HW week 12

w203: Statistics for Data Science

Tako Hisada

11/27/2017

OLS Inference

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

```
Data = read.csv('videos.txt', header=TRUE, sep='\t')
summary(Data)
```

```
##         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
```

```
(n = nrow(Data))
```

```
## [1] 9618
```

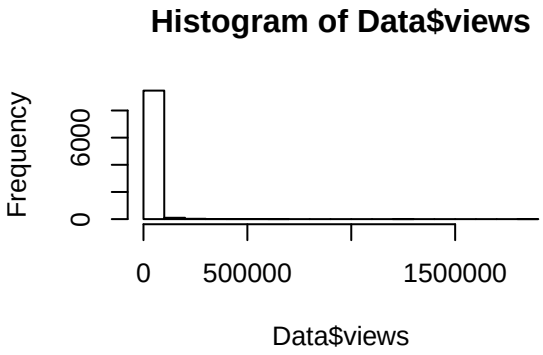
1. Fit a linear model predicting the number of views (views), from the length of a video (length) and its average user rating (rate).

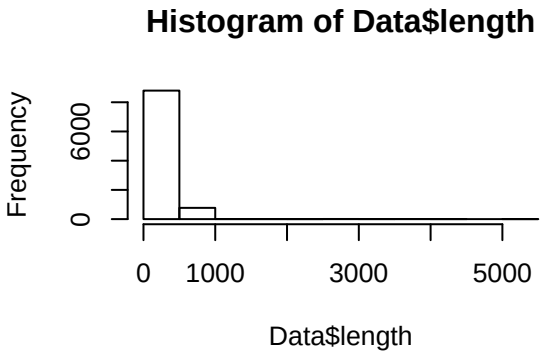
We can formulate our model as below:

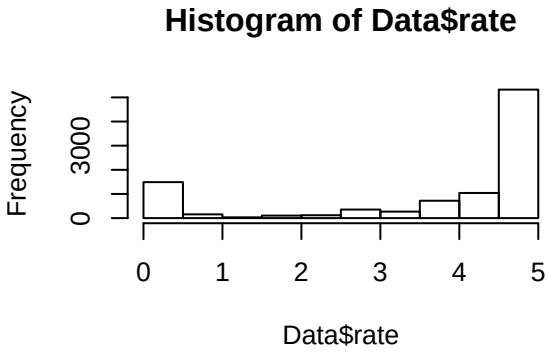
$$views = \beta_0 + \beta_1 length + \beta_2 rate + u$$

We'll first analyze histograms of the 3 variables Data$views, Data$length and Data$rate.

```
par(mfrow=c(2,2))
hist(Data$views)
hist(Data$length)
hist(Data$rate)
```

