Making Movies:
Are films shorter when the script-writer directs?

I. Gordon Blackadder

(Dated: 10 December 2015)

Introduction

What really sets going to the movies apart from watching TV is the big screen. From landscapes that stretch for miles to close-ups that show every detail of a characters face, films have the power to succinctly convey complex emotions and situations. With the added benefit of a captive audience, a movie can explain an idea in moments that might take minutes of dialogue on a TV show or several pages in a novel.

So who decides how a film should express its story? There are actors, producers, make-up artists, editors etc, but the two people principally responsible are the writer and the director. The writer creates or adapts the story which the director then brings to life on the screen. It is easy to imagine that these two people might have different strategies for expressing ideas, with the director preferring a quick visual to paint the scene and the writer choosing dialogue and actions.

However sometimes a film is written and directed by the same person in which case there would be only one approach. In these cases I would expect the films to be shorter than average as the movie would be more likely to express ideas visually. Of course it would be easy to argue the other way so what is needed is a thorough analysis.

In this paper I present a data analysis project that sought to answer whether films are shorter when the script-writer and the director are the same person.

The next section describes obtaining and munging the dataset. After that I explain the model used to analyze it. Finally there will be a discussion of the results.

Data

A screen scraping algorithm was written to extract data on about the first million entries of the Internet Movie Database (IMDb.com). First the algorithm checked that the entry was indeed a movie and not a TV show. After that it extracted the title, the year the movie was released, its length (in minutes), the countries where it was made, the languages spoken in it and the genres that best describe it. Finally the algorithm extracted the names of the writers and directors. Note that frequently there were several writers and directors listed. I defined a movie as having the same writer and director if there was any overlap between these two lists.

Next the dataset was munged. A large number of entries were labeled with the genre "Short". These films had runtimes of less than about 45 minutes. While short films are an important part of cinema, this analysis is focused on feature length movies and so all short films were removed. So too were films that don't have a writer in the traditional sense. This meant removing documentaries, adult movies and news reels.

This left the dataset with approximately 200,000 entries.

Analysis

The dataset was first analyzed by year and then the trend over time was found. For each year group the natural log of the runtimes was found so that the runtimes follow an approximately normal distribution. After the analysis the runtimes were converted back so that all results are given in minutes.

I wanted to find the difference in runtimes when the writer and director were the same person compared to when they were not, but I also wanted to look at how that might vary between countries, languages and genres. To do this I assumed that each film drew its runtime from different normal distributions. These distributions had a mean(μ) that was equal to a constant (μ_0 called the global average), that was the same for all movies, plus deviations from that constant depending on the movie. There were two deviations for each country, language and genre. One deviation (δ^{same}) for movies where the writer and director were the same person and the other deviation (δ^{diff}) when they were different.

If a movie (M) had the same person write as direct then this was denoted by $S_M = 1$.

If the movie was not directed by one of the writers then $S_M = 0$. Furthermore if a movie was made in country i then this was indicated by $c_{M,i} = 1$ otherwise $c_{M,i} = 0$. The same convention was followed if the film used language j (identified by $l_{M,j}$) and if the film was made in genre k (given by $g_{M,k}$). Thus the mean of the normal distribution for each movie was given by

$$\begin{aligned}\mu_M = \mu_0 + \frac{1}{d_M}\Bigg(S_M \Bigg[\sum_i^{\text{countries}} c_{M,i}\,\delta_i^{\text{same}} + \sum_j^{\text{languages}} l_{M,j}\,\delta_j^{\text{same}} + \sum_k^{\text{genres}} g_{M,k}\,\delta_k^{\text{same}} \Bigg] \\ + (1 - S_M) \Bigg[\sum_i^{\text{countries}} c_{M,i}\,\delta_i^{\text{diff}} + \sum_j^{\text{languages}} l_{M,j}\,\delta_j^{\text{diff}} + \sum_k^{\text{genres}} g_{M,k}\,\delta_k^{\text{diff}} \Bigg] \Bigg)\end{aligned} \tag{1}$$

where the total number of descriptors that described the film was given by

$$d_M = \sum_i^{\text{countries}} c_{M,i} + \sum_j^{\text{languages}} l_{M,j} + \sum_k^{\text{genres}} g_{M,k} \tag{2}$$

To ensure that the δ's were merely deviations from the global average (μ_0), the model demanded that the total, weighted sum of the deviations was zero

$$\sum_i^{\text{countries}} (n_i^{\text{same}}\delta_i^{\text{same}} + n_i^{\text{diff}}\delta_i^{\text{diff}}) = 0 \tag{3}$$

Here the weight n_i^{same} was the total number of movies made in country i which had the same person write and direct. Similarly n_i^{diff} was the total number in which the writer was not the same person who directed. Identical demands were made of the language and genre deviations.

