MACS 30200 | Methods & Initial Results
YIQING ZHU

** Some parts asked in this assignment are not in my methods/initial results section, so I include some fractions of other sections, which may be confusing and I'm sorry for that.*

Research Question

How hard are movies with different gender of top billing cast get success respectively by time?

1. Introduction

Acting a prominent role in popular culture, films have been frequently scrutinized to discern how different genders are perceived and evaluated. Presumably, mainstream motion pictures largely reflect prevailing cultural attitudes about gender roles, norms, attitudes, and expectations (Haskell, 1987; Rosen, 1973), and besides mirroring the sociocultural images, films also create them (Millburn, Mather, & Conrad, 2000). Film industry exhibits what it itself believes (Simonton, 2004) through cinematic products and share it with the society at large. The worldwide dominance of Hollywood "blockbusters" has enlarged and extended this effect. Thus, detecting the biases and misconceptions especially regarding to gender in the motion picture industry has been an essential issue.

While most former researchers relevant to this topic focus on the portrayal of different genders in the motion picture industry, this research will empirically examine the relationship between gender of leading cast and cinematic success over time, and will try to fit in a logit model to define the gender of top billing cast of a film concerning its cinematic success. More specifically, this research wants to explore if gender acts as an impact factor of cinematic success in aspect of art and business, and if this impact varies among genres and changes over time.

2. Literature Review

2.1 Cinematic Success

Beginning in nickelodeons as lucrative venture in popular entertainment, film is transformed into a serious art form which partly relied on film scholars and critics who could discuss the medium independent of marketing and box office (Baumann, 2001). Thus, art and business are two antithetical categories films have been often assessed in, in other words, cinematic success consists of aesthetic success and commercial success (Landes, 2002; Hammad Afzal, 2016), which provides basic guidance for the evaluation of films.

Among empirical studies, several investigations have addressed the predictors of critical acclaim or movie awards (e.g., Zickar, Slaughter, 1999; Simonton, 2002, 2004c) and a very large literature has been focused on the factors indicating box office success of a film (e.g., De Vany & Walls, 1999; Dodds & Holbrook, 1988; Litman & Kohl, 1989; Prag & Casavant, 1994; Wallace, Seigerman, & Holbrook, 1993). A few researchers also examined both questions simultaneously (e.g., Simonton, 2005, 2009), however, no research has been conducted directly relevant to the relationship between gender and cinematic success.

2.1.1 Aesthetic success

"The aesthetic evaluation of artworks is, and always has been, a very controversial exercise." (Ginsburgh & Weyers, 2005) There are three broadly acknowledged methods to evaluate beauty: one decomposes an artwork into attributes and rate each (De Piles, 1708; Beardsley, 1958; Vermazen, 1975; Dickie, 1988, 1997), and some art philosophers "locate the ground of judgments of taste, not in some object which is the target of the judgment, but in the maker of the judgment" (Shiner, 1996), and others take "test of time" and "test of space" to examine beauty (Hume, 1757; Dickie, 1988, 1997; Savile, 1982; Budd, 1995).

Instead of analyzing the aesthetic characteristics or examining the temporal and spatial spread of a film, taking judgment from judges as the predictor of aesthetic success is a more available method. In the context of motion picture industry, the judges would be film critics and consumers with the appearance of the film ratings.

"Philosophers typically put the burden of proving quality on experts, while economists often argue that the actual choices made by consumers are a better measure." (Ginsburg, 2003) Actually, the

views of consumers correlate positively with the opinions of film critics (Boor, 1992; Holbrook, 1999; Wanderer, 1970). Moreover, since individual judge may be prone to judgement errors and short-sighted and consumers largely outnumber critics and consumer ratings displays a longer accumulating period, both during theatrical run and post-theatrical period, the consumer ratings appears to be valid as well as more reliable and would be the predictor of aesthetic success of films in this research.

2.1.2 Commercial success

Rather than artistic expression, financial performance has been the goal of "film industry" since its emergence, which is even intensified with the booming of the highly profitable "blockbusters". A quite large number of films are little more than elaborate "get rich" schemes – "products replete with movie stars and special effects but sadly lacking in plot, dialogue, and characterization" (Simonton, 2005b).

The commercial success can be evaluated by the actual profit a film made, however, due to proprietary information including cost of production and budget, the actual profit of a film is unavailable to public in most cases (Litman & Ahn, 1998) and researchers have to estimate the financial performance of a film by various criteria, for example, gross box office earnings or receipts (e.g., Sochay, 1994; Pat Topf, 2010), first weekend gross (e.g., Basuroy, Chatterjee, & Ravid, 2003; Simonton, 2005), the total length of the theatrical run (Sochay, 1994), distributor rental revenue. Though the film revenue encompasses several parts including box-office revenue, DVD versions, and television showings, the box-office revenue appears to be the best way to estimate the commercial success of a film at this time because the information is readily available and movie theaters are still accepted as the major source of revenue for a particular film.

2.2 Gender and Film

2.2.1 Gender schema and Social learning theory

2.2.2 Gender stereotypes in films

3. Method

3.1 IMDb Data

Since the United States film industry dominates the world market (Acheson & Maule, 1994), and cinematic products do not transport well across linguistic and culture boundaries (Lee, 2006), we will confine the research among the American films. The currency difference, inflation, and different development of the commercialization of motion picture also restrict the validity of the evaluation commercial success of multinational films.

The dataset obtained is from Internet Movie Database (IMDb)[1], which can be accessed as compressed plain text files from the ftp sites or extracted using the Unix command-line interface tools[2].

