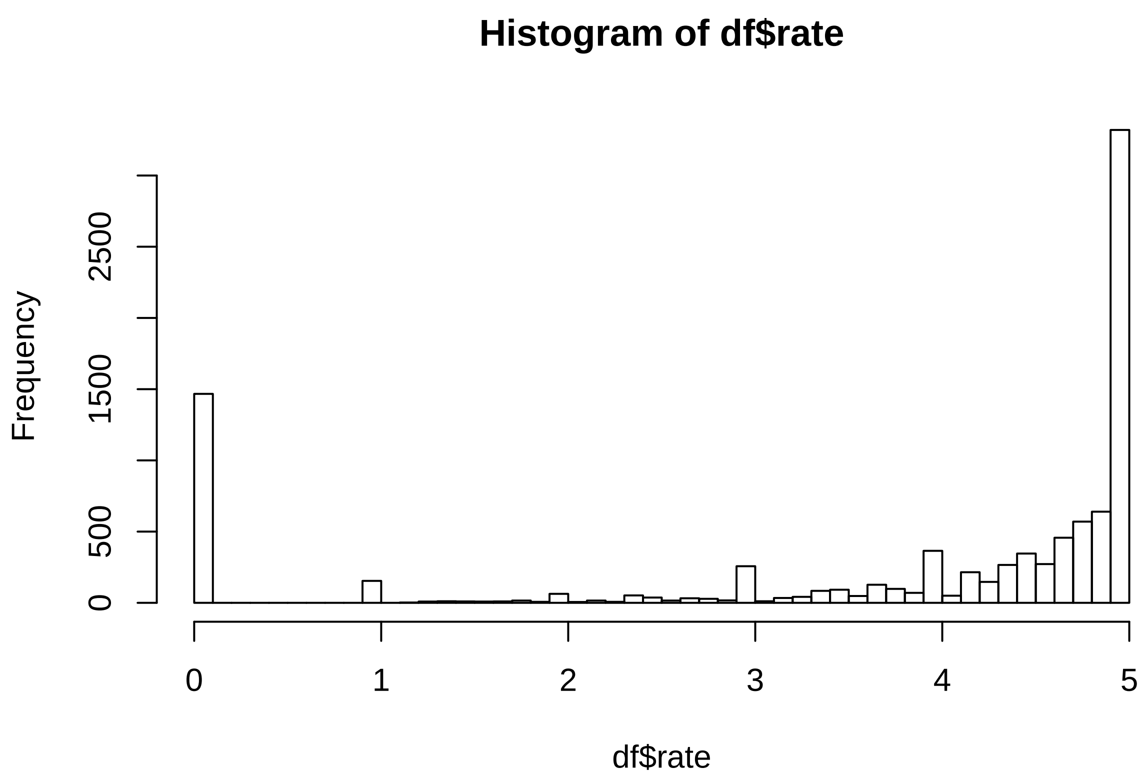
```r
hist(df$length, breaks = 50)
```



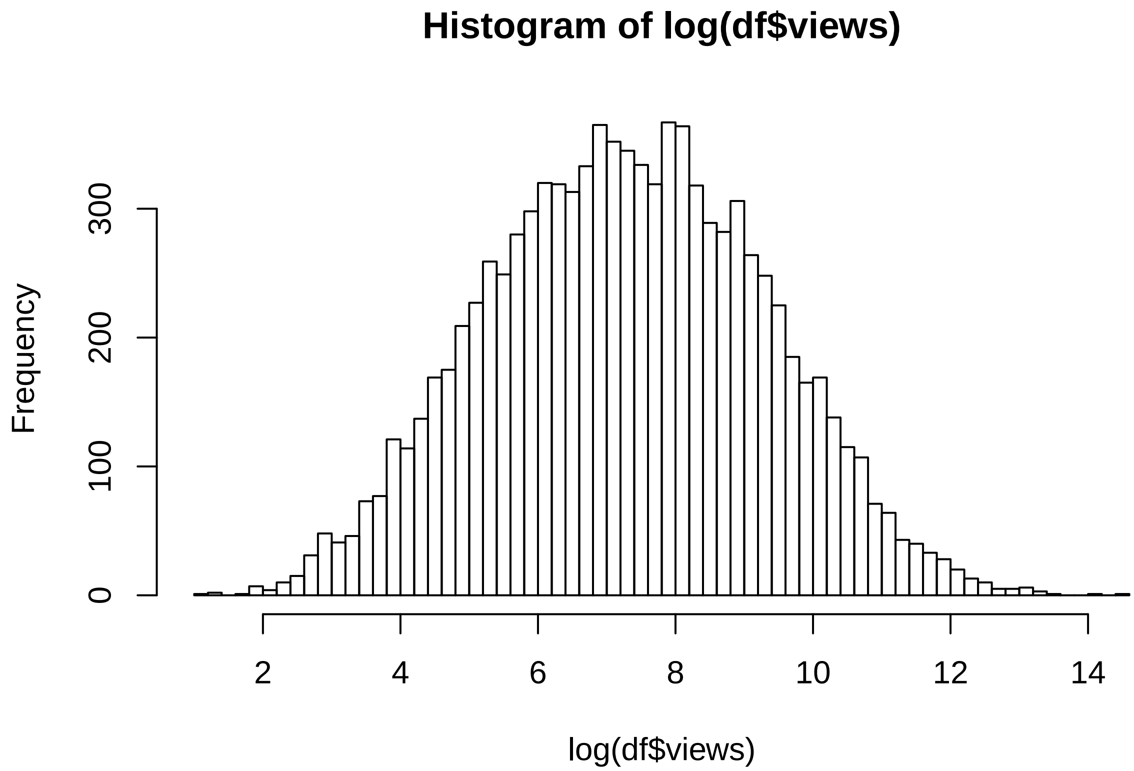

```r
summary(df$rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   3.400   4.670   3.746   5.000   5.000       9
```

