HW week 12

w203: Statistics for youtb_data Science

w203 teaching team

OLS Inference

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

```
youtb_data = read.delim("./videos.txt",header=TRUE , sep="\t")
summary(youtb_data)
```

```
##         video_id                uploader          age        
##  #NAME?     : 129   Pan93bn         :  56   Min.   :   0  
##  -0yS9zc_290:   1   nikodora        :  28   1st Qu.: 920  
##  -0z5PEZt_Wk:   1   gar6301         :  22   Median :1115  
##  -0Zkx9Sh6DU:   1   WWEOfficialPPVs:  22   Mean   :1045  
##  -1PT00GVE7k:   1   dermayon        :  20   3rd Qu.:1226  
##  -1RjRtQRoEc:   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
```

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

```
## Warning: package 'lmtest' was built under R version 3.3.3
```

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

```
## Warning: package 'zoo' was built under R version 3.3.3
```

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

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

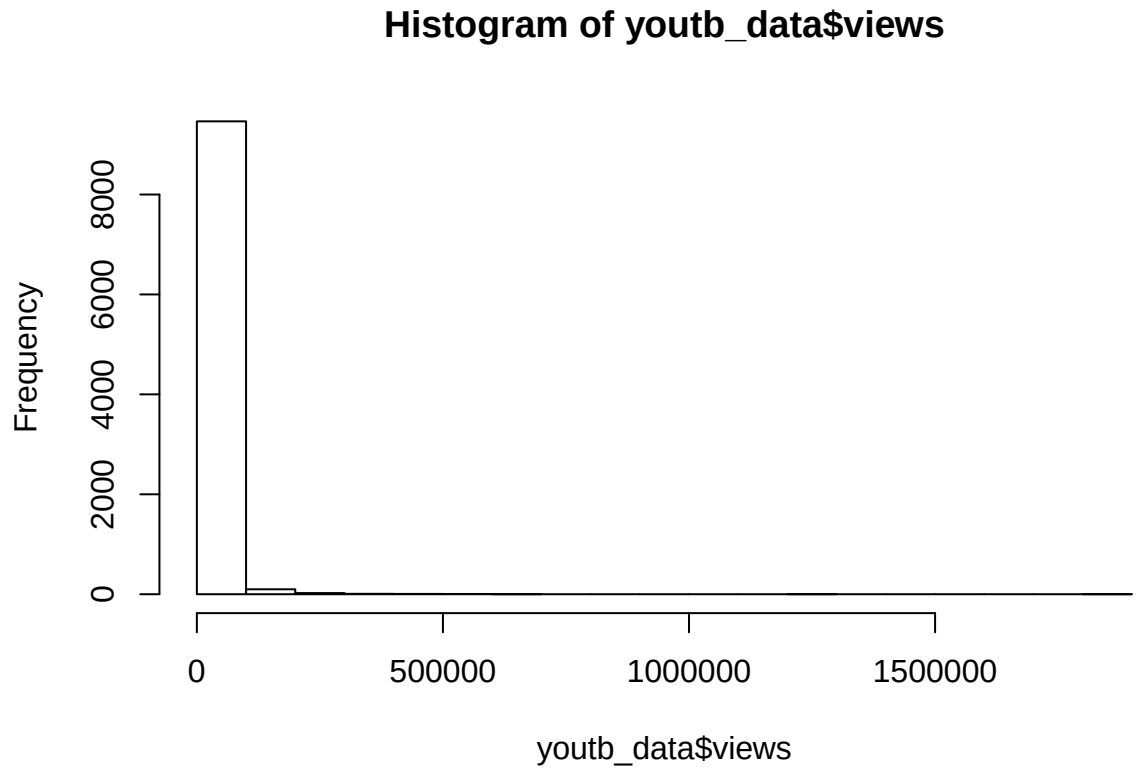
```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.3.3
```

