HW12: CLM Practice

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

2022-07-25

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:

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

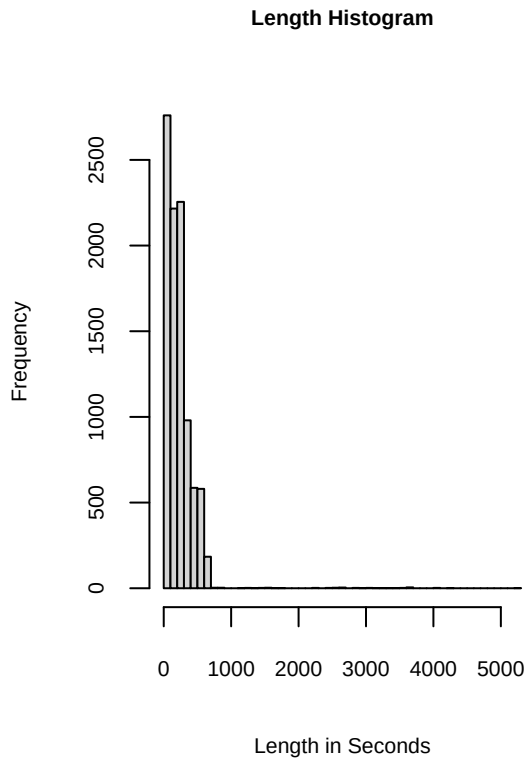
The variables are as follows:

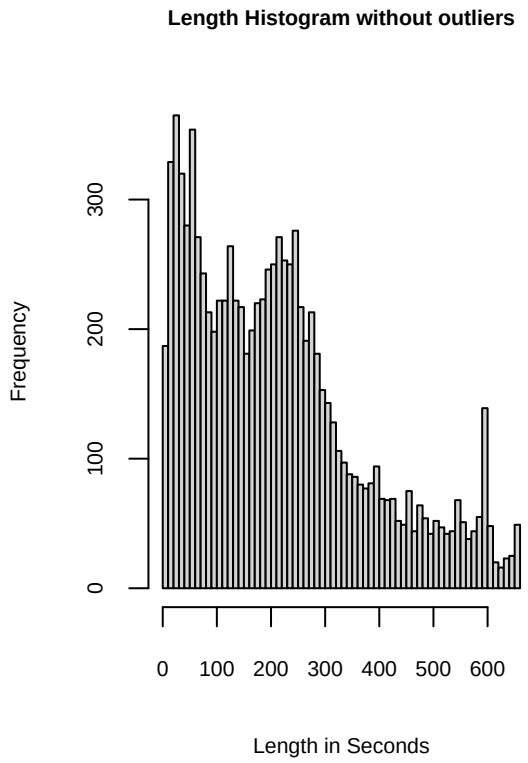
- `views`: the number of views by YouTube users.
- `rate`: This is the average of the ratings that the video received. You may think of this as a proxy for video quality. (Notice that this is different from the variable `ratings` which is a count of the total number of ratings that a video has received.)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

Response:

A. Summary of EDA Work

- We performed extensive EDA work to understand the sample collection process and the dataset. During our data wrangling process, we took the following actions to improve the data quality and model interpretability. The original dataset has 9618 observations, and our final dataset has 9558 observations after removing 60 observations.
  - Removed 9 observations with NA in rate and length field
  - Removed 51 observations with video length more than 11 minutes (considered as outliers) to address the multimodal distribution issues. Based on the histogram of the length variable, there are a small number of videos (51 videos, ~0.5% of the total videos) that are longer than 11 minutes (outliers) which suggests multimodal distribution (a distribution with more than one mode), which may indicate violation of identical distribution. In our EDA work, we considered removing the 51 videos from the final dataset as these videos may have a different distribution, which improves the sample data quality for the model. See below the chart that compares the video length histogram with and without the outliers.
  - For 1490 rows with 0 value in the rate field, we believe that this is likely due to viewers not assigning a rating and not because viewers assigning a 0 rating. Therefore, we replaced the 0 rating with the mean rating of the dataset





B. Evaluate five CLM Assumptions

We run the linear regression model on the final dataset of *9558* observations. See our assessment below.

