Unit 12 HW: The Classical Linear Model Q1.1

#Loading in the data

```
videos <- read_delim("videos.txt", delim = "\t")
videosfilteredout <- videos %>% filter(!is.na(videos$views))

#glimpse(videosfilteredout)
#summary(videosfilteredout)
```

#creating the model

```
#model_one <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
videos_model <-lm(log(views) ~ rate + length, data=videosfilteredout)

videos_modelnotfitted <-lm(views ~ rate + length, data=videosfilteredout)
```

- Q1.1 I.I.D. data

  To assess IID data, we need to understand the sampling process used to collect the data. From the videos.txt documentation, we learned that the videos were selected initially from the set of videos included in "Recently Featured", "Most Viewed", "Top Rated", and "Most Discussed" from "Today", "This Week", "This Month" and All Time" on February 22nd, 2007. This totalled 189 unique videos followed by a "crawl" and this process was followed for the remaining data collected through 2008. The data collected was by video ID and includes the variables uploader, age, category, length, views, rate, ratings, comments, and related IDs extracted from the YouTube API. There are several reasons why this data collection process might not result in IID data, below.

  The primary focus of this data collection was on successful videos by collecting from the "Recently Featured", "Most Viewd", "Top Rated", and "Most Discussed" categories. This likely indicates that the data is not distributed as the population of all YouTube videos and is heavily weighted towards successful videos.

  Clustering may also be a factor in this data given that videos in the top categories are likely related to one another. They could be videos uploaded by the same user or very similar content.

  Given the findings above, we believe the IID assumption is not met.