Unit 12 Homework

More regression analysis of YouTube dataset

You want to explain how much the quality of a video affects the number of views it receives on social media. **This is a causal question.**

You will use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations about 9618 videos shared on YouTube. Please see this link for details about how the data was collected. You will use the following variables:

- views: the number of views by YouTube users
- rate: the average rating given by users
- length: the duration of the video in seconds

A) Perform a brief EDA on the data to discover patterns, outliers, or wrong data entries and summarize your findings.

Import videos dataset with first row as the header. Total of 9489 rows.

```
videos <- read.table("videos.txt",sep="\t",header=TRUE)
```

Delete rows for which key values are blank. 9 rows eliminated.

```
vid <- na.omit(videos)
```

Summarize key variables.

```
summary(vid$views)
```

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

```
summary(vid$rate)
```

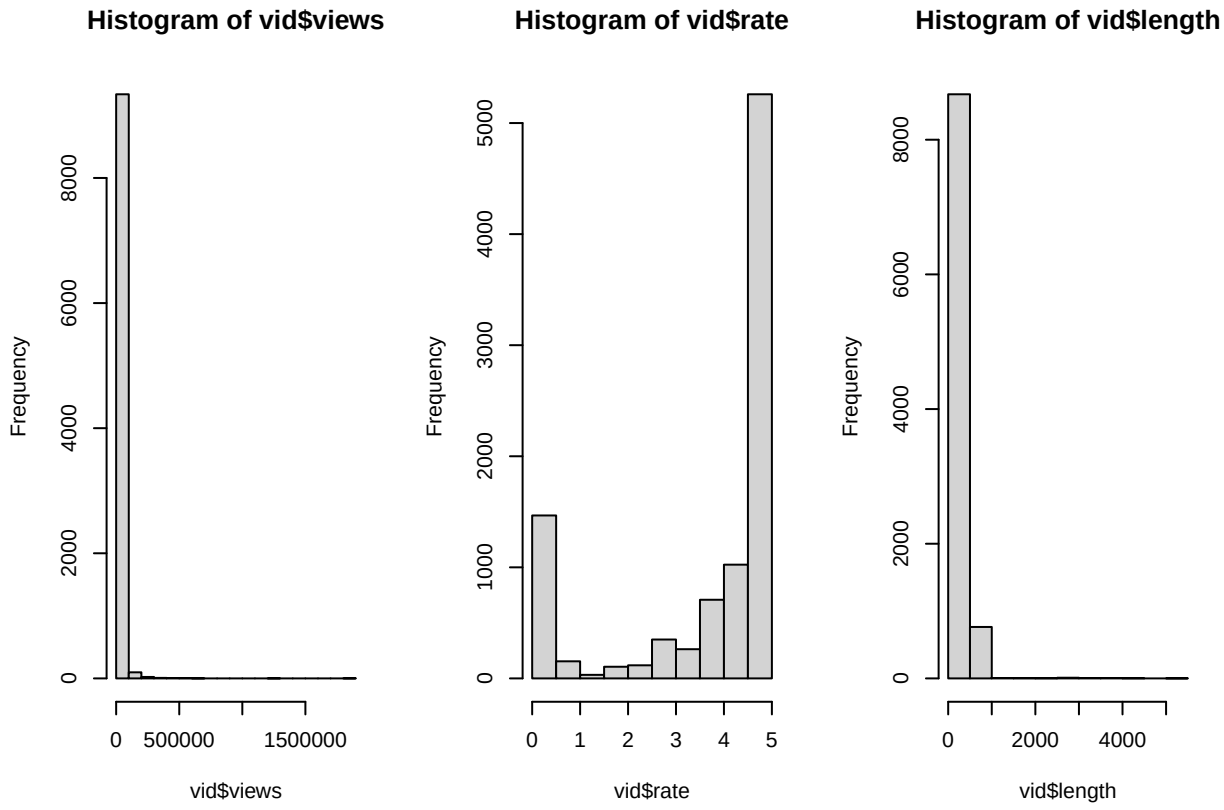
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.400   4.670   3.746   5.000   5.000
```

```
summary(vid$length)
```

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

Histogram of key variables.

