The Effect of Lead Actor's Gender and Movie Title on Gross Earnings

Group 4 - Ronald Lee, Lawrence Jiang, Adam Weintraut, Victor Ramirez

April 4, 2022

Contents

1. Introduction **1**
1.1 Motivation . . . 1
1.2 Research Question . . . 2

2. Description of the Data and Research Design **2**
2a. Data Characteristics . . . 2

3. Modeling **6**
3.1 Base Model: Female indicator and Gross Earnings . . . 6
3.2 Second Model: Adding Runtime . . . 6
3.3 Third Model: Adding Title Uniqueness . . . 6

4. Results **7**

5. Limitations of your Model **8**
5.1 Large-Sample Assumptions . . . 8
5.1.1 Independent and Identically Distributed (I.I.D.) . . . 8
5.1.2 Unique BLP Exists . . . 8
5.2 Omitted Variables . . . 9
5.2.1 Intentional Omitted Variables . . . 9
5.2.2 Unintentional Omitted Variables . . . 9

6. Conclusion **10**

1. Introduction

1.1 Motivation

Movies have been in existence since 1888. Since then, watching movies has been one of our favorite pastimes. Movies are a form of blending visual and sound communication to tell a vivid story.

The movie entertainment industry is a high profile multi-billion-dollar global industry. The movie industry has exponentially grown over the past decades. With this growth there has come a windfall of revenue generating profits. As technology evolves and enhances the viewing experience, you can watch movies from various devices. People can enjoy movies in the comfort of their homes or while traveling.

The ability to predict movie revenue can be a very insightful opportunity. With the predicted revenue information, movie makers can intelligently plan their movie budget. Movie budget line items including star salaries, production, and distribution costs can be smartly negotiated and set.

1.2 Research Question

The success or failure of a movie depends on a variety of different factors: star cast, cast gender, budget, and title. With the breadth of data available today, making accurate revenue predictions is extremely difficult. However, we do have many data science tools and methodologies at our disposal to help in attempting to make movie revenue predictions.

Our research question is:

How the Lead actor's gender affect a movie's gross revenue?

In this study we will explore the relationship between the following features, cast gender, budget, runtime, and title. First, to harness the power of the regression testing harness we engineered three linear regression models to predict the movie revenue. Second, we engineer the model using various data features. We iterated over different features including the following extracted features, cast gender, budget, runtime, and title. Third, we collected various types of datasets from different freely open sources. We then joined and sanitized the data for inspections and prediction modeling.

2. Description of the Data and Research Design

We use two datasets for our analysis:

IMDb dataset (url): For this dataset, the data that we are interested in are the movie title, budget, and gross. Budget and Gross are in US dollar.

Movie roles by gender (url): This dataset has the gender info for movie actors. The source has instructions on how to execute R code to parse the original data file, which has data in JSON format, and convert it into CSV format that can be imported to R studio. The values of gender is either a value of 2 for "Male", or 1 for "Female".

With this two datasets, we then perform an inner-join on them by their movie lead actor name to generate a single dataset that has all the movie info plus lead actor's gender info.

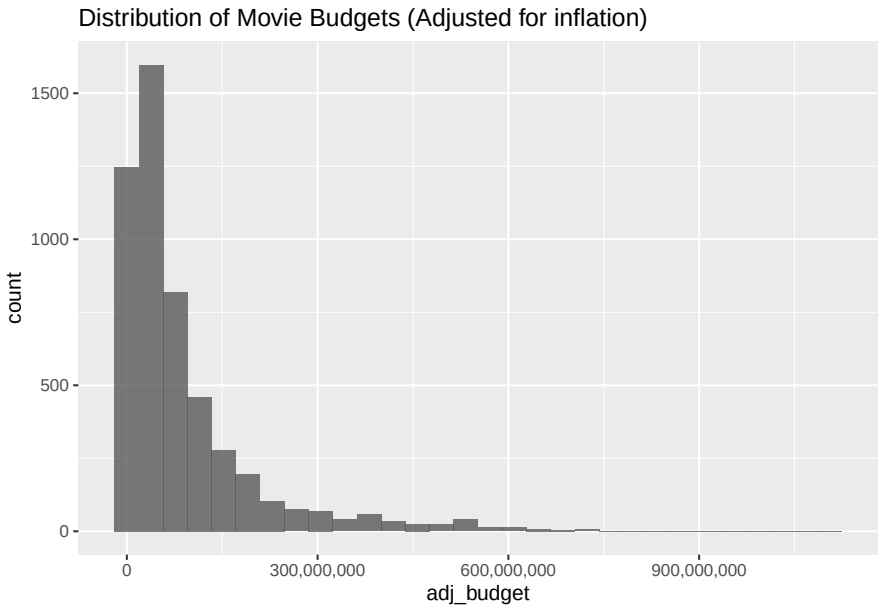
Before we perform our analysis, we also clean up the data by removing all the entries that have empty values for the fields that we are interested (budget, gross, runtime). The final dataset has over 5000 rows of movie data.

