Estimating Revenue of Horror Movies

DATASCI 203 Lab 2 - Jonathan Phan, Jasmine Teo, Spencer Zezulka

Introduction

Film budgeting has always played a major role in the filming process cycle. The success of a film is oftentimes measured by the return on investment (ROI). In other words, how much profit is a film able to make after spending costs on casting, filming, marketing, and so on? Another important metric to consider is the amount distributors are willing to pay, which is calculated by the total budget divided by the total sales corollary. For example, if two completely identical films are created, where one was made on a $5 million budget and the other was made on a $20 million budget, then the one that used the $20 million budget for the film would sell for 4 times the amount of the $5 million budget one.

Since there are various film genres, our team decided to focus only on horror movies, which has always been known to be one of the highest profitable film genres, despite high budgeting spent on creating special effects and CGI. However, budgets for various horror movies can vary greatly, typically ranging from $5 million to $20 million, for those released in the U.S. between 2000 and 2016. Interestingly enough, box-office winners are not necessarily the movies that spend most on production and filming. Because revenue and budget are both important components when determining total profit, we wanted to explore how budget might affect revenue for horror movies.

Thus, our research question is as follows: "Does budget affect revenue for horror movies?" By focusing solely on the horror movie genre, we are hoping to use statistical methods to explore and identify a clearer relationship between budgeting and profit. Additionally, we know from both background knowledge and information gathered from movie budgeting articles that budget is a feature that has a significant impact on how much revenue is generated.[1] Here, revenue is chosen as a metric of success because high revenue usually reflects a high sales count of movie tickets. We run three sets of regression models in order to provide an estimate for revenue.

Data and Methodology

