HW week 11

w203: Statistics for Data Science

w203 teaching team

Regression analysis of YouTube dataset

YouTube is the most popular free video-sharing website today. Use a linear regression framework to analyze how the number of views is affected by video quality and video lenght.

Dataset

- videos.txt (contains 9618 records describing YouTube videos. You can pull some of these by placing appending the `video_id` to URL https://www.youtube.com/watch?v= , but many videos are no longer available in YouTube)

Dependent Variables

- views (count or frequency of views by YouTube users)

Regressors

- rate (a proxy for video content quality, an average of user's ratings.)
- length (duration of the video in seconds)

(a) Perform a brief exploratory data analysis on data to discover patterns, outliers, or wrong data entries and summarize your findings.

(b) Estimate the following model to investigate the relationship between views and regressors further.

$$f(views) = \beta_0 + \beta_1 g(rate) + \beta_3 h(length)$$

f(.), g(.), and h(.) any functional form such as log().

(c) Using diagnostic plots, background knowledge, and statistical tests, assess all six assumptions of the CLM. When an assumption is violated, state what response you will take.

- When you check zero conditional mean assumption, identify one omitted variable that is not within this dataset, and estimate whether they are biasing the effect you measure towards zero or away from zero.

(d) Display all your model specification in a regression table and interpret and explain your result in terms of statistical and practical significance.