```
par(mfrow=c(1,3))
hist(vid$views)
hist(vid$rate)
hist(vid$length)
```



EDA Summary

The videos dataset was imported with 9489 observations of 9 variables - video ID, uploader, video age, category, length, number of views, average viewer rating, number of viewer ratings, and number of comments. The dataset contained 9 observations with blank values for one or more of the three variables of interest - length, views, and average viewer rating. These 9 rows were removed.

A summary was performed on each of the 3 variables of interest, and the lack of warnings or errors indicates that all of the data imported as integers or doubles as expected. A histogram of the 3 variables of interest was generated to observe the distribution their values. Using these two analytical tools revealed the following:

- The *views* variable is always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers. This indicates that this variable is a good candidate for a logarithmic transform.
- The *rate* variable is always positive, spans from 0 to 5, and clusters near the extremes.
- The *length* variable is always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers. always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers

B) Based on your EDA, select an appropriate variable transformation (if any) to apply to each of your three variables. You will fit a model of the type:

$$f(views) = \beta_0 + \beta_1 g(rate) + \beta_3 h(length)$$

Transformations:
- *views* : Log Transformation
- *rate* : No Transformation
- *length* : Log transformation

```r
lviews = log(vid$views)
rate = vid$rate
llength = log(vid$length)
```

Generate Linear Model

```r
model = lm(lviews ~ rate + llength)
summary(model)
```

```
##
## Call:
## lm(formula = lviews ~ rate + llength)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5782 -1.2792 -0.0122  1.2629  6.6768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.00882    0.09264  54.069  < 2e-16 ***
## rate         0.46740    0.01068  43.754  < 2e-16 ***
## llength      0.10537    0.01840   5.726 1.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 9477 degrees of freedom
## Multiple R-squared:  0.1891, Adjusted R-squared:  0.1889
## F-statistic:  1105 on 2 and 9477 DF,  p-value: < 2.2e-16
```

C) Using diagnostic plots, background knowledge, and statistical tests, assess all five assumptions of the CLM. When an assumption is violated, state what response you will take. As part of this process, you should decide what transformation (if any) to apply to each variable.

I.I.D

The data was collected using a crawler that, once it catalogued a video, it then catalogued all videos that appeared in the "Related videos" tab. This indicates that the data is not i.i.d, as one video sampled will likely impact the next video that is sampled. However, this dataset is only a small portion of the tens of thousands of videos crawled and thus can be treated as i.i.d within that population.

No Perfect Collinearity

There is no perfect collinearity between the three variables used in this model. This assumption is met.

```r
cor(lviews, rate)
```

```
## [1] 0.4316351
```

```r
cor(llength, rate)
```

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

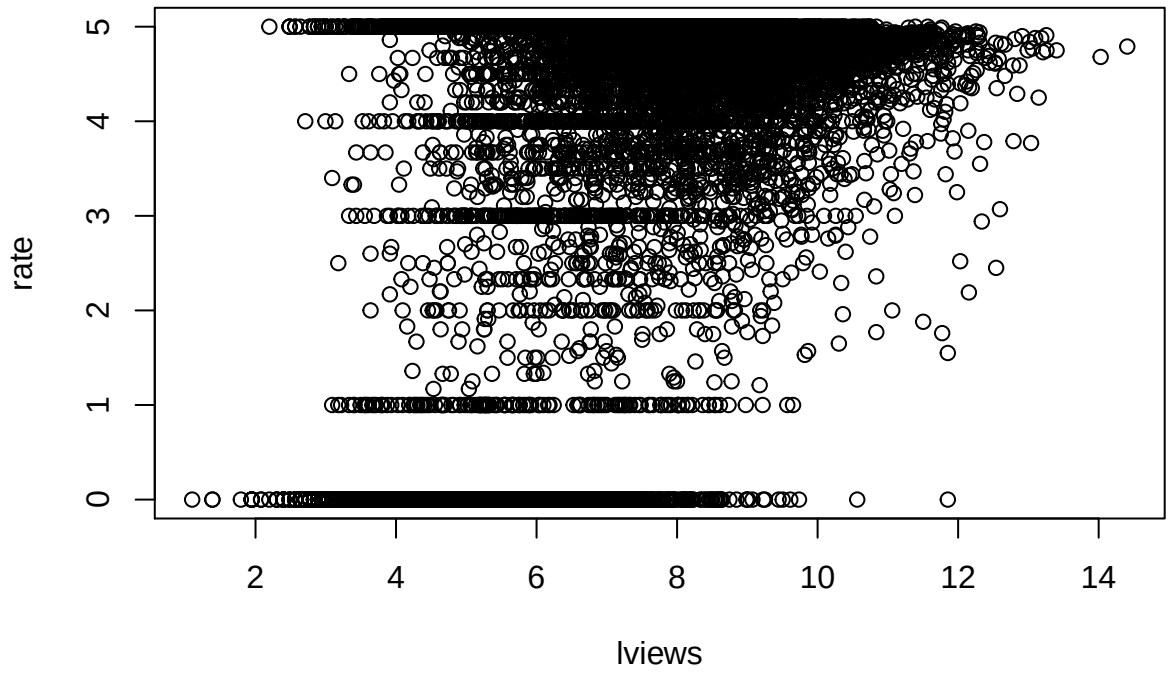
```r
cor(lviews,llength)
```

