HW week 12

w203: Statistics for Data Science

Ted Pham

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

OLS Inference

```
names(data)
```

```
## [1] "video_id" "uploader" "age"      "category" "length"   "views"
## [7] "rate"     "ratings"  "comments"
```

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).

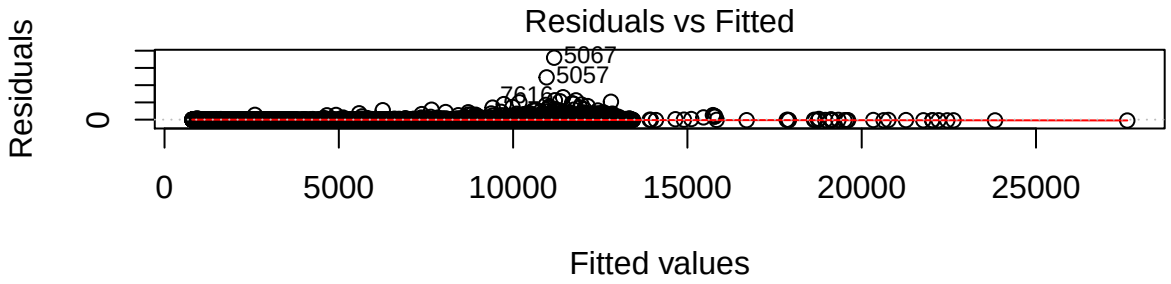
```
# model 1 without any transformation
model1 = lm(views~length + rate, data)

# model 2 with log
model2 = lm(log(views)~length +rate,data)
```

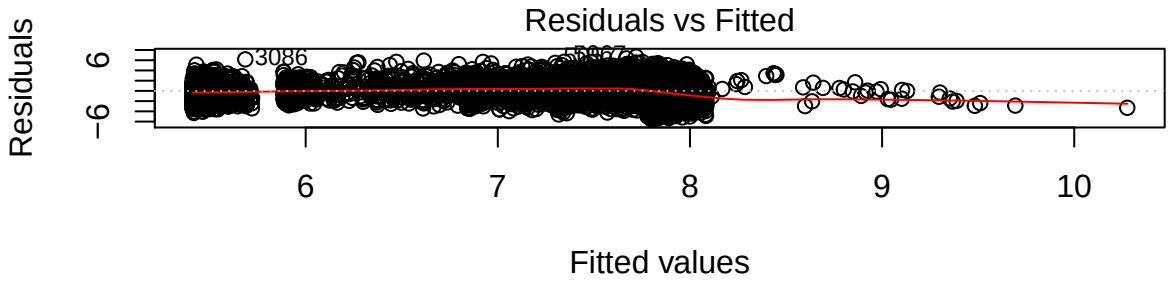
2. Using diagnostic plots, background knowledge, and statistical tests, assess all 6 assumptions of the CLM. When an assumption is violated, state what response you will take.

CLM1: Linearity in Parameters

Both models are ok

```
par(mfrow=c(2,1))
plot(model1, which = 1)
plot(model2, which =1)
```

