HW 12: CLM Practice

Brian Tung, Kevin Cahillane, Meng-Kang Kao, Nic Brathwaite

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.

0. Data Wrangling

Upon reviewing Youtube's video length limit in 2008, we discovered that at the time only the "verified" accounts can upload videos more than 10 minutes. There are only 232 rows, compared with overall 9618 in the dataset, we believe those videos should be excluded from our CLM assumption evaluation.

We also removed 9 rows that do not have any views from the dataset.

Lastly, there are 1473 rows that have rate value as 0. Upon reviewing how the rate is calculated, value 0 means the video does not have rate information. We decided to use the average rate for all videos with valid rate information (4.42) to replace those rate with 0 values.

1. Evaluate the **IID** assumption.

The IID assumption for justifying the use of a classic linear regression model mandates that each video in the dataset was independently selected from one another and had the same probability distribution of being included in the data as all other Youtube videos. This sample of Youtube videos violates the IID assumption as explained in the Cheng, Liu and Dale's paper[1]. When explaining the web crawler they used to sample youtube videos, they write "we defined the initial set of a list of IDs, which the crawler reads in to a queue at the beginning of the crawl. When processing each video, [the crawler] checks the list of related videos and adds any new ones to the queue" [1185]. As a result, the videos were not sampled independently of one another but were actually clustered by topics and keywords. This would be similar to sampling all individuals in a randomly selected household where individuals in larger families would be more likely to be selected.

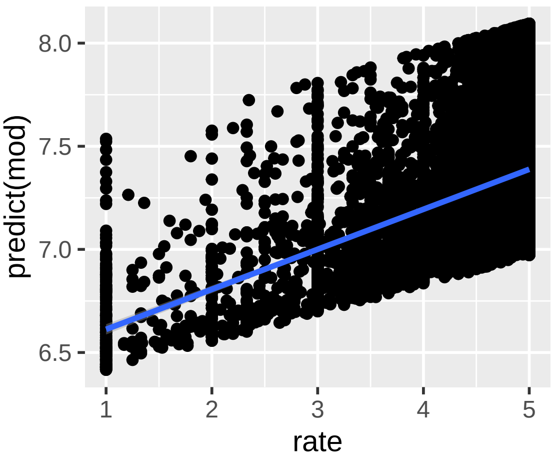
2. Evaluate the **No perfect Collinearity** assumption.

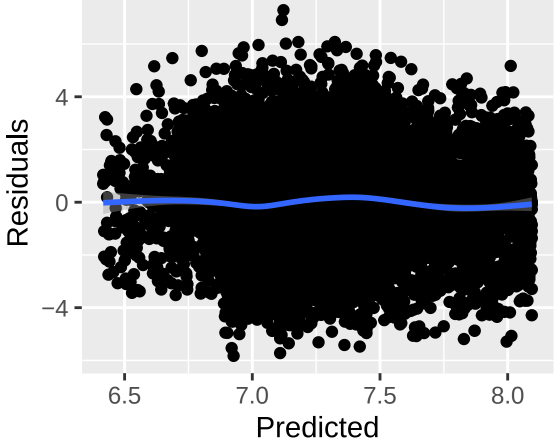
The no perfect collinearity assumption states that there isn't an exact linear relationship between 2 or more input variables. Using the vif function, we can confirm that there is no risk of perfect collinearity as the vif coefficients of our input variables are very low at 1.025.

[1] Cheng, Liu and Dale (2013), Understanding the Characteristics of Internet Short Video Sharing: A YouTube-Based Measurement Study https://www2.cs.sfu.ca/~jcliu/Papers/UnderstandingCharactiristics.pdf

