HW Week 12

w203: Statistics for Data Science

Adam Yang

OLS Interface

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

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

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

```
model <- lm(views~length+rate, data = videos)
```

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.

CLM.1 Linear population model

Any population distribution could be represented as a linear model plus some error. We don't have to worry about this assumption at the moment because we haven't constrained the error term.

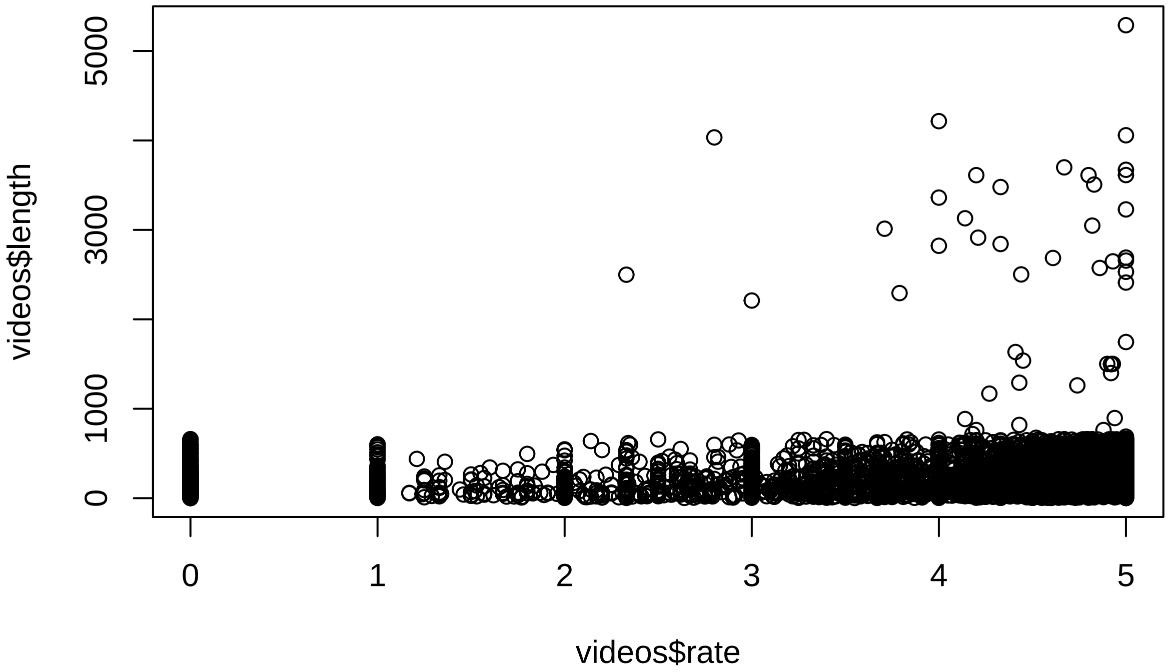
CLM.2 Random Sampling

To check random sampling, we need background knowledge of how the data was collected. Unfortunately, we do not know how these data values were scraped from Youtube so we cannot claim that the sample is totally random.