In addition, we also generate a "title_unique_score" using a simple logic. We take all the words from all movies and count the number of occurrences for each word. Then for each title we sum the counts. This means that the lower the score, the more unique the title. The score might also be affected by the title length itself, since a short title will also tends to get a low score since it has less words.

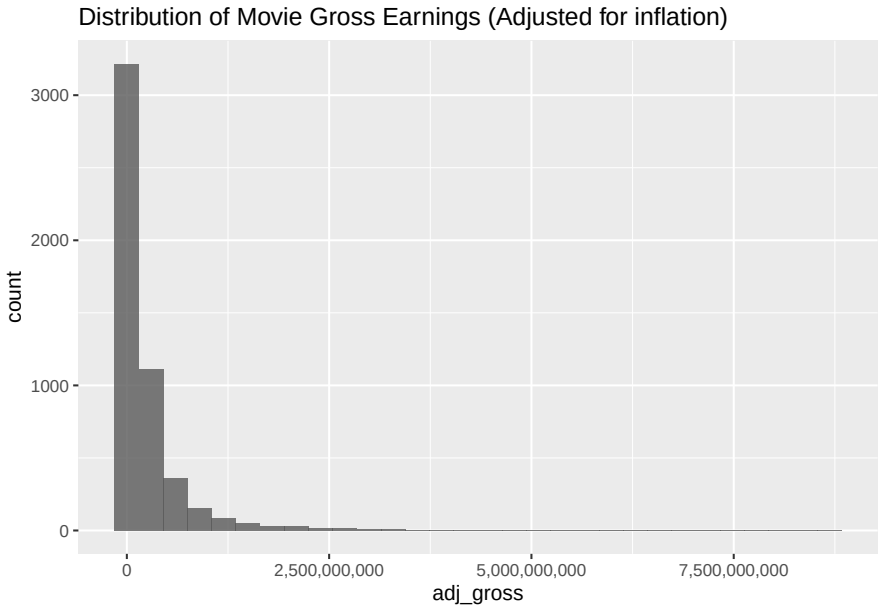
Finally, we uses a R package called 'quantmod' to retrieve CPI data for inflation adjustment for our Gross and Budget data. Then we perform adjustment to the Gross and Budget data in our main dataset. We use 1980 as the base year for our inflation adjustment, then perform inner-join with the main dataset by the years. Additional calculations to adjust Gross and Budget is performed afterwards.

2a. Data Characteristics

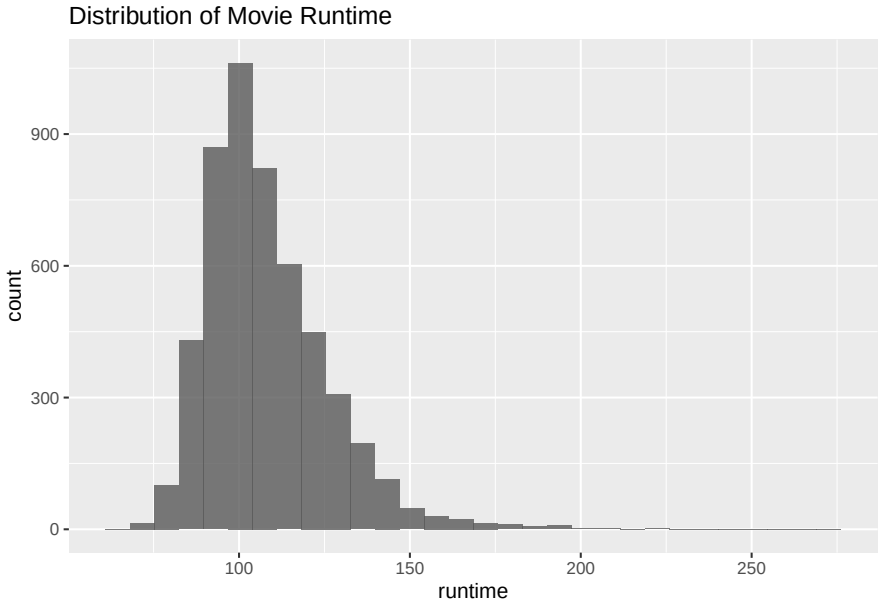
In this section we take a look at the data characteristics for our variables that we are interested and see if there is any need to perform cleanup or transformation before we build our models.



