Estimating Movie Box Office Revenue

Datasci 203: Lab 2 Report

Erik Sambrailo, Jenna Sparks, Bailey Kuehl, and Roshni Tajnekar

04-18-2023

Introduction

Understanding the key factors that contribute to the success of a movie is critical for movie producers, investors, and distributors in order to maximize their impact on the movie industry. Revenue is a key indicator of movie success for such stakeholders, warranting a thorough investigation into which variables contribute significantly.

Our company, Data Science Consulting™, believes that explanatory models are a powerful tool we could use to gain insight into such effects. For our expertise with explanatory models, we have been contracted by Big Fat Movie Productions™ (BFM) to create a revenue prediction model and help them understand key factors influencing movie revenue. By examining variables such as production budget, Internet Movie Database (IMDb) votes, country of origin, runtime, IMDb rating, and age certification, we hope to advise BFM on important factors for increasing revenue. We will utilize the classic linear model (CLM) to answer the following question:

How does the spending (budget) for a movie affect the revenue it earns?

The relationship between the predictor variable and the outcome outlined in the research question is displayed in our initial casual path diagram (Figure 1). We predict that increasing the budget will increase the revenue it earns.

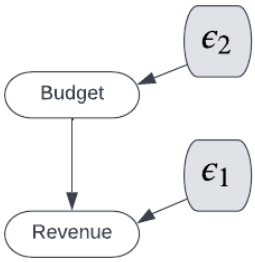


Figure 1: Causal Diagram

Data and Methodology

The observational dataset that we found is from the popular film database site (IMDb)[1].
The dataset consists of the most popular films for a given year and includes information about budget, revenue, age certifications, cast/crew, IMDb votes, runtime, and other movie attributes. We chose to narrow our research to films that were released in the years 2016-2019. These are the most recent years available in the dataset that were not heavily affected by the COVID-19 pandemic. This narrowed timeframe also reduces time-based IID concerns.

We randomly split our base dataset into training (30% of original data) and confirmation (70% of our original data) datasets. This data splitting was used to reduce the potential overfitting of our models in the training stage, thus promoting more reliable coefficients in our final regressions.

Operationalized Variables

The below table shows the variables we chose to represent our concepts:

Table 1: Operationalized Variables

Concept	Operation
The financial success of a movie. (Y)	IMDb's Reported Movie Gross Revenue Value ($)
The cost of making a movie. (X)	IMDb's Movie Budget Estimate ($)
Movie inclusivity by age. (Covariate)	Movie Age Certification
runtime, country of origin, IMDb votes (Covariates)	As-is

For all of our models our outcome variable (Y) is movie gross revenue. Gross revenue may give an inflated perception of financial success without a proper understanding of the associated taxes and fees to be deducted. On the contrary, these gross revenue numbers, reported by IMDB, only represent theater box office sales, and only for the US and Canada, and therefore could lead to an under-representation of true movie revenue.

For our primary predictor variable (X) we are using IMDb's budget values. As indicated in an article from Forbes[2], there is some discretion as to what is included in a defined movie budget. IMDb also discloses similar ambiguity in the budget information collected, and they identify their reported numbers as estimates.

Key Modeling Decisions

Through exploratory data analysis, we investigated several variables in our dataset, including runtime, the total number of IMDb votes (regardless of score), country of origin, and recommended age advisory rating for the movie (age certification). Some variables, such as IMDb score were removed because they could also be considered outcome variables. For others, like movie director and writer, we aimed to create folded categorical variables but were unsuccessful finding the proper data to do so. There were also concerns around the causal relationships between these variables and budget. Lastly, we explored the effect of genre in our models, but found it insignificant and concerning regarding IID due to the subjective nature of specific genres (i.e. romantic comedy). This left us with six columns for both the training and confirmation datasets. This workflow and corresponding datapoint count updates are shown below in Table 2.

Table 2: Accounting Table - Sample Removal

Reason for Removal	Explanation	# Samples Removed	# Samples Remaining
Start	original dataset	0	7668
Year Consolidation	reduce years to 2016-2020	6868	800
Null Values	nulls in budget, revenue, age certification	253	547
Not Rated	records w/o age certification	5	542
Folding	country (us/other) & age cert (G w/ PG)	0	542
Training Data	30%		163
Confirmation Data	70%		379

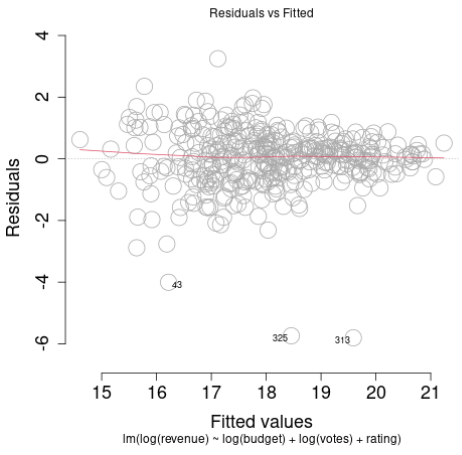
Initial analysis found that the revenue, budget, and votes variables were positively skewed. To optimize the performance of our models, we performed a natural logarithm transformation of these three variables. We executed these transformations to address skewness of the variables, reduce the heteroscedasticity of the variance in the data, and achieve better linear relationships between the outcome and treatment variables.

