HW week 11

w203: Statistics for Data Science

Kamaljeet Ghotra

Contents

Regression analysis of YouTube dataset

In this project, I use regression analysis to explain how much the quality of a video affects the number of views it receives on social media. Note that this is **This is a causal question.**

I use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations of about 9618 videos shared on YouTube. Please see this link for details about how the data was collected.

The data set contains 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.

I start by reading the data into R and exploring its structure.

```
movies_data <- read_tsv("videos.txt")

glimpse(movies_data)
```

```
## Rows: 9,618
## Columns: 9
## $ video_id <chr> "9QR1tni70fo", "l1DCSqAJ740", "ZES_o3XYGjM", "4I8b40cViDE", "~
## $ uploader <chr> "BHJJYP", "musicalrox", "tessaceleste", "booloveswondergirls"~
## $ age      <dbl> 1131, 1236, 1243, 1237, 1252, 1236, 1053, 1240, 1237, 1187, 1~
## $ category <chr> "Comedy", "Music", "Entertainment", "Entertainment", "Comedy"~
## $ length   <dbl> 126, 243, 105, 278, 26, 252, 162, 37, 166, 139, 361, 243, 167~
## $ views    <dbl> 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  <dbl> 2, 11, 81, 24, 8, 2, 6, 0, 1, 24, 1, 3, 52, 30, 114, 0, 1, 17~
## $ comments <dbl> 1, 4, 36, 13, 17, 3, 6, 0, 0, 17, 1, 1, 50, 17, 119, 101, 9, ~
```

```
head(movies_data) %>%

        kbl(., booktabs = TRUE,
```

```
            caption = "First 6 Rows of the Data") %>%

        kable_classic(full_width = FALSE, latex_options = "hold_position", font_size = 8)
```

Table 1: First 6 Rows of the Data

video_id	uploader	age	category	length	views	rate	ratings	comments
9QR1tni70fo	BHJJYP	1131	Comedy	126	204	3.00	2	1
l1DCSqAJ740	musicalrox	1236	Music	243	1652	3.91	11	4
ZES_o3XYGjM	tessaceleste	1243	Entertainment	105	898	4.48	81	36
4I8b40cViDE	booloveswondergirls	1237	Entertainment	278	928	5.00	24	13
Elp6Bf0HJIM	Fizz101Productionz	1252	Comedy	26	392	1.50	8	17
VPuKu7aU9GY	slytherin66	1236	Entertainment	252	318	5.00	2	3

I check the data for missing values and duplicates. The variables `uploader`, `age`, `category`, `length`, `views`, `rate`, and `ratings` and `comments` have 9 missing values each. These missing values are not a significant number for a dataset with 9618.

```
sapply(movies_data, is.na) %>%

        colSums() %>%

        tibble(variables = names(movies_data), missing = .) %>%

        arrange(desc(missing)) %>%

        kbl(., booktabs = TRUE, caption = "Missing Data") %>%

        kable_classic(full_width = FALSE, latex_options = "hold_position", font_size = 8)
```

Table 2: Missing Data

variables	missing
uploader	9
age	9
category	9
length	9
views	9
rate	9
ratings	9
comments	9
video_id	0

The data set has no duplicated observations.

```
movies_data %>%

        filter(duplicated(.))
```

```
## # A tibble: 0 x 9
## # ... with 9 variables: video_id <chr>, uploader <chr>, age <dbl>,
## #   category <chr>, length <dbl>, views <dbl>, rate <dbl>, ratings <dbl>,
## #   comments <dbl>
```

Next, I visualize the data in a pairs plot (Figure 1).

