# Homework 12

*Joanna Yu (W203 Tuesday 4pm Fall 2018)*

*12/4/2018*

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

```
library(stargazer)
```

```
## 
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

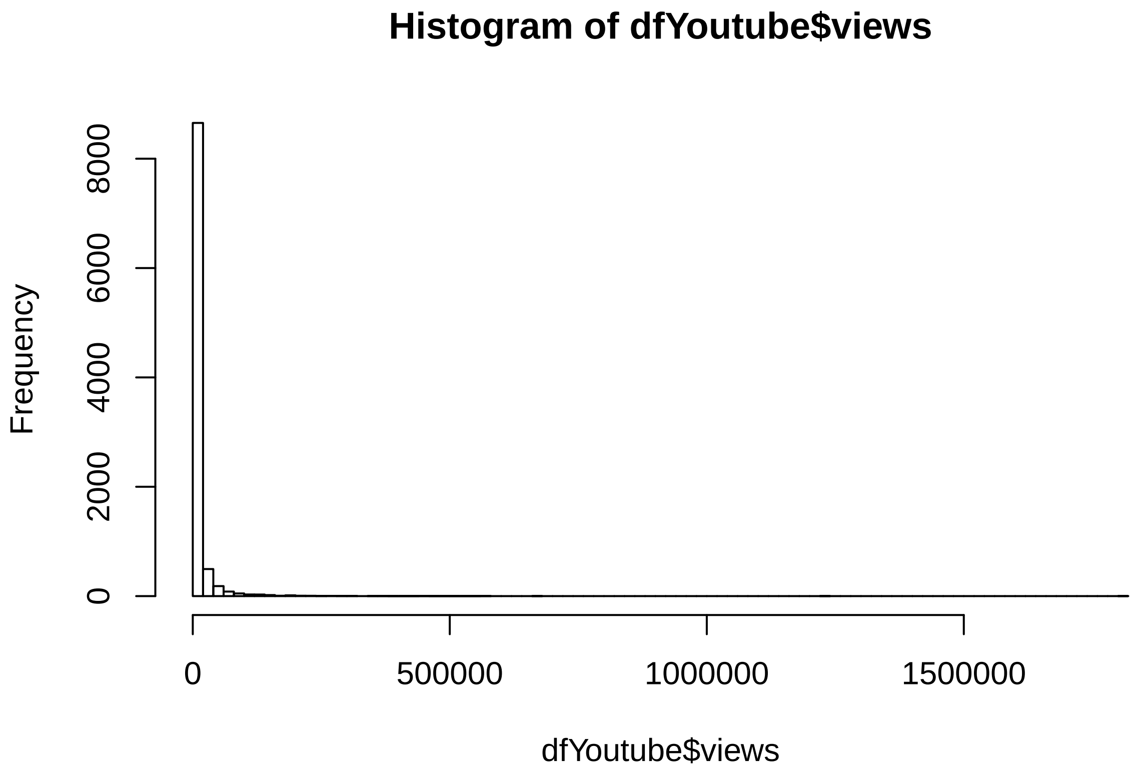
```
## 
## Attaching package: 'zoo'
```

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

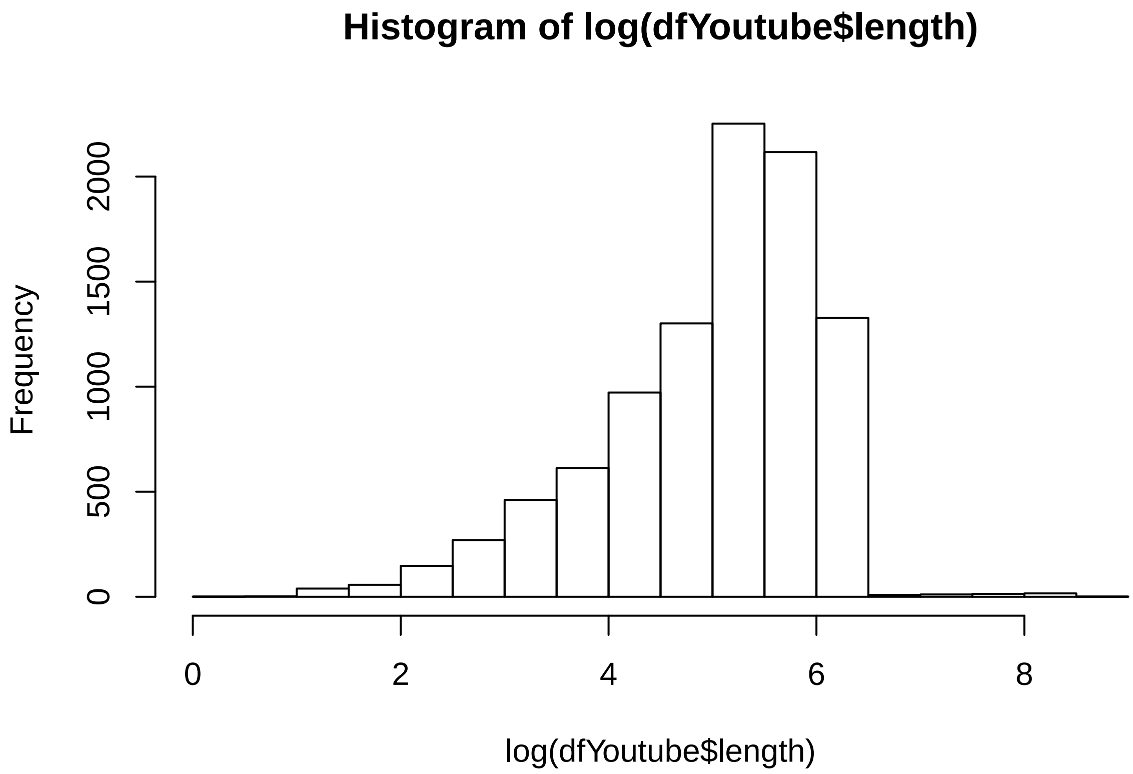
```
library(sandwich)
dfYoutube = read.delim("videos.txt")
summary(dfYoutube)
```