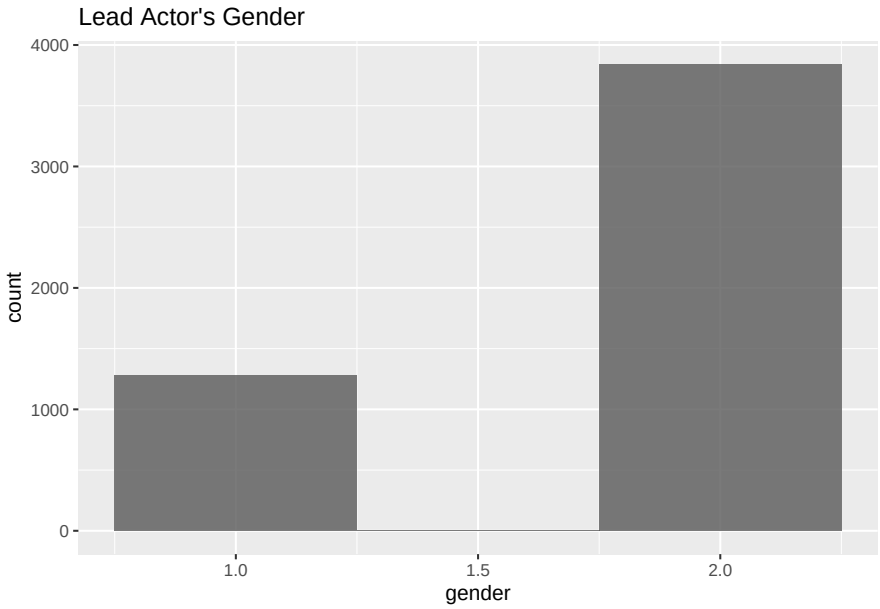
Movie Budget is highly skewed to the left which means most movies have relatively low budget, and there are only a few movies that have really high budget.



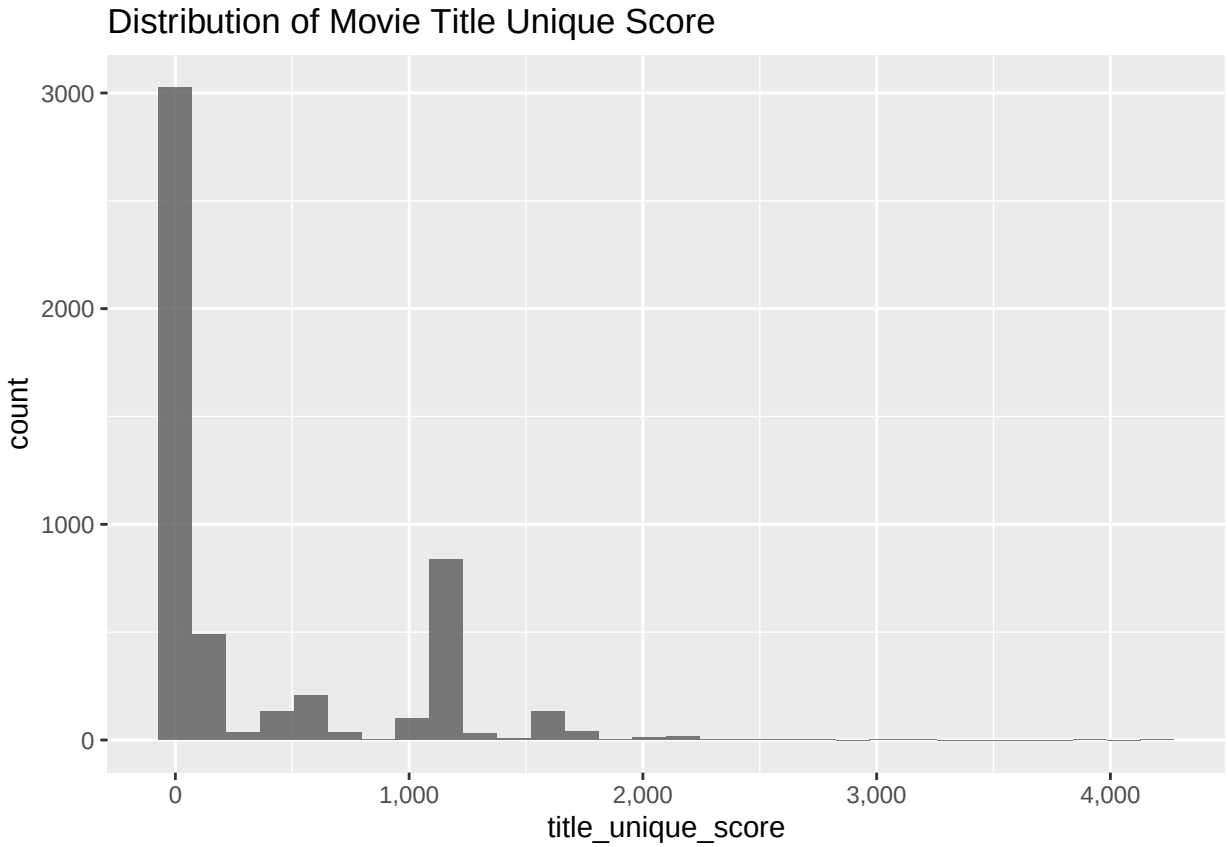
Movie Gross Earnings has similar characteristics as the budget, also highly skewed to the low end.



