Question 1.1: Independently and Identically Distributed

DATASCI 203 Homework: Classical Linear Model Practice

Wesley Chang

Evalutate the IID Assumption:

After viewing the description of the data collection for the dataset (Statistics and Social Network of Youtube Videos), it is clear that the intended analysis does not seek to gather independently distributed data. The study makes the assumption that all Youtube videos can be connected together in directed graph structure, based on the first 20 related videos. The crawler algorithm attempts to build a database of all Youtube videos by implementing breadth-first search using the related videos to find connecting nodes. When the algorithm finds a video not previously gathered, it adds that video to the current list of videos.

The method is an obvious violation of independence as Youtube compiles the list of related videos based on perceived relevance to the current video such as category, topic, title, tags, channel, and numerous other factors. The website indicates that videos are still being added to the list (73000 new videos per day on average), so it is a safe assumption that the data is not complete. If we consider that the videos currently gathered can all be linked through related videos and that the data is not complete, we can expect a systematic bias towards videos more similar to the initial starting videos and that videos with fewer shared characteristics with the original ones will be underrepresented.

Therefore, given the seemingly incomplete data and intentionally clustering prone collection method, I conclude that the data does not meet IID.