```
## [1] 0.1590971
```

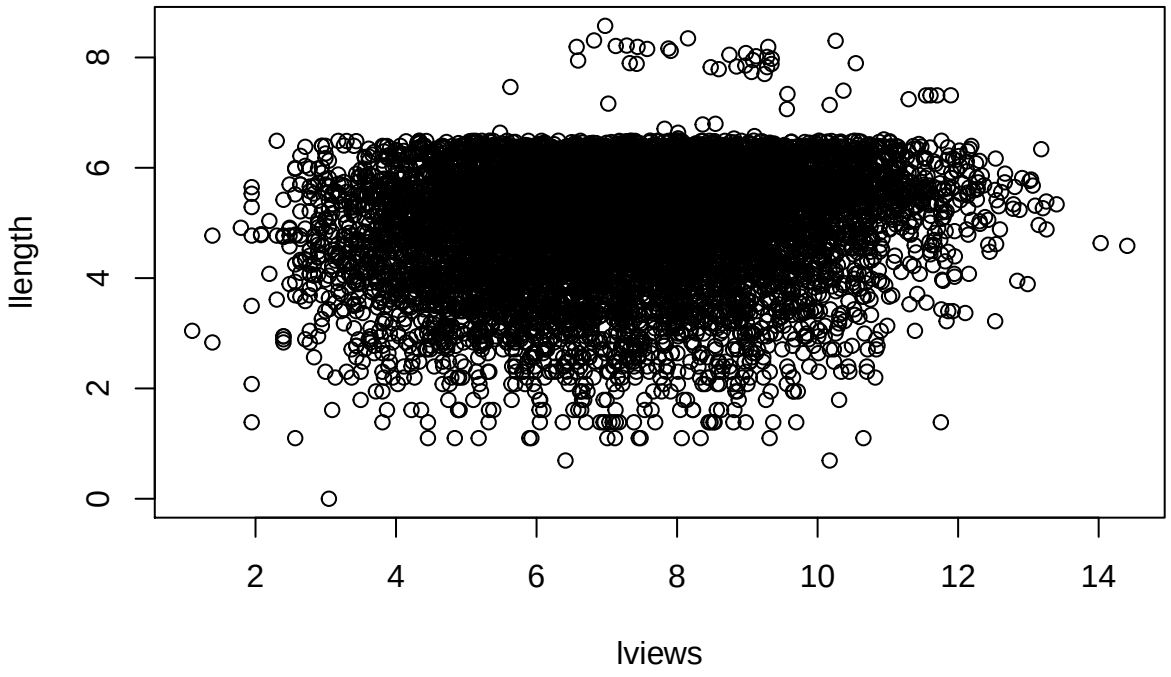
Linear Conditional Expectation

The relationship between rating, log transformed views, and log transformed length is very messy. There is not clear evidence of a linear relationship, but also no clear evidence of any other higher order relationship. The residual plots below seem to confirm this. As a result, this assumption is considered met.