Most movies have runtime under 150 minutes.



In this histogram that shows the number of Male versus Female Lead Actor. The value "2.0" represent 'Male', and you can see that there are much more Male lead actors in our movie dataset.



The title unique score is skewed to the low end, which means most titles are pretty unique, since the titles are composed of words that has low occurrences.

3. Modeling

3.1 Base Model: Female indicator and Gross Earnings

In order to answer our research question, we build our first model with 'Gross earnings' as our outcome variable and 'Female Indicator' as our feature variable. A female indicator is a binary flag which has the value of 0 or 1. When it is '1', it means that the lead actor of that movie is a female. We generate this indicator by processing the 'gender' data and add a new column in the dataset to store the resulting indicator. We also include the 'Budget' variable since it has strong correlation with Gross.
Moreover, since the Gross earning is highly skewed, we also apply Log transformation to it in our first model. So the model becomes this:

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation)

3.2 Second Model: Adding Runtime

For the second model we are trying to improve the fitness of the model by introducing a feature variable that has a strong correlation with our outcome varable which is gross earnings. From our EDA we generated a correlation heatmap to show the correlations between variables. We identify that the 'budget' variable has the strongest correlation with gross earnings. This also makes sense in reality, since a bigger budge movies are expected to have higher gross earnings in order to generate profits for the production companies. Of course big budget doesn't guarantee a big payout. Success of a movie still depends on other factors, like the story itself.

The Budget data is also highly skewed, so we apply Log transformation to it in our second model as well:

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation) + Runtime

3.3 Third Model: Adding Title Uniqueness

We are also interested to find out if the movie title has effects on gross earning. First, we build a simple model that uses only 'title_unique_score' as the feature variable against gross earnings.

So we will introduce this new feature into our 3rd model:

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation) + Runtime + Log(Title_Unqiue_Score)

4. Results

Table 1:

	Dependent variable:		
	log(adj_gross)		
	(1)	(2)	(3)
female_indicator	0.124***	0.129***	0.131***
	(0.047)	(0.047)	(0.047)
log(adj_budget)	1.026***	1.014***	1.012***
	(0.020)	(0.020)	(0.020)
runtime		0.003**	0.003**
		(0.001)	(0.001)
log(title_unique_score)			0.026***
			(0.007)
Constant	−0.064	−0.155	−0.199
	(0.360)	(0.362)	(0.362)
Observations	5,118	5,118	5,118
R^2	0.493	0.493	0.495
Adjusted R^2	0.493	0.493	0.494
Residual Std. Error	1.404 (df = 5115)	1.404 (df = 5114)	1.402 (df = 5113)
F Statistic	2,485.789*** (df = 2; 5115)	1,660.426*** (df = 3; 5114)	1,250.942*** (df = 4; 5113)
Note:			*p<0.1; **p<0.05; ***p<0.01

In the first model, we see that the p-value of the "Female_indictor" is less than 0.01 which means there is significant evidence that female gender has effects on "Log(Gross)". The coefficient is a positive value, meaning that a movie with a female lead actor actually brings more gross earnings than a movie with a male lead actor. The coefficient value of -0.124 means that it will be 12% more on gross if a movie uses a female lead actor. The adjusted R-squared is 0.493, mainly because the variable 'Budget' we added to this model can explain almost half of the model.