Each δ was itself assumed to be drawn from a normal distribution centered on 0 (all except one which was fixed based on the constraint of equation 3). The global average was taken from a normal distribution centered on (the natural log of) 100 minutes.

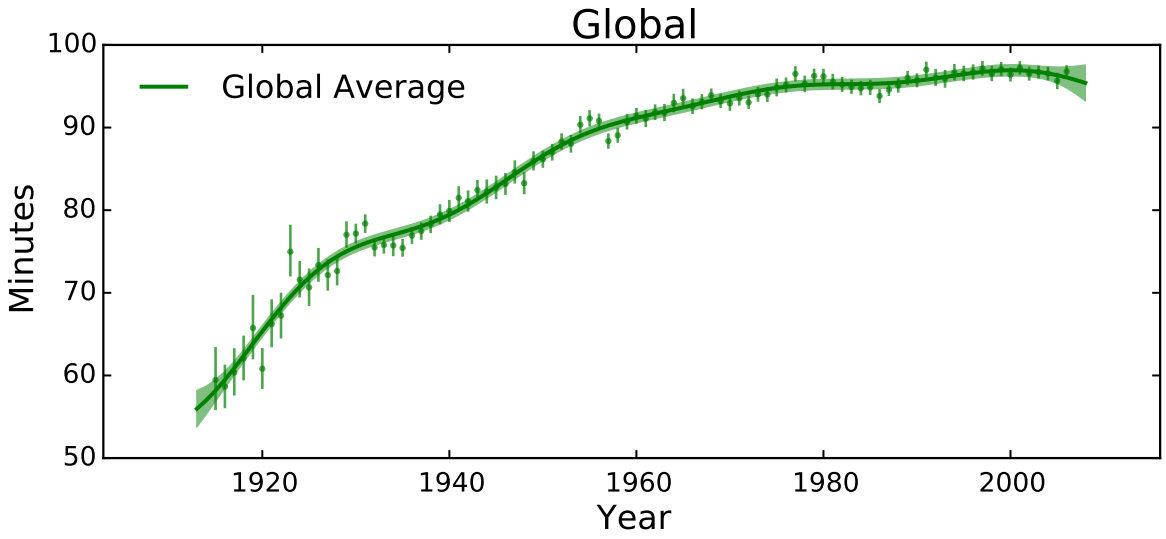
For this model a Markov Chain Monte Carlo simulation was executed (using the python module pyMC2). The MCMC was iterated 300,000 times. The results consisted of means, standard deviations and 95% confidence intervals for the global average and each deviation. This was done for each year in the data.

After these were plotted a gaussian process regression was used to find the trend in the data over time. It had a length scale of 10 years meaning that it found the "slow trend" in the results.

Note that while every country, language and genre that appeared in the dataset was analyzed by the MCMC only a subsample were plotted. The subsample consisted of those countries, languages and genres that had at least fifty years worth of data with twenty films in each year. Many countries and languages were sparsely represented in the original dataset which was distinctly skewed towards European and North American movies.

Results

Let's first consider the global average runtime. In the plot below we see a rise in the length from less an hour in early cinema to approximately 95 minutes since the wide spread adoption of television.

