Leading Ladies and Lost Revenue: A Causal Analysis of Female Representation and Box-Office Returns

Lizzie Healy[1]

[1]Georgetown University,

Corresponding author: Lizzie Healy, emh201@georgetown.com

Abstract

This work will investigate the impact of gender bias in the film industry pertaining to economic outcomes. Specifically, it will establish a causal link between a film casting a female actress in the leading role and the resulting box-office revenue. This will be accomplished utilizing propensity weighting, which will match movies based on the perceived similarity of their characteristics. These predictor variables will include the year, season of release, genre, runtime, director and writers, star power level of the cast, MPAA rating, IMDb Metascore, IMDb Votes, number of awards won, country of release, language, film description, the production budget, the aspect ratio, the color, the countries of origin, filming locations, production companies, and tagline. To deal with the variables that are non-numeric the following steps will be taken. Firstly, a network analysis will be performed on the cast to produce a measure of centrality of the actors/actresses. Secondly, a sentiment analysis will be performed on the film description and tagline. The primary outcome variable will be the box-office number measured in US dollars, measured as the gross value worldwide. The IMDb score will be employed as an additional outcome measure to be used as a robustness check. A secondary robustness check may be employed in which the primary variable of interest will be whether the film passes the Bechdel test, indicating true female representation in the film. The initial hypothesis is that films that opt to feature a female in the leading role will experience a decrease value in the box office revenue.

Plain Language Summary

Earthquake data for the island of La Palma from the September 2021 eruption is found …

1 Introduction

Source: Article Notebook

1.1 Causality

My previous paper: Behind the Box Office: Directorial Influence on Film Revenue in the United States Entertainment Industry attempted to analyze the link between director quality and box-office success of a film. The paper created two novel measures of director quality; a summation of all box-office revenue earned by and the director's films and the accumulated number of critical awards from the fifteen years leading up to the film in question. The main dependent variable was domestic box-office revenue and a robustness check was implemented changing the dependent variable to the IMDb rating earned.

Table 11 Directorial Effect on Domestic Gross with Director Controls.

VARIABLES	(1) lnDomesticGross	(2) lnDomesticGross	(3) lnDomesticGross	(4) lnDomesticGross	(5) lnDomesticGross	(6) lnDomesticGross
ProductionBudget	2.16e-08***	2.10e-08***	1.87e-08***	2.29e-08***	2.18e-08***	1.94e-08***
	(1.70e-09)	(1.69e-09)	(1.67e-09)	(1.73e-09)	(1.71e-09)	(1.68e-09)
lnTotalGrossPre	0.0365***	0.0289***	0.0307***			
	(0.00758)	(0.00791)	(0.00861)			
tot_awards_dir				-0.00103	0.000459	0.0200
				(0.0228)	(0.0246)	(0.0266)
Film Controls	Yes	Yes	Yes	Yes	Yes	Yes
first_time_dir1		-0.470**	-0.590**		-0.626***	-0.696***
		(0.219)	(0.265)		(0.212)	(0.265)
castdir		0.311*	0.283		0.452**	0.424**
		(0.176)	(0.177)		(0.176)	(0.178)
preact		-0.463*	-0.612**		-0.427*	-0.617**
		(0.251)	(0.258)		(0.257)	(0.263)
dir_domestic			-0.230*			-0.180
			(0.127)			(0.125)
dir_female			-0.563**			-0.569**
			(0.265)			(0.265)
filmschool			-0.0643			-0.0639
			(0.105)			(0.106)
years_in_industry			-0.0150*			-0.0106
			(0.00839)			(0.00867)
dir_age			-0.000280			0.000932
			(0.00819)			(0.00821)
Constant	11.81***	11.97***	12.63***	11.79***	12.00***	12.65***
	(0.455)	(0.464)	(0.550)	(0.461)	(0.467)	(0.553)
Observations	1,484	1,484	1,349	1,484	1,484	1,349
R-squared	0.416	0.421	0.428	0.407	0.417	0.423

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 12 Directorial Effect on IMDb Rating with Director Controls.

VARIABLES	(1) imdbRating	(2) imdbRating	(3) imdbRating	(4) imdbRating	(5) imdbRating	(6) imdbRating
ProductionBudget	-3.75e-10	-4.52e-10	-4.38e-10	-5.72e-10	-6.18e-10	-5.84e-10
	(8.08e-10)	(8.08e-10)	(7.21e-10)	(8.00e-10)	(8.02e-10)	(7.23e-10)
lnTotalGrossPre	-0.00403	-0.00395	-0.00516			
	(0.00381)	(0.00404)	(0.00445)			
tot_awards_dir				0.0155	0.0183*	0.0270**
				(0.0105)	(0.0106)	(0.0123)
Film Controls	Yes	Yes	Yes	Yes	Yes	Yes
first_time_dir1		0.0286	0.250***		0.0544	0.247**
		(0.0876)	(0.0966)		(0.0844)	(0.0967)
castdir		0.0526	0.00789		0.0301	-0.0209
		(0.0876)	(0.0863)		(0.0856)	(0.0844)
preact		-0.103	-0.0738		-0.125	-0.0991
		(0.0968)	(0.0958)		(0.0976)	(0.0966)
dir_domestic			0.0579			0.0419
			(0.0563)			(0.0555)
dir_female			-0.0824			-0.0788
			(0.0904)			(0.0903)
filmschool			0.116**			0.113**
			(0.0513)			(0.0514)
years_in_industry			0.0103***			0.00735*
			(0.00393)			(0.00400)
dir_age			-0.0151***			-0.0153***
			(0.00389)			(0.00387)
Constant	4.026***	4.018***	4.608***	4.050***	4.036***	4.658***
	(0.222)	(0.225)	(0.260)	(0.222)	(0.225)	(0.262)
Observations	1,468	1,468	1,333	1,468	1,468	1,333
R-squared	0.333	0.333	0.347	0.333	0.334	0.348

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The paper found an increase in director financial quality yielded between a 0.0289% and 0.0307% increas in domestic gross and no impact on IMDb rating. Conversely, director quality in terms of critical acclaim yielded no significant impact on domestic gross, but betweeen 0.01803 and 0.0270 point increase in IMDb rating. The paper also discovered a statistically significant decrease in domestic gross for demale directors as compared to male directors.