Through the exploratory stage of our analysis, we found that our second model demonstrated residuals that were fairly equally distributed around zero. Figure 2 on the right shows a residual plot performed on the 70% confirmation set of our BLP model. This model also had a relatively high adjusted R^2 value and statistically significant coefficients. Thus, we chose to model revenue in the following regression:

$$\widehat{ln(\text{revenue})} = \beta_0 + \beta_1 \cdot ln(\text{budget}) + \beta_2 \cdot ln(\text{votes}) + \beta_3 \cdot \text{age certification} + \epsilon$$

where ϵ represents the error term that is not captured by the regression model. With our main predictor variable as budget, we evaluate the following hypotheses. Our null hypothesis is that budget has no effect on how much revenue a movie earns ($H_0 : \beta_1 = 0$) and our alternate hypothesis is that budget does have an effect on how much revenue a movie earns ($H_A : \beta_1 \neq 0$)

Figure 2: Residuals Plot



Results

The stargazer table (Table 3) shows the estimated coefficients, standard errors, and p-values for the relationships among our test variables.

Based on the data, all three models demonstrate a statistically significant coefficient ($p<0.001$) for the natural logarithmic transformation of budget. Therefore, we will reject our null hypothesis that budget has no effect on how much a movie earns. Point estimates for this key variable are all positive and range from 0.433 to 0.892 among the three models. In practice, this signifies that increasing the budget allocated to a movie is predicted to increase the amount of revenue it draws in. In our second model (2), for example, the

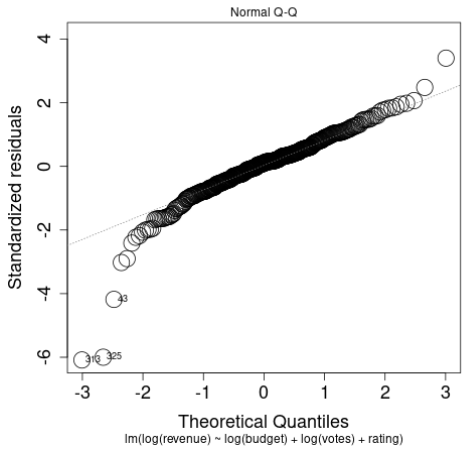
Table 3: Regression Results for Movie Revenue

	Dependent variable:		
	ln(revenue)		
	(1)	(2)	(3)
ln(budget)	0.892^{***} (0.056)	0.433^{***} (0.055)	0.526^{***} (0.055)
ln(votes)		0.778^{***} (0.058)	0.812^{***} (0.061)
rating: PG-13		-0.758^{***} (0.122)	-0.589^{***} (0.144)
rating: R		-1.312^{***} (0.157)	-1.115^{***} (0.156)
runtime			-0.016^{***} (0.006)
country: United States			0.220^{*} (0.130)
Constant	2.759^{***} (0.981)	2.809^{***} (0.874)	2.341^{***} (0.805)
Observations	379	379	379
R^2	0.431	0.655	0.680
Adjusted R^2	0.430	0.651	0.675
Residual Std. Error	1.230 (df = 377)	0.962 (df = 374)	0.928 (df = 372)
Note:			*p<0.1; **p<0.05; ***p<0.01

coefficient of 0.433 indicates that a 10% increase in budget is expected to lead to approximately 4.2% increase in revenue.[3] To put this into perspective, the budget for "Jumanji: Welcome to the Jungle" was approximately \$90 million, with the revenue being approximately \$962 million. If Jumanji's production company, Columbia Pictures, had increased the budget to \$99 million, we would expect their revenue to have increased to \$1 billion. While this method could be quite effective for increasing revenue, it is important to consider the size and scope of the production company and whether they could accommodate such an increase in upfront spending. Because our relationship is represented as the elasticity of the effect of budget on revenue, it does not provide enough information to make practical significance claims. For instance, for a movie like Jumanji, where the revenue is 10x the budget, the predicted return on investment on additional budget spend may have been beneficial. On the contrary, if a movie's revenue is only 2x its budget, then in hindsight, the increased budget would not have been beneficial.

Surprisingly, we also found that the age certification of a movie can have a statistically significant impact on its revenue. Movies targeted more towards older audiences (PG-13 & R) are predicted to have 53.14% and 73.07% less revenue, respectively, than those that are inclusive of younger audiences (G & PG). Less surprisingly, we note that models (2) and (3), which evaluated the effect of IMDb votes, had a positive impact on revenue. This was the expected effect, as a large number of votes indicates that many people have watched the movie and contributed to the gross revenue. However, this correlation does not offer much practically, as our model does not take into account what prompts consumers to vote, or whether votes/ticket purchases were motivated by advertisement spent within our budget.

