HW week 11

w203: Statistics for Data Science

Da Qi Ren

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:

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

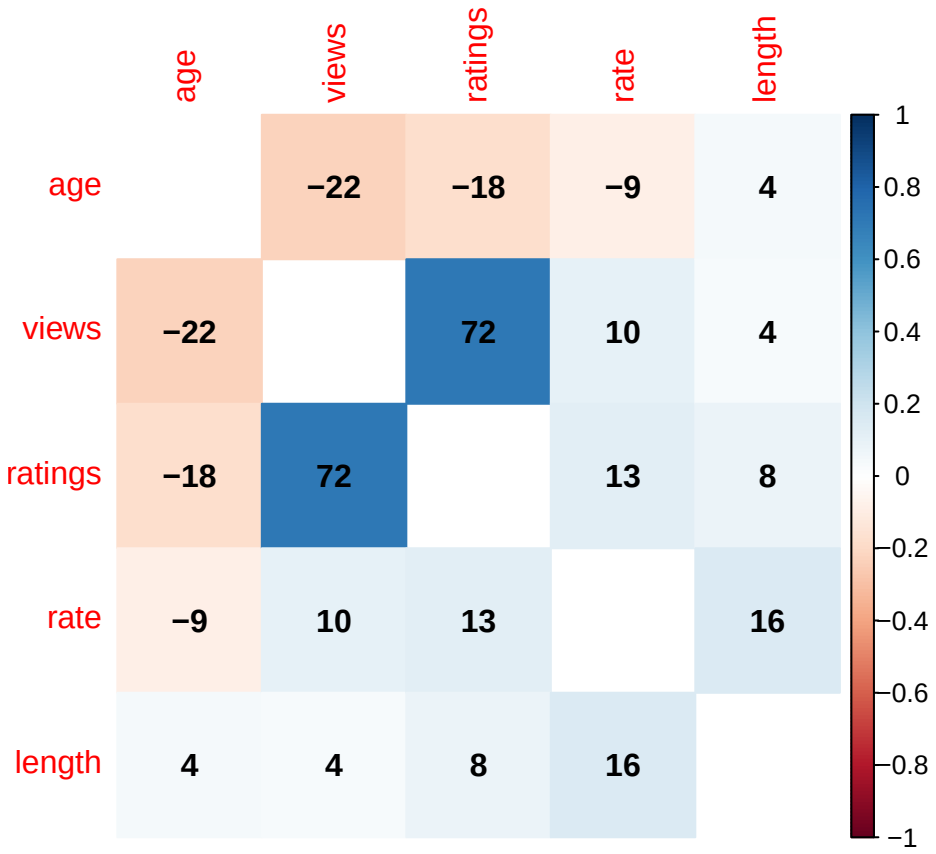
a. Name an omitted variable that you think could induce significant omitted variable bias. Argue whether the direction of bias is towards zero or away from zero.

ANSWER:

I firstly imported data from the csv file, and did cleaning and checking up by using summmary() and corrplot:

```
## [1] 9609
```

```
##       age            rate           views             length      
##  Min.   :   0   Min.   :0.000   Min.   :      3   Min.   :   1  
##  1st Qu.: 920   1st Qu.:3.400   1st Qu.:    348   1st Qu.:  83  
##  Median :1115   Median :4.670   Median :   1453   Median : 193  
##  Mean   :1045   Mean   :3.744   Mean   :   9346   Mean   : 227  
##  3rd Qu.:1226   3rd Qu.:5.000   3rd Qu.:   6179   3rd Qu.: 299  
##  Max.   :1258   Max.   :5.000   Max.   :1807640   Max.   :5289  
##     ratings       
##  Min.   :   0.00  
##  1st Qu.:   1.00  
##  Median :   5.00  
##  Mean   :  20.66  
##  3rd Qu.:  15.00  
##  Max.   :3801.00
```



I then answer this question in 2 ways:

(Method 1)

I name an omitted variable that is not in the given data set, called "recommendation", representing the status if the video is recommended by the YOUTUBE system.

Therefore,

$$\text{views} = 789 + 2103\ \text{rate} + 3.00\ \text{length} + \beta \times \text{recommendation} + u$$

and,

$$\text{recommendation} = \alpha 0 + \alpha 1 \times \text{rate} + u$$

most likely,

$$\beta > 0 \text{ and } \alpha 1 > 0, \text{ then OMVB} = \beta \times \alpha 1 > 0$$

.

And the coefficiency of rate is 2103 >0, therfore the direction of bias is away from zero.

(Method 2)

Using the data that already in videos.csv file. I found one omitted variable "age" that the direction of bias away from zero. I break down the components of the omitted variable bias below.

$$\text{views} = 789 + 2103\ \text{rate} + 3.00\ \text{length} - 36.87\text{age} + u$$

and,

$$\text{age} = \alpha 0 - 10.60\text{rate} + u$$

Therefore,

$$\text{OMVB} > 0$$

.