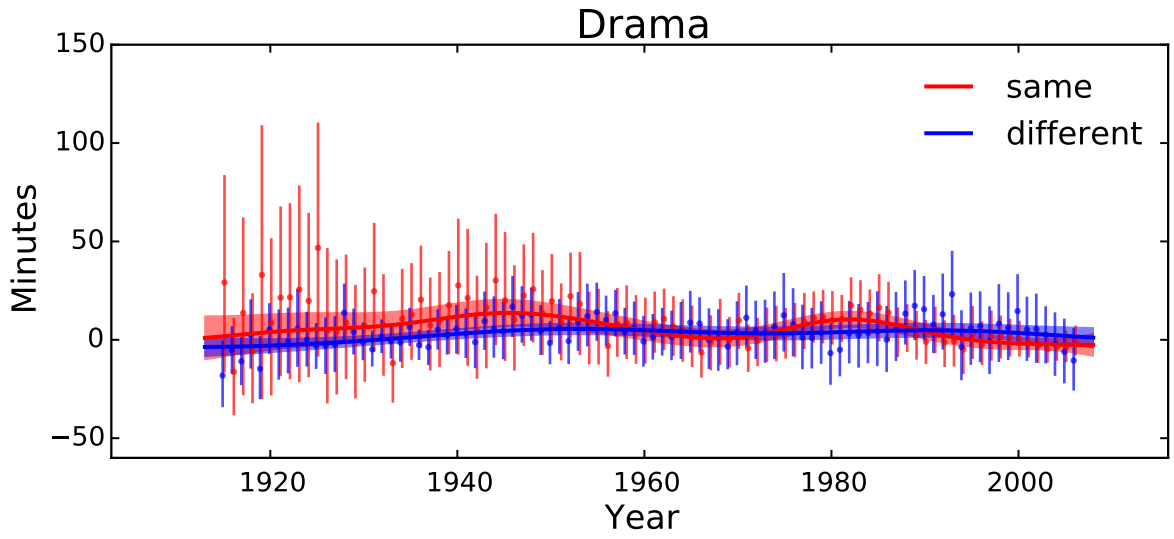


Note that in general, the size of the error bars are inversely proportional to the number of movies made that year. We will see ink almost all of the plots that the error bars are larger at earlier times.

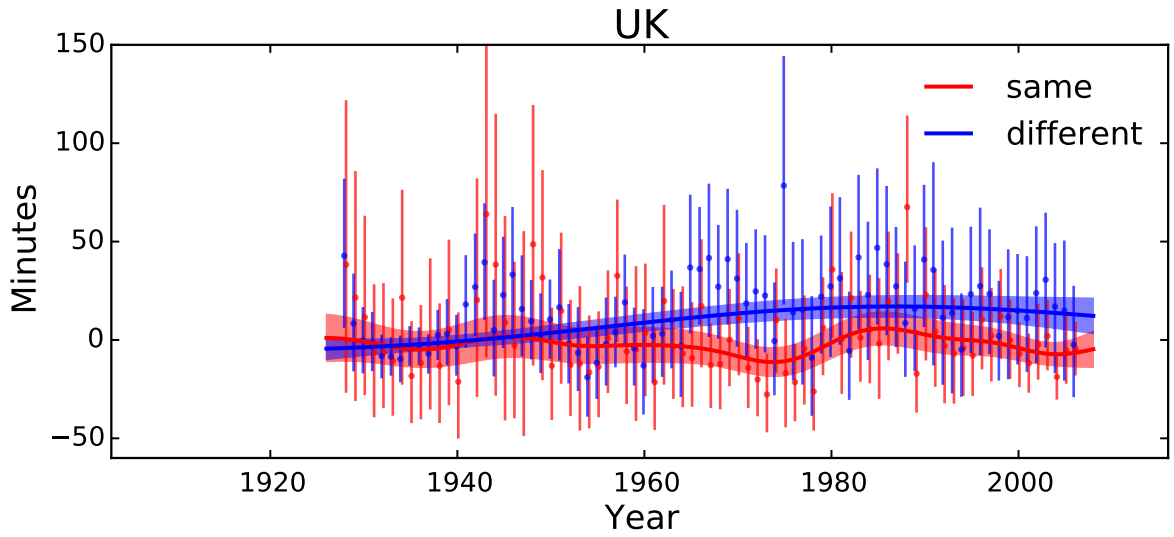
Now let's look at the deviation plots. In the figures that follow, movies with the same person writing as directing at labeled as "same" and are plotted in red. All other movies are labeled as "different" and are plotted in blue.