After introducing the runtime into our second model, we don't see any improvement on R-square. Coefficients and P-values of other variables also don't seem to be affected by this new variable.

For our third model, it seems that the "title uniqueness" feature doesn't really make much difference to the model. Coefficients of the other feature variables don't change much, so as the R-square value. The title uniqueness score has a very low p-value less that 0.01. So it seems that title uniqueness does affect the gross earnings. The coefficient of the 'Log(title_unique_score)' is positive, meaning a high score will help increase gross earnings. A high title uniqueness score means that titles are less unique actually. This translates to that a movie with a more common title tends to bring in more gross earnings.

Based on these statistics, we see that using a female lead actor in a movie does help bringing more gross earnings in a significant way. The effect on gross earnings is about 13%, which is substantial.

5. Limitations of your Model

5.1 Large-Sample Assumptions

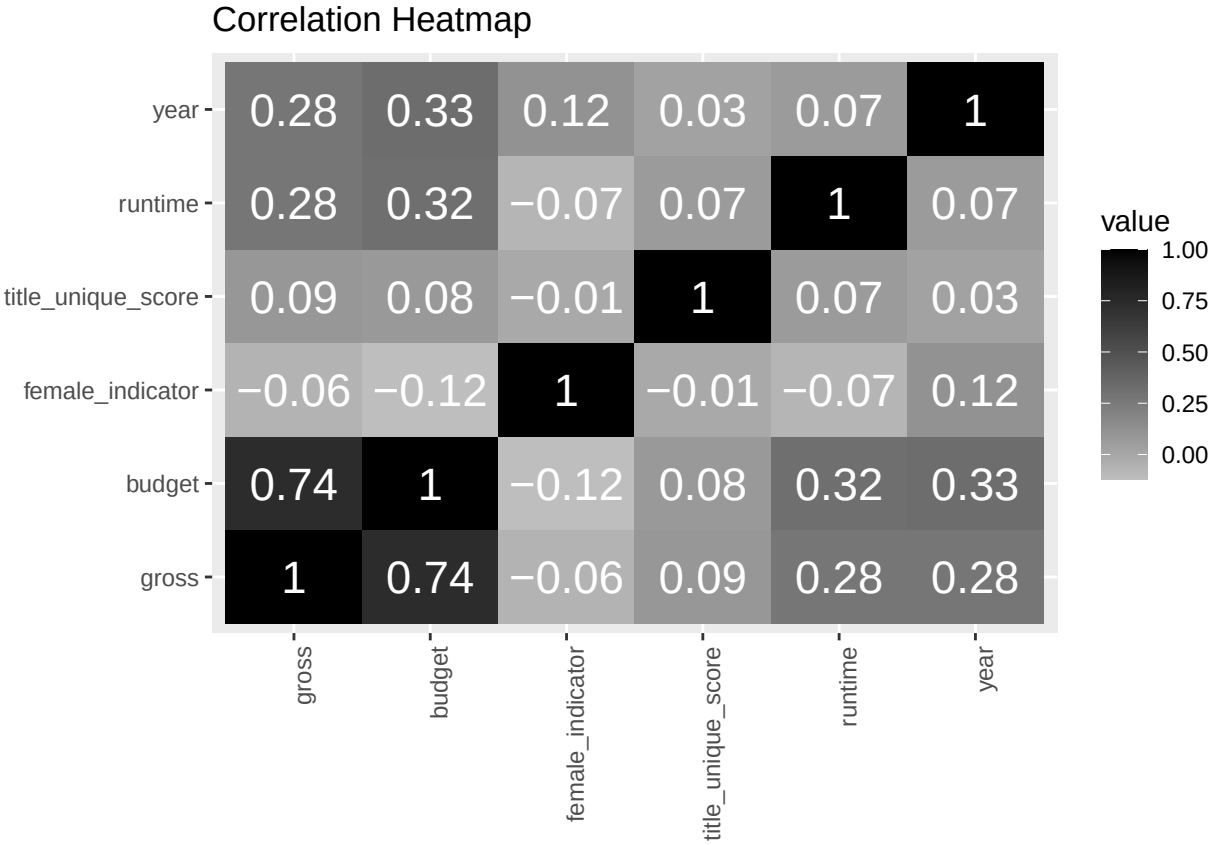
Since our analysis is based on a dataset with more than 5000 entries, we can use the large-sample assumptions.

