Evaluating the Classical Linear Model Assumptions

Q1.1 Independent and Indentically Distributed

To assess if data is IID we need to analyze the design and sampling process. Details about how the data was collected can be found at https://netsg.cs.sfu.ca/youtubedata/.

The data was collected using a breadth-first crawler algorithm. The initial set of 0-depth videos was selected as all unique videos from the "Recently Featured", "Most Viewed", "Top Rated" and "Most Discussed" lists, for "Today", "This Week", "This Month" and "All Time" on February 22nd, 2007. For each video in the 0-depth set, the algorithm scanned the first 20 entries on the related videos list and added any new videos to queue. The crawl then proceeds to the next depth and repeats.

From this description alone we expect the videos may not be independent.

- New videos are collected by scanning the related videos list. Videos from this list are not randomly selected and have some relation to the original video such as common creators or descriptive keywords (clustering). Content from the same creator may receive similar views.
- Some clustering based on Youtube metrics ("Recently Featured", "Most Viewed", "Top Rated" and "Most Discussed") is also present which restricts the reference population to only videos found within these groups and those related to them. Thus, some videos in the total population will not have the same probability of appearing in the crawls.
- Successful videos may be imitated by other content creators (strategic effect).
- The data may not fully represent all Youtube videos, as the description states that new videos are still being added every day. If this is true, the current data will have been gathered based on related videos. Therefore, videos with less characteristics in common with the initial videos will be underrepresented.

For these reasons we conclude that the assumption of IID data is questionable.