```
plot(lviews,rate)
```



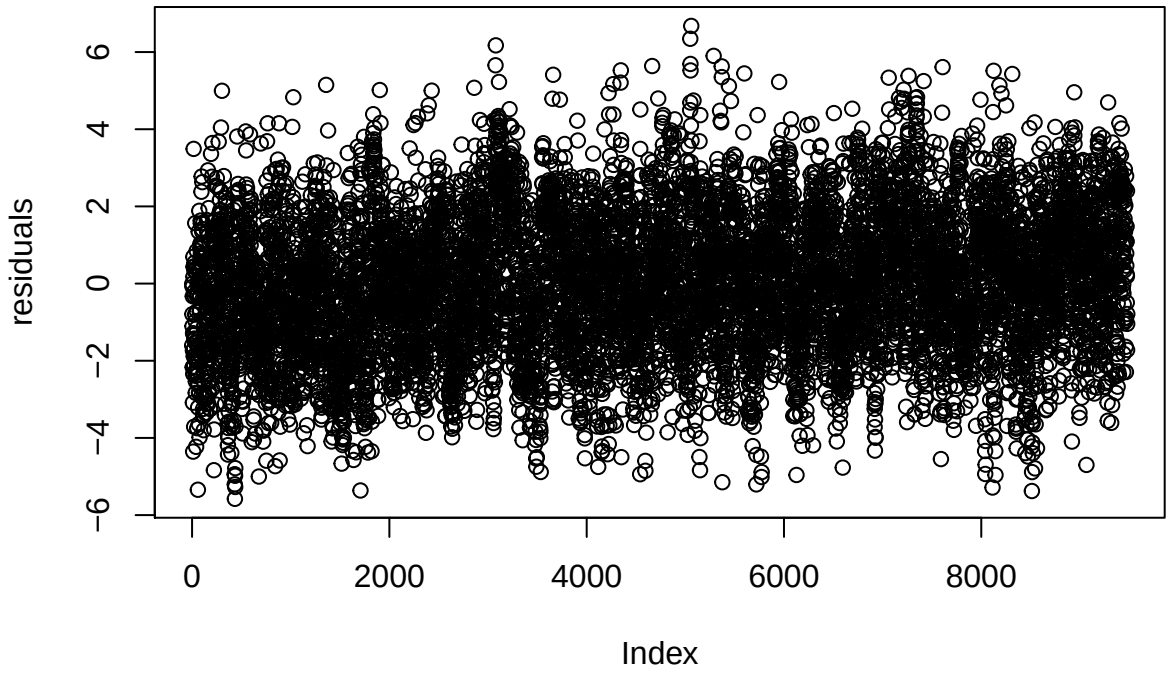

```
plot(lviews,llength)
```



Homoskedastic Errors

The residual plot below shows no higher order pattern in the residuals of the model generated. This condition is met.

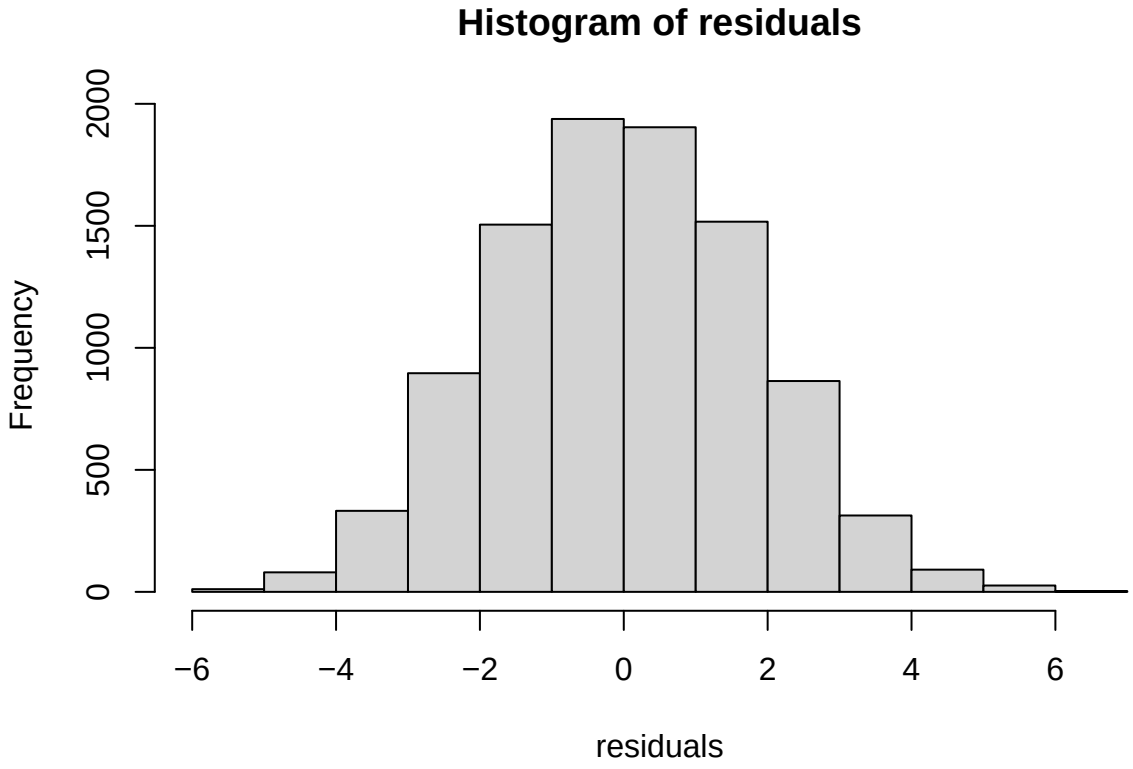
```
residuals = resid(model)
plot(residuals)
```



Normally Distributed Errors

the histogram of the residuals shows a normal distribution centered at zero. This condition is met.

