The Effect of Popular Movie Characters on Baby Naming Trends

Ben Graf

STA 6233 – Advanced R Programming, Spring 2020, Dr. Matthew Martinez
The University of Texas at San Antonio, San Antonio, TX 78249

OVERVIEW

Popular movies affect many facets of modern society and culture, including how we speak, what we eat, what we wear, and even how we think. Does their impact go so far as to affect what we name our children? **This project examines whether character names from popular movies (as determined by box office revenue) affect trends in baby naming.**

I combine box office totals from Box Office Mojo, baby name counts from the Social Security Administration, and scraped movie character names from the Internet Movie Database to create a dataset to investigate this question. I examine individual names extracted from the top-billed movie characters from the top five movies in each year going back as far as 1921. I employ the Chow statistical test to determine whether the movie's release represents a structural break in each name's time series.

[1]

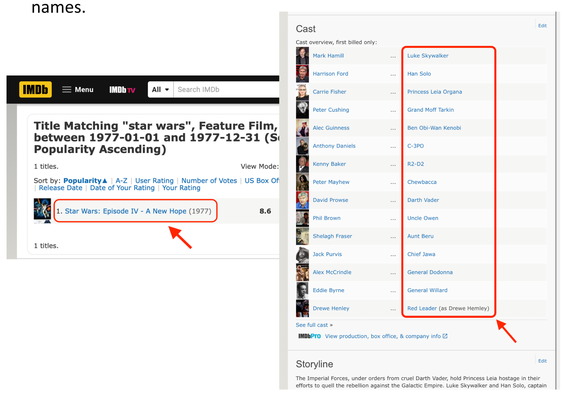
METHODOLOGY

Answering this question requires three separate datasets:

- Box Office Mojo lifetime domestic gross of 16,542 feature films (downloaded file previously scraped by others)[2]
- Social Security Administration (SSA) name counts by year, from 1880 to 2018, from Social Security applications, 1,957,046 observations (downloaded)[3]
- Internet Movie Database (IMDb) character names for applicable films (scraped by me)[4]

I use the Box Office Mojo data to identify the name and release year for the **top 5 movies (by lifetime domestic gross)** for all years in the dataset. Then I scrape the character names for these films from IMDb in two steps:

- First, I adapt the URL for IMDb's Advanced Search results page to get the link to the specific film's page. My initial pass does not find the film's page for roughly 8.5% of the movies, so I revise those search terms to successfully acquire their URLs.
- Second, I scrape each film's page for all of the top-billed character names.



METHODOLOGY (continued)

For the **top 5 billed characters for each film**, I break apart the full character names into individual names (e.g., "Princess Leia Organa" into "Princess", "Leia", and "Organa"), correcting for special characters (á), non-names ("The"), and common honorifics ("Princess").

I then pull the SSA baby name data for each of these names for the **10 years before and after the movie's release**.

For my analysis, I utilize the **Chow test**. This statistical test detects a structural break at a particular point in a time series analysis. Separate linear regressions are performed on the data before the event and after the event. The null hypothesis is that the coefficients of the two regressions are the same, and the test statistic follows an F distribution.[5]

A p-value is generated for each individual name having enough associated trend data. (43% of individual names do not.) I consider a p-value below a Type-I error rate of α=0.01 to be significant, a **"name effect"**.

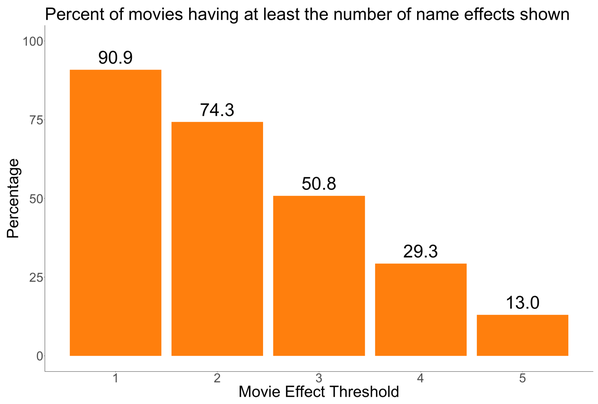
I define a **"movie effect"** on baby name trends as occurring when a movie's characters (collectively) meet a certain threshold of "name effects". I test several such thresholds, as shown in the Results section.

RESULTS

My analysis includes 2,515 individual names from 1,347 characters from 307 movies. Below is example data for a single movie, *Gladiator* (2000). Note the No Data column, which counts how many of the 21 years examined did not contain data for that name. The "Movie Effect" threshold used in this case was 2, hence the result of TRUE.

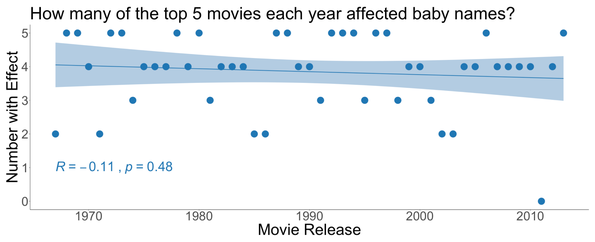
Individual Name	Character	Movie	Year	No Data	Chow p-value	"Name Effect"	"Movie Effect"
Maximus	Maximus	*Gladiator*	2000	7	0.00019	TRUE	TRUE
Commodus	Commodus	*Gladiator*	2000	21	NA	NA	TRUE
Lucilla	Lucilla	*Gladiator*	2000	16	0.26599	FALSE	TRUE
Proximo	Proximo	*Gladiator*	2000	21	NA	NA	TRUE
Marcus	Marcus Aurelius	*Gladiator*	2000	0	0.00838	TRUE	TRUE
Aurelius	Marcus Aurelius	*Gladiator*	2000	5	0.01812	FALSE	TRUE