Overall, the results were thought-provoking, however, the methodology used was lacking in the causality department. This paper, if anything, worked towards establishing a weak association due to it statical analysis going only so far as a simple

ordinary-least-squares regression and controlling for confounding variables. While, the variables were considered and included in the regression equation, they were all treated equally as controls, thus a more complex analysis is warranted.

Furthermore, I wanted to investigate the conclusion of gender bias further and shifted this analysis to examine actors instead of directors.

Moving forward, the work to get to causality includes introducing causality instead of just controlling for all covariates.

Thus, this paper will investigate the impact of gender bias in the film industry pertaining to economic outcomes. Specifically, it will attempt to establish a causal link between a film casting a female actress in the leading role and the resulting box-office revenue by employing propensity score matching.

Need to argue that there is sufficient common support between the treatment and control groups in a dataset in order to use propensity scores.

Data collection and preparation is discussed in Section 2.

Methodology and propensity scoring is discussed in Section 3

Results and anlysis are disucssed in Section 4

Conluding remarks, limitations, and future work are discussed in Section 5

2 Data

Data was collected from IMDb utilizing two separate APIs: OMDB and TMDB.

https://www.omdbapi.com/
https://www.themoviedb.org/

The datasets were merged on the movie title

Cleaning:

- dropped unimportant columns
- dropped missing values and zeros
- converted to correct data types
- one-hot encoded categorical variables
- Big Production Company Variable
- Top 25 Director
- Top 20 Writer
- English/Other Language
- Domestic/International

2,816 movies

61 columns

Variables:

- year
- runtime
- budget
- Month of release
- Day of Release**
- Genre (Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film Noir, History, Horror, Music, Mystery, Romance, Sci-Fi, Sport, Thriller, War, Western w/ Musical excluded)
- MPAA Rating (G, GP, M, M/PG, NC-17, Not-Rated, PG, PG-13, R, TV-MA, Unrated w/ Accepted excluded)
- Top Production Company

- Top Director
- Top Writer
- Domestic (w/ International excluded)
- English Language (w/ Other excluded)
- Sentiment of Description
- Sentiment of Tagline
- Starpower Metric

outcomes:

- box office
- imdb rating

Add in some descriptive statistics of the dataset.

Table 1: Male versus Female Director

Name	Year
Female Leads	488
Male Leads	2328

Table 1

3 Methodology

3.1 Sentiment Analysis of tagline and description

BERT model for sentiment analysis
created labels of 0/1 for tagline and movie description

3.2 Creation of the starpower variable

Collected lists of A-list and B-list actors/actresses
Added 2 points if one of the cast members was A-list
Added 1 point if one of the cast members was B-list
Averaged by dividing score by 3 (number of cast members)

actor1	actor2	actor3	starpower
Mark Wahlberg	Tyrese Gibson	André 3000	0.666667
Jamie Bell	Andy Serkis	Daniel Craig	1.333333
Ryan Reynolds	Blake Lively	Peter Sarsgaard	0.333333
Marc Singer	Tanya Roberts	Rip Torn	0.000000
Tom Hiddleston	Samuel L. Jackson	Brie Larson	1.000000
Jeremy Renner	Ed Helms	Jake Johnson	0.000000
Frankie Muniz	Amanda Bynes	Paul Giamatti	1.000000
Ben Barnes	Skandar Keynes	Georgie Henley	0.000000
Jason Bateman	Charlie Day	Jason Sudeikis	1.000000
Jack Black	Ana de la Reguera	Héctor Jiménez	0.333333

3.3 Creation of the variable that indicates female in leading role

Collected list of all female actresses
Took the first cast member named
Merged and created female_lead variable
1 = female lead

Title	First Actor	Female Lead
The Family	Robert De Niro	0
The Shack	Sam Worthington	0
The Dead Zone	Christopher Walken	0
The Ref	Denis Leary	0
Flyboys	James Franco	0
ATL	Tip 'T.I.' Harris	0
Like a Boss	Tiffany Haddish	1
Enemy Mine	Dennis Quaid	0
Proud Mary	Taraji P. Henson	1
Valmont	Colin Firth	0

3.4 Propensity score matching

StandardScaler Propensity scores calculated

3.5 Robustness Checks

Implemented the IMDb score as the outcome variable as a robustness check
Does a female in the lead role impact the critical success of the film?

4 Results

4.1 Box Office Results

Female-lead had higher box-office average

4.2 IMDb Rating Results

Female-lead had lower IMDb scores

Gender of Lead Role	Box Office	IMDb Rating
Female	62,708,871.36	6.329
Male	61,076,707.45	6.601

5 Conclusion

5.1 Limitations:

1. More data
2. Network Analysis
3. Robustness Checks with Bechdel test
4. Inacurracy from matching female leads & first actor not always the lead
5. Unbalanced Dataset (is this an issue?)

References

Source: Article Notebook