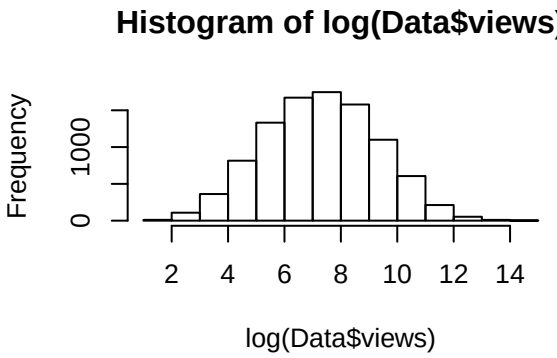


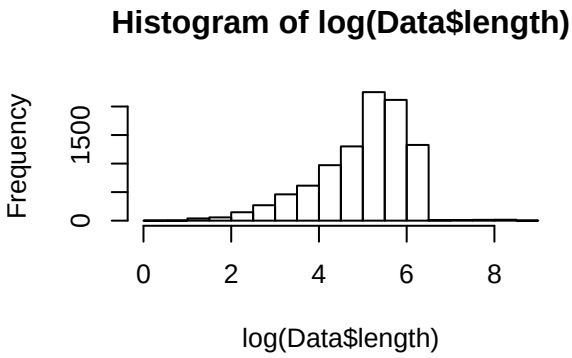


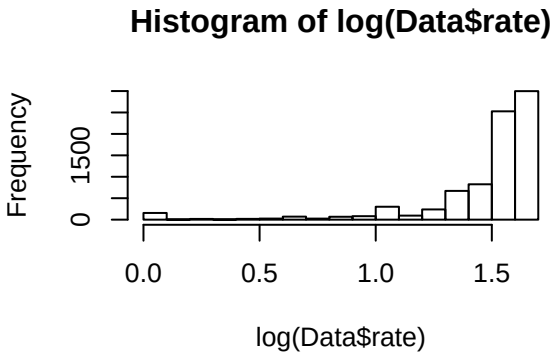


Data*views*, *Data*length are positively skewed while Data$rate has upticks on the both ends of the X-axis with a drop in the middle.

```
par(mfrow=c(2,2))
hist(log(Data$views))
hist(log(Data$length))
hist(log(Data$rate))
```







Applying log(), the distributions for Data$views and Data$length have become much more normal. However, log(Data$rate) is still negatively skewed. Also, Data$rate is an ordinal variable and contains a number of 0 values which actually have a meaning so it does not make sense to apply log().

From this, we will modify our model as below:

$$log(views) = \beta_0 + log(\beta_1 length) + \beta_2 rate + u$$

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

```
##
## Call:
## lm(formula = log(views) ~ log(length) + rate, data = 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 ***
## log(length)  0.10539    0.01826    5.77 8.17e-09 ***
## rate         0.46708    0.01059   44.10  < 2e-16 ***
## ---
## 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
```

We are seeing that our p-values are statistically significant which is a good sign.

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.

MLR.1 Linear in Parameters

Our model is defined as below which is a linear model:

$$log(views) = \beta_0 + log(\beta_1 length) + \beta_2 rate + u$$

MLR.2 Random Sampling

The data provided is scraped from youtube.com. It has not been made clear how exactly the data was collected and hence is difficult to say if it is randomly sampled or not. In case our sampling method is not random, we could employ methods such as bootstrapping to achieve random sampling as our n is sufficiently large.

MLR.3 No Perfect Collinearity

We will analyze the VIF:

```
library(car)
vif(model1)
```

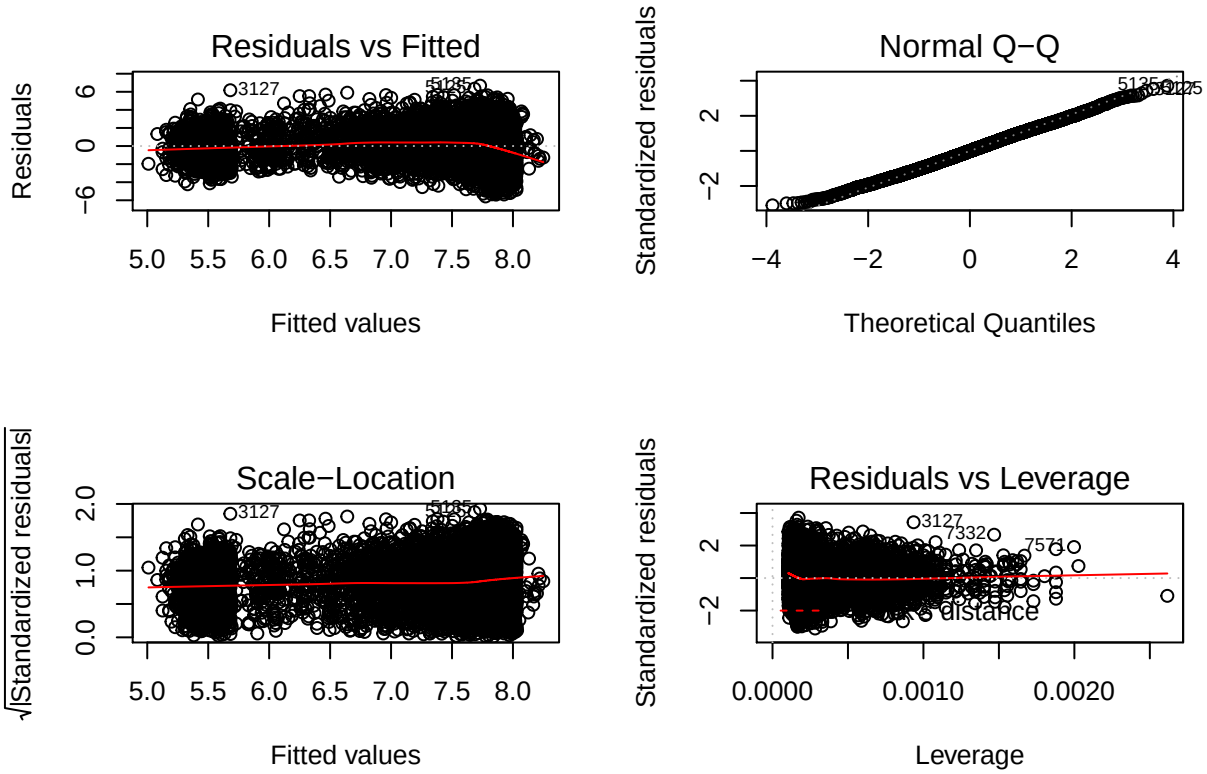
```
## log(length)        rate
##     1.06594     1.06594
```