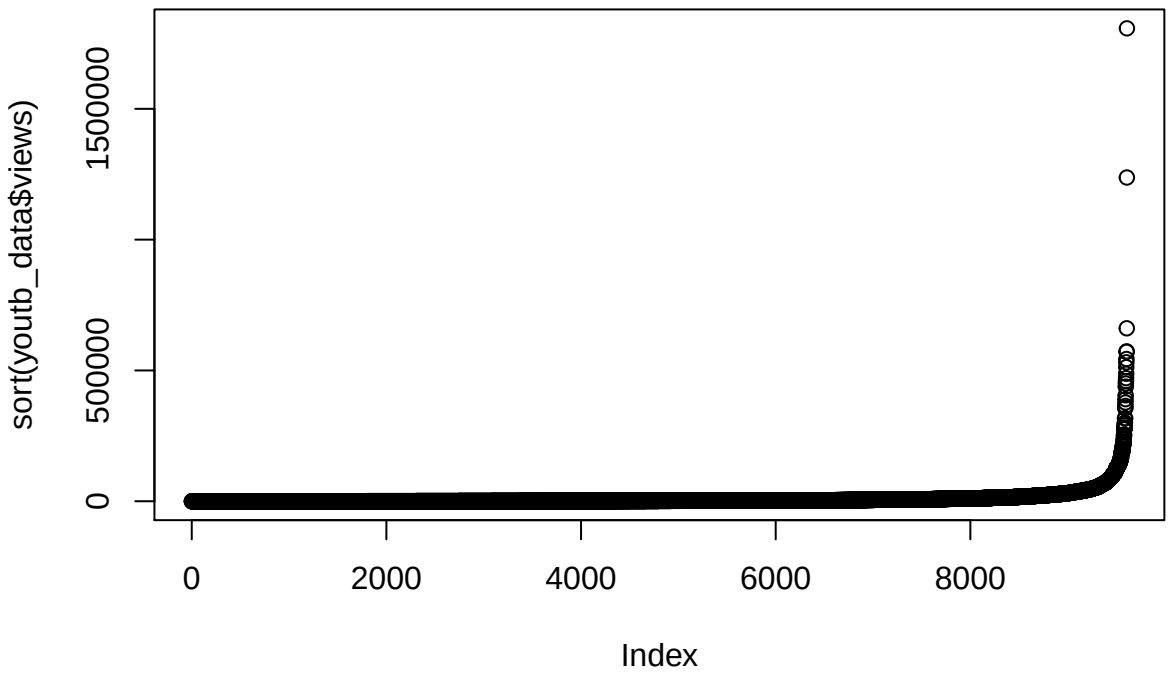
```
library(stargazer)
```

```
##
## 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
```