Figure 3: QQ Norm Plot



Limitations

In order to gain meaningful information from the CLM approach, we must validate the five assumptions required for the CLM against our best linear model (model 2). One of these assumptions is independently and identically distributed (I.I.D) data points. The dataset we have chosen was curated by querying IMDb's movie database. However, the movies selected were sorted based on ascending "popularity" according to IMDb, which is a potential concern for I.I.D observations, which means there were likely existing trends in the data, violating the independence assumption.

Normal distribution of errors is another requirement for the CLM in order for the Ordinary Least Squares (OLS) to produce consistent, accurate coefficients. Numerically, we found a high kurtosis value (7.1) and visually, we observed heavy tails in the Normal Quantile-Quantile plot (Figure 3). Thus, normality is a potential concern with our proposed models.

We found no major violations to perfect collinearity among variables, demonstrated by variance inflation factors (VIF) of approximately 1.2. Similarly, we found no significant problems with the assumptions of homoskedastic errors or linear conditional expectation.

Structurally, our models are limited by the variables available in the dataset, leaving room for omitted variable bias in our estimates. One variable that was not included in the dataset but which would have ideally been incorporated into our model is "star power" (i.e. a measure of how many "star" actors are in the film). While we were able to join our dataset with a secondary dataset[4] to derive "star power," the variable was limited in scope and difficult to interpret, so it was not considered in our models. We expect that star power would be positively correlated with our primary treatment variable (budget), given that larger budgets can afford higher-profile actors. Furthermore, we expect star actors to draw in a larger audience, and thus that star power would have a positive correlation with the revenue of a movie. Thus, we would predict the effect of this omitted variable bias to be away from zero, or, an overestimation of our coefficients. We would expect similar behavior if we had included a variable for the current status of the economy, such as the Gross Domestic Product (GDP).

Figure 4: Causal Diagram with Omitted Variables

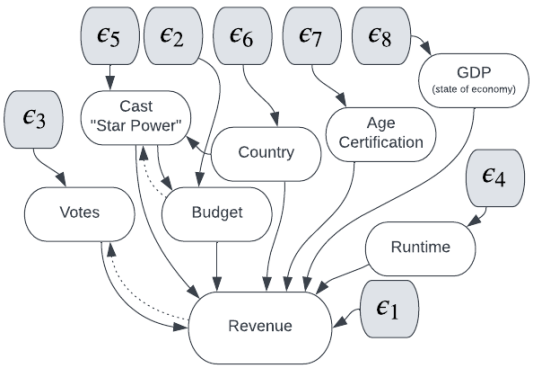


Figure 4, shows an updated causal path diagram that demonstrates conceptual relationships between each of the predictor variables and the outcome. Dotted lines indicate reverse causality, which occurs between the votes and revenue. Higher revenue also suggests increased popularity leading to more viewing and votes which indicates a positive correlation between revenue and votes.

Conclusion

In our study, we explored the use of explanatory linear regression models to evaluate the effect of certain movie production components on revenue produced. Our highest confidence model predicts that a 10% increase in budget will result in a 4.2% increase in revenue, all other variables are kept constant. Secondarily, our model predicts that a 10% increase in the number of IMDb votes signifies a 7.7% increase in revenue, and a PG-13 or R rating will see 53.14% and 73.07% decreases in revenue respectively, compared to their younger-audience inclusive counterparts (G & PG). However, we do not recommend that BFM utilize our findings regarding the relationship between IMDb votes and revenue, as we do not have enough information about this relationship, and believe that reverse causality inflated our coefficients.

From a product perspective, we feel that the increase in budget required to make a substantial return on investment in revenue is likely not worth the cost. Given this, along with the limited size of our dataset and potential statistical concerns, we would not recommend that Big Fat Movie Productions (BFM) increase their budget to increase revenue until further research is conducted. We would, however, recommend that BFM consider producing movies geared toward younger audiences.

Based on our analysis, we believe a strong direction for future research would be to study what factors lead to an increase in the number of IMDb votes, as this seemed to positively impact revenue. Additionally, future research could explore the effect that different marketing tactics (social media, TV commercials, etc.) have on a movie's revenue output. We recommend that production companies interested in further analysis utilize explanatory models and pay close attention to the five main assumptions of the CLM to optimize the predictability of the results.

References

1. GRIJALVA, DANIEL. (2021). Movie Industry. Kaggle. https://www.kaggle.com/datasets/danielgrijalvas/movies
2. Moore, S. (2021, December 10). Why film budgets are important, beyond the cost of production. Forbes. Retrieved April 13, 2023, from https://www.forbes.com/sites/schuylermoore/2019/04/13/the-importance-of-film-budgets/?sh=668e219b27f5
3. Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group. https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/
4. IMDB-Editors. (2016, December 7). Top 100 of 2016. IMDb. Retrieved April 3, 2023, from https://www.imdb.com/list/ls066347533/
5. Apte, N., Forssell, M., & Sidhwa, A. (2011). Predicting movie revenue. CS229, Stanford University.