The VIF is < 4 which is consistent with R not flagging perfect multicollinearity.

MLR.4 Zero Conditional Mean

n is 9618. Our n is sufficiently large.

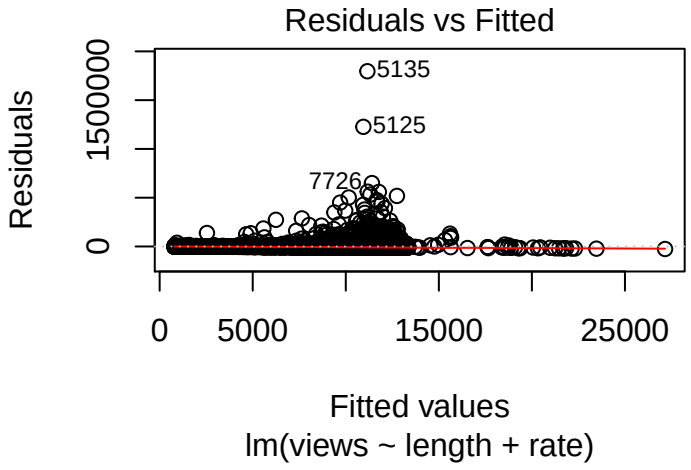
```
par(mfrow=c(2,2))
plot(model1)
```



Looking at the Residuals vs Fitted plot, the red line is pretty flat around the x-axis except towards the right-hand-side of the x-axis where it starts to point downwards which might be influenced by some of the extreme outliers. One thing we can do is to omit extreme outliers.

We could also transform the model by not applying log() as below:

```
model2 <- lm(views ~ length + rate, data = Data, na.action = na.omit)
plot(model2, which = 1)
```



The red line is completely flat on the x-axis and achieves zero conditional mean.

MLR.5 Homoskedasticity

Looking again at the Residuals vs Fitted plot, the band of the plot is relatively even although it it a little bit heavier on the right-hand-side and lighter in the middle. As we can see from the plot above, the band is more even compared to the model without log().

We may want to look into using heteroskedasticity-robust standard errors as it is not completely even.

MLR.6 Normality of Errors

Looking at the Normal Q-Q plot, we see that the points are on the line, suggesting we have normality. Also our n is considerably larger than 30 and hence we can also use CLT to assume that our OLS coefficients have normal distributions.

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.

```
(result <- summary(model1))
```

```
##
## Call:
## lm(formula = log(views) ~ log(length) + rate, data = 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 ***
## log(length)  0.10539    0.01826    5.77 8.17e-09 ***
## rate         0.46708    0.01059   44.10  < 2e-16 ***
## ---
## 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
```

The p-values suggest our variables log(length) and rate are statistically significant at the 0.1% significance level.

In terms of practical significance, $log(length) = 0.10539$. This means 1% increase in length will result in 0.1% increase in Views which seems relatively trivial. On the other hand, an incremental increase in Data$rate will result in $0.46708 = 47\%$ increase in views which is quite significant.