The following graph depicts the percentage of movies that demonstrate a "movie effect" for a variety of movie effect thresholds. **Over 91% of movies tested have at least one "name effect", and over 50% have at least three.**

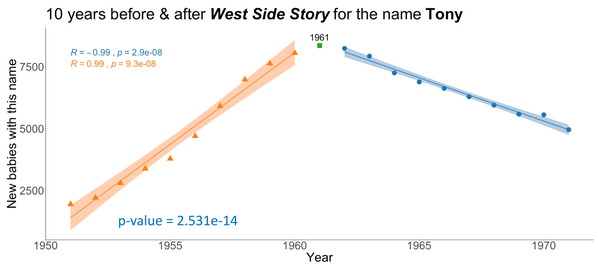


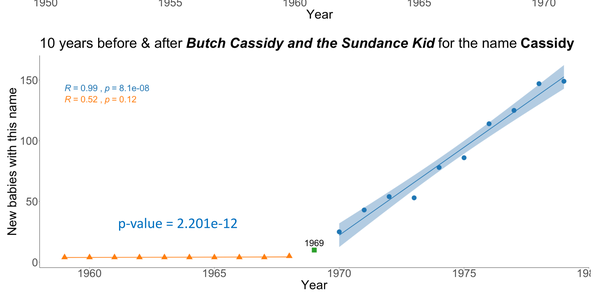
Next, looking at the number of "movie effects" each year (out of the top 5 films in that year) does not show a statistically significant correlation. I cannot say the effect of popular movies on baby name trends has gone up or down over the decades.

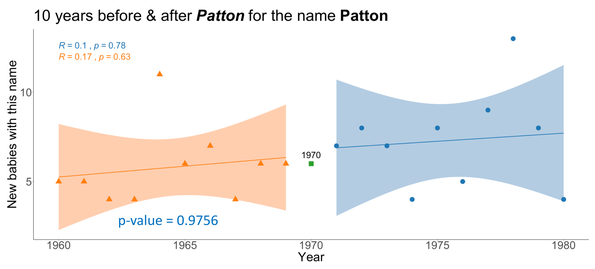
RESULTS (continued)

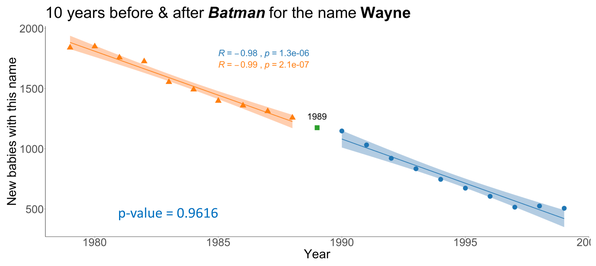


Below are two examples of extremely low p-values (high "name effects") and two examples of extremely high p-values (low "name effects").



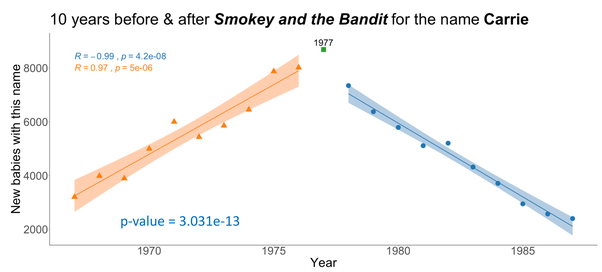






LIMITATIONS

A structural break may exist but may not be in that exact year, and/or it may not be caused by the movie. In the example below, the name "Carrie" from *Smokey and the Bandit* (1977) shows a precipitous drop beginning around the time the movie released, and the Chow test found a substantial effect (p-value = 3e-13). However, in this case, another movie is the more likely cause of the tide turning so dramatically against the name "Carrie": Stephen King's *Carrie* came out the year before, in 1976.



The character names on IMDb may not be listed in order of importance. *Gone with the Wind* (1939) is an example of this. The male romantic lead, Rhett Butler, is not only not in the top 5 billed characters, he does not even appear in the top 15 characters shown on the movie's main page (and scraped). Anecdotally, this appears to be unusual, and further investigation might reveal that it is more common in older films, where the cast was often listed in order of appearance, rather than by prominence.

CONCLUSIONS

- Despite the fact that many factors likely affect baby naming trends, **my results support the idea that popular movies do have an effect.**
- Those effects are not always positive. Villains and other characters with negative associations may drive down the popularity of a baby name.
- Further research might focus on subcategories of movies and/or characters to determine if certain genres, genders, or sequels, for example, are more likely to have an effect. Similar studies might examine baby name trend effects from other sources, such as celebrity names, politician names, television, or books.

REFERENCES

1. Jonnie Andersen, "6 month twins." Licensed under CC BY-NC-ND 2.0 (https://creativecommons.org/licenses/by-nc-nd/2.0/legalcode). https://www.flickr.com/photos/johnnyvintage/25558077263/
2. Elias Dabbas, "Boxofficemojo Alltime Domestic Data." (September 2, 2018). Distributed by data.world. https://data.world/eliasdabbas/boxofficemojo-alltime-domestic-data
3. Social Security Administration, "Baby Names from Social Security Card Applications - National Data." (November 27, 2019). Distributed by Open Data. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data
4. Ben Graf, "IMDb Movie Characters." (March 2, 2020). Dataset scraped from Internet Movie Database. https://www.imdb.com
5. "Chow test." Wikipedia.com. https://en.wikipedia.org/wiki/Chow_test (accessed February 24, 2020).

ACKNOWLEDGEMENTS

Thank you to Dr. Wenbo Wu and Dr. Keying Ye for their discussion and advice regarding the appropriate statistical test to use.