```
##         video_id                uploader          age      
##  #NAME?     : 129   Pan93bn        :  56   Min.   :   0  
##  __zVzDy4MOM:   1   nikodora       :  28   1st Qu.: 920  
##  _-TUODhKgcs:   1   gar6301        :  22   Median :1115  
##  _-VVIFAn7xw:   1   WWEOfficialPPVs:  22   Mean   :1045  
##  _0FCaXY42Yw:   1   dermayon       :  20   3rd Qu.:1226  
##  _0LdlpFQfa8:   1   wishinonastar07:  20   Max.   :1258  
##  (Other)    :9484   (Other)        :9450   NA's   :9     
##              category        length          views               rate      
##  Music           :2676   Min.   :   1   Min.   :      3   Min.   :0.000  
##  Entertainment   :2240   1st Qu.:  83   1st Qu.:    348   1st Qu.:3.400  
##  People & Blogs  : 811   Median : 193   Median :   1453   Median :4.670  
##  Film & Animation: 810   Mean   : 227   Mean   :   9346   Mean   :3.744  
##  Comedy          : 621   3rd Qu.: 299   3rd Qu.:   6179   3rd Qu.:5.000  
##  Sports          : 568   Max.   :5289   Max.   :1807640   Max.   :5.000  
##  (Other)         :1892   NA's   :9      NA's   :9         NA's   :9      
##     ratings           comments       
##  Min.   :   0.00   Min.   :   -2.00  
##  1st Qu.:   1.00   1st Qu.:    1.00  
##  Median :   5.00   Median :    3.00  
##  Mean   :  20.66   Mean   :   19.99  
##  3rd Qu.:  15.00   3rd Qu.:   13.00  
##  Max.   :3801.00   Max.   :13211.00  
##  NA's   :9         NA's   :9         
```

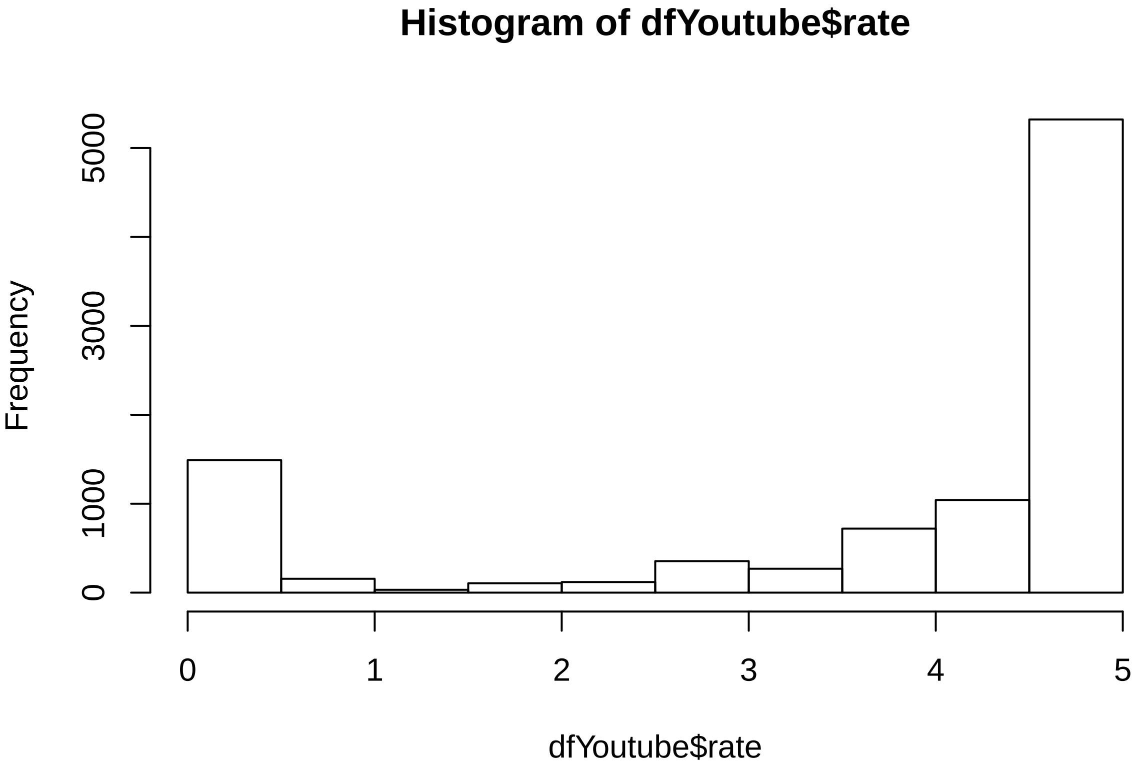
```r
hist(dfYoutube$views, breaks=100)
```