```
hist(residuals)
```



More regression analysis of YouTube dataset

You want to explain how much the quality of a video affects the number of views it receives on social media. **This is a causal question.**

You will use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations about 9618 videos shared on YouTube. Please see this link for details about how the data was collected. You will use the following variables:

- views: the number of views by YouTube users
- rate: the average rating given by users
- length: the duration of the video in seconds

A) Perform a brief EDA on the data to discover patterns, outliers, or wrong data entries and summarize your findings.

Import videos dataset with first row as the header. Total of 9489 rows.

```
videos <- read.table("videos.txt",sep="\t",header=TRUE)
```

Delete rows for which key values are blank. 9 rows eliminated.

```
vid <- na.omit(videos)
```

Summarize key variables.

```
summary(vid$views)
```

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

```
summary(vid$rate)
```

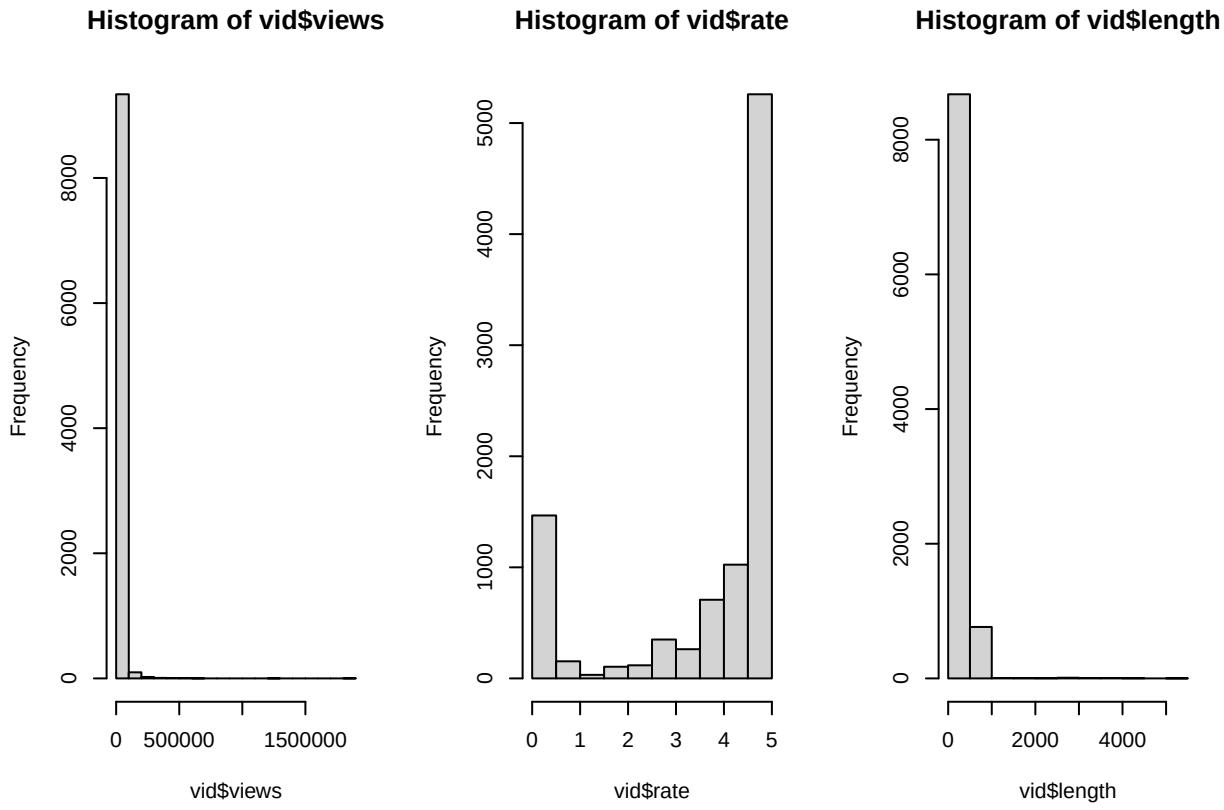
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.400   4.670   3.746   5.000   5.000
```

```
summary(vid$length)
```