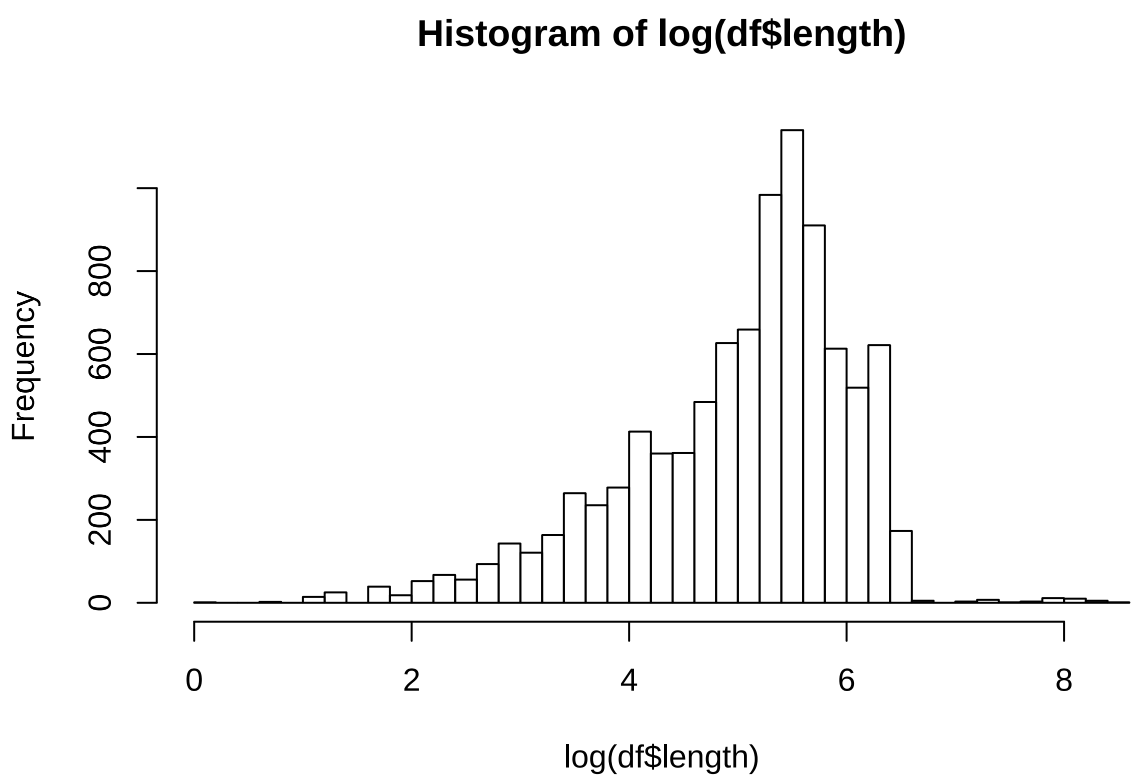
```r
hist(df$rate, breaks = 50)
```



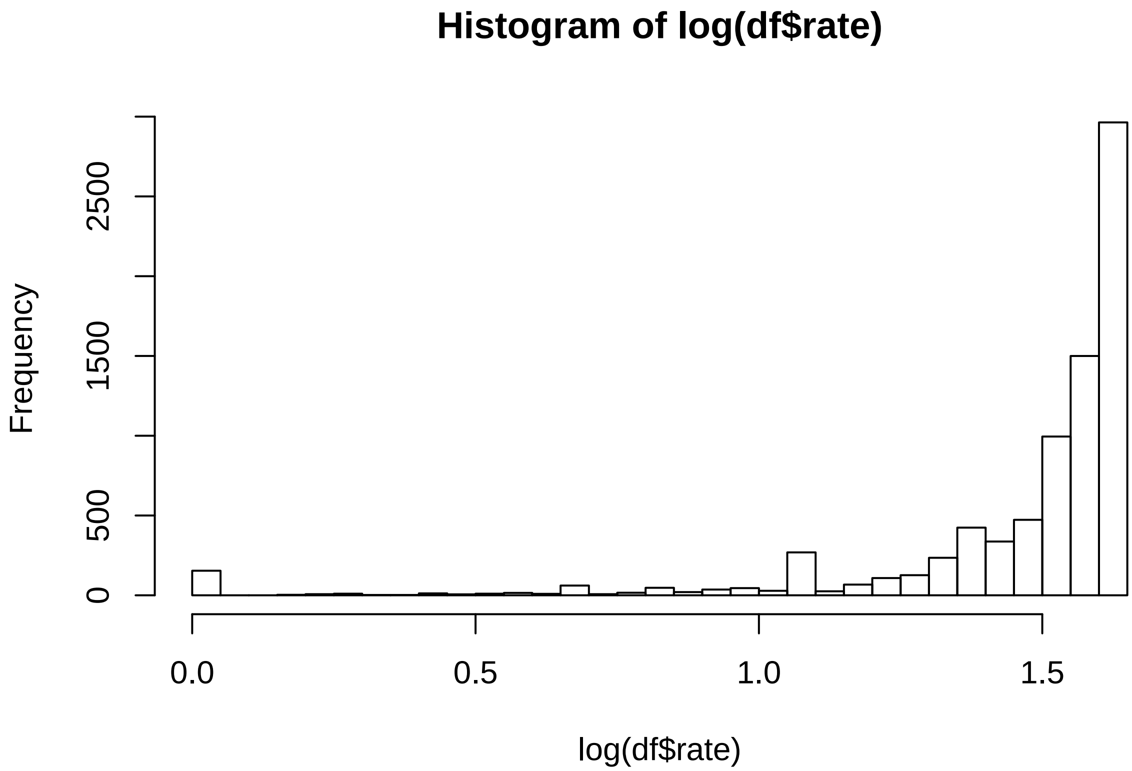
```r
hist(log(df$views), breaks = 50)
```




```r
hist(log(df$length), breaks = 50)
```



```r
hist(log(df$rate), breaks = 50)
```