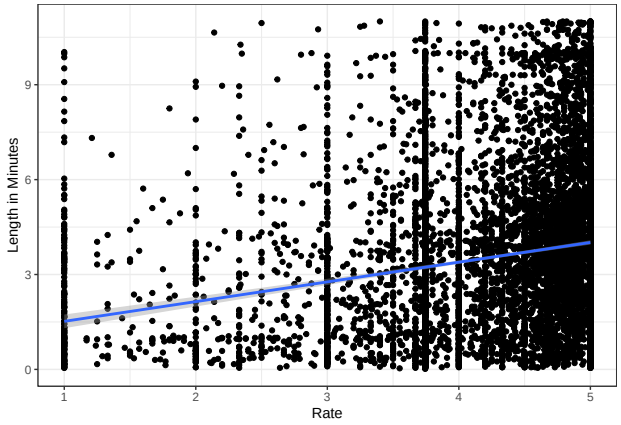
1. Evaluate the **IID** assumption

- Assessing the IID assumption requires an analysis of the sample selection design process. Based on our understanding of the selection process, the list of videos was selected from YouTube using a crawling algorithm which 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 uses this list to find more related videos. This sampling process does not general independent samples, and we believe that the videos in this dataset have clustering issues and are related to each other in content or in time sequence. For example, if the sample time frame is around election time, we would expect that the initial list of "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 other 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 of the classical linear model, one of the mitigating solutions to consider is get new data by using a new random sampling process.

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

   In order to assess nearly perfect colinearity, we use our background knowledge to evaluate the relationship between rate and length, examine the scatter plot of the two input 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.
   - The regression model has not dropped rate or length variable automatically which means that there is no perfect colinearity.
   - We examined the scatter plot of rate vs length below which shows that rate and length do not have strong relationship. Please note in our EDA work, we observed for 1490 videos with 0 rating, which is very likely because viewers did not rate the videos (and not because viewers assigned a 0 rating). Therefore we chose to replace the 0 rating with the mean rating (3.744057) of the dataset, to improve the data interpretability.



- Since there are only two input variables, We can use Pearson correlation test to check whether the pair (rate and length) has near perfect correlation. The test shows that the estimate correlation between rate and length is 0.19 (or 0.16 prior to the EDA data cleanup noted in #1 above). This means that there is low correlation between these two variables.
- We also performed the Variance Inflation Factor (VIF) test below. When VIF is less than 5, there is no evidence of the problem of multicollinearity among the input variables (rate and length).

