

HW12: CLM Practice

The Principal Components - Ed Brown, Daphne Lin, Linh Tran, Lisa Wu

Part 2 - CLM Practice

For the following questions, your task is to evaluate the Classical Linear Model assumptions. It is not enough to say that an assumption is met or not met; instead, present evidence based on your background knowledge, visualizations, and numerical summaries.

The file `videos.txt` contains 9618 observations of videos shared on YouTube. It was created by Cheng, Dale and Liu at Simon Fraser University. Please see this link for details about how the data was collected.

You wish to run the following regression:

$$\ln(\text{views}) = \beta_0 + \beta_1 \text{rate} + \beta_3 \text{length}$$

The variables are as follows:

- **views**: the number of views by YouTube users.
- **rate**: This is the average of the ratings that the video received. You may think of this as a proxy for video quality. (Notice that this is different from the variable **ratings** which is a count of the total number of ratings that a video has received.)
- **length**: the duration of the video in seconds.

Response:

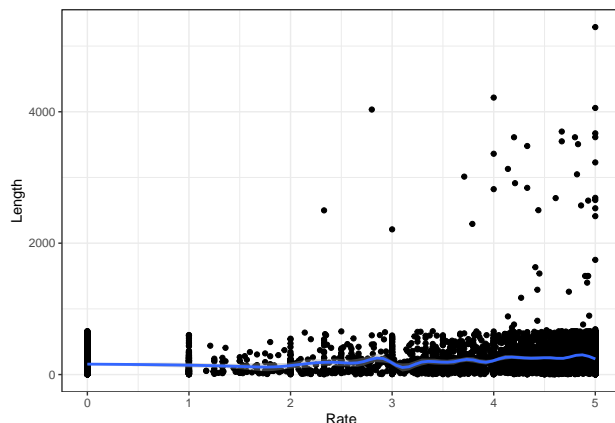
1. Evaluate the **IID** assumption

- Assessing the IID assumption requires an analysis of the sample selection design process. Based on our research of the selection process of the video data, the list of video was selected from YouTube using a crawling algorithm in which it starts with a set of videos from the list of “Recently Featured”, “Most Viewed”, “Top Rated” and “Most Discussed”, for “Today”, “This Week”, “This Month” and “All Time” and then the algorithm uses this list to find more related videos. Given this process, we believe that the videos in this dataset are not independently sampled. For example, if the sample time frame is around election time, we would expect that “Recently Featured” or “Most Discussed” videos are more likely to be related to the topic of election for “Today” or “This Week”. In addition, the crawl algorithm adds videos to the list by finding videos that are directly related to the initial set of videos. Therefore, by the nature of the sampling process, this dataset does not meet the **IID** assumption.
- To address this violation of the IID assumption, the researchers need to get new data by using a new random sampling process, or adjust measures of uncertainty to reflect the clustered nature of the data generation process.

2. Evaluate the **No perfect Colinearity** assumption.

In order to assess nearly perfect colinearity, we use our background knowledge to evaluate the input variables (rate and length), review the scatter plot of the two variables, and perform correction test and VIF test.

- Based on our background knowledge, the length of a video may affect a viewer's rating of the video, but we don't expect near perfect correlation between rate and length, as the content of the video also plays a key role in viewer's rating.
- We examined the scatter plot of rate vs length below which shows that rate and length has no obvious relationship.



- We performed the Pearson correlation test which shows that the estimate correlation between rate and length is 0.156 (CI: 0.1372389 0.1762458, $p = 2.2e-16$). There is not enough evidence of a problem of nearly perfect collinearity.
- We also performed the Tolerance and VIF test below. When Tolerance is close to 1 and VIF is less than 5, there is no evidence of the problem of nearly perfect collinearity.