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

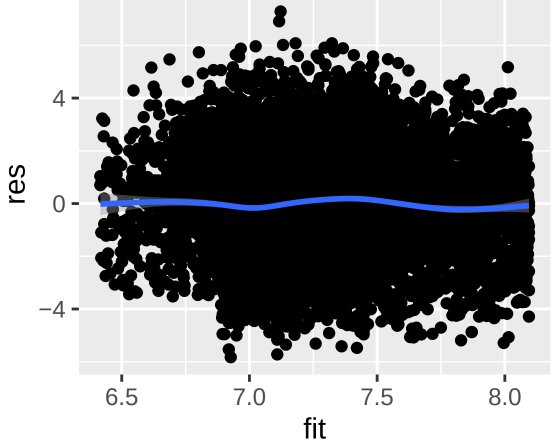
When assessing for any linear conditional expectations within the data our x values will show dependency with our y values by offsetting the residuals and predictions. Since the data set in use is looking at video rates and length we have plotted a graph for the rate vs our model's predicted length and predictions vs residuals to visualize if there are any conditional expectations between the two. Our rates vs the model's prediction clearly visualizes a predicted linear regression line that captures the association between our variables. Our residuals and predictions plot shows a steady balance where our predictions account for a majority of the residual values. Based off these plots and transformation of data we can conclude that the condition for Linear Conditional Expectation is met. The residual and prediction graph is evenly split between the residuals range in our graph.

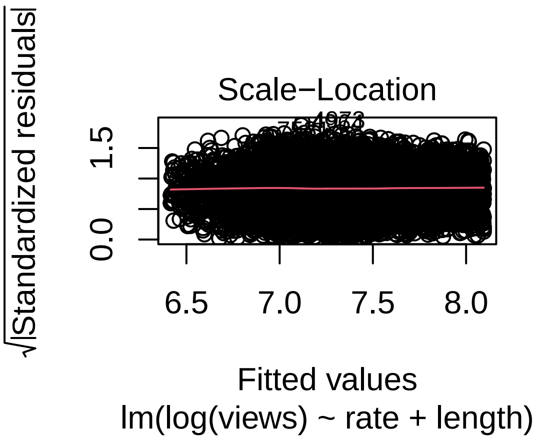




4. Evaluate the **Homoskedastic Errors:** assumption.

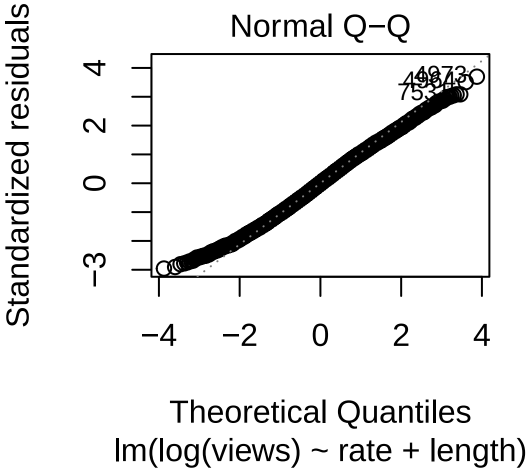
The homoskedastic assumption states that the variance of residuals in the regression model, while unknown, remains constant for all predictors. Homoskedasticity of a regression model can be determined through the Breusch-Pagan statistical test, where the null hypothesis is homoskedastic error and the alternative hypothesis is heteroskedastic error. This test reveals a p-value of 0.03, so we can reject the null hypothesis. Although the scale-location visualization shows a flat smoothing curve, we decided to form our conclusion on the homoskedastic assumption based on the statistical test rather than a visualization estimate. As a result, we conclude that the variance of residuals is not constant and is thus, heteroskadistic.

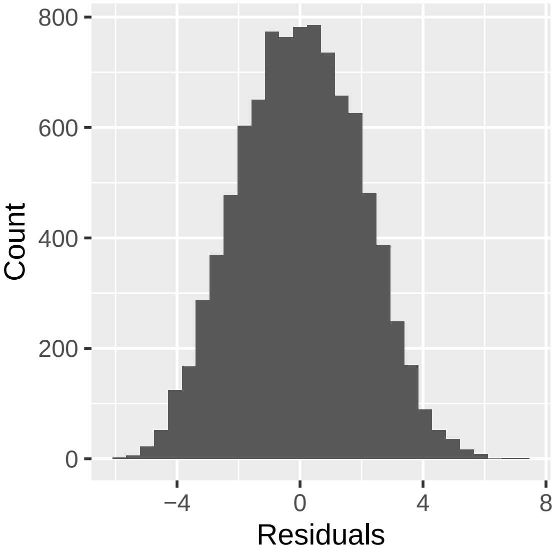




5. Evaluate the **Normally Distributed Errors:** assumption.

Upon visually inspecting the QQ plot and histogram on the model and residuals, we believe the errors are normally distributed.


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(views) ~ rate + length)


Count
Residuals