And the coefficiency of rate is 2103 >0, therefore direction of bias is away from zero.

```
##
## ================================================================================
##                                     Dependent variable:
##               ---------------------------------------------------------------
##                   views       rate      length       age          views
##                     (1)         (2)       (3)         (4)           (5)
## --------------------------------------------------------------------------------
## age            -37.799***   -0.001***  0.047***                  -36.837***
##                  (2.766)    (0.0001)   (0.009)                    (2.718)
##
## rate                                                -10.596***  1,672.690***
##                                                      (1.071)      (111.365)
##
## length                                                            4.949***
##                                                                    (1.217)
##
## Constant      48,829.320*** 4.491***  177.654*** 1,084.218*** 40,437.700***
##                (3,178.563)   (0.075)   (9.509)     (4.347)     (3,000.482)
##
## --------------------------------------------------------------------------------
## Observations      9,609       9,609     9,609       9,609         9,609
## R2                0.049       0.008     0.002       0.008         0.057
## Adjusted R2       0.049       0.007     0.002       0.007         0.057
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

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.

ANSWER

- Yes, there is a reverse causal pathway (from the number of views to the average rating), refer to the model analysis below.
- The story is: the more people watch the video, the more ratings the video will have. This is very resonable.
- Bias directions: from the model, I find the coefficiency of views is 0.001 >0, direction of bias is away from zero.

$$\text{ratings} = 7.11 + 0.001 \text{ views} + u$$

- Finding: Though it is of positive direction, but the coefficiency is really a small number, which means increasing the number of views will not significantly impact the number of ratings. usually 0.1 percent of people will give a rate after watching the video.

```
##
## Call:
## lm(formula = df$ratings ~ df$views)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1472.88    -7.44    -5.76    -0.73  1408.03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.1103069  0.5478719   12.98   <2e-16 ***
## df$views    0.0014495  0.0000143  101.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.08 on 9607 degrees of freedom
## Multiple R-squared:  0.5169, Adjusted R-squared:  0.5169
## F-statistic: 1.028e+04 on 1 and 9607 DF,  p-value: < 2.2e-16
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                ratings
## ------------------------------------------------
## views                           0.001***
##                                (0.00001)
##
## Constant                        7.110***
##                                 (0.548)
##
## ------------------------------------------------
```

```
## Observations                   9,609
## R2                             0.517
## Adjusted R2                    0.517
## Residual Std. Error      52.084 (df = 9607)
## F Statistic         10,280.550*** (df = 1; 9607)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01

## [1] "a1 is 1"
```

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

ANSWER

Adding the new variable `rating` would significantly improve the measurement of the model:

I compare the measurement outcomes from 2 different models using summary() and anova() functions. The output of summary(oldmodel) shows multiple R-squared = 0.011, and the output of summary(newmodel) shows multiple R-squared = 0.5178. which means newmodel fits the data better than oldmodel.

The function of anova(oldmodel, newmodel) compares models statistically. The anova() function will take the model objects as arguments, and return an ANOVA testing whether the more complex model is significantly better at capturing the data than the simpler model. If the resulting p-value is sufficiently low (usually less than 0.05), we conclude that the more complex model is significantly better than the simpler model, and thus favor the more complex model. If the p-value is not sufficiently low (usually greater than 0.05), we should favor the simpler model. Here the newmodel has p value < 2.2e-16 , which mean the newmodel is significantly better than the oldmodel.

```
##
## Call:
## lm(formula = df$views ~ df$rate + df$length)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -26084  -10242   -6542    -828 1796480
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  789.033    906.424   0.870    0.384
## df$rate     2103.880    213.447   9.857   <2e-16 ***
## df$length      2.996      1.599   1.874    0.061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36960 on 9606 degrees of freedom
## Multiple R-squared:  0.01123,    Adjusted R-squared:  0.01103
## F-statistic: 54.57 on 2 and 9606 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = df$views ~ df$rate + df$length + df$ratings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -441737   -3407   -1904    -103 1117644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1665.344    633.058   2.631 0.008536 **
## df$rate      345.894    150.084   2.305 0.021206 *
## df$length     -4.332      1.119  -3.872 0.000109 ***
## df$ratings   356.725      3.551 100.459  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25810 on 9605 degrees of freedom
## Multiple R-squared:  0.5178, Adjusted R-squared:  0.5177
## F-statistic:  3439 on 3 and 9605 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: df$views ~ df$rate + df$length
## Model 2: df$views ~ df$rate + df$length + df$ratings
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1   9606 1.3124e+13
## 2   9605 6.3997e+12  1 6.7242e+12 10092 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```