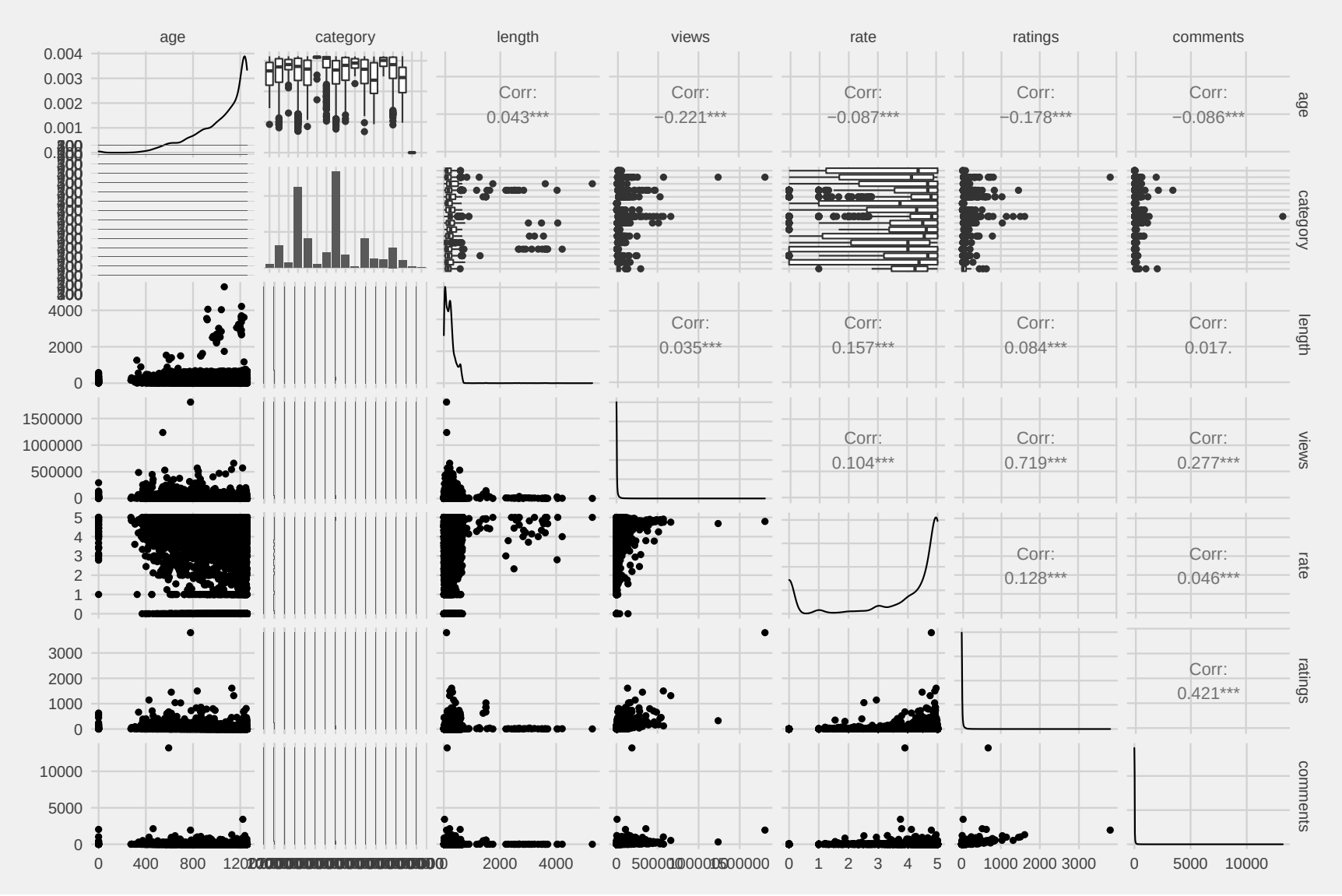

Figure 1: Pairs Plots of Variables

What is notable is the high degree of skewness among the numeric variables and the substantial variations in correlation among the variables.

Figure 2 focuses on the correlation among the numeric variables in the data set. There is a high correlation between `ratings` and views of `0.72`. Other variables with significant correlation with views include `comments` (0.227), `age` (-0.221), and `rate` (0.104).

```
movies_data %>%

        select(where(is.numeric)) %>%

        na.omit() %>%

        cor() %>%

        corrplot(type = "lower", diag = FALSE,

                 title = NULL)
```

Figure 2: Correlation Matrix: YouTube Videos Data

In the next section, I delve into the project questions.

