The Effect of Movie Lead’s Gender on Movie’s Revenue

University of Southern California

Fall 2022

Introduction

In the past 5 years, there is a trend in the film industry that more films are produced with a single female character as a leading role. This study intends to identify a relationship between the gender of the leads and box office revenue. The hypothesis stands that the gender of movie leads has no impact on movie revenue when we control all the other factors. The ideal experiment is to create a series of two movies with the same qualities (genres, scripts, directors, etc.) but different gender of the leading role to compare their revenues.

Historically, male-lead movies have been generating more revenue. Without controlling all other factors though, the historical number is not decisive. We need to use data and modeling to control other objective factors and examine the possible relationship between revenue discrepancy and gender of the leading role.

Data

The project includes two primary datasets: one includes the movie names, revenue and other related factors[1]; the other one includes the gender of the leading roles.[2]

After cleaning the name strings and dropping duplicates, the two datasets were combined by the name of the lead actor/actress with a one-to-one relationship. The final dataset includes 3891 movies with **Gender** as the main independent variable and **Gross revenue** as the outcome variable.

[1] www.kaggle.com/datasets/danielgrijalvas/movies

[2] www.github.com/taubergm/HollywoodGenderData/blob/master/all_actors_movies_gender_gold.csv

Model

Model1: ln(gross) ~ ln(budget) + year + score + rating + genre

The first model excludes gender effects. The coefficients of dependent variables are all significant at 5% level. It indicates that all controlled variables have impacts on revenue.

Model2: ln(gross) ~ ***gender*** *+ ln(budget) + year + score + rating + genre*

Gender is added to the second model, with "1" being Male and "0" being Female. The coefficient of gender (male) is -0.1230 with a p-value of 0.001(statistically significant), which indicates a result that when controlling for other variables, male-led movies produce 11.57% less revenue compared to female-lead movies on average. Surprisingly, the result is against the common perception that male-led movies are more profitable than female-lead movies (excluding impacts from other variables). Compared to the first model, Adjusted-R increases slightly from 0.440 to 0.441, indicating gender cannot predict the revenue well.

Model3: ln(gross) ~ gender + ***gender * ln(budget)*** *+ ln(budget) + year + score + rating + genre*

We are curious about how other factors affect revenue based on the value of Gender. In the third model, an interaction term, **gender*ln (budget)** is added to investigate if the change in revenue with respect to budget depends on the value of gender. Per the regression results, the coefficient of the interaction term gender*ln(budget) is 0.1117. It indicates that, when other variables are controlled and regardless of the effect of gender on revenue, with a 1% increase in budget, male-lead movies will produce 0.1117% extra revenue on average.

However, the coefficient of gender(male) is negative (-0.5145). To better understand the model, we generated a graph, which demonstrates the relationship between budget and gross revenue: as the budget increases, the expected revenue of male-lead movies increases higher than female-lead movies, which explains why the coefficient of the interaction term is positive. To be exact, male-lead movies will generate more revenue than female-lead movies when the budget is $99.97 million or higher, whereas female-lead movies tend to generate more revenue when the budget is below $99.97 million.

We further examine our sample and find that $99.97 million represents 86th percentile of our sample. Thus, we further conclude that male-lead blockbuster movies (high budget movies) tend to achieve higher revenue, such as the Avengers, Spiderman, etc. However, among movies with extremely high budgets, only few of them are female-lead movies. Lack of sample might be the reason why female-lead movies are expected to produce less revenue than male-lead with high budgets. Therefore, we have some reservation in our initial conclusion and are aware that the lack of sample in high budget female-lead movies is one of the limitations of our model.

Model4: ln(gross) ~ gender + ***gender * score*** *+ ln(budget) + year + score + rating + genre*

We consider gender may affect revenue through the quality of movies. In other words, we think scores might have different impacts on revenue based on the value of gender. Therefore, we add the interaction term **gender*score** into our fourth model. The coefficient of gender(male) equals -1.2699, indicating a 71.91% decrease in revenue for male-lead movies than female-lead movies on average, controlling for other variables and when budget equals to 1. The coefficient of interaction term gender*score equals 0.1813, which indicates that male-lead movies generate 19.88% higher revenue on average with 1 point increase in score, regardless of the effect of gender on revenue.

Model4 has similar results as Model3. When scores are lower than 7, female-lead movies are expected to generate higher revenue than male-lead movies. Oppositely, if scores are higher than 7, male-lead movies tend to achieve higher revenue. Per calculation, male-lead and female-lead movies will generate the same revenue when scores are 7, which represents 72th percentile of our sample. Therefore, we conclude that well-made movies whose leading characters are male tend to produce more revenue.

We further examine our sample and have some insights from it. By constructing the distribution of scores, we find that male-lead movies tend to be rated higher than female-lead movies. According to the T-test based on scores of female-lead and male-lead movies, the result (p-value = 0.00) shows that male-lead movies have higher rated scores than female-lead movies on average. Therefore, we suspect that people tend to give male-lead movies a higher score than female-lead movies.

Model5: ln(gross) ~ gender + ***gender * year*** *+ ln(budget) + year + score + rating + genre*

Model6: ln(gross) ~ gender + ***gender * genre*** *+ ln(budget) + year + score + rating + genre*

At last, we add two other interaction terms to see whether audience preference towards gender changes by year or genre. However, the p-values of **gender*year** and **gender*genre** are statistically insignificant (>0.05), which implies that audience preferences towards gender do not change based on the year or genre.

Limitation

Due to the limitation of our dataset, our model doesn't include factors like cast list, director, producer, composer, cinematographer, and distributor, which can also have an impact on movie revenue. Score and budget can be a proxy for other factors. For example, a movie with a higher budget can have a better chance of attracting a famous director and a star actor/actress. However, it's very difficult to measure the effect independently.

We also notice that our dataset doesn't include a meaningful number of samples of blockbuster movies with female-lead. For example, male-lead blockbuster movies include Titanic, Avengers, Pirate of the Caribbean and so on, but we only have a few female-lead blockbuster movies like Star Wars 7-9.

Besides, it's difficult for us to take into account social media factors, such as YouTube, Instagram, TikTok, etc., which can also affect the popularity of movies and thus the box office.

Conclusion

In conclusion, we believe that gender does affect the revenue with significant p-value. Controlling all the other variables, we can conclude that movies with male leads seem to generate less revenue than movies with female leads. However, after we add the interaction term gender*budget, we notice that males will generate more revenue than females when the budget is $99.97 million or higher, and when the budget is lower, female-lead movies outperform male-lead movies. According to our model, the slope of the interaction term gender*score is also significant. In consideration of the above two interaction terms, we can draw a conclusion that male-led movies tend to generate more revenue with a higher budget or a higher score.