```r
hist(log(dfYoutube$length))
```



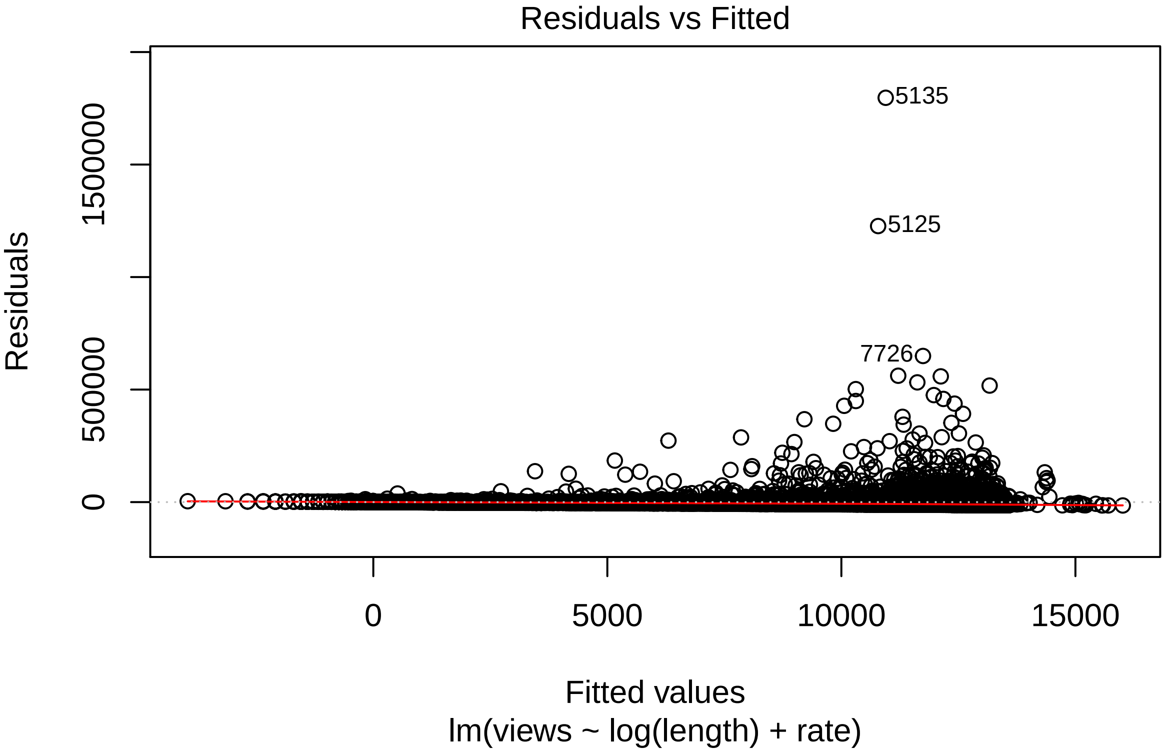
```
hist(dfYoutube$rate)
```

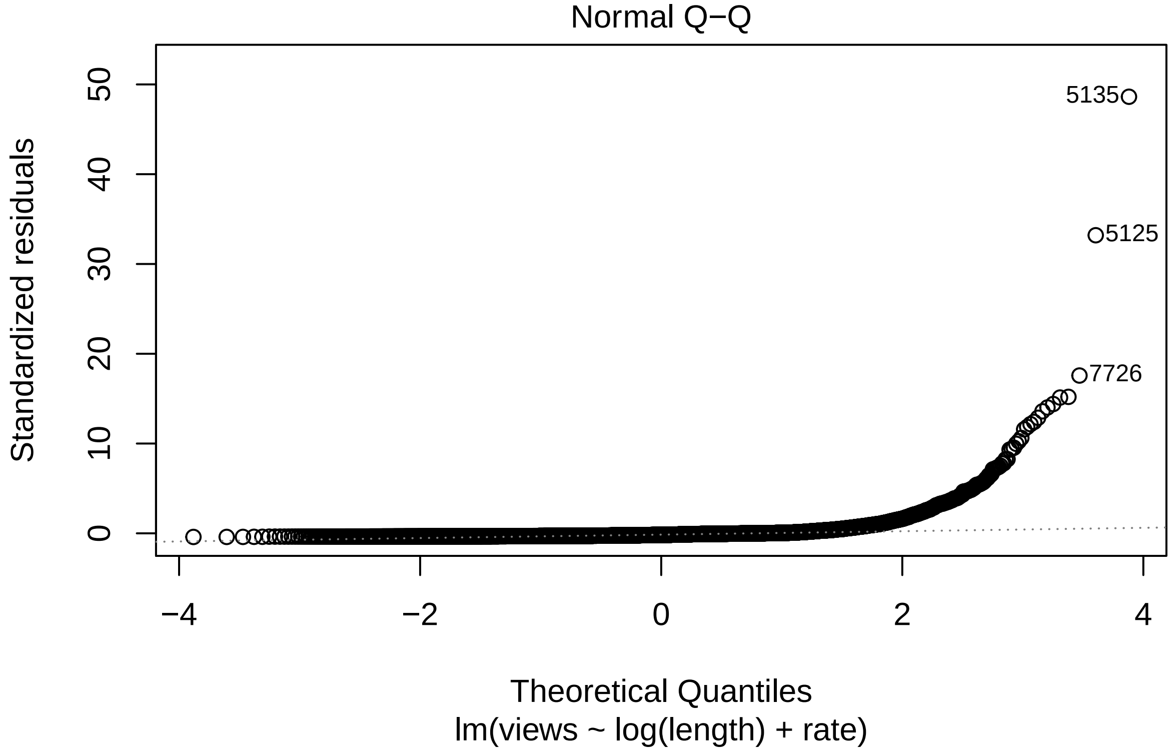