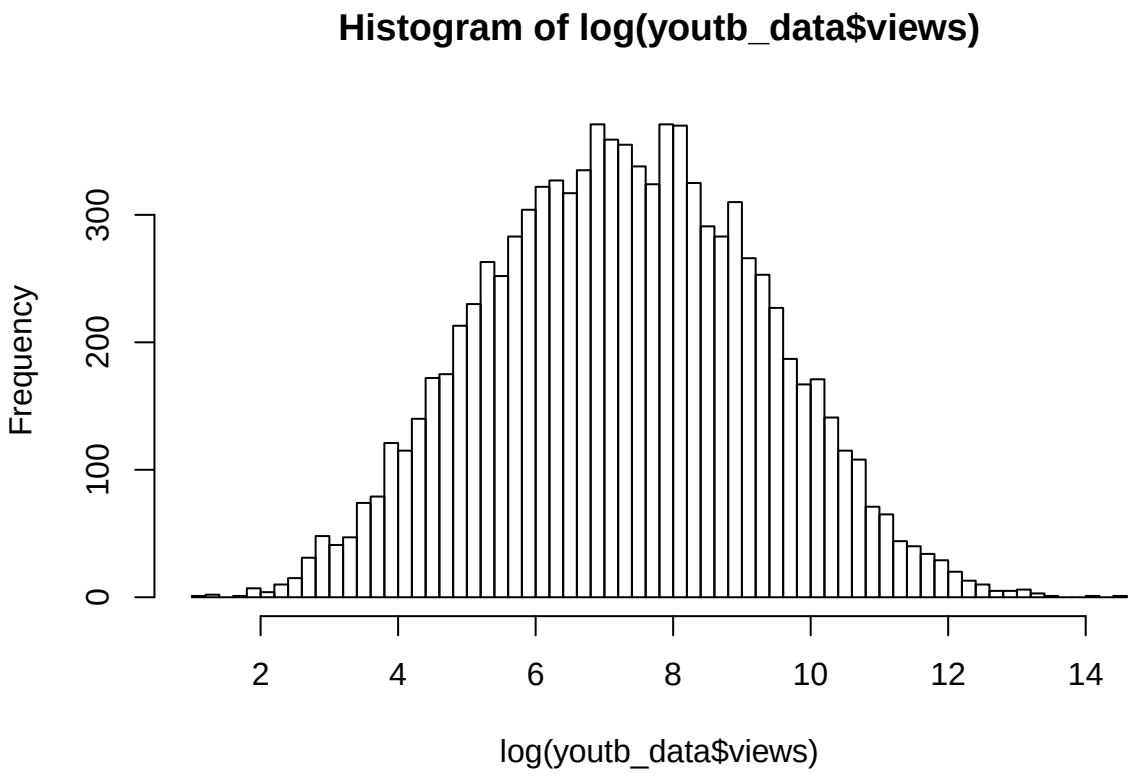
```
hist(youtb_data$views)
```



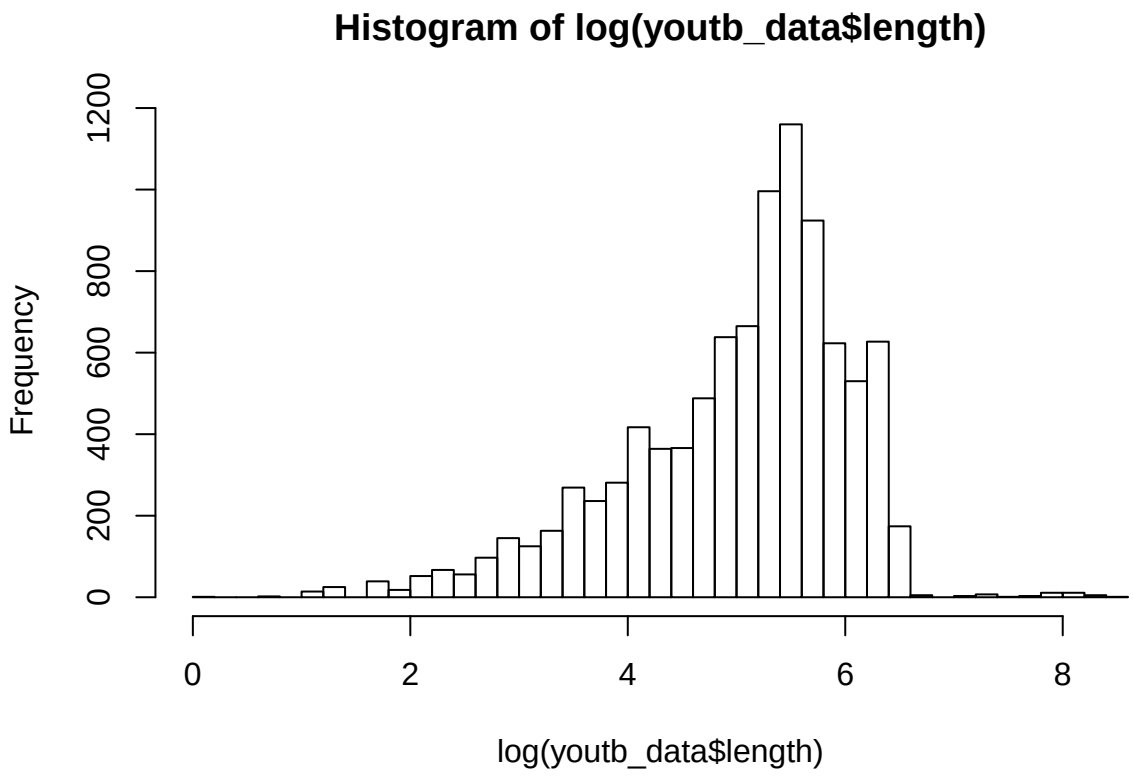

```
plot(sort(youtb_data$views) )
```



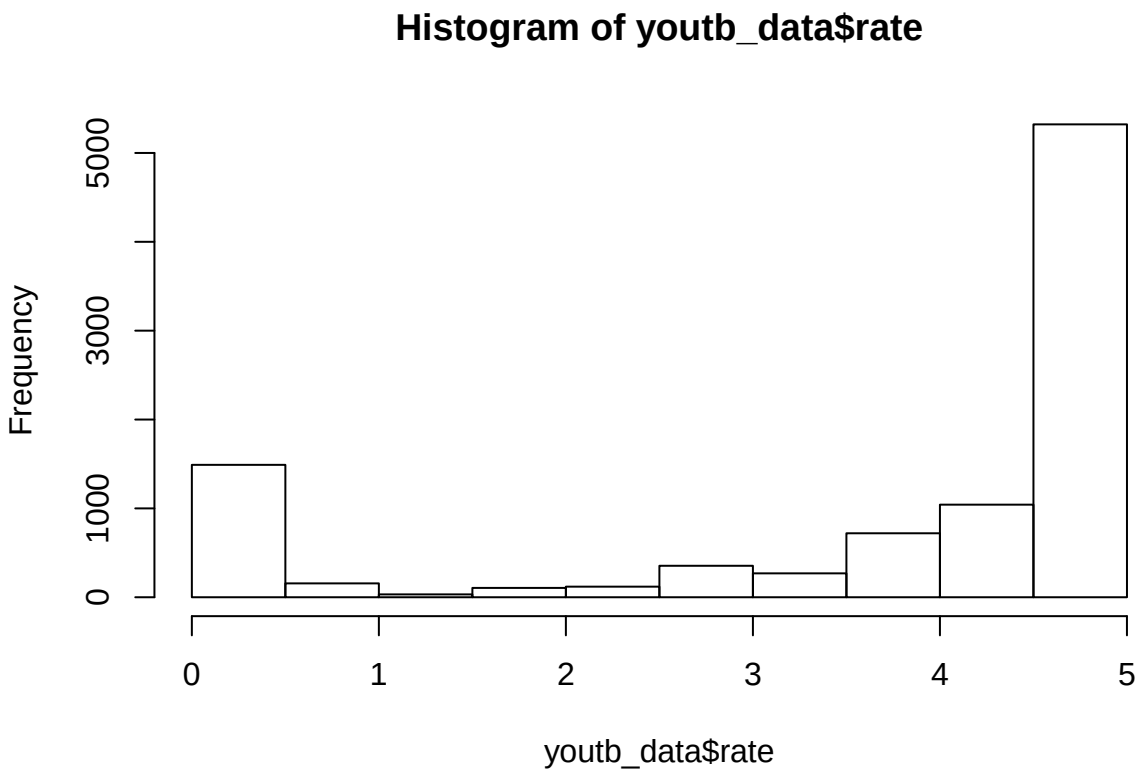

```
## From the plots we can see that the views are large variations in data.
##Also they have some outliers.

## We will be take the log of views and log of length for this model
## Let us look at the histogram of log(view), log(length) and rate
hist(log(youtb_data$views),breaks=50)
```




```
hist(log(youtb_data$length),breaks=50)
```




```
hist(youtb_data$rate)
```