Log(views) and Log(length) seem to have reasobable shape, better than with out log().

```r
model1 <- lm(log(views) ~ log(length) + rate, data = df)
```

2. Using diagnostic plots, background knowledge, and statistical tests, assess all 6 assumptions of the CLM. When an assumption is violated, state what response you will take.

a. Linear population model

We don't have to check the linear population model, because we haven't constrained the error term, so there's nothing to check at this point.

b. Random Sampling

To check random sampling, we need to understand how the data was collected. Independence of the sample data can also be an issue, for example, users that watch a video that already has an average of 5 star review might tend to rate the videos higher.

c. No perfect multicollinearity

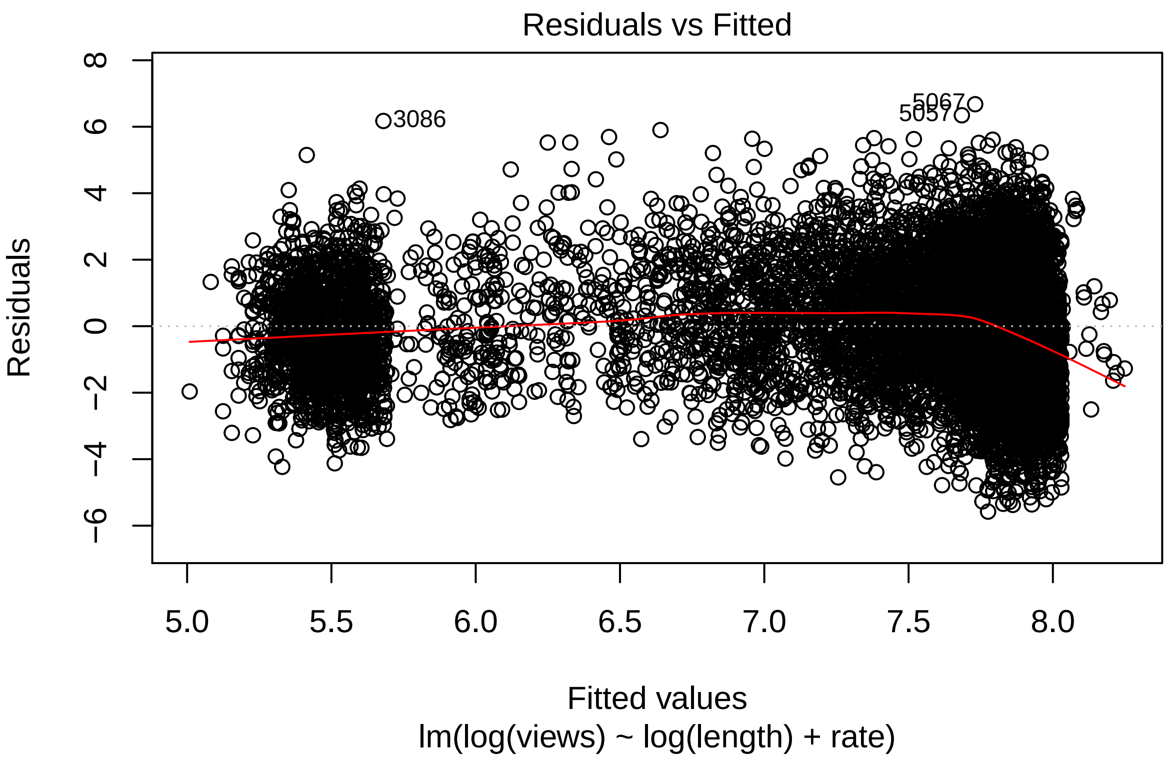
```r
cor(df$rate, log(df$length), use = "complete.obs")
```

```
## [1] 0.2497783
```

