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.

   One omitted variable that I can think of is the age of the users. It would be useful to have the age of the users as another coefficient. As for whether the direction of bias is towards zero or away from zero, it could go either way depending on the age cohort.

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.

   There might be a reversal causal pathway from the number of views to average ratings. Because if the number of views is low then the impact of each user rating might affect the average rating more than if the number of views were high, it can also increase the chances for outliers. For example, a video with fewer views may have very high or very low ratings which may not be enough to determine the quality of the video based on the model. The direction of bias could go either way because having more views could cause low rating or high rating.

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

   I believe that `views` and `ratings` have a causal relationship. That is, the more the views the more the number of ratings. Therefore, we may have an outcome variable on the right side if `ratings` is included in our model which can result in a poor regression estimate for our model.