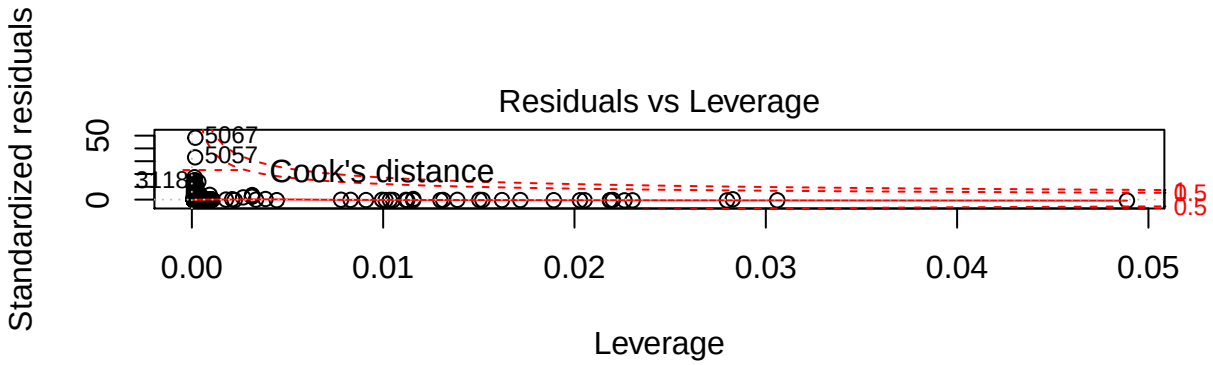


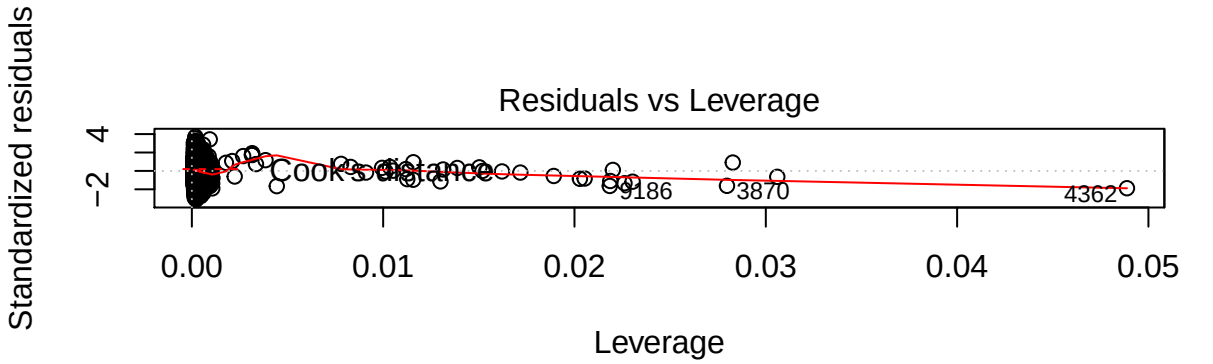


CLM2: Random Sampling

Model 1 has 2 variables near Cook's distance Model 2 seems better

```
par(mfrow=c(2,1))
plot(model1,which=5)
plot(model2, which =5)
```





