W203 Lab 02 Report - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

0.1 Introduction

The movie industry is a multi-billion dollar business, and the budget of a film has become a crucial factor in determining its success, particularly in terms of revenue. Movie production firms, who have control over the budget, are highly motivated to enhance their movie revenue. This study aims to explore the relationship between movie budget (X concept) and revenue (Y concept) using a comprehensive dataset containing metadata for over 700,000 movies listed in the TMDB Dataset.

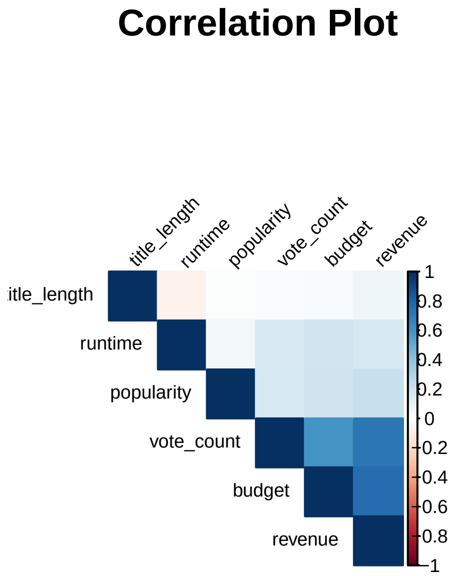
The dataset provides a unique opportunity to analyze the impact of budget on movie revenue, with both the X and Y concepts operationalized in terms of dollar amounts. However, there may be other omitted variables, such as the movie's storyline and competition from other movies, that could also impact revenue. The results of this study will have important implications for movie production firms, who can use the findings to make informed decisions about their investments in movie budgets.

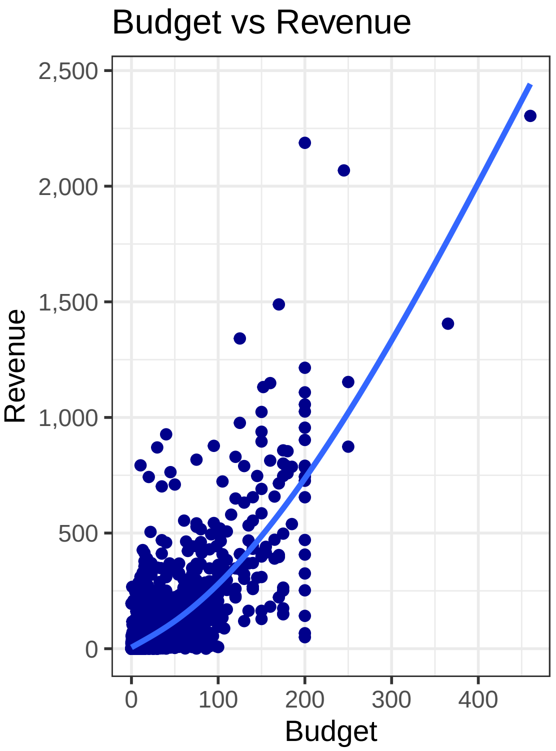
Future research could focus on generating new datasets to estimate the value of specific types of movies, such as different genres or production companies. Overall, the goal of this line of research is to provide accurate tools for movie production firms to make informed decisions about their investments and reduce uncertainty in the movie industry.

0.2 Data and Methodology

We gathered the dataset from Kaggle. The dataset comprises metadata for more than 700,000 movies listed in the TMDB Dataset. It has 722986 rows and 20 columns. It is relevant and provides an opportunity to analyze the relationship between the budget and the revenue of the movies. The dataset contains unique information on films, with each movie being distinct from the others and identified by an id column in the dataset. Additionally, the distribution of the X and Y concept is identical, satisfying the assumptions of IID. Because each movie is not subject to multiple samplings, the dataset is regarded as cross-sectional.Each row in the data represents a movie. I performed all exploration and model building on a 30% subsample of the data. The remaining 70%, totaling 9832 rows, was used to generate the statistics in this report.

