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:

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

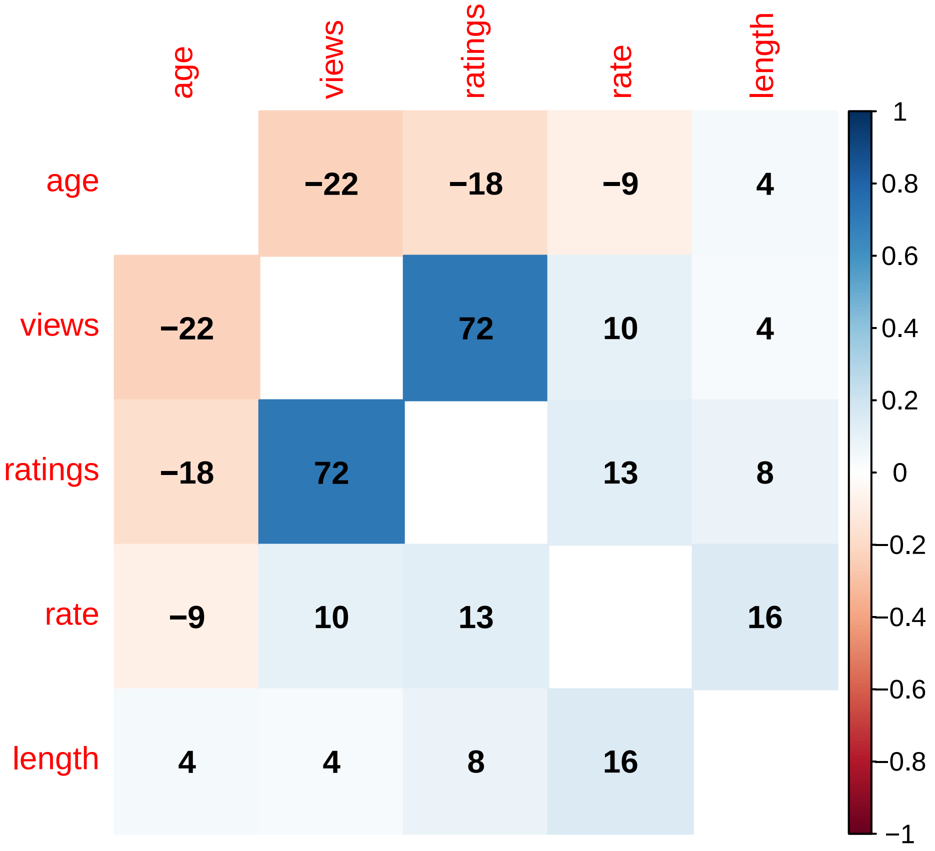
a. Name an omitted variable that you think could induce significant omitted variable bias. Argue whether the direction of bias is towards zero or away from zero.

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, the OLS coefficient on views will scaled away from zero (more positive) gaining statistical significance.

(Method 2)

Using the data that already in videos.csv file. I found 2 omitted variables: (1) "ratings" the direction of bias scaled away from zero;
(2) "age" the direction of bias towards to zero.

```
##
## =======================================================================
##                                 Dependent variable:
##              ----------------------------------------------------------
##                 views      rate     length       age          views
##                  (1)       (2)       (3)         (4)           (5)
## -----------------------------------------------------------------------
## rate                                                         196.576
##                                                             (204.174)
##
## length                                                       -3.306*
##                                                              (1.798)
##
## ratings      356.622***  0.003***  0.267**    -0.516***    348.548***
##               (48.358)   (0.001)   (0.117)     (0.168)      (53.238)
##
## age                                                        -16.198***
##                                                              (3.204)
##
## Constant     1,979.117** 3.681*** 221.466*** 1,055.205*** 19,080.330***
##               (890.094)  (0.027)   (3.305)     (3.822)     (3,385.898)
##
## -----------------------------------------------------------------------
## Observations    9,609     9,609     9,609       9,609         9,609
## R2              0.517     0.016     0.007       0.032         0.526
## Adjusted R2     0.517     0.016     0.007       0.031         0.526
## =======================================================================
## Note:                                       *p<0.1; **p<0.05; ***p<0.01
```

Estimator Positively Biased Away from Zero In this case, we have an estimator that is biased in the positive direction. Since the coefficient that it is associated with is positive as well we would say it is biased away from zero. We break down the components of the omitted variable bias below.

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

```
## 
## 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

```
## Analysis of Variance Table
##
## Response: df$views
##             Df     Sum Sq    Mean Sq F value  Pr(>F)
## df$rate      1 1.4431e+11 1.4431e+11 105.629 < 2e-16 ***
## df$length    1 4.7968e+09 4.7968e+09   3.511 0.06099 .
## Residuals 9606 1.3124e+13 1.3662e+09
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Analysis of Variance Table
##
## Response: df$views
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)
## df$rate       1 1.4431e+11 1.4431e+11   216.5921 < 2.2e-16 ***
## df$length     1 4.7968e+09 4.7968e+09     7.1993  0.007306 **
## df$ratings    1 6.7242e+12 6.7242e+12 10091.9994 < 2.2e-16 ***
## Residuals  9605 6.3997e+12 6.6629e+08
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## 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
```