CLM3: No Perfect Multicollinearity

vif's are small so no perfect multicollinearity

```
print(car::vif(model1))
```

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

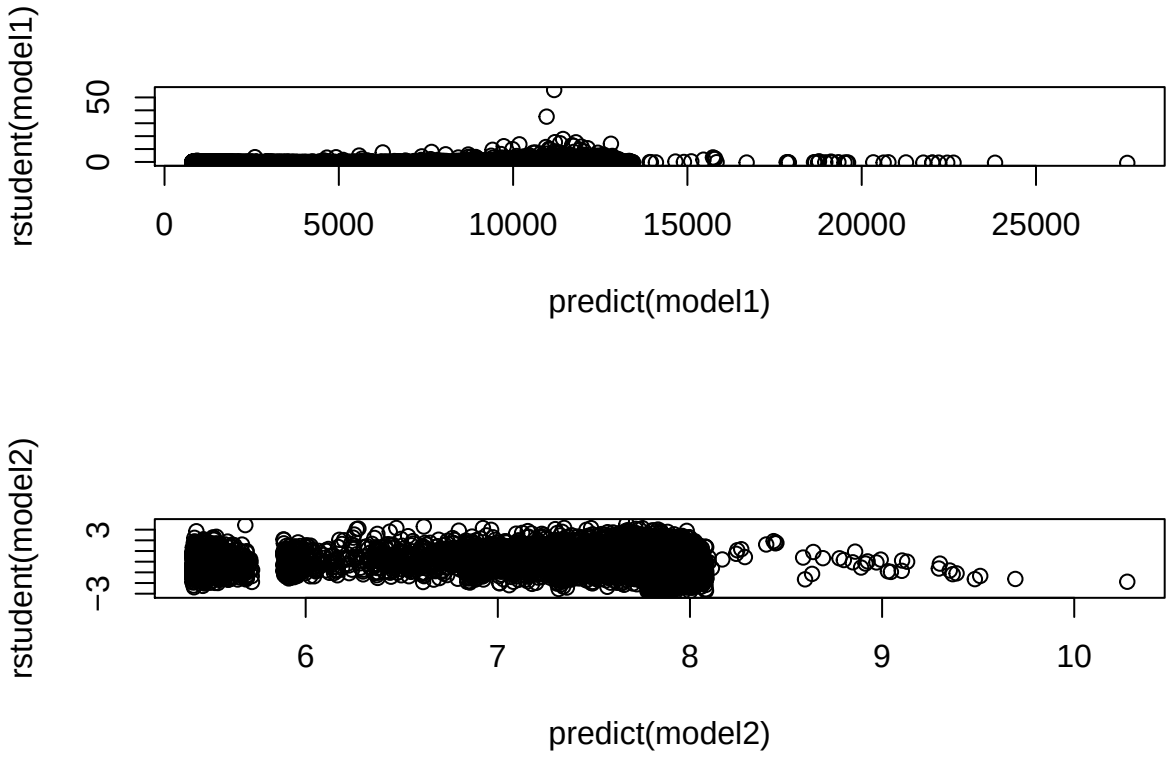
```
print(car::vif(model2))
```

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

CLM4: Zero Conditional mean

Model 2 satisfies this assumption with values across zero line.

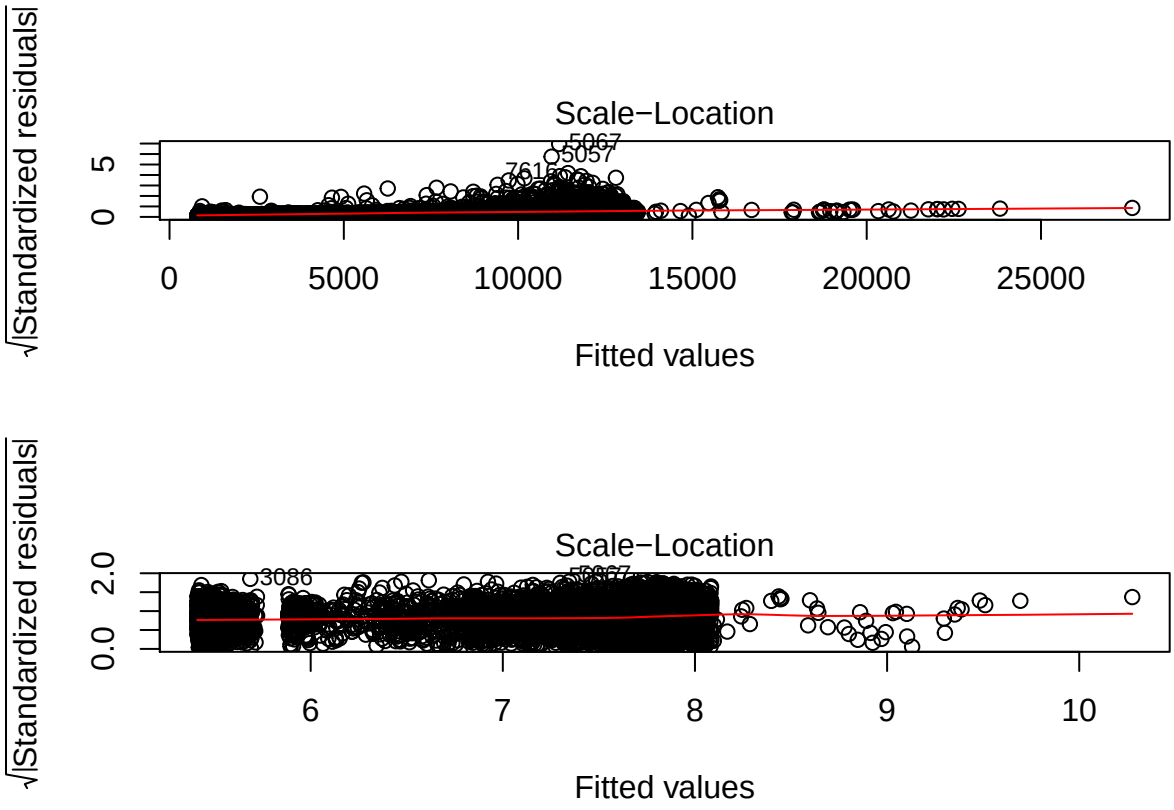
```
par(mfrow=c(2,1))
plot(predict(model1),rstudent(model1))
plot(predict(model2), rstudent(model2))
```



CLM5: Homoskedasticity

As the fitted values increase, the stdized residuals increases for model1 but not model2. Therefore, model1 shows signs of heteroskedasticity. While the Scale-location plot did not.