To determine the factors that affect movie revenue, we need to extensively analyze data and conduct research .To analyze the factors that affect movie revenue, we have defined inclusion criteria that could potentially impact the dependent variable "revenue" (Y), and selected "budget" as the independent variable. We are focusing on movies that have been officially released to the public, and to refine our analysis, we are excluding movies with budgets or revenues exceeding $999 and budget values of $5 billion and $800 million. After filtering the data, we are left with a sample size of r filtered_rows observations.





I am interested in analyzing the relationship between movie revenue and budget, and identifying the factors that impact this relationship. To determine the covariates that impact the dependent variable, we have chosen the following variables:

- runtime: An important aspect of the movie-watching experience that has a correlation with revenue.
- vote_count: A proxy for popularity and an indicator of the level of engagement a movie has received from the public.
- popularity: A comprehensive metric that takes into account various factors such as user ratings, page views, and social media mentions.
- title_length: A factor that can impact audience interest and expectations, with shorter titles being easier to remember and share.
- release_season:The release season of a movie can impact its revenue. For example, summer and winter are considered peak movie seasons, and movies released during these seasons tend to have higher revenue.Many film studios release their biggest and most anticipated movies during this time to take advantage of the increased audience turnout and box office revenue potential.
- release_language:The language in which a movie is released can also impact its revenue. For example, movies released in widely spoken languages such as English have a higher potential audience and therefore may generate higher revenue. By including release language as a covariate, the model can capture this effect and make more accurate revenue predictions.

Our research proposal aims to gain insights into the factors that affect movie revenue and provide recommendations for filmmakers to optimize their marketing and production strategies. We have chosen covariates based on their potential impact on the dependent variable and their recognition in the industry. By including these covariates in our analysis, we can better understand the key factors driving movie revenue.

To best fit the relationship observed in exploratory plots, we operationalize the variables Budget(X) and Revenue(Y) in terms of amount. Figure 1 plots revenue as a function of budget. Before running the regression models, we conducted a correlation plot to understand the correlation between the dependent and independent variables. We further checked the collinearity between every predictor by running a variation inflation factor across all the predictors.

To meet large sample assumptions, we ensured that our dataset was independent and identically distributed from the population and had a unique BLP. We also checked for homoscedasticity between the residuals of the model and the predictions. Since evidence of heteroscedasticity was found, we applied robust standard errors to our models. Additionally, to make sure our data met the unique BLP assumption, we applied log transformations to both revenue and budget, which were heavily right-skewed.

From scatter plots, we observed a linear relationship between budget and revenue. To further understand the impact of different factors, we considered two categorical factors - the language of the movie when released and the month when the movie was released.

Our exploratory plots suggested that both language and release month could impact movie revenue. We believe that movies released in widely accepted languages have better revenue-generating potential. Additionally, we consider only December and January as holiday seasons due to their higher likelihood of generating more revenue.

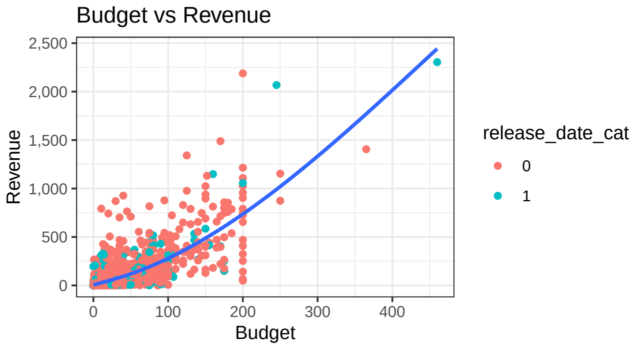
We created three regression models to understand these factors. The first model studied the correlation between budget and revenue. In the second model, we gradually added covariates such as vote_count and runtime. Finally, in the third model, we added more covariates such as popularity, title_length, release_date_cat, and lang_cat.

$$\widehat{revenue} = \beta_0+\beta_1\cdot Budget+\beta_2\cdot votecount+\beta_3\cdot titlelength+\beta_4\cdot runtime+\beta_5\cdot releasemonth+\beta_6\cdot releaselanguage+\mathbf{Z}\gamma$$

where β_1 represents the budget coefficient, β_2 represents the coefficient of vote count on the movies, β_3 represents the coefficient of vote count on the movies, β_4 represents the coefficient of title length on the movies, β_4 represents the coefficient of movie's runt time, change in the value increase for each year that passes after the remodel, β_5 represents the coefficient of movie's release month, β_6 represents the coefficient of movie's release language, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

To consolidate the categories and obtain actionable insights, I categorized the movies released in the months of December and January as holiday movies. Additionally, I segregated the language of the movies into two categories: English and non-English.