Table 1: Tolerance and VIF Table

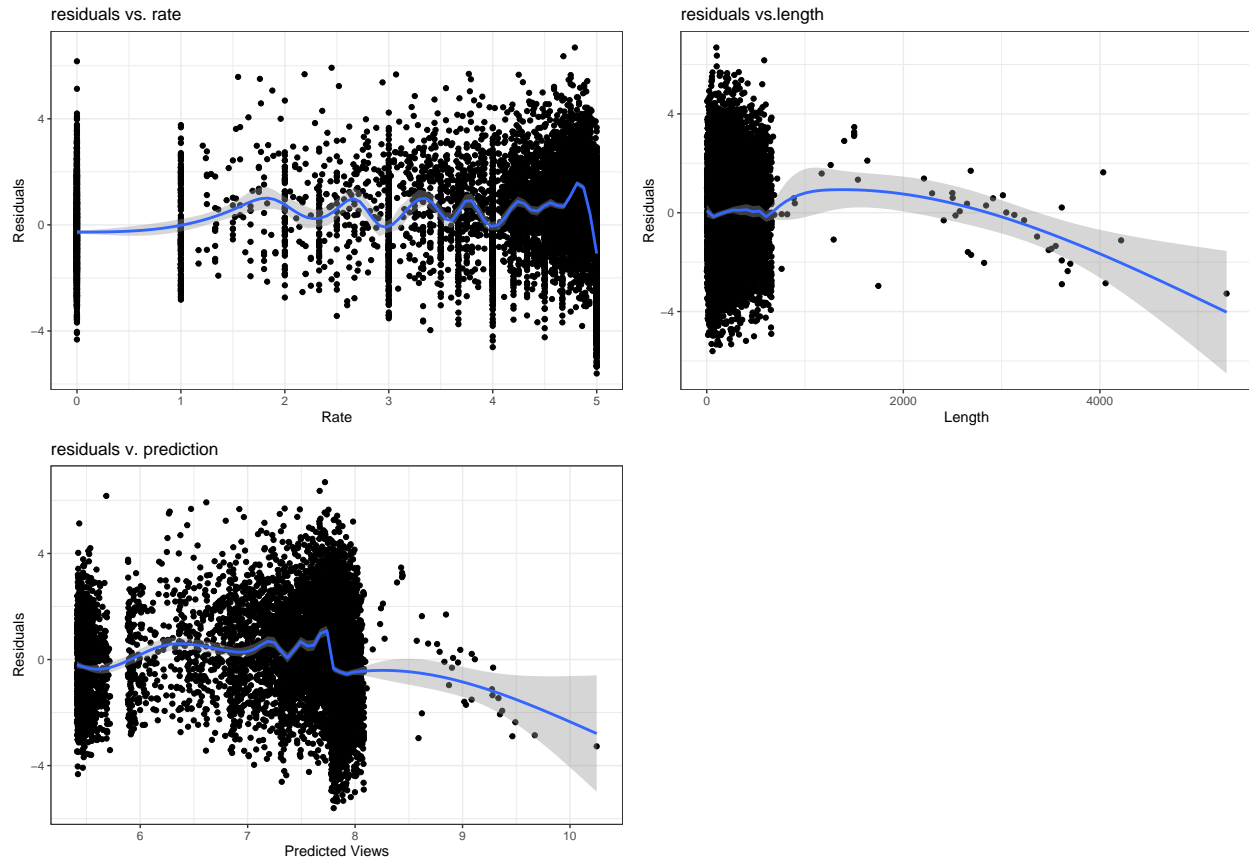
Statistic	N	Mean	St. Dev.	Min	Max
Tolerance	2	0.9754	0.0000	0.9754	0.9754
VIF	2	1.0252	0.0000	1.0252	1.0252

- Based on the above assessment, we believe this data meets the **No perfect Colinearity** assumption.

3. Evaluate the **Linear Conditional Expectation:** assumption.

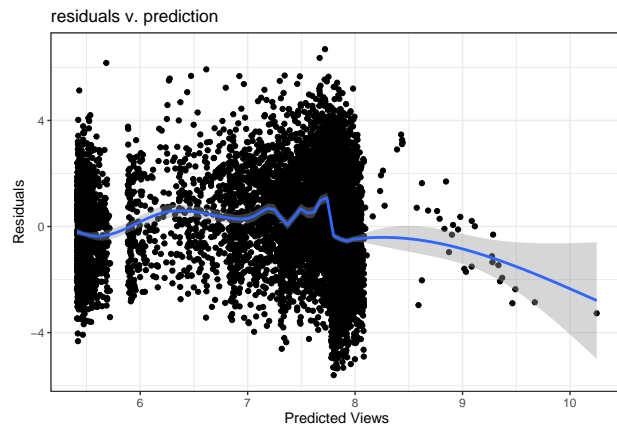
- This is an assessment whether the conditional expectation of Y given X exists and has the linear form, which also means the expected error term is zero. Based on our background knowledge, views of videos are complex observational data and may not simply reflect a linear relationship with rate and length of the video. we further plotted the residuals (the error term) against the input variables (rate and length) and the predicted views, to assess whether residuals are close to zero with respect to input variables and the outcome variable.
- Looking at the graph of residual versus rate (the left graph below), we see that the line of residual average is oscillating around zero but is not flat. The line of residual versus length (the middle graph below) has a downward curvature (deviate from zero) as length increases. This means that there is non-linear relationship between the outcome variable (log(views)) and the input variables. This non-linear relationship is likely causing the line of residual versus predicted values (the right graph) to curve downward as the predicted views increase.