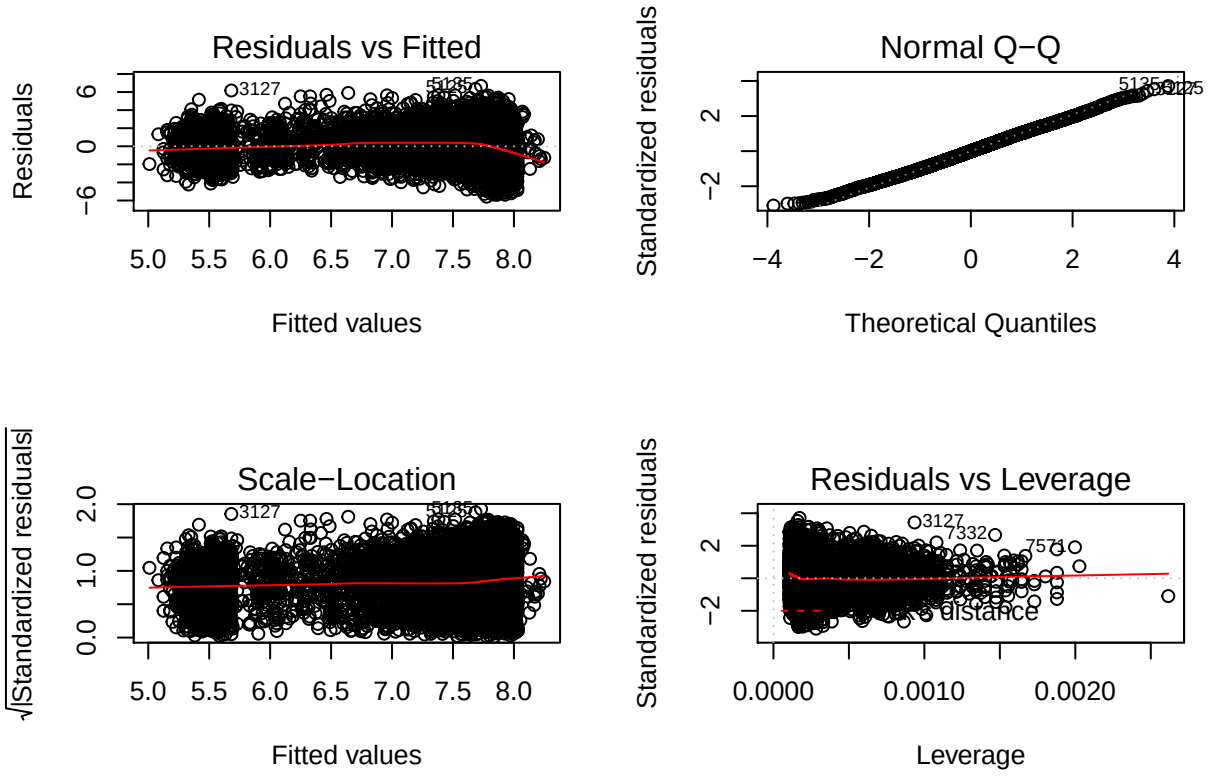


1. Fit a linear model predicting the number of views (views), from the length of a video (length) and its average user rating (rate).

```
model1 = lm(log(views)~rate + log(length) , data = youtb_data, na.action = na.omit)
summary(model1)
```

```
##
## Call:
## lm(formula = log(views) ~ rate + log(length), data = youtb_data,
##     na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5778 -1.2714 -0.0172  1.2604  6.6771
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.00991    0.09199   54.46  < 2e-16 ***
## rate         0.46708    0.01059   44.10  < 2e-16 ***
## log(length)  0.10539    0.01826    5.77 8.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.799 on 9606 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1894, Adjusted R-squared:  0.1892
## F-statistic:  1122 on 2 and 9606 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model1)
```




```
## Looking at the Residuals vs Leverage plots we can see that even though
##there are many outliers they are not influential.
```

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.

```
## ASSUMPTION 1:Linear population model
## We can see that the plot is linear in the coefficents due to
##its desgin.

##ASSUMPTION 2: Random Sampling
## Since the data is as is from the source(Youtube), we are safe
##to assume random sampling

##Assupmtion 3:No perfect multicollinearity
##We will be using Vif as well as vcovHC to see if there are
#any variables with mutlicollinearity
vif(model1)
```

```
##        rate log(length)
##     1.06594     1.06594
```

We don't see any variables with vif value more that 4

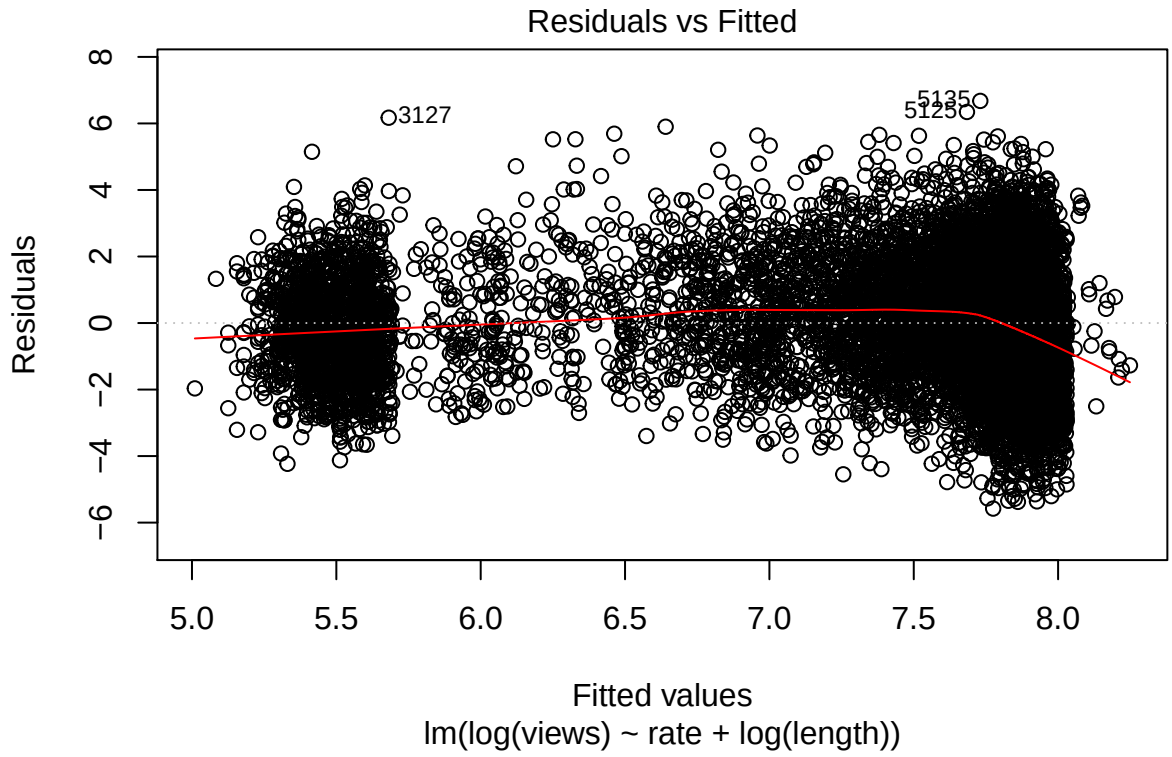
```
## We us also can check the variance-covariance matrix
vcovHC(model1)
```

```
##                (Intercept)         rate   log(length)
## (Intercept)  0.0077214129 -8.68789e-05 -0.0014292408
## rate        -0.0000868789  9.20233e-05 -0.0000457230
## log(length) -0.0014292408 -4.57230e-05  0.0003192099
```