5.1.1 Independent and Identically Distributed (I.I.D.)

Our dataset doesn't satisfy the IID assumption. Our dataset has movies between year 1980 and 2020 scraped from IMDb. For movies that are actually in the same series, like 'Star Wars' or "Mission Impossible", their data will not be independent. Our analysis is based on robust standard errors which can adjust for dependencies issue in our dataset.

5.1.2 Unique BLP Exists

We want to make sure that no variables are perfectly collinear with any one of the other variables. From the correlation heatmap below, we can see that perfect collinearity only happens at the diagonal of the matrix which has all 1's and no where else. Also, while building our model, R didn't drop any variables considered to cause perfect collinearity with any other variables.



5.2 Omitted Variables

5.2.1 Intentional Omitted Variables

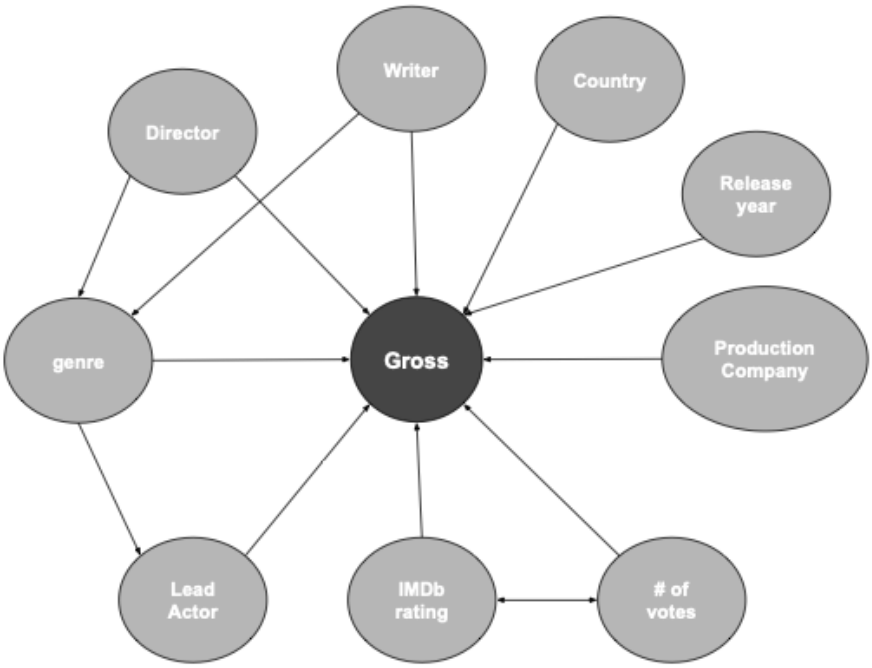


Figure 1: Intentional Omitted Variables

The variables that are available in our dataset but are omitted in building our models are:

- Genre
- Director
- Writer
- Production Company
- Lead Actor
- Country
- Release year
- IMDb rating
- Number of votes in IMDb

Our EDA shows that these variables have low correlation with Gross earnings, so we simply omit them in our models. For "Release year", since we already adjusted our Gross and Budgets with inflation, so we didn't include year as a variable that could affect the model outcomes.

5.2.2 Unintentional Omitted Variables

There are other variables not available in our dataset that might have causal effects on a movie's gross earnings.

- Film distribution platforms
- In theater duration
- Story of the movie
- Acting skills of the cast

6. Conclusion

In conclusion, our analysis shows strong evidence that a female lead actor brings more gross revenue to a movie than a male lead actor. In this case, it is about 13% more on average. The "Title Unique Score" also seems to have effects on the Gross, but not as much. Our model also performs Log transformation on the title unique score, the percentage change on Gross will be about +2.6%. The percentage change will be much lower if we take out the log from title unique score (~0.01%).