CLM.3 No perfect multicollinearity

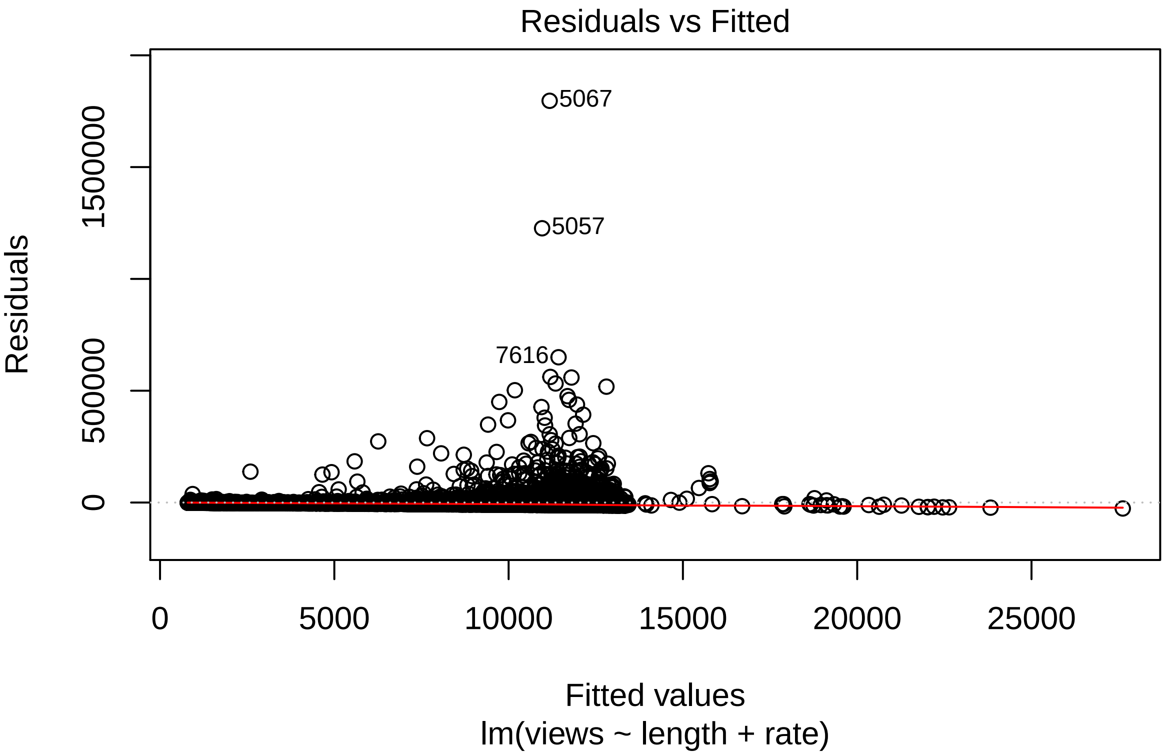
Our two variables do not have perfect multicollinearity according to the graph shown below. R would've also alerted us if there was multicollinearity in our model, which it didnt.

```
plot(videos$rate, videos$length)
```



CLM.4 Zero-conditional mean

```r
plot(model, 1)
```

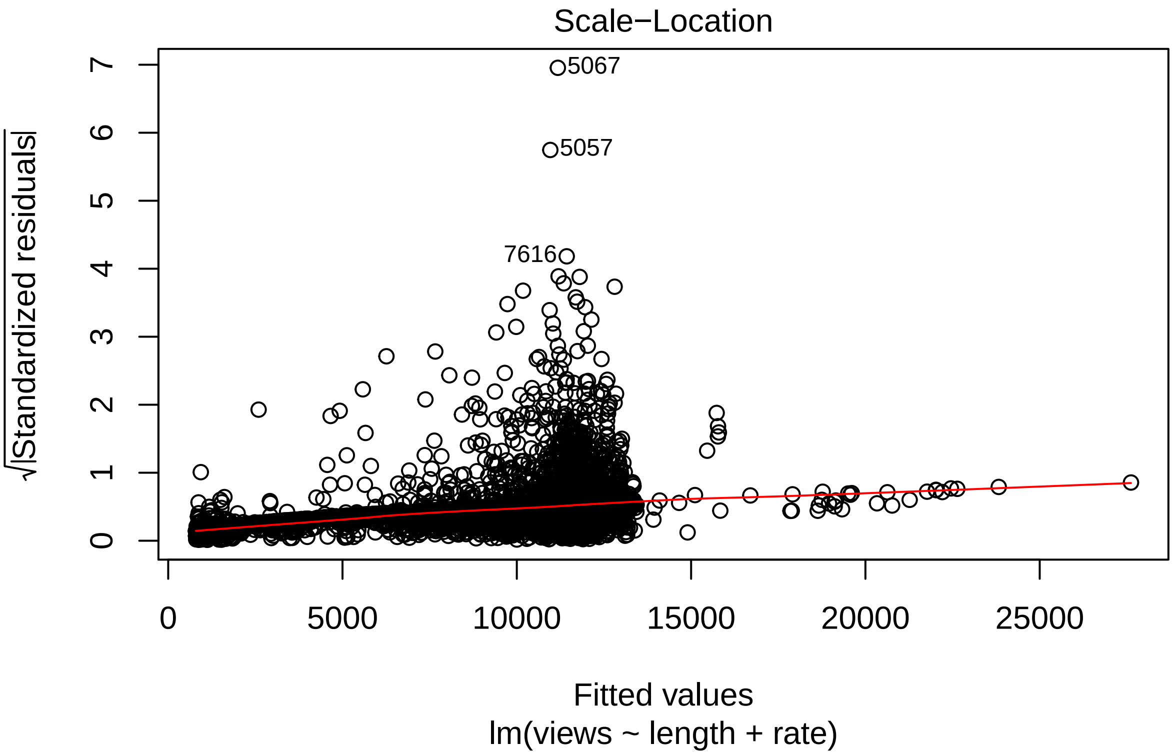


Judging by the Residuals vs Fitted graph above, there does not seem to be a clear deviation from the zero conditional mean for any fitted value as the red fitted line sticks pretty close to 0. Notice the clear deviation from zero conditional mean, indicated by the parabolic shape. This means that our coefficients will be biased. There are three different approaches to resolving this issue.