0.3 Results



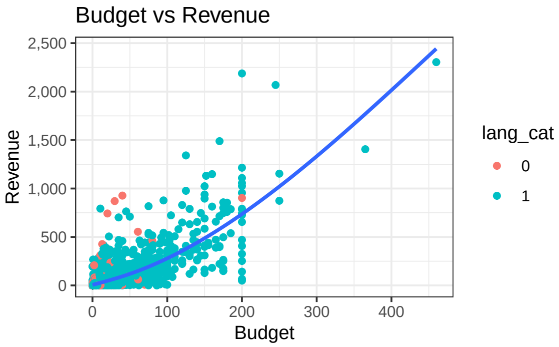


Table 1 shows the results of three regression models. The co-efficient of the X-concept variable "Budget" is having high statistical significane in all the three models studied. Its point estimates ranges from 0.78 to 0.91.

0.4 Limitations

After closely observing the movies dataset we can assume that the data is not strictly I.I.D. as two movies can be having the same cast & crew or same production company and exactly same release dates. Also by

Table 1: Estimated Regressions

	Output Variable: Revenue of the movie		
	(1)	(2)	(3)
Budget	0.910***	0.783***	0.775***
	(0.008)	(0.009)	(0.010)
Vote Count		0.0002***	0.0002***
		(0.00001)	(0.00001)
Run time		0.007***	0.007***
		(0.001)	(0.001)
Popularity			0.001**
			(0.0003)
Movie Title Length			0.128***
			(0.036)
Release season			0.172***
			(0.049)
Release Language			0.150**
			(0.051)
Constant	0.519***	−0.311***	−0.837***
	(0.026)	(0.071)	(0.127)
Observations	6,880	6,880	6,880
R^2	0.625	0.670	0.672
Adjusted R^2	0.625	0.670	0.672
Residual Std. Error	1.738 (df = 6878)	1.631 (df = 6876)	1.626 (df = 6872)

Note: HC_1 robust standard errors specified in parentheses. Holiday season is Dec/Jan and rest of the months are considered as non-holiday season. Language Category is either English or Non-English movies.

the scatterplot we can observe a strong co-relation between revenue vs budget and revenue vs runtime, thus we can say that columns in the movie dataset are not independent of each other. The Dataset has a chance of introducing Sampling Bias as the data about the movies is self reported on TMDB which is a user-generated content platform. Thus, the dataset may represent some movies but not be representative of all movies that have been produced globally. The data may also be biased towards movies that are more popular then the other. The dataset does not mention any data about Motion Picture Association of America(MPAA) Ratings hence dataset does not gives us any information about the anticipated age groups of the viewers. The dataset has lot of NA or zero values for key columns such as revenue, budget and vote_count, after the data cleaning process the size of actual dataset has reduced considerably. The dataset represents some extreme outliers such as $0 value for budget, revenue and runtime of more then 1000 minutes (16.66 hr), these outliers can influence the overall statistical analysis of the dataset. The dataset may not represent true picture of current movie industry/trends since it contains information of movies which has a release date before 2013, movie industry is a fast paced world and the trends get outdated very soon moresoever their is no consideration of inflation when concluding the statistical analysis between covariates such as budget & revenue.

0.5 Conclusion

This study estimated the economic value of movie production, specifically examining the relationship between movie budget and revenue.We also found that several covariates, such as vote_count, runtime, popularity, title_length, release_date_cat, and language_cat, have a significant impact on movie revenue. Additionally, we identified the holiday season and language as important categorical factors that can impact the success of a movie. Our hope is that this line of work will provide filmmakers with accurate tools to plan their investments and optimize their production strategies, reducing uncertainty in the film industry. Future research could examine the value of specific production decisions, such as marketing strategies or casting choices.