Unit 11: Homework

John Andrus

November 13, 2020

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:

$$views = 789 + 2103rate + 3.00length$$

1. 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 could be a potential source of omitted variable bias is the number of subscribers that that video's YouTube channel has. Subscriber count could both contribute to the video's ratings (for example, subscribers might tend to give the video a higher rating) and to the number of video views (subscribers get notifications when a video is posted, making them more likely to view it).

   Given that there is a positive relationship between views and rate, and that there is likely a positive relationship between rate and subscribers, the omitted variable bias is likely away from zero.

2. Provide a story of 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 can be a tendency for people to dislike pieces of media on the grounds of their popularity rather than the grounds of their content. Consider the idea of bands "selling out" by becoming too popular to the disappointment of their original fans, or people who switch hobbies once theirs becomes mainstream. It could be that as YouTube videos garner more views, they tend to receive more negative ratings from viewers who might otherwise like them. Another possibility is that as videos become more popular, they begin to be viewed by more people outside of the video's original intended audience, leading to those individuals rating the videos lower because they don't like content that wasn't intended for them in the first place. Either of these cases would result in an omitted variable bias towards zero, as this relationship would eat into the positive rate coefficient in the model.

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

   This new variable could add an important control to the model. One way that videos are promoted on YouTube is through rating engagement, and accounting for the number of ratings a video has could reduce a potentially large source of omitted variable bias. There is likely a positive relationship between number of ratings and number of views (although we can debate which one causes the other) that creates bias away from zero in this model.