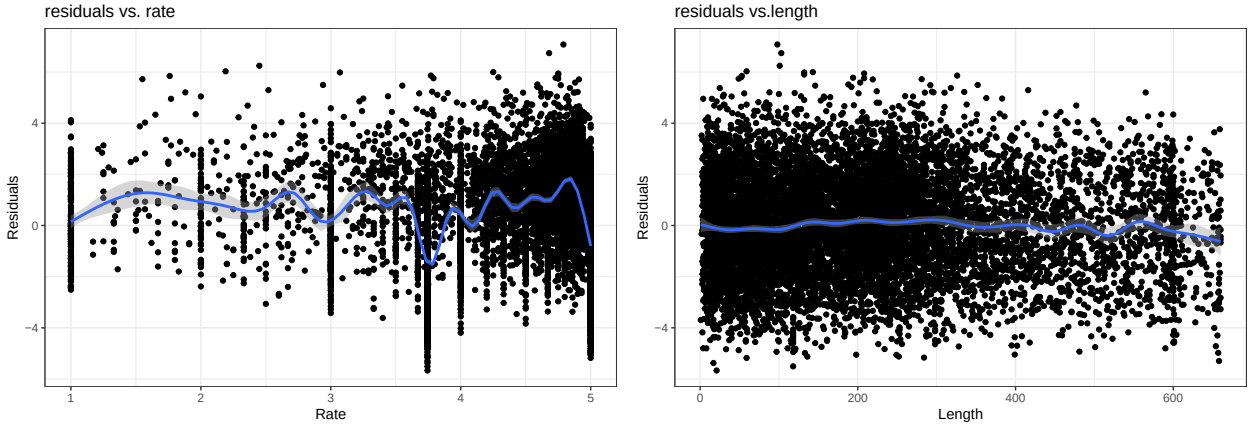
Table 1: Tolerance and VIF Table

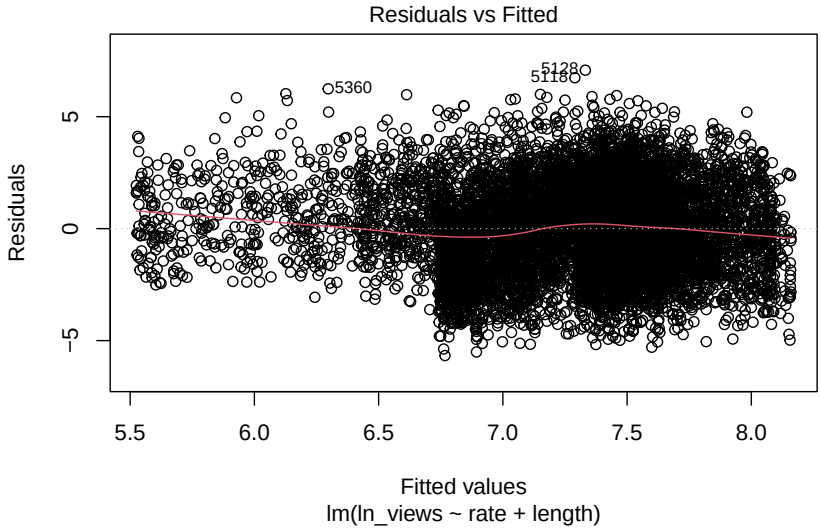
Statistic	N	Mean	St. Dev.	Min	Max
Tolerance	2	0.9621	0.0000	0.9621	0.9621
VIF	2	1.0394	0.0000	1.0394	1.0394

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

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

- This is an assessment of 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 each input variable (rate and length) and the outcome variable (predicted views) to assess whether residuals are close to zero with respect to input variables and the outcome variable.
- Using our final 9558 observations, we plotted the residuals with our input variables (rate and length) and outcome variables (log views). Looking at the graph of residual versus rate (the left graph below), we see that the line of residual average is oscillating around zero. The line of residual versus length (the right graph below) is relative flat, which has improved from the original chart if we were to use the length data without removing the outliers (videos longer than 11 minutes). The line of residuals versus fitted values (the bottom graph) is generally around zero. Please note that when we generated the model using the original 9618 observations without removing the 60 outliers, we observed the line of residuals versus fitted values were not flat and curved downward (a violation of the linear conditional expectation).

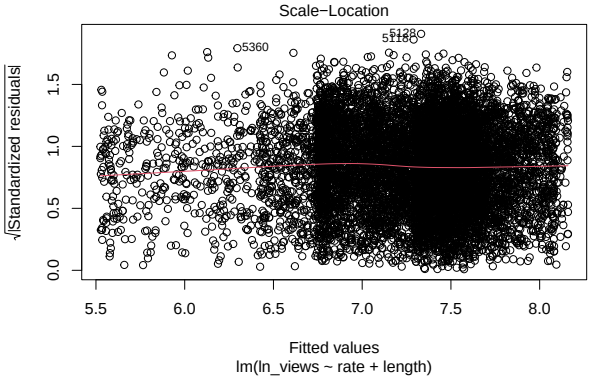




- With our EDA data wrangling work, our final model meets the Linear Conditional Expectation assumption. Note that this assumption would have been violated if we were to use the original 9618 observations for OLS regression.