We don't see any co-variance more that .75

AssumptionZero-conditional mean

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

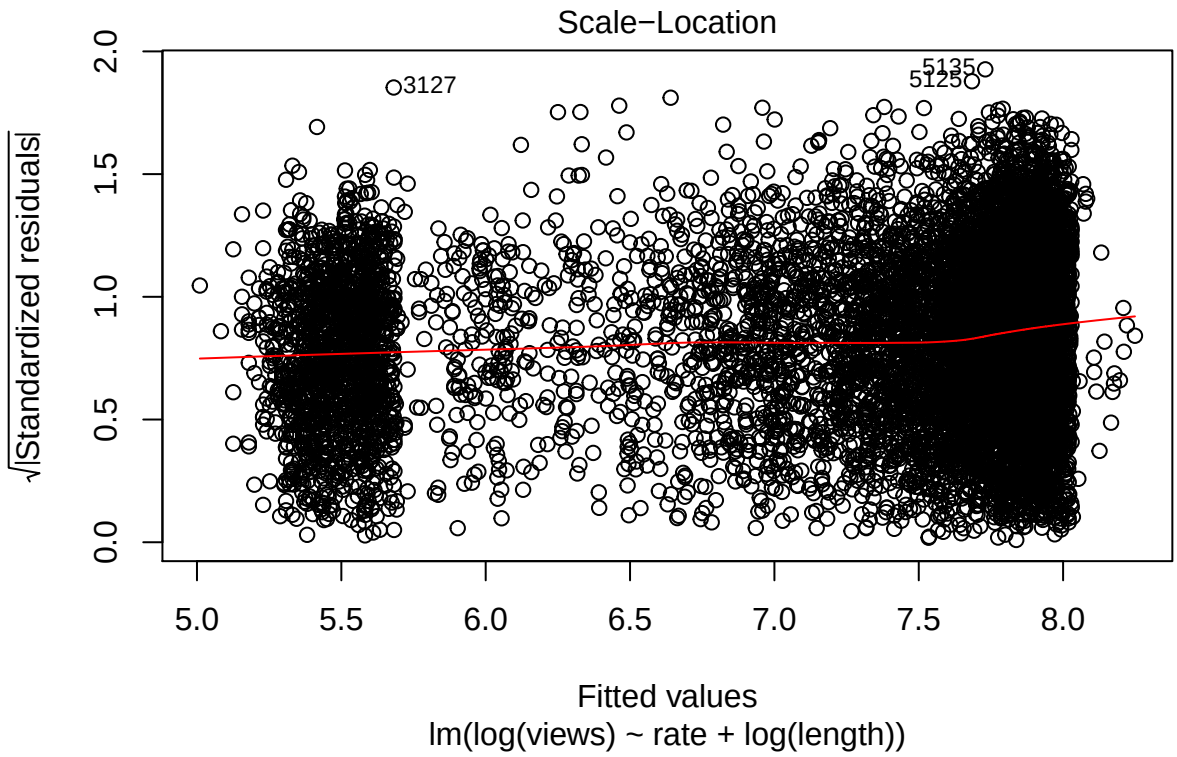



```
mean(model1$residuals)
```

```
## [1] -9.069296e-16
```

We can see that the line is almost straight and around zero. Also the mean of the residuals is very close to zero. So we can satisfy the zero conditional mean assumption.

```
## using the the scale-location plot and Breusch-Pagan test we
## will be testing for Homoskedasticity
plot(model1,which = 3)
```