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

Histogram of key variables.

```
par(mfrow=c(1,3))
hist(vid$views)
hist(vid$rate)
hist(vid$length)
```



EDA Summary

The videos dataset was imported with 9489 observations of 9 variables - video ID, uploader, video age, category, length, number of views, average viewer rating, number of viewer ratings, and number of comments. The dataset contained 9 observations with blank values for one or more of the three variables of interest - length, views, and average viewer rating. These 9 rows were removed.

A summary was performed on each of the 3 variables of interest, and the lack of warnings or errors indicates that all of the data imported as integers or doubles as expected. A histogram of the 3 variables of interest was generated to observe the distribution their values. Using these two analytical tools revealed the following:

- The *views* variable is always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers. This indicates that this variable is a good candidate for a logarithmic transform.
- The *rate* variable is always positive, spans from 0 to 5, and clusters near the extremes.
- The *length* variable is always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers. always positive, spans multiple orders of magnitude, cluster near zero, and have several high outliers

B) Based on your EDA, select an appropriate variable transformation (if any) to apply to each of your three variables. You will fit a model of the type:

$$f(views) = \beta_0 + \beta_1 g(rate) + \beta_3 h(length)$$

Transformations:
- *views* : Log Transformation
- *rate* : No Transformation
- *length* : Log transformation

```r
lviews = log(vid$views)
rate = vid$rate
llength = log(vid$length)
```

Generate Linear Model

```r
model = lm(lviews ~ rate + llength)
summary(model)
```

```
##
## Call:
## lm(formula = lviews ~ rate + llength)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5782 -1.2792 -0.0122  1.2629  6.6768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.00882    0.09264  54.069  < 2e-16 ***
## rate         0.46740    0.01068  43.754  < 2e-16 ***
## llength      0.10537    0.01840   5.726 1.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 9477 degrees of freedom
## Multiple R-squared:  0.1891, Adjusted R-squared:  0.1889
## F-statistic:  1105 on 2 and 9477 DF,  p-value: < 2.2e-16
```

C) Using diagnostic plots, background knowledge, and statistical tests, assess all five assumptions of the CLM. When an assumption is violated, state what response you will take. As part of this process, you should decide what transformation (if any) to apply to each variable.

I.I.D

The data was collected using a crawler that, once it catalogued a video, it then catalogued all videos that appeared in the "Related videos" tab. This indicates that the data is not i.i.d, as one video sampled will likely impact the next video that is sampled. However, this dataset is only a small portion of the tens of thousands of videos crawled and thus can be treated as i.i.d within that population.