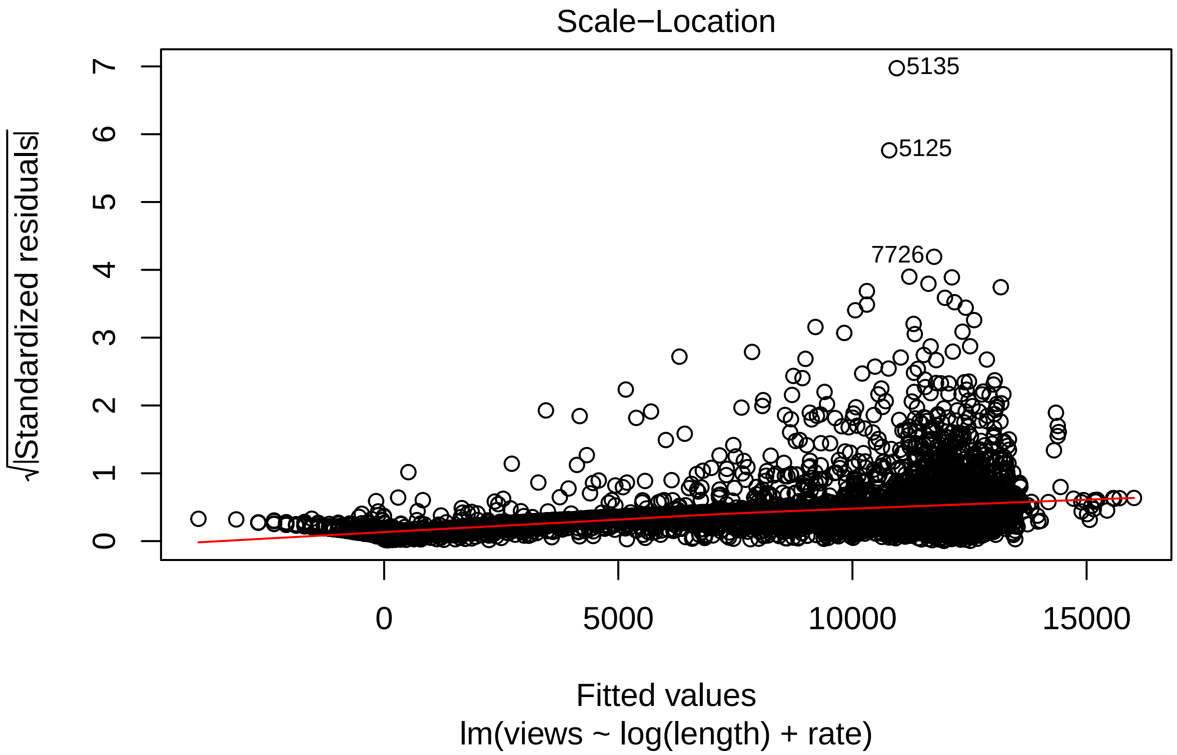
```
model_view_len_rate = lm(views ~ log(length) + rate, data = dfYoutube)
```

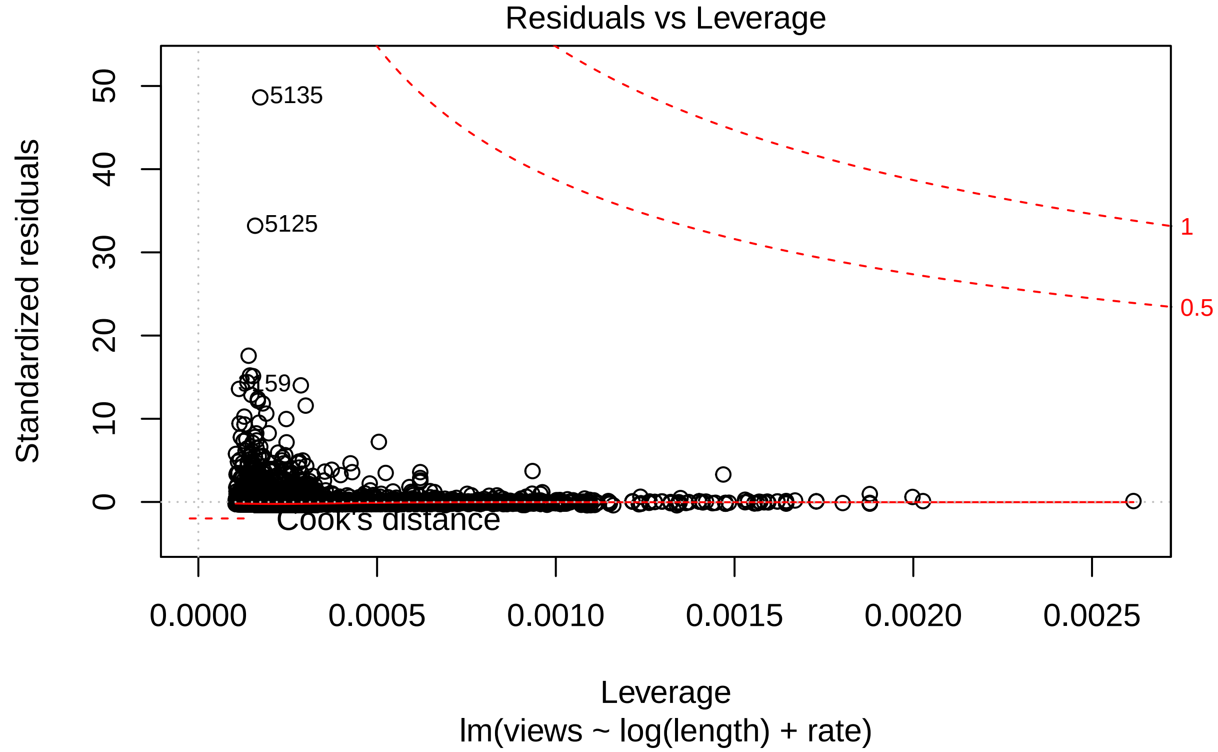
2. Using diagnostic plots, background knowledge, and statistical tests, assess all 6 assumptions of the CLM. When an assumption is violated, state what response you will take.