Interestingly the analysis shows that there is almost no disparity between the same and the different across the genres. For example the drama figure, below, appears to show no

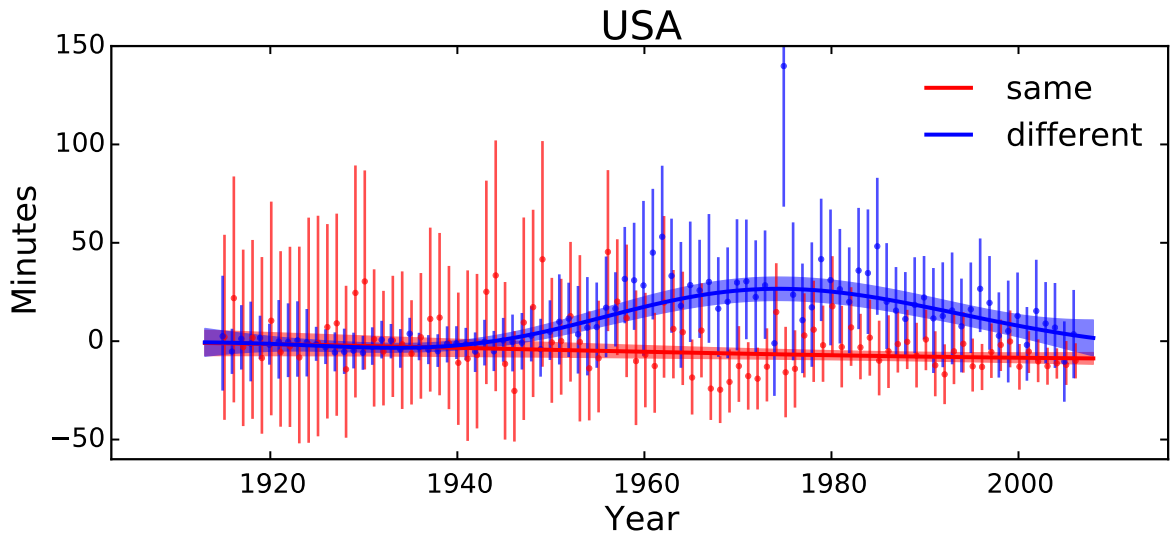
clear separation at any point in history between the two. The is representative of the results from all genres and would seem to disagree the hypothesis (that films where the writer was also the directer are shorter on average).



Now consider deviations by country. The next two figures show the deviations for the UK and the USA. The US is the country most extensively represented in the dataset so we might expect that it should have the clearest results.



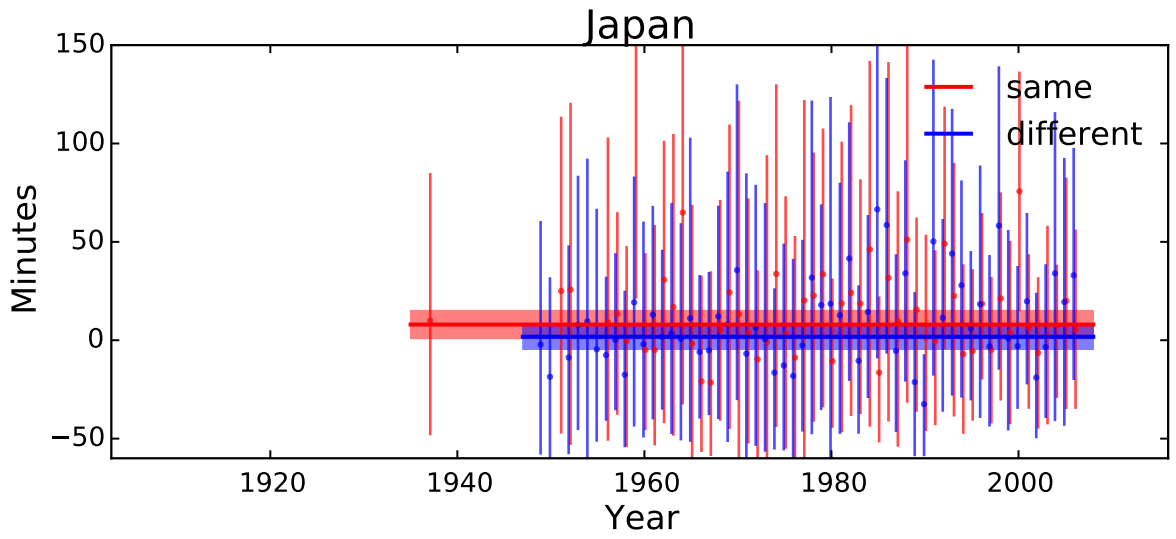
And indeed we can see in both the UK and US results that there is a clear gap between in the runtimes. While there is overlap between "same" and "different" in the early history of cinema we can see that, as movie runtimes have on average increased, a clear gap has emerged with films written and directed by the same person up to tens of minutes shorter.