CLM.5 Homoskedasticity

From the Residuals vs Fitted graph in CLM.4, we can see that there is a some heteroskedasticity because the data points form a cone shape, starting narrow and becoming wider.

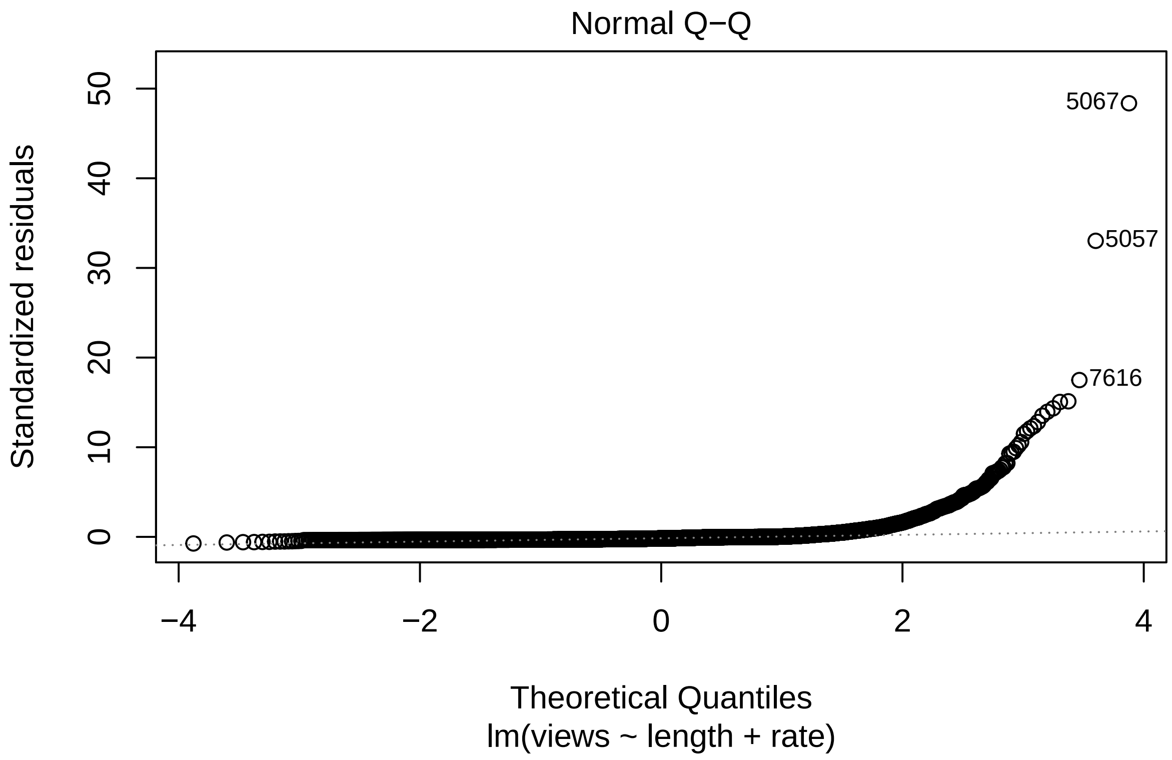
```
plot(model, 3)
```



In the Scale-Location graph above, the red fitted line has a positive slope which also shows that there is some heteroskedasticity to our model.