IMDb is currently the world's most popular and authoritative source for movie, TV and celebrity content, offering a searchable database of more than 185 million data items including more than 3.5 million movies, TV and entertainment programs and 7 million cast and crew members. IMDb is a user-contributed encyclopedia, with information mainly gathered by people in the industry and entertainment fans around the world and constantly verified with studios and filmmakers through on-screen credits, press kits, official bios, autobiographies, and interviews.

The IMDb data is regarded as information source about both films and netizens and has been analyzed in a large number of literatures including researches on popular geopolitics of films (Dodds, 2006; Jung, 2012), recommendation systems (Lamprecht, 2015), and online social networking (Fatemi, Maryam, & Tokarchuk, 2012).

3.2 Model & Measurement

The IMDb dataset contains various film information, among which this research will focus on the information about title, cast, aesthetic success indicator consumer rating, commercial success indicator box office gross, production year, and genre.

[1] http://www.imdb.com/
[2] http://www.imdb.com/interfaces

The logit model this research tries to fit in is in the formula as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_1 X_{4i}$$

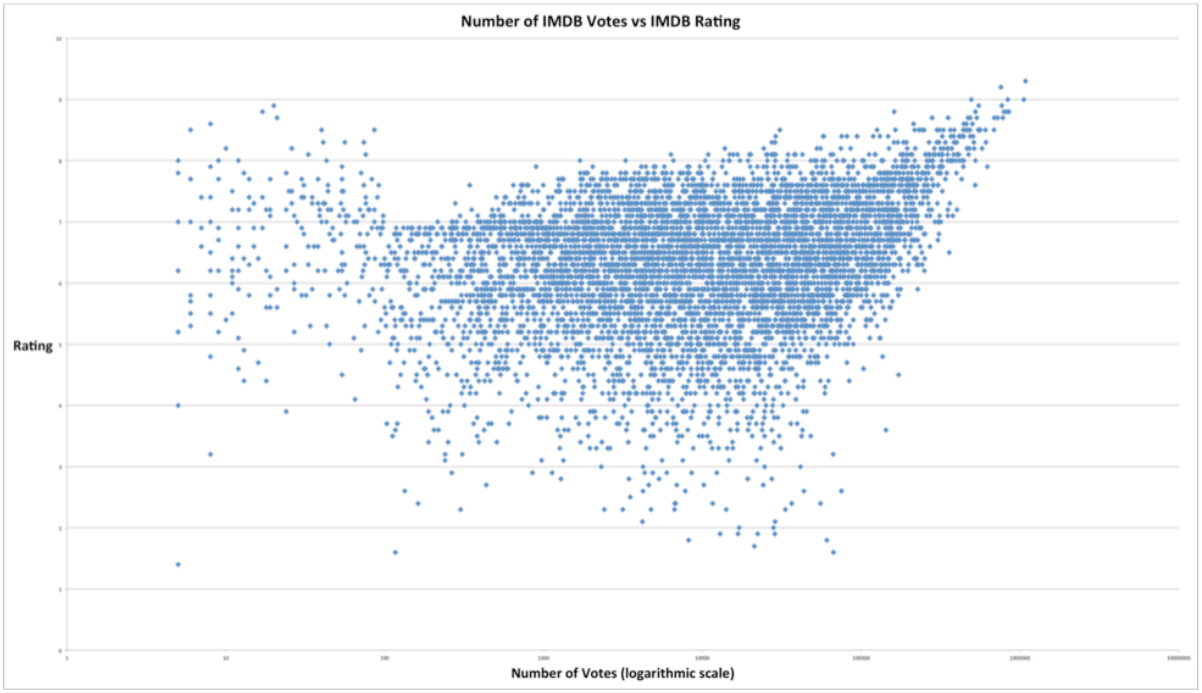
where Y_i is a binary variable, 0 indicating female leading film and 1 indicating male leading film; β_1 is a categorical variable indicating the production year of a film; β_2 is a categorical variable indicating the genre of a film; β_3 is a numerical variable indicating the weighted consumer rating; β_4 is a numerical variable indicating the standardized box office gross.

After dropping years with less than 100 films, the IMDb dataset collected in this research contain 364932 U.S. films from 1894 to 2016 with all information on title, top-billing actor, consumer rating, number of raters, box office gross, production year and genre complete.

**The dataset distribution (counting) by genre over time: stacked histogram*

The billing order of credits generally signifies their importance. The actors whose names appear first on credit are said to have "top billing", who usually play the principal characters in the film and have the most screen time (Wikipedia). Thus, a film with top-billing actor can be regarded as male-leading film and one with top-billing actress can be regarded as female-leading film. This can be identified with cast gender and billing order of each film in the IMDb dataset.

The IMDb rating is the weighted vote average on a scale of one to ten. IMDb has applied various filters which are not disclosed to the raw data in order to avoid ballot stuffing, and IMDb claims that the weighted vote average is a more accurate vote average. Each IMDb rating comes with both a numeric rating and the number of votes cast. Concerned that less well-known films with fewer votes would be unreliably rated, a close glance at the rating data can be observed as below.



A strong pattern shows that movies with more than 100 votes trend towards a higher rating with more votes. However, with fewer than 100 votes, there is little structure to the data. Based on this, movies with fewer than 100 votes are removed from the dataset.

**The dataset distribution by rating (mean, median, min, max, dev) over time (before/after drop): box plot*

The box office information is heavily affected by inflation and commercialization of motion picture industry thus are not comparable across years. To overcome this difficulty, the box office of the film with highest income in each specific year is normalized to 100, and the other box office grosses are computed accordingly.

**The dataset distribution by box office (mean, median, min, max, dev) over time (before/after standardization): box plot*