Problem set 4: Movie ratings

S470/670 Spring 2020

Upload your submission to Canvas by 11:59 pm, Thursday 27th February.

Do longer movies tend to get higher IMDB ratings, after accounting for their year of release?

Get the data

Download the following two files from `https://datasets.imdbws.com/`: `title.ratings.tsv.gz` and `title.basics.tsv.gz`. (You may need to download software to unzip the file, e.g. 7Zip.) The unzipped files are in tab separated value form. (Warning: the latter file is about half a gigabyte unzipped.)

Unzip the files and read them into R. Using `read_tsv` or `read_delim` is recommended, e.g.:

```
read_tsv("title.ratings.tsv", na = "\\N", quote = '')
```

(Those are supposed to be straight quotes. You'll get a few tens of thousands of warnings, but these pertain to the variable `endYear`, which is irrelevant for our purposes.) Merge the two data sets by the unique identifier `tconst`. We only want movies, so only keep data for which `titleType` is "movie." After doing this, I ended up with about 240,000 movies.

Questions

1. Fit a model to predict a movie's IMDB rating (variable `averageRating`) by year (`startYear`) and length (`runtimeMinutes`.) You will have to make a number of modeling choices:

   (a) Do you need any transformations?

   (b) Should you fit a linear model or something curved?

   (c) Is an additive model adequate?

   (d) Do you need to filter out or downweight tail values to prevent the fit from being dominated by outliers?

   (e) Should you weight by number of votes?

   Some of these choices are clear-cut, while others will be a matter of preference. You must justify all your choices. You'll be graded on the justification, not the choice (unless the choice is really bad.)

   Note that computational concerns will also drive your modeling choices. For example, you will not be able to put all the data into a loess unless you have a supercomputer.

2. Draw ONE set of faceted plots to display the model — either condition on year or length, whichever seems to you to be more interesting. Choose a sensible number of panels. Briefly describe what this set of plots shows you.

3. Draw a raster-and-contour plot (or other "3D" plot of your choice) to further display your model. The plot should show predictions for the majority of movie years and lengths (you don't have to show outliers.) Briefly describe what, if anything, this plot shows you that your plot for question 2 didn't.

4. Answer the substantive question: Do longer movies tend to get higher IMDB ratings, after accounting for their year of release? (The answer will likely be more complicated than "yes" or "no.")

What to submit

Submit an R/Markdown code file to reproduce your fit and plots, and a write-up containing:

- The line of R code you used to produce your model.
- A justification of your modeling choices.
- TWO graphs and brief comments on those two graphs. (You should draw many more than two graphs when deciding what model to fit, but only include two graphs in your submission.)
- An answer to the substantive question.

You don't have to upload the data.