Rate and length show small correlation, which is allowed by MLR.3

d. Zero-conditional mean

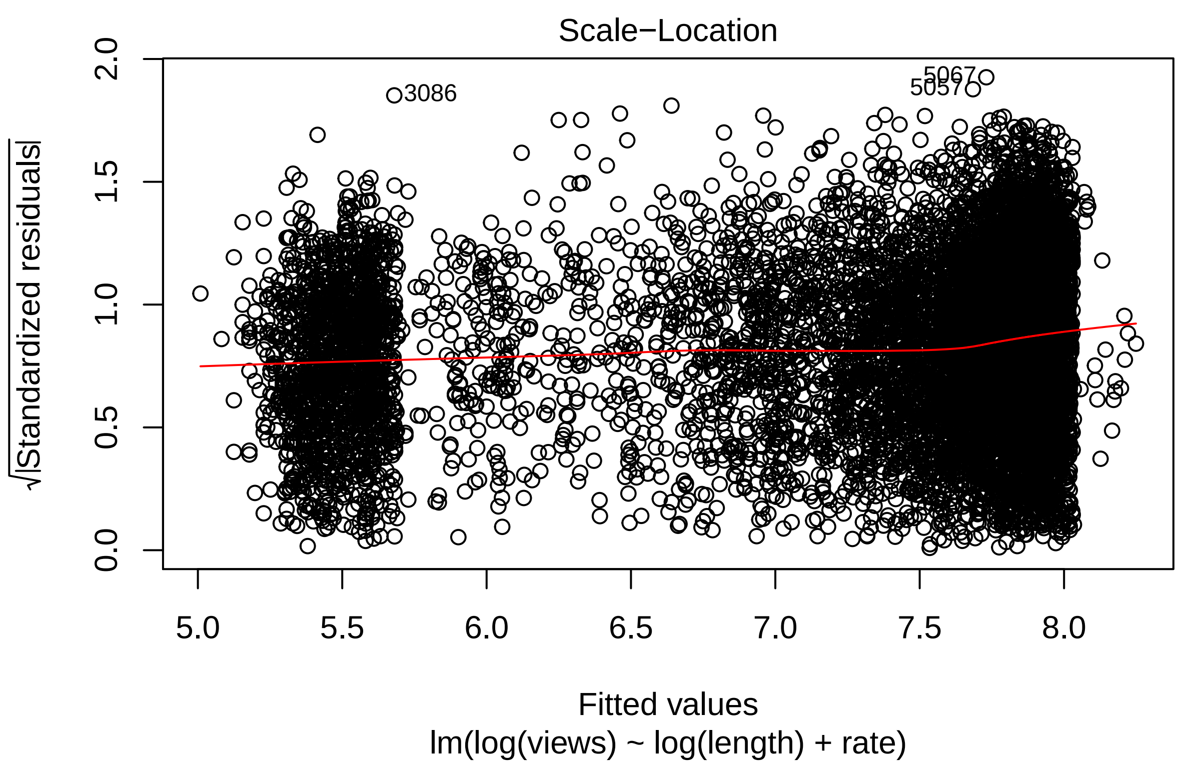
```r
plot(model1, which = 1 )
```



Overall, the conditional mean of residuals stay close to 0. Although we see some outliers around higher fitted values, it could be just due to a lack of data poitns around there.