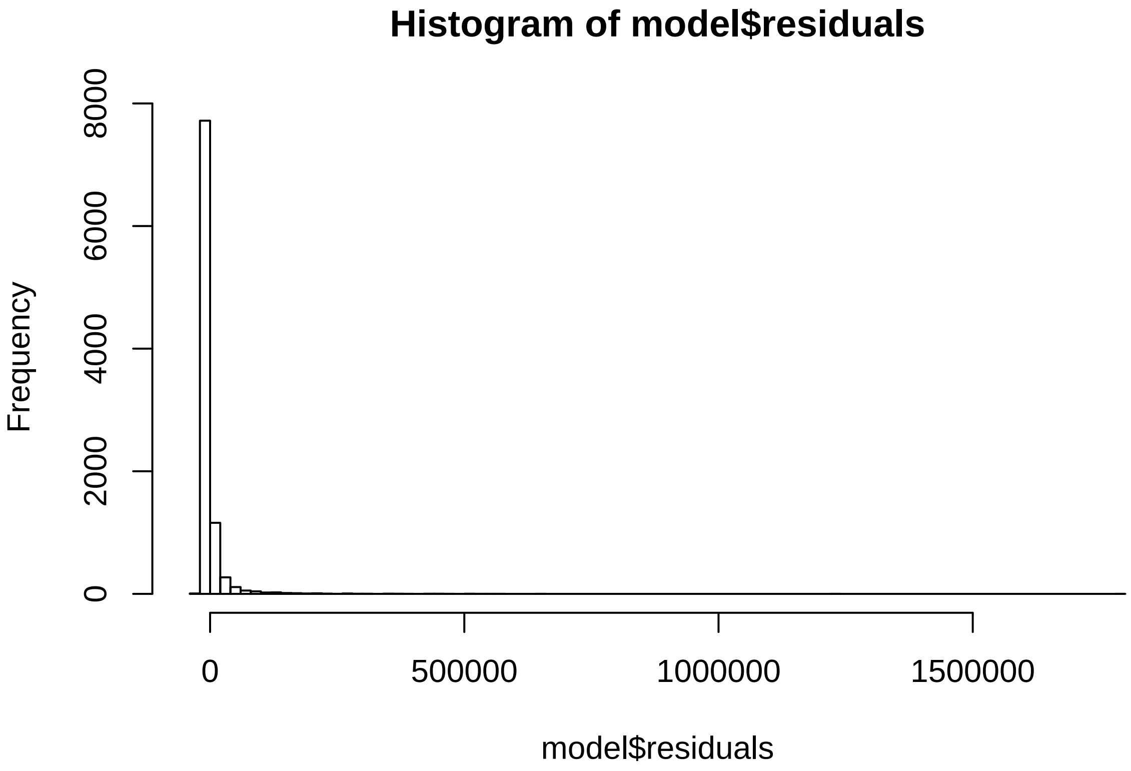
CLM.6 Normality of Errors

```
plot(model, 2)
```



From the Q-Q plot above, we see that maybe at low values, the residuals are normal, but there seems to be a strong deviation from normality at the right side of the Q-Q plot. That suggests a pretty strong positive skew in our residuals.

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



To confirm this, we can plot the histogram of our residuals, and we do see a very strong positive skew in our residuals. The CLM says if our sample size is large enough, we can assume our estimators have a normal sampling distribution. The rule of thumb is that the CLM can be applied when the sample size is greater than 30. This isn't always true, however, with strong positive skews. In our case, we have a sample size of 9489 which is much larger than 30 so I guess it might be okay to assume the CLM holds.

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.

```
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 (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

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

```
vif(model)
```

```
##   length     rate
## 1.025714 1.025714
```

From the variance inflation factors calculated above, it seems like the lenghth and rate variables do not have very strong colinearity with each other. Therefore, we do not have to worry too much about the nonperfect multicolinearity.

```
# using robust standard errors because we have heteroskedasticity
coeftest(model, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  789.6825   281.1757  2.8085  0.004987 **
## length         3.0822     1.2515  2.4628  0.013804 *
## rate        2105.4545   128.1371 16.4313 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(videos$length)
```

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

We can see that both length and rate are statistically significant variables. The more significant value is the rating of the video. By increasing the rating of the video by 1, it will result in 2105 more views. By increasing the length of the video by 1 second (the average is 226.7 which seems to be too high to be minutes for videos on Youtube) gains 3 more views.

```r
stargazer(model, type = "text", omit.stat = "f",
          se = sqrt(diag(vcovHC(model))),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## =================================================
##                          Dependent variable:
##                     -----------------------------
##                                 views
## -------------------------------------------------
## length                          3.082
##
##
## rate                          2,105.454
##
##
## Constant                      789.683**
##                               (281.176)
##
## -------------------------------------------------
## Observations                    9,480
## R2                              0.011
## Adjusted R2                     0.011
## Residual Std. Error    37,145.040 (df = 9477)
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

For a better view of the coefficients, a stargazer table is presented above.