- Therefore, the **Linear Conditional Expectation** is not met. This model version assumes that data is linear (the model is still the best linear predictor), but the estimated coefficient does not match the relationship of the data. In order to meet this assumption, we may need to consider other families of nonlinear models as the linear model does not fully model the complexity of the data.



4. Evaluate the **Homoskedastic Errors:** assumption.

- To evaluate this assumption, we plotted the residuals against the predicted views and evaluated whether the conditional variance is constant. From the graph below, residuals start around the range of $(-4, 4)$ and widen to the range of $(-6, 6)$ as the predicted views increase to 8, a strong evidence that the conditional variance is not constant.

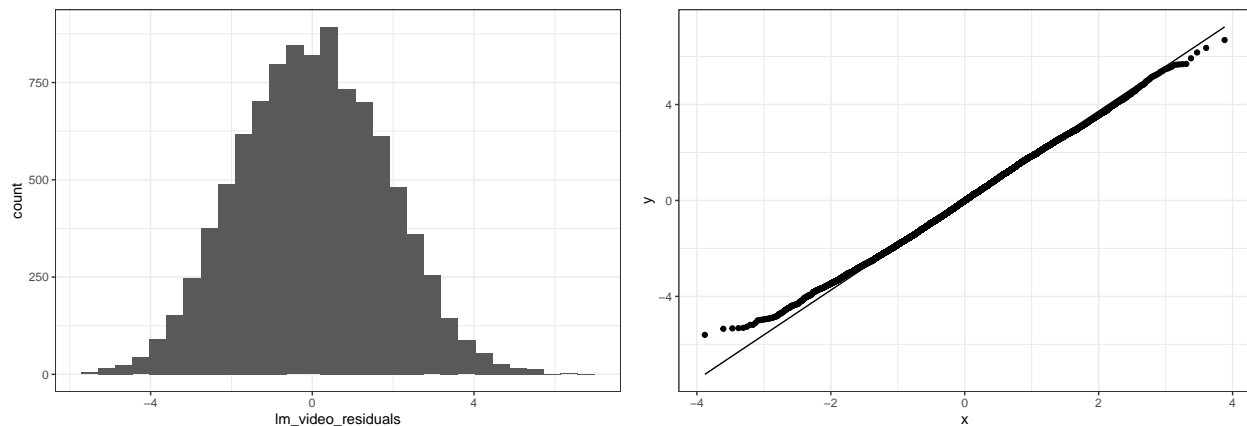


- Additionally, we performed the Breusch-Pagan (BP) test where the null hypothesis is that homoscedasticity is present. With BP value of 128.39 and p values less than $2.2\text{e-}16$ means, we reject the null hypothesis and conclude that there is strong evidence that heteroskedasticity exists in the regression model. Given the heteroskedasticity problem, we should use the robust standard error to solve in our model. While the problem of heteroskedasticity is evident here, one of the key causes could be this model is not correctly specified (as noted in our response to the first three assumptions) which requires more care and corrections.

```
##
## studentized Breusch-Pagan test
##
## data: lm_video
## BP = 128.39, df = 2, p-value < 2.2e-16
```

5. Evaluate the **Normally Distributed Errors**: assumption.

- From the histogram and Q-Q plots of residual below, we observe that the errors are close to normal distribution, with slight thin tails (platykurtic). We also measured skewness and kurtosis of the errors, with skewness of 0.05738289 (within ± 0.5 of zero, with zero being normal distribution) and kurtosis of 2.784654 (within ± 0.5 of 3, with 3 being normal distribution), which further confirms our observation from the graphs. We conclude that there is no strong evidence that this model violates the **Normally Distributed Errors** assumption.



““