

HW week 11

Arisa Nguyen, Ayman Bari, Emanuel Mejía, Jorge Bonilla

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.

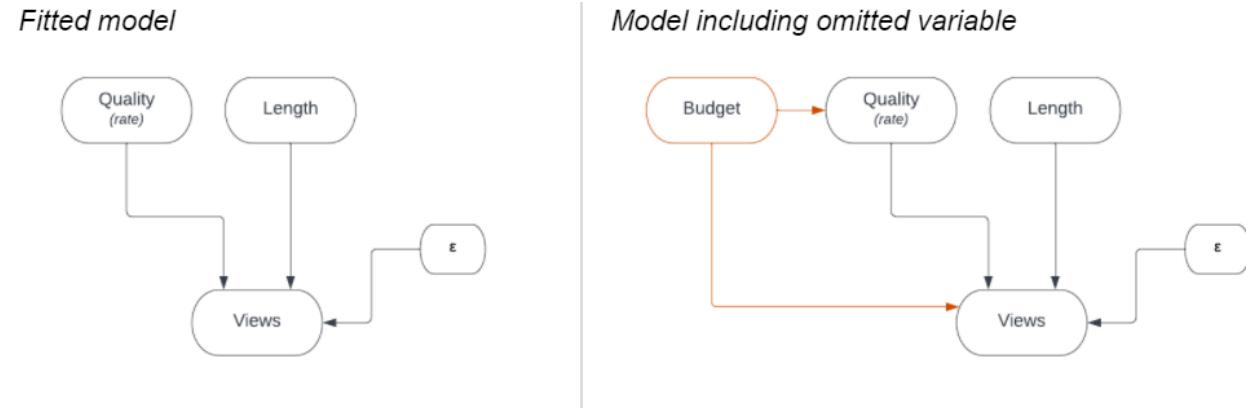
You want to use the `rate` variable as a proxy for video quality. You also include `length` as a control variable. You estimate the following ols regression:

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

- 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.

Production budget is a variable that may induce significant omitted variable bias. Production budget generally refers to the entire budget of a video or film project (including equipment, staff and marketing). We can imagine a causal link between budget and quality (the rate variable in this case), assuming that creators with higher budgets afford better quality recording equipment and editing software.

We can also assume a causal link between budget and views, through the portion of the budget attributed to marketing. Our assumption is the more invested in marketing the video project, the more views the video will have.



We expect the direction of the bias would be away from zero.

Fitted model	$V = \beta_0 + \beta_1 Q + \beta_2 L + \varepsilon$
Including term for budget	$V = \beta_0 + \beta_1 Q + \beta_2 L + \beta_3 B + \varepsilon$
Omitted Variable Bias	$bias = \beta_3 \delta_1$

Where V = Views , Q = Quality (rate), L = Length,
 B = Budget, δ_1 = the relationship between B and Q

Our variable of interest here is 1 (the coefficient of the quality predictor). The bias is given by $\beta_3 \delta_1$. While we do not have either of these values, we assume that β_3 is positive (that higher budget results in more views for reasons explained above), and that δ_1 is also positive (that the relationship between budget and quality is also positive). As both values are positive, we will have a positive bias, shifting the result further 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.

There might be a reverse causal pathway from the number of views to the average rating in the form of a positive feedback where the direction of the bias is toward zero. In this one-equation structural model, views is the response variable and rate is the treatment variable used to explain the variation in the number of views. However, it can be the case that as the number of views increase YouTube content creators earn more money which they can spend to improve the quality video. In this violation of the one-equation structural model, the reverse causality dictates that more views can lead to higher rates due to higher production values.

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

The relationship between ‘ratings’ and ‘views’ is positive, because videos with many views will have more people rating the video. The relationship between ‘ratings’ and ‘rate’ is positive, because a video that is rated highly will typically draw more people to it, who will in turn rate the video. Finally, the relationship between ‘rate’ and ‘views’ is also positive, because a video that is rated highly will draw more viewers to it. Thus, the bias of ‘ratings’ as an omitted variable is away from 0. Adding ‘ratings’ will affect the model so that it better captures ‘views’. Adding ‘ratings’ may also decrease the coefficient of ‘rate’ to account for the coefficient of ‘ratings’ also being positive.

$$\text{estimate} = \text{true parameter} + \text{omitted variable bias}$$

$$\alpha_1 = \beta_1 + \beta_2 \delta_1$$