HW week 11

w203: Statistics for Data Science

Da Qi Ren

Regression analysis of YouTube dataset

You want to explain how much the quality of a video affects the number of views it receives on social media. **This is a causal question.**

You will use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations about 9618 videos shared on YouTube. Please see this link for details about how the data was collected.

You will use the following variables:

- views: the number of views by YouTube users.
- rate: the average rating given by users.
- length: the duration of the video in seconds.

You want to use the `rate` variable as a proxy for video quality. You also include `length` as a control variable. You estimate the following ols regression:

$$\text{views} = 789 + 2103\ \text{rate} + 3.00\ \text{length}$$

a. Name an omitted variable that you think could induce significant omitted variable bias. Argue whether the direction of bias is towards zero or away from zero.

```
dat<-read.csv('videos.csv', sep='')

summary(dat)
```

```
##     video_id        uploader          age               category
##  0      :  592   0      :  888   Min.   :   0               :3239
##  1      :  409   1      :  422   1st Qu.: 920   Music       :2676
##  2      :  288   2      :  273   Median :1115   Entertainment:2240
##  3      :  207   3      :  187   Mean   :1045   People      : 811
##  5      :  152   4      :  124   3rd Qu.:1226   Film        : 810
##  4      :  146   5      :  120   Max.   :1258   Comedy      : 621
##  (Other):11054   (Other):10834   NA's   :3239   (Other)     :2451
##      length           views           rate            ratings
##         :3239            :3239   Min.   :   0.00   Min.   :     0
##  &      :3230   Blogs    : 811   1st Qu.:   4.38   1st Qu.:     3
##  30     :  44   Animation: 810   Median :   5.00   Median :    16
##  230    :  33   Style    : 426   Mean   :  83.46   Mean   :  2700
##  252    :  32   Politics : 364   3rd Qu.:  70.00   3rd Qu.:   287
```

```
##  180    :  31   Animals  : 251   Max.   :4216.00   Max.   :531004
##  (Other):6239   (Other)  :6947   NA's   :3239      NA's   :3239
##     comments
##  Min.   :    0.00
##  1st Qu.:    1.00
##  Median :    4.30
##  Mean   :   15.61
##  3rd Qu.:    7.00
##  Max.   :13211.00
##  NA's   :3239
```

```
df<-dat %>%
  select('rate', 'views', 'length')
```

b. Provide a story for why there might be a reverse causal pathway (from the number of views to the average rating). Argue whether the direction of bias is towards zero or away from zero.

c. You are considering adding a new variable, `rating`, which represents the total number of ratings. Explain how this would affect your measurement goal.