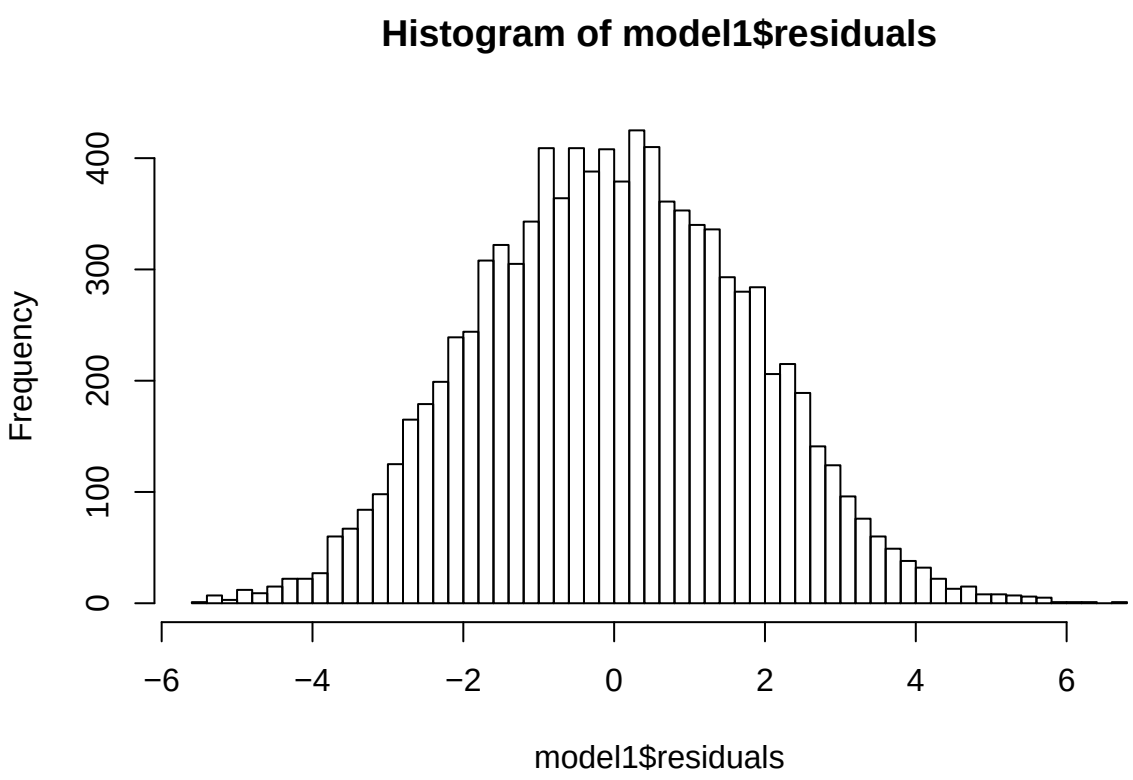



```
bptest(model1)
```

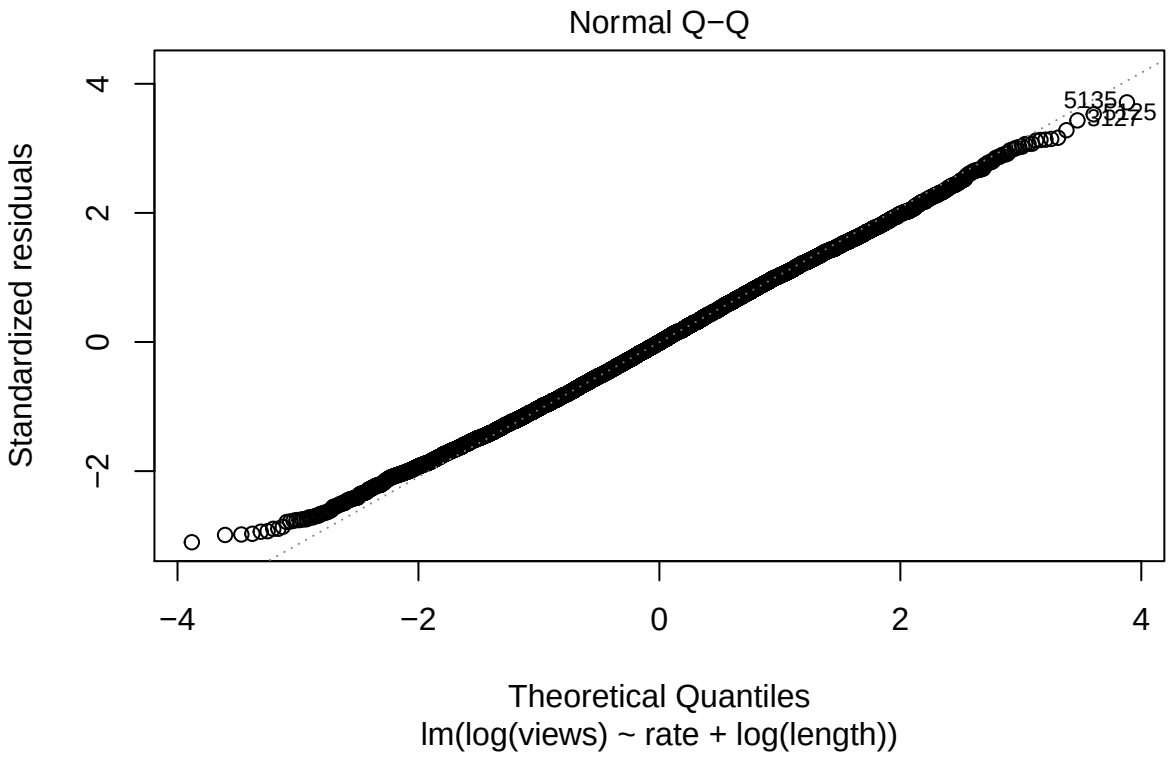
```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 123.66, df = 2, p-value < 2.2e-16
```

From the scale-location plot is looks like the variance is less in the lower values and increase for higher values. Also, the p-value from Breusch-Pagan test is below 0.05 so we reject the null hypothesis, so the data is heteroskedasitic. Going forward we will be using heteroskedasitic robust coefficients

```
## Lets do check  for normaility of the residuals
## We will be looking at qq-plot and a histogram of the errors
hist(model1$residuals,breaks=50)
```




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



the plots looks quite normal, but the qq-plot looks little off at the edges. Let us try the Shapiro wik test Since Shapiro test can only take 5000 record we will be running the test 1000 times with different samples of the data(sets of 5000) and looking at the mean p-value.

```
## Let us test it using the Shapiro-Wilk

sampler <-function(full_youtb_data){
  return (sample(full_youtb_data,5000))
}
s=rep(shapiro.test(sampler(model1$residuals))$p.value,1000)

i=1
for (i in 1:1000){
  s[i]=shapiro.test(sampler(model1$residuals))$p.value
}

mean(s)
```

```
## [1] 0.003614861
```

Here we see that the p-value is 0.0034 which is less than the threshold of 0.05 we can reject the null hypothesis and say it is not normal.

But the sample size is huge and so we can apply CLT which implies that OLS coefficients have a normal sampling distribution.

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

```
(coeftest(model1, vcov = vcovHC))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 5.0099076  0.0878716 57.0140 < 2.2e-16 ***