e. Homoskedasticity

```
plot(model1, which = 3)
```



```r
bptest(model1)
```

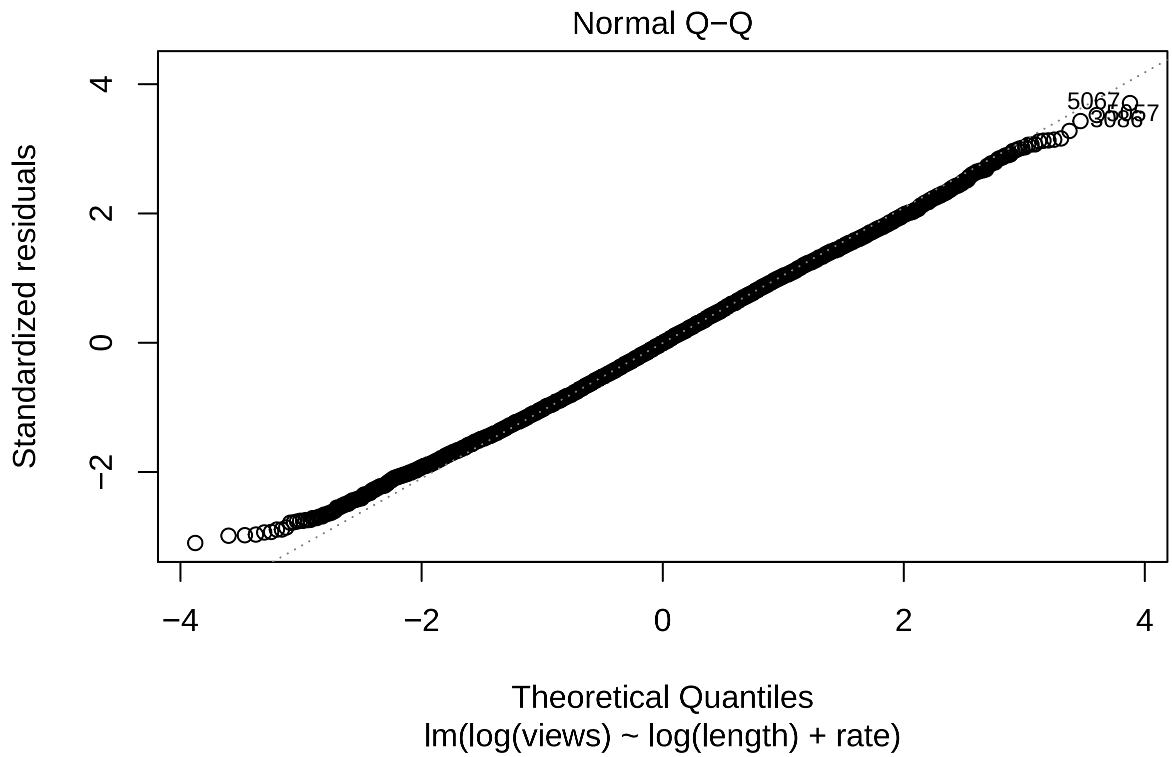
```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 122.69, df = 2, p-value < 2.2e-16
```

According to the Scale-Location graph, the variance seems pretty close across different fitted values. This implies homoskedasticity for our linear model.

However, the Breusch-Pagan test results show strong statistical significance, rejecting the null hypothese os homoskedasticity. This could be caused by the large sample size but we should be cautious about the this assumption when testing our parameters.

f. Normality of Errors

```r
plot(model1, which = 2)
```



The QQ plot of the residuals suggest normality of errors for our linear model

3. Generate a printout of your model coefficients, complete with standard errors that are valid given your diagnostics. Comment on both the practical and statistical significance of your coefficients.

Since we aren't sure about the homoskedasticity of our linear model, we should use the

```r
# To address heteroskedasticity, we use robust standard errors.
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 5.0088242  0.0884376 56.6368 < 2.2e-16 ***
```

```
## log(length) 0.1053702  0.0179951  5.8555 4.914e-09 ***
## rate        0.4673962  0.0096753 48.3084 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the t test, we see statistical significance of the intercept and slope parameter of "rate". More specifically, if we hold "length" constant, then an increase of 1 in the average rating is associated with ~ ~ 188% increase in the views, which is practically significant.