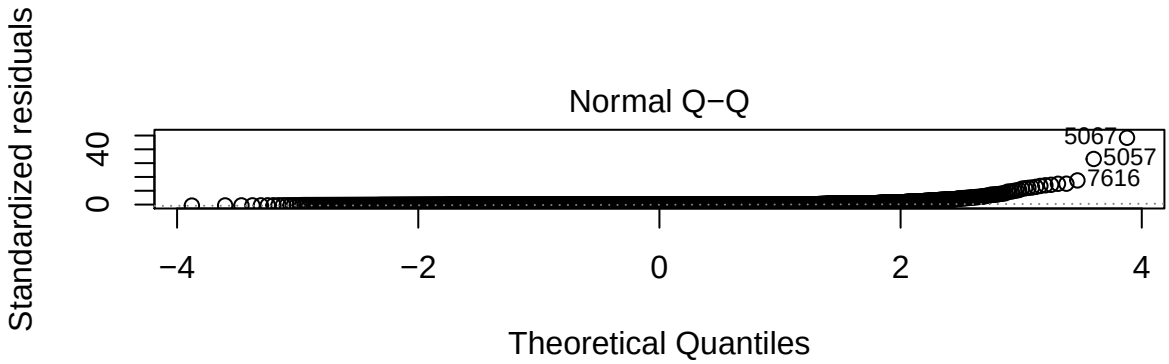
```
par(mfrow=c(2,1))
plot(model1,which=3)
plot(model2, which =3)
```

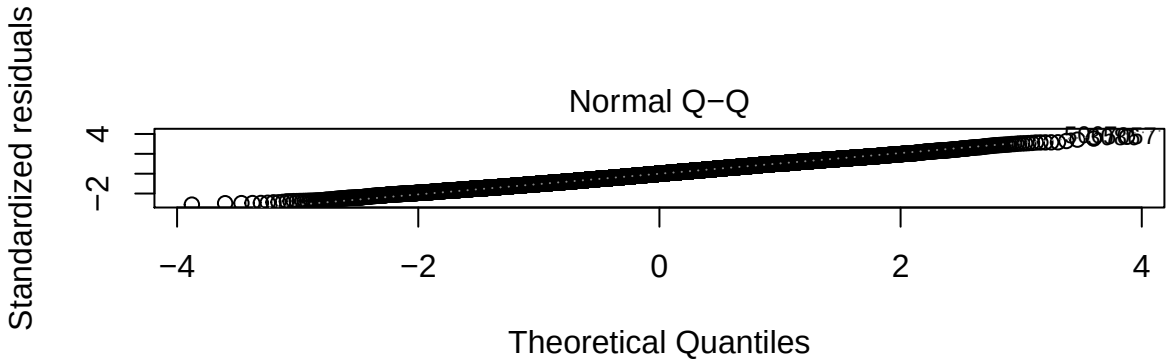


CLM6: Normality in errors

Model 2 fairs much better with this assumptions.

```
par(mfrow=c(2,1))
plot(model1,which=2)
plot(model2, which =2)
```





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.

Model 1 (without transformation) the coefficients for length and rate are 3.082 and 2105 respectively. The length coefficient isn't statistically significant while for rate it is with 2105 more views with 1 unit of rate increase which is practically significant.

Model 2 with log transform of view, both coefficients are statistically significant. With Doubling length or rating results in 1.001 and 8.83 times more viewers. So Rating coefficient is practically significant.

```
stargazer::stargazer(model1, model2, type = "text", omit.stat = "f",
                     title = "Linear Models Predicting Youtube Views",
                     add.lines = list(c("AIC", round(AIC(model1),0) , round(AIC(model2),0))),
                     star.cutoffs = c(0.05, 0.01, 0.001),
                     column.labels = c("no transformation", "log transformed"),
                     model.names = F)
```

```
##
## Linear Models Predicting Youtube Views
## ================================================================
##                                       Dependent variable:
##                                   ---------------------------
##                                         views         log(views)
##                                   no transformation log transformed
##                                          (1)              (2)
## ----------------------------------------------------------------
## length                                  3.082          0.0005***
##                                        (1.628)          (0.0001)
##
## rate                                  2,105.454***       0.473***
##                                        (216.082)         (0.010)
```

```
##
## Constant                                789.683         5.409***
##                                        (917.714)        (0.044)
##
## ------------------------------------------------------------------
## AIC                                     226416           38058
## Observations                             9,480            9,480
## R2                                       0.011            0.189
## Adjusted R2                              0.011            0.189
## Residual Std. Error (df = 9477)      37,145.040           1.801
## ==================================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```