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

- To evaluate this assumption, we plotted the standardized residuals against the fitted values. Our visual observation is that the variances stay constant (in the range of 0 to 1.5 for all fitted values), no strong evidence of heteroskedasticity. Please note that when we used the *original 9618* observations, we observed strong evidences that the variance increased as fitted value increased, which suggested the problem of heteroskedasticity. Our final model does not show evidence of the problem of heteroskedasticity.

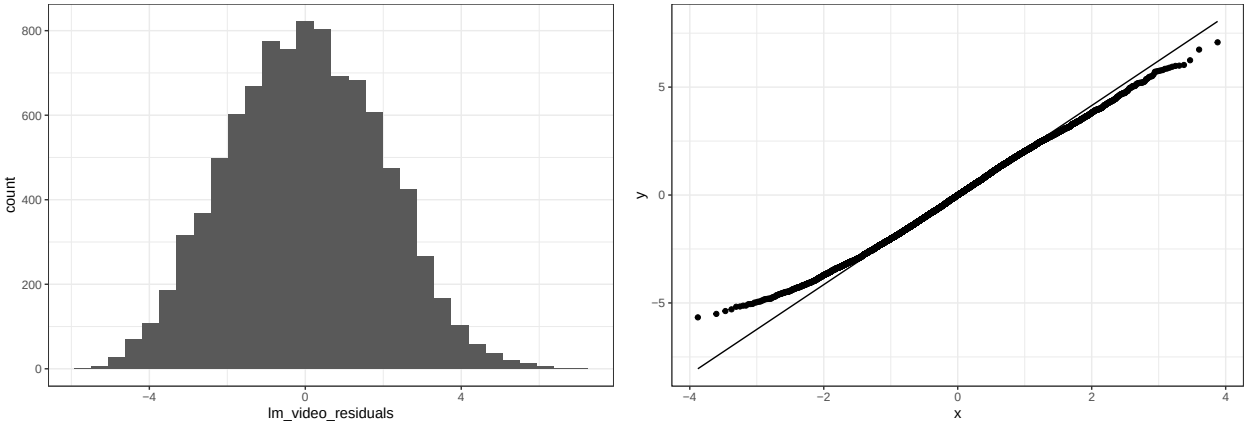


- Additionally, we performed the Breusch-Pagan (BP) test where the null hypothesis is that homoscedasticity is present. with BP value of 3.16 and p-value of 0.20, we fail to reject the null hypothesis of Homoskedastic Errors. Please note that with the original dataset of 9618 observations, the model would have BP value of 128.39 and p values less than 2.2e-16, which shows strong evidence that heteroskedasticity exists in the regression model. In summary, our final model meets the **Homoskedastic Errors** assumption.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_video
## BP = 3.1602, df = 2, p-value = 0.206
```

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

- From the histogram and Q-Q plots of residuals below, we observe that the residuals' distribution has a shape of a normal distribution, with slight thin tails (platykurtic). We also measured skewness and kurtosis of the errors, with skewness of 0.05519 (within +-0.5 of zero, with zero being normal distribution) and kurtosis of 2.6385 (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.




```
## [1] "Skewness of Residuals is: 0.0551892327383251"
```

```
## [1] "Kurtosis of Residuals is: 2.63852906588865"
```

C. Final Regression Model Coefficients and Test Results

Table 2: Video Views Model

	Dependent variable:
	Views (natural log)
Rate	0.4440***
	(0.0229)
Length(in minutes)	0.0013***
	(0.0001)
Constant	5.0772***
	(0.1009)
Observations	9,558
R^2	0.0526
Adjusted R^2	0.0524
Residual Std. Error	1.9439 (df = 9555)
F Statistic	265.0835*** (df = 2; 9555)
Note:	*p<0.05; **p<0.01; ***p<0.001

The coefficient test results are below. All coefficients are statistically significant.

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 5.07717221 0.10558225  48.087 < 2.2e-16 ***
## rate        0.44401340 0.02439733  18.199 < 2.2e-16 ***
## length      0.00130643 0.00012693  10.293 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```