James_Nguyen_hw12

James Nguyen

April 8, 2017

Home Work Week 12

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

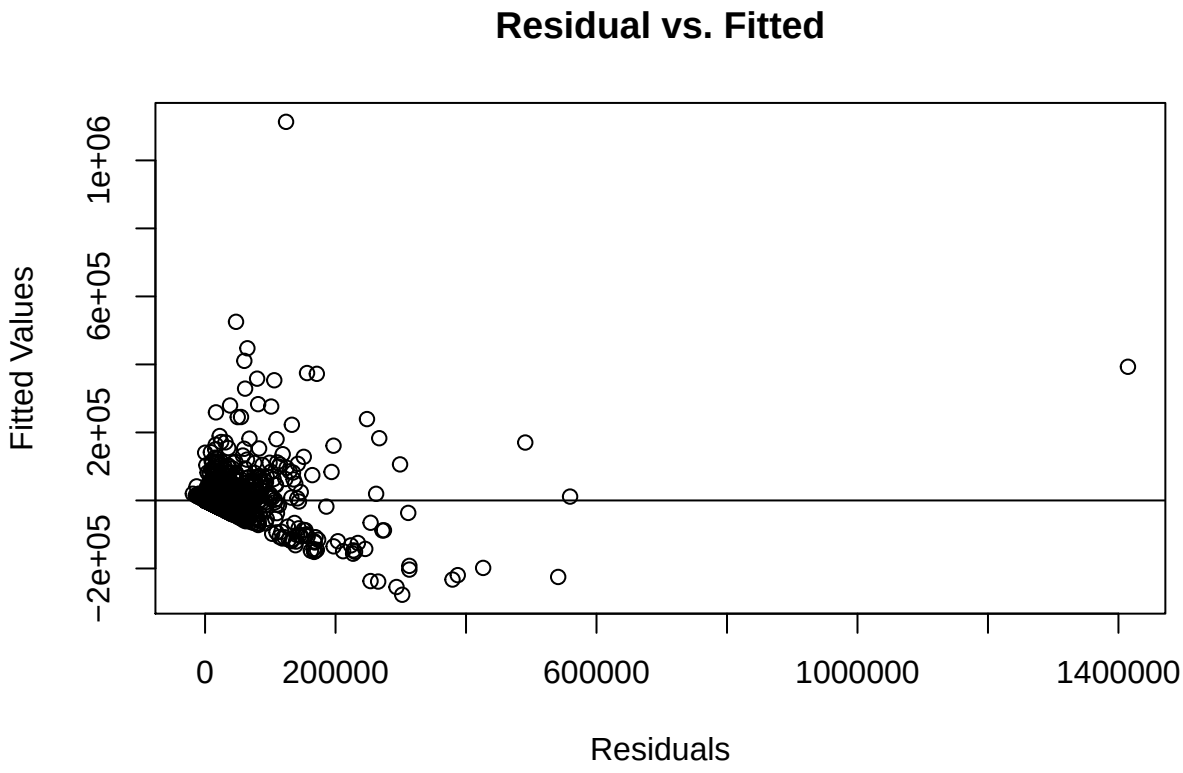
In model building, I removed na data for the relavant variables.

```
videos = data[(!is.na(data$views))&(!is.na(data$ratings))&(!is.na(data$length)),]
model<- lm((views)~ratings+length, data = videos)
```

2. Testing 6 assumptions of the CLM

a. Linear Population Model

```
plot( model$fitted.values,model$residuals, xlab = "Residuals", ylab = "Fitted Values", main ="Residual v
abline(0,0)
```