No Perfect Collinearity

There is no perfect collinearity between the three variables used in this model. This assumption is met.

```r
cor(lviews, rate)
```

```
## [1] 0.4316351
```

```r
cor(llength, rate)
```

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

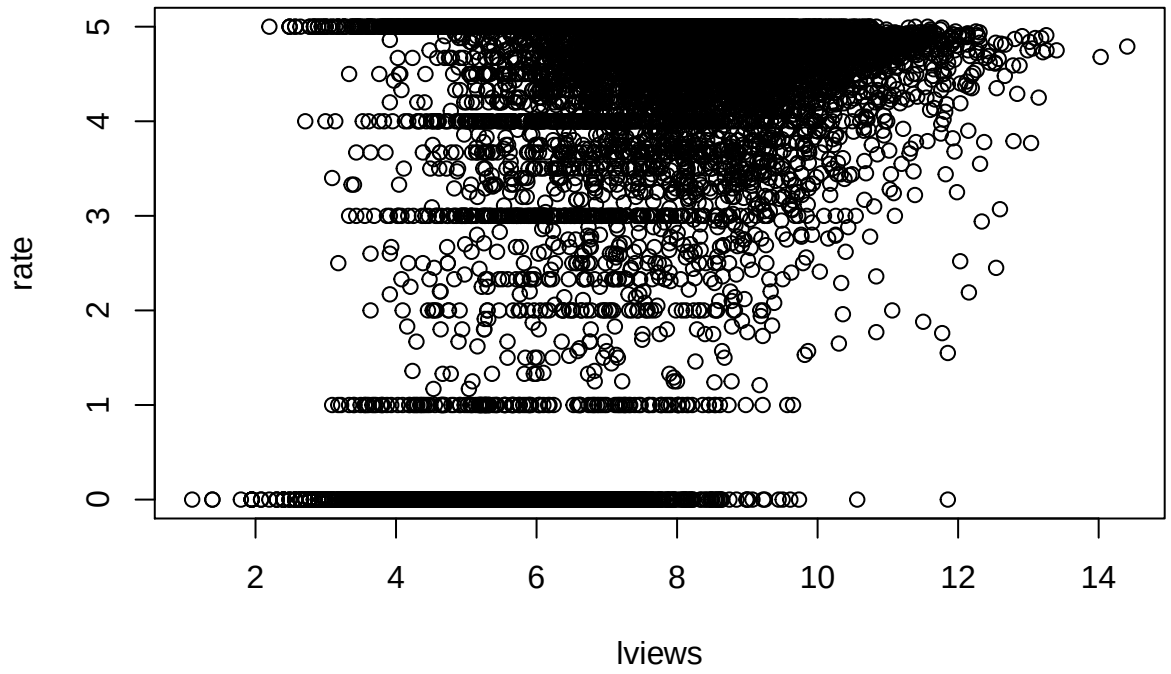
```r
cor(lviews,llength)
```

```
## [1] 0.1590971
```