```
plot(model_view_len_rate)
```

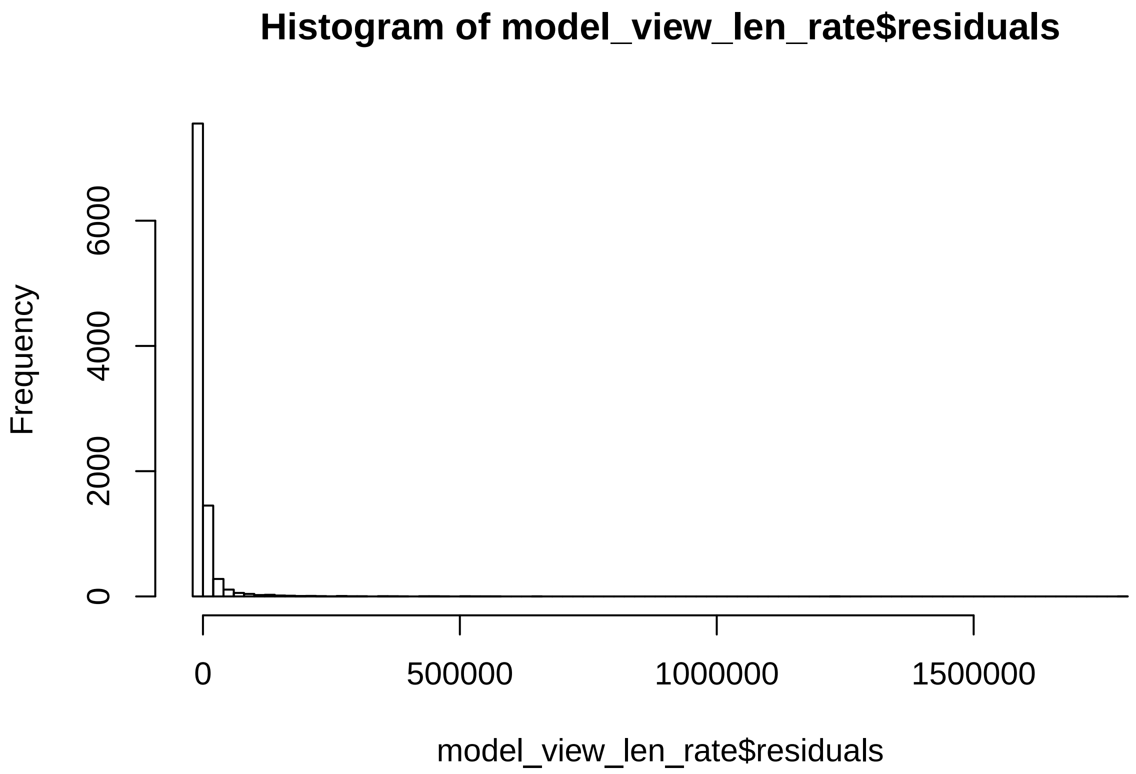

Residuals vs Fitted
Residuals
0
500000
1500000
0
5000
10000
15000
Fitted values
lm(views ~ log(length) + rate)
5135
5125
7726


Normal Q−Q
Standardized residuals
0
10
20
30
40
50
−4
−2
0
2
4
Theoretical Quantiles
lm(views ~ log(length) + rate)
5135
5125
7726





