HW week 11

w203: Statistics for Data Science

w203 teaching team

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.

- **Answer** An omitted variable that would induce a significant omitted variable bias would be the age variable. The age variable would be a highly positive correlated with the rate and highly positive correlated with the views. The age variable should be included in the model because it is a bias away from zero. The coefficient estimate for rate is biased.

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.

- **Answer** A possible reverse casual pathway could involve views and average rating. Our initial observation would be that number of views tend to increase the average rating. Thus, we may assume that number of views causes an increase to the average rating. However, it is possible that we could be getting this backwards and average rating actually causes people / users to view the video.

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

- **Answer** An especially important effect of adding a new variable to the model would be to better adjust to confounding, this would allow for a better understanding of casual interpretation. With the added variable we could identify a closer effect on the output views.