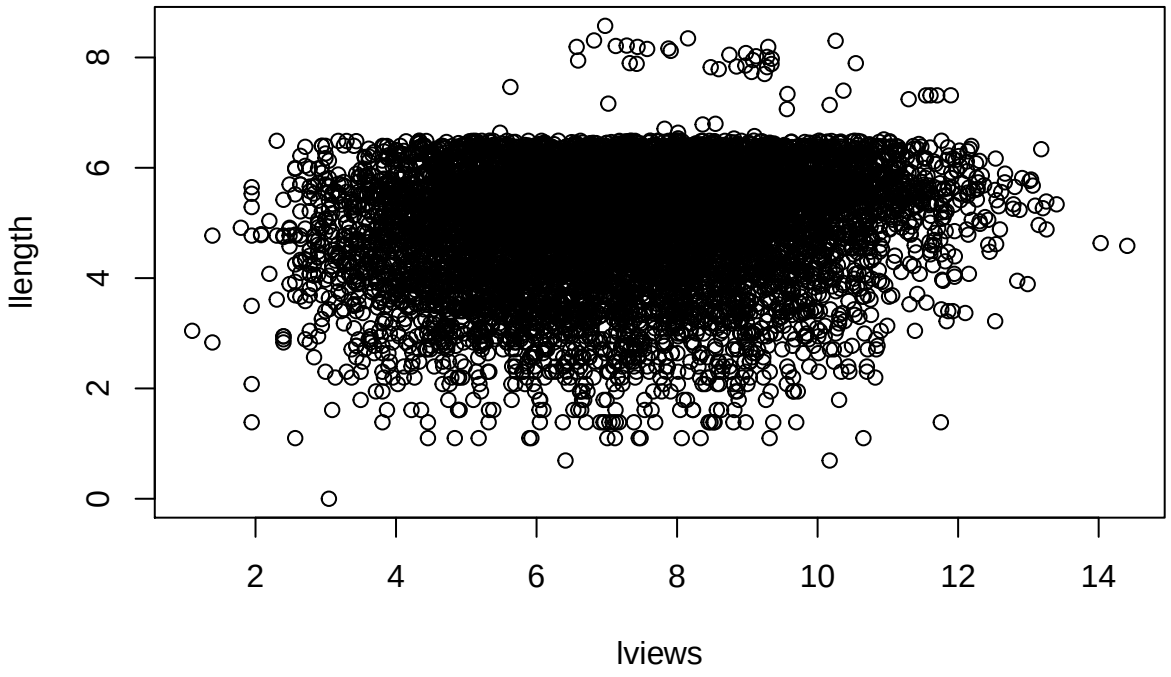
Linear Conditional Expectation

The relationship between rating, log transformed views, and log transformed length is very messy. There is not clear evidence of a linear relationship, but also no clear evidence of any other higher order relationship. The residual plots below seem to confirm this. As a result, this assumption is considered met.

```
plot(lviews,rate)
```



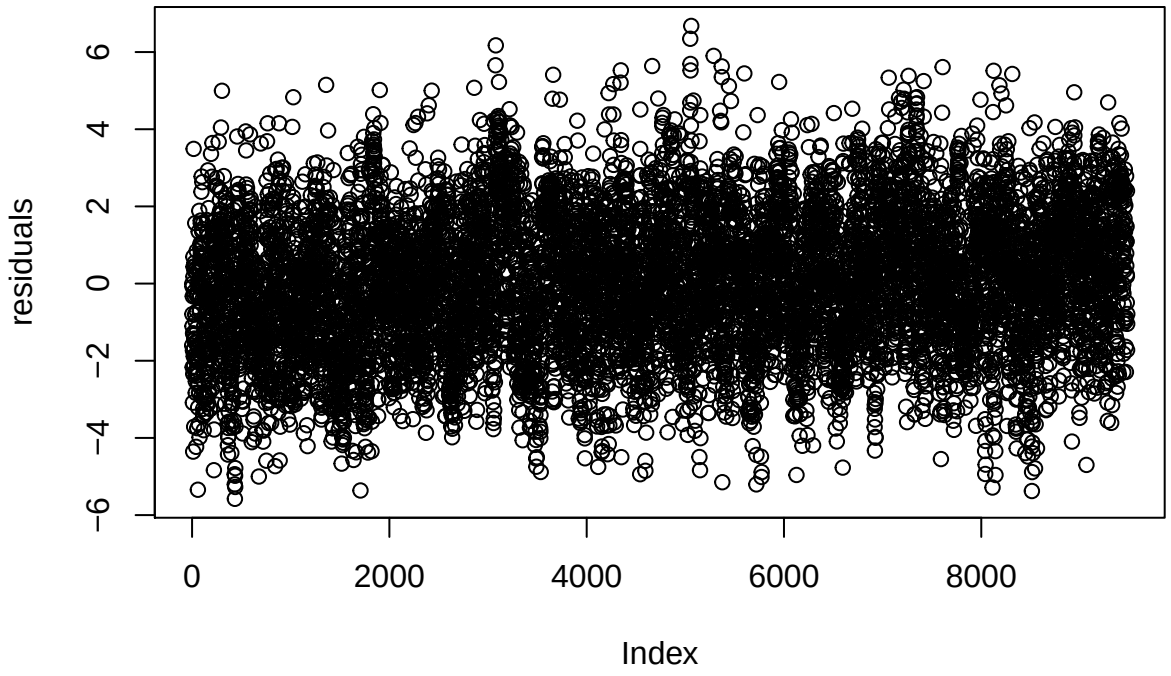

```
plot(lviews,llength)
```



Homoskedastic Errors

The residual plot below shows no higher order pattern in the residuals of the model generated. This condition is met.

```
residuals = resid(model)
plot(residuals)
```



Normally Distributed Errors

the histogram of the residuals shows a normal distribution centered at zero. This condition is met.

```
hist(residuals)
```