```
hist(model_view_len_rate$residuals, breaks=100)
```



```r
summary(model_view_len_rate)
```

```
##
## Call:
## lm(formula = views ~ log(length) + rate, data = dfYoutube)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -14941  -10202   -6442    -729 1796693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3965.8     1889.9  -2.098  0.03589 *
## log(length)    1164.6      375.2   3.104  0.00192 **
## rate           1998.6      217.6   9.186  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36950 on 9606 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.01186,    Adjusted R-squared:  0.01166
## F-statistic: 57.67 on 2 and 9606 DF,  p-value: < 2.2e-16
```

```r
m = data.matrix(subset(dfYoutube, select = c("length", "rate")))
(cor = cor(m))
```

```
##        length rate
## length      1   NA
## rate       NA    1
```

CLM assumptions:

1) Linearity - the model is a linear function.

2) Random Sampling - it's unclear if the sample is random. From the summary of the "uploader" variable, the sample contains 56 videos from one of the users. With so many Youtube users, it seems unlikely that a random sample will pick 56 videos from a single user. But it could happen if the data is drawn from the early days of Youtube when there were fewer users. If random sampling is violated, there will be bias in the data.

3) Multicollinearity - the two independent variables are not perfectly correlated.

4) Zero-Conditional Mean - based on the Residuals vs Fitted plot, we can see that the spline curve is a straight line along 0. We have zero conditional mean.

5) Homoskedasticity - from the Scale-Location plot, we can see that heteroskedasticity is present.

6) Residual Normality - based on the residual plot, we can see that the residuals are not normally distributed. Based on the Normal QQ plot, the errors does not have a normal distribution. We have a violation of the normality of the errors. However, since the sample size is pretty big, we can still rely on asymptotics.

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.

```
se.model_view_len_rate = sqrt(diag(vcovHC(model_view_len_rate)))

stargazer(model_view_len_rate, type = "text", omit.stat = "f",
          se = list(se.model_view_len_rate), star.cutoffs = c(0.05, 0.01, 0.005) )
```

```
##
## ==================================================
##                           Dependent variable:
##                      ------------------------------
##                                 views
## --------------------------------------------------
## log(length)                 1,164.624***
##                               (255.241)
##
## rate                        1,998.631***
##                               (127.689)
##
## Constant                   -3,965.785***
##                              (1,150.226)
##
## --------------------------------------------------
## Observations                    9,609
## R2                              0.012
## Adjusted R2                     0.012
## Residual Std. Error    36,950.620 (df = 9606)
## ==================================================
## Note:                *p<0.05; **p<0.01; ***p<0.005
```

Based on the p values, the coefficients seem statistically significant. I dont think this has high practical significance because the model is too naive with too many omitted variables. Also it's unclear if there is random sampling.