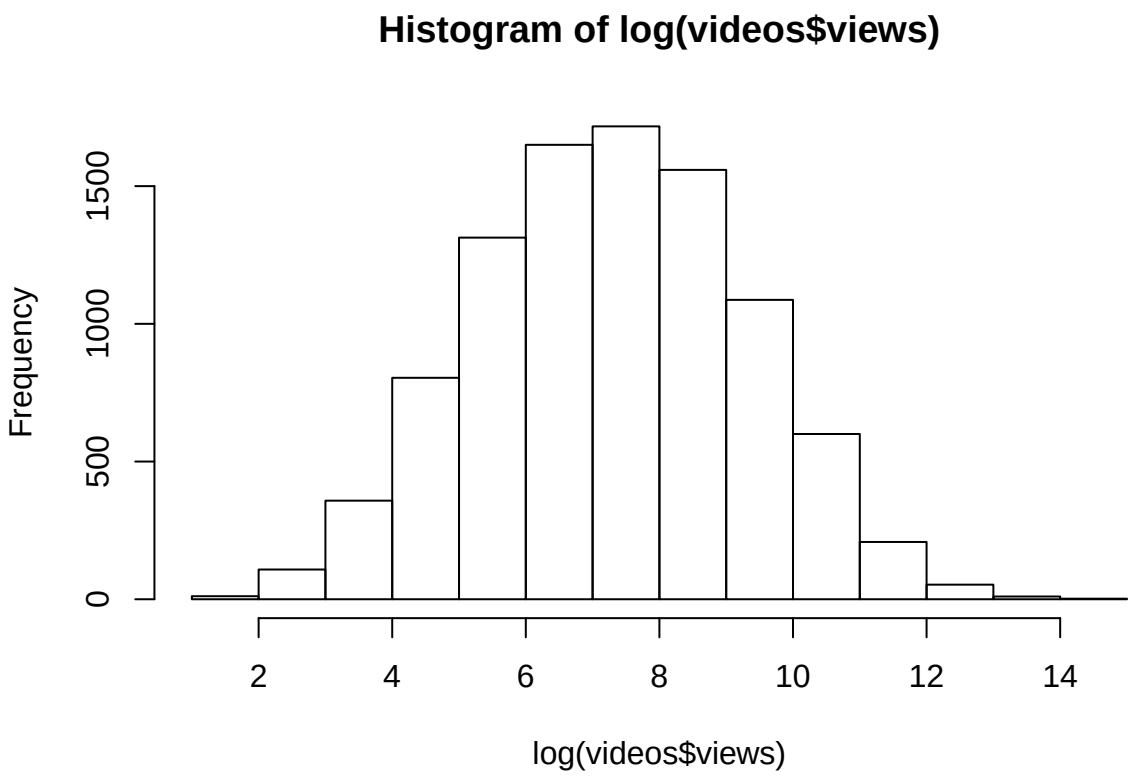
From the Residual vs. Fitted, the distribution of points are not well symmetrical. This indicate that the underlying populuation model is not linear. ###b. Random sampling

We do not have data about the collection process but by looking at the summary of key information bellow, for example the distribution of category is reasonable, there's not too much concentration on one particular uploader, the dstribution of view (log) is pretty normal, I think that the sampling is random

```
summary(videos)
```

```
##         video_id               uploader         age        
##  -0yS9zc_290:   1   Pan93bn        :  55   Min.   :   0.0  
##  -0z5PEZt_Wk:   1   nikodora       :  28   1st Qu.: 918.8  
##  -0Zkx9Sh6DU:   1   gar6301        :  22   Median :1115.0  
##  -1PT00GVE7k:   1   WWEOfficialPPVs:  22   Mean   :1044.2  
##  -1RjRtQRoEc:   1   dermayon       :  20   3rd Qu.:1226.0  
##  -2kpyJcyzEE:   1   wishinonastar07:  20   Max.   :1258.0  
##  (Other)    :9474   (Other)        :9313                   
##              category        length           views        
##  Music           :2639   Min.   :   1.0   Min.   :      3  
##  Entertainment   :2207   1st Qu.:  83.0   1st Qu.:    348  
##  Film & Animation: 801   Median : 193.0   Median :   1454  
##  People & Blogs  : 798   Mean   : 226.7   Mean   :   9374  
##  Comedy          : 613   3rd Qu.: 298.2   3rd Qu.:   6207  
##  Sports          : 561   Max.   :5289.0   Max.   :1807640  
##  (Other)         :1861                                     
##       rate          ratings           comments       
##  Min.   :0.000   Min.   :   0.00   Min.   :   -2.00  
##  1st Qu.:3.400   1st Qu.:   1.00   1st Qu.:    1.00  
##  Median :4.670   Median :   5.00   Median :    3.00  
##  Mean   :3.746   Mean   :  20.56   Mean   :   19.84  
##  3rd Qu.:5.000   3rd Qu.:  15.00   3rd Qu.:   13.00  
##  Max.   :5.000   Max.   :3801.00   Max.   :13211.00  
##                                                      
```

```
hist(log(videos$views))
```



###c. No perfect multicollinearity

```
library(car)
vif(model)
```

```
##  ratings   length
## 1.007496 1.007496
```

none of the VIF for ratings and length is more than 4. So we can say there's no perfect multicollinearity effect here.