The Model and Project Questions

I use the `rate` variable as a proxy for video quality while also include `length` as a control variable. I estimate the following OLS regression:

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

Question 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.

The variable `ratings` could significantly affect the regression model, yet it is absent from the model. First, there is a high and significant correlation between ratings and views (0.71), as Figure 1 and Figure 2 show. To see why this is the case, I also visualise video ratings against views in Figure 3 below. Hence, Ratings could have significant explanatory power over the number of views for a YouTube video.

Omitted variables are one of the causes of `endogeneity` where the error term correlates highly with one or more independent or explanatory variables, violating the OLS assumptions. `Endogeneity` leads to inconsistent estimates and hence raises the absolute value of bias. Hence, the bias is likely to be away from zero as the model may have a weak predictive capability, especially when applied to a new dataset.

```
movies_data %>%

        ## Add a small number to eliminate zero ratings
        mutate(ratings = ratings + 0.1) %>%

        ggplot(mapping = aes(x = ratings,

                             y = views,

                             col = views)) +

        geom_point(alpha = 0.5) +

        ## Log both the X and Y scales
        coord_trans(x = "log10", y = "log10") +

        scale_color_gradient(low = "blue", high = "red") +

        labs(x = "Ratings- Log Scale", y = "Views- Log Scale",

             title = "Views versus Ratings of YouTube Videos") +

        theme(legend.position = "bottom") +

        scale_y_continuous(labels = scales::comma_format()) +

        theme(axis.text.x = element_text(angle = 90))
```