The fact that this exists in the country deviations but not the genre deviations suggests that the difference is down to cultural factors. Such gaps are not obvious in the results from other countries. For example considering films made in Japan "same" and "different" movies are very close to each other in length and exhibit a slight preference towards longer runtimes when the writer also directs (although it is not clear exactly how significant that is).

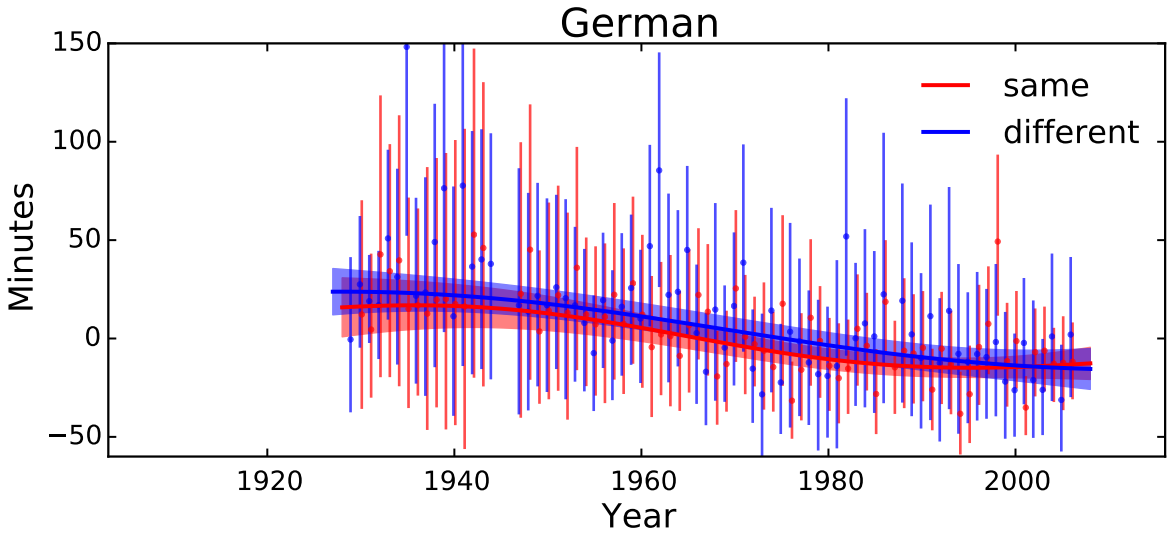
Notice that there is a clear outlier in the plots for "different" in both the UK and, especially, the USA results at 1975. This outlier is matched by an opposing outlier in the results of English language movies. In this instance it appears that the model found a local minimum by differing strongly from the global average one way in the country deviations which was balanced by a large language deviation. Although the use of gaussian process regression prevents this outlier from stingily affecting the results, it does highlight a broader problem with the model. Clearly there is a correlation between the countries in which a film is made and the languages spoken in it. A future iteration of this analysis should take this into account (although it would also be iterating to see if nested sampling, instead of

Markov Chain Monte Carlo, might have prevented this problem).



Finally let's consider deviations by language. It is instructive to consider films made in German, which is easier to consider than films made in the country Germany which was split and then unified over the course of the 20th century. Here we see no difference between "same" and "different" although interestingly we do see a downward trend in both as German films have fallen in runtime relative to the global average. This result again suggests that differences in the runtimes of films written and directed by the same person, as opposed to all other films, are due to cultural reasons and may not be found in all of cinema.

Further figures were plotted and are available online at github.com/igblackadder/FilmProject/plots


German
same
different
Minutes
Year
150
100
50
0
−50
1920
1940
1960
1980
2000