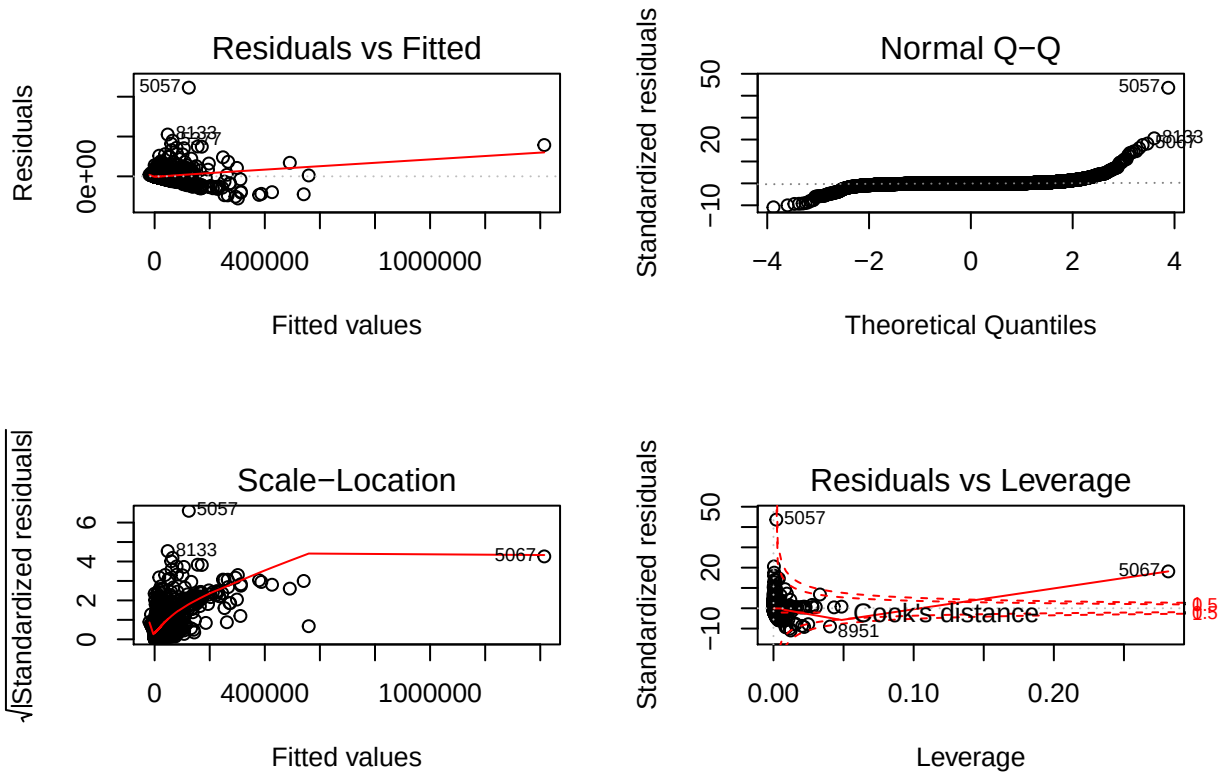
d. Zero-conditional mean

```
mean(model$residuals)
```

```
## [1] 2.529997e-12
```

The mean of residual is extremely close to zero so we can say that this assumption holds true for this model.
e. Homoskedasticity

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



The top left Residuals vs. Fitted look pretty flat, the Scale-Location except for an outlier value also look quite flat. We can accept that disturbances are homoscedastic and this assumption holds true

f. Normal distribution of errors

From Normal Q-Q chart, most of the points lie in the line except for the two ends which is expected. We can accept that the residuals are normally distributed

3

```
library(lmtest)
```

```
## 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
```

```r
coeftest(model, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2718.3313   645.3338  4.2123 2.551e-05 ***
## ratings      371.5513    44.9251  8.2705 < 2.2e-16 ***
## length        -4.3353     1.7680 -2.4521   0.01422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Statistically, both ratings and length are significant in the t test. Pratically, ratings have strong impact on the number of views with 371 increase in view with a one point increased in rating. For length, it seems the impact is insignificant with just 4 view drop when the video length is 1 minute longer. This in addition to std. error of 1.76 we can conclude that it's practiccally insignificant.