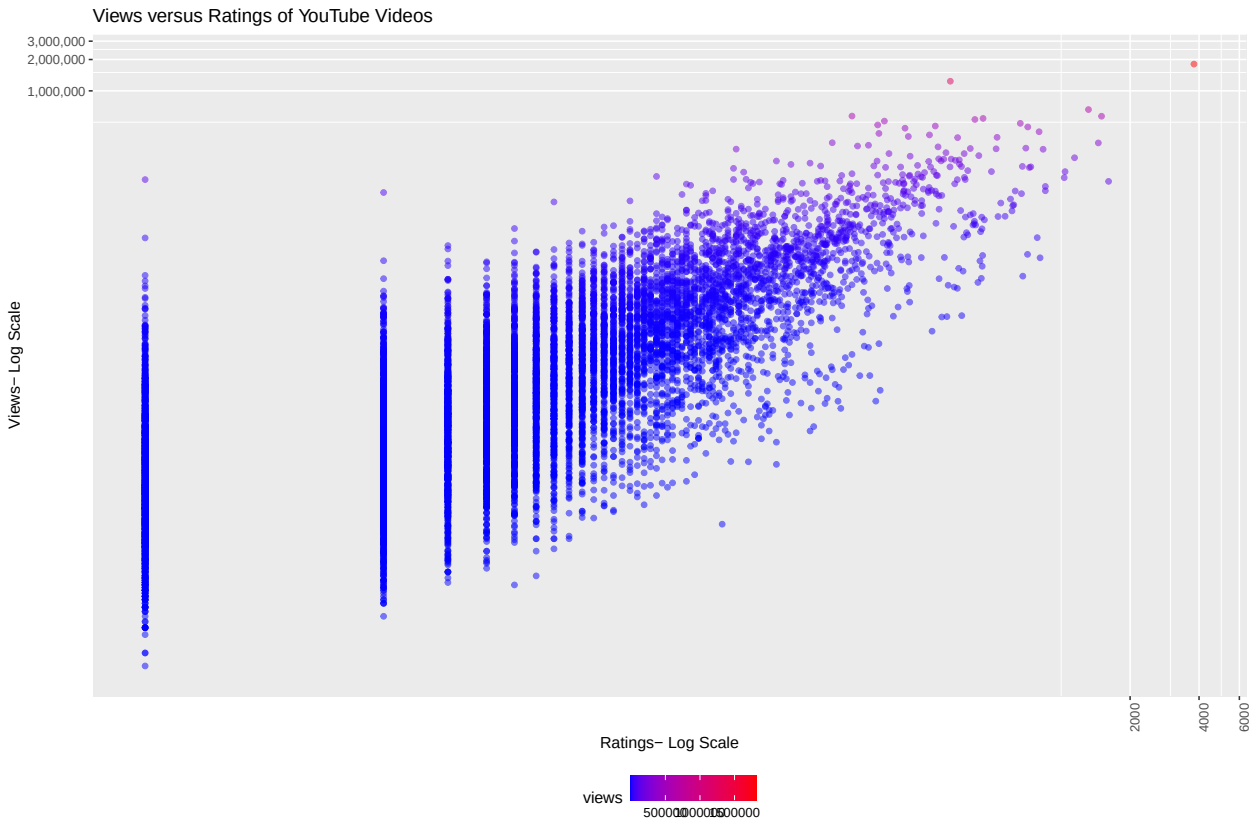

Figure 3: Video Ratings versus Views

Question 2

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.

Reverse causality refers to either a direction of causality that goes against common assumptions or a two-way loop-like causality (Guan and Tena 2021). Even a substantial correlation between variables does not imply causation - that one variable causes the other. Nonetheless, in the case of our model, the average rating (`rate`), being one of the independent variables, should explain the number of video `views`.

However, the number of `views` that a video receives could also drive the average rating. Many recommendation engines will suggest to viewers videos of a particular genre that have received the highest number of views, a proxy for video popularity. There is a high likelihood that the YouTube recommendation engine will suggest to viewers videos that these viewers will like, further driving up or sustaining high average ratings. It is essential to have a basic understanding of how recommendation engines for social media sites such as `YouTube` work to grasp this issue properly.

> A recommendation system suggests a good or service to a customer based on their previous revealed preferences. Recommendation systems come in three versions: `Collaborative filtering recommendation systems` make recommendations to users based on the activities of other similar users. `Content filtering recommendation Systems` suggest content based on the contents/characteristics of previous goods or services purchased. `Hybrid Systems` use both `collaborative filtering` and `content filtering` to make suggestions (Cui et al. 2020).

Thus, videos with many views are likely to be recommended to viewers who like such videos, which in turn sustains the high average ratings. A more viable possibility is the existence of simultaneity where high average ratings drive up views, and more views drive up average ratings.

I would argue that the bias is away from zero, noting that the actual and predicted values are close for a model with no bias. However, reverse causality causes models to be inaccurate due to endogeneity. Under endogeneity, the error terms for the model correlated with one or more independent variables, which in turn affects the model's predictive power. One primary cause of endogeneity is reverse causality/ simultaneity which leads to inconsistent estimates of model parameters (Stone and Rose 2011).

Question 3

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 model measurement goal would change, given that we are adding a new variable to the model. The new variable would contribute to predicting the number of views. In this case, the coefficients of the `ratings` variable would be interpreted as follows. Holding `rate` and `length` of videos constant, how much does a unit change in the number of ratings affect the number of views a video receives? Notably, adding the variable `ratings` could help improve the model. As shown in Figure 2, `ratings` have the highest correlation with views. However, as in the previous exercise, simultaneity and reverse causality are likely to lead to inconsistent model parameters that affect causal inference.

References

Cui, Zhihua, Xianghua Xu, XUE Fei, Xingjuan Cai, Yang Cao, Wensheng Zhang, and Jinjun Chen. 2020. "Personalized Recommendation System Based on Collaborative Filtering for IoT Scenarios." *IEEE Transactions on Services Computing* 13 (4): 685–95.

Guan, Jing, and Juan de Dios Tena. 2021. "Does Sport Affect Health and Well-Being or Is It the Other Way Around? A Note on Reverse-Causality in Empirical Applications." *Journal of Sports Economics* 22 (2): 218–26.

Stone, Susan I, and Roderick A Rose. 2011. "Social Work Research and Endogeneity Bias." *Journal of the Society for Social Work and Research* 2 (2): 54–75.