## rate        0.4670827  0.0095929 48.6906 < 2.2e-16 ***
## log(length) 0.1053878  0.0178664  5.8986 3.789e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(se.model1 = sqrt(diag(vcovHC(model1))))
```

```
## (Intercept)        rate log(length)
## 0.087871570 0.009592877 0.017866447
```

```
stargazer(model1, type = "text", report = "vc",
          header = TRUE,
          se = se.model1,
          title = "Linear Models Predicting Youtube views",
          keep.stat = c( "rsq", "n","aic","ser"),
          omit.table.layout = "n")
```

```
##
## Linear Models Predicting Youtube views
## ========================================
##                       Dependent variable:
##                       -------------------
##                           log(views)
## ----------------------------------------
## rate                         0.467
##
## log(length)                  0.105
##
## Constant                     5.010
##
## ----------------------------------------
## Observations                 9,609
## R2                           0.189
## Residual Std. Error  1.799 (df = 9606)
## ========================================
```

From the coeftest command we can see that all the coefficients are statistically significant.

The coefficients for rate is 0.467. That means a point increase in rate increases exp(1)*.467 = 1.269 views. At the same time the difference between the Q1 and Q3 is only 1.6 and also rate can only vary from 1-5.Thus even though rate is statistically significant, rate is not practically significant.

But for length every 100 seconds increase in length increases views by 100 * .105 = 10.5. A change in 100 seconds of video length is very likely given this context. So length is statistically and practically significant.