HW Week 11

w203: Statistics for Data Science

Adam Weintraut, Ronald Lee, Lawrence Jiang, Victor Ramirez

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.

> Number of Shares could influence the number of views and induce significant omitted variable bias. There should be a positive correlation between shares and views, since more shares will result in more views. The correlation between shares and ratings is less clear, as there could be cases when a video is shared because it is controversial, and therefore the rating may decrease as a result. In this case, there could be a negative correlation between shares and ratings. However, we would argue that this is an edge case, not the norm, and in most cases the correlation between shares and ratings would be positive. This suggests that users share videos that are good, and therefore would increase the average rating. Therefore, since both the omitted variable's correlations with the covariates and the outcome are positive, the direction of the bias is away from zero.

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.

> Because Youtube uses recommendation algorithms, users could be directed to see a video if it is viewed by other similar users (based on likes, ratings, and similar video history). In this case, the algorithm could create a reverse causal pathway, since the video could be promoted to new viewers who would be interested in this video, and therefore increase the average rating. The direction of the bias should be away from zero in this case, since there would be a positive correlation between the views and the average rating, assuming that the recommendation algorithm is effective.

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

Adding in ratings could change the effective bias in the average rating coefficient, and potentially push the coefficient for average rating closer to zero. Views and ratings should logically be positively correlated, but there could be cases in which a video gains views not because of how great it is, but because of how bad or controversial it is. For example, if a video has a middling average rating, but many people who rate it, average rating should have less influence on the number of views than the number of ratings. Ratings should help to handle these edge cases and reduce some of the biases from average/poorly rated videos that garner a large number of views.