HW week 11

w203: Statistics for Data Science

Baughman, S., Kostelc, K., Rejniak, M., Williams, S.

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:

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

First, let us load and preview the data.

```
## Rows: 9,489
## Columns: 9
## $ video_id <chr> "9QR1tni70fo", "l1DCSqAJ740", "ZES_o3XYGjM", "4I8b40cViDE", "~
## $ uploader <chr> "BHJJYP", "musicalrox", "tessaceleste", "booloveswondergirls"~
## $ age      <int> 1131, 1236, 1243, 1237, 1252, 1236, 1053, 1240, 1237, 1187, 1~
## $ category <chr> "Comedy", "Music", "Entertainment", "Entertainment", "Comedy"~
## $ length   <int> 126, 243, 105, 278, 26, 252, 162, 37, 166, 139, 361, 243, 167~
## $ views    <int> 204, 1652, 898, 928, 392, 318, 749, 10, 115, 617, 37, 266, 45~
## $ rate     <dbl> 3.00, 3.91, 4.48, 5.00, 1.50, 5.00, 3.00, 0.00, 2.00, 4.67, 5~
## $ ratings  <int> 2, 11, 81, 24, 8, 2, 6, 0, 1, 24, 1, 3, 52, 30, 114, 0, 1, 17~
## $ comments <int> 1, 4, 36, 13, 17, 3, 6, 0, 0, 17, 1, 1, 50, 17, 119, 101, 9, ~
```

Let's evaluate the model.

```
videos_df$y_int = 789
video_model <- lm(views ~ y_int + I(2103 * rate) + I(3.00 * length), data=videos_df)
video_model
```

```
##
## Call:
## lm(formula = views ~ y_int + I(2103 * rate) + I(3 * length),
##     data = videos_df)
##
## Coefficients:
##    (Intercept)           y_int  I(2103 * rate)   I(3 * length)
##        789.683              NA           1.001           1.027
```

```
v_test <- coeftest(video_model)
v_test
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    789.68253  917.71427  0.8605  0.38954
## y_int                NA         NA      NA       NA
## I(2103 * rate)   1.00117    0.10275  9.7438  < 2e-16 ***
## I(3 * length)    1.02738    0.54269  1.8931  0.05837 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

An omitted variable that could induce significant omitted variable bias is `age`. As `age` represents the count of the days a given video has been available to view and, therefore, submit a rating, the variable is independent of and precedes views and ratings. Since more time would allow for more ratings to be submitted, it figures that `age` would move the bias 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 is a reverse causal pathway between views and average rating. Since ratings can only follow views, the greater number of views can predicate higher ratings. Furthermore, if the ratings are high, they are likely to be shown in suggested videos or viewers would be more inclined to pick them. Conversely, if there are few views, then the video may only have a smaller–and less representative–number of ratings.

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

```
videos_df$y_int = 789
video_ratings_model <- lm(views ~ y_int + I(2103 * rate) + I(3.00 * length) + ratings, data=videos_df)
video_ratings_model
```

```
##
## Call:
## lm(formula = views ~ y_int + I(2103 * rate) + I(3 * length) +
##     ratings, data = videos_df)
##
## Coefficients:
##    (Intercept)           y_int  I(2103 * rate)   I(3 * length)         ratings
##      1683.4003              NA          0.1433         -1.5573        370.6992
```

```
v_test <- coeftest(video_ratings_model)
v_test
```

```
##
## t test of coefficients:
##
##                    Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)    1683.400260  631.114352   2.6673  0.007658 **
## y_int                   NA          NA       NA        NA
## I(2103 * rate)    0.143303    0.071145   2.0142  0.044014 *
## I(3 * length)    -1.557337    0.374022  -4.1638 3.158e-05 ***
## ratings         370.699232    3.606250 102.7935 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, `ratings`, interestingly, has a statistically significant p-value–as does `length`–indicating a significant effect on the model. However, it should be noted that the `length` also produces a large, statistically significant p-value. Intuitively, `ratings` representing the count of ratings submitted becomes the denominator that divides the sum of submitted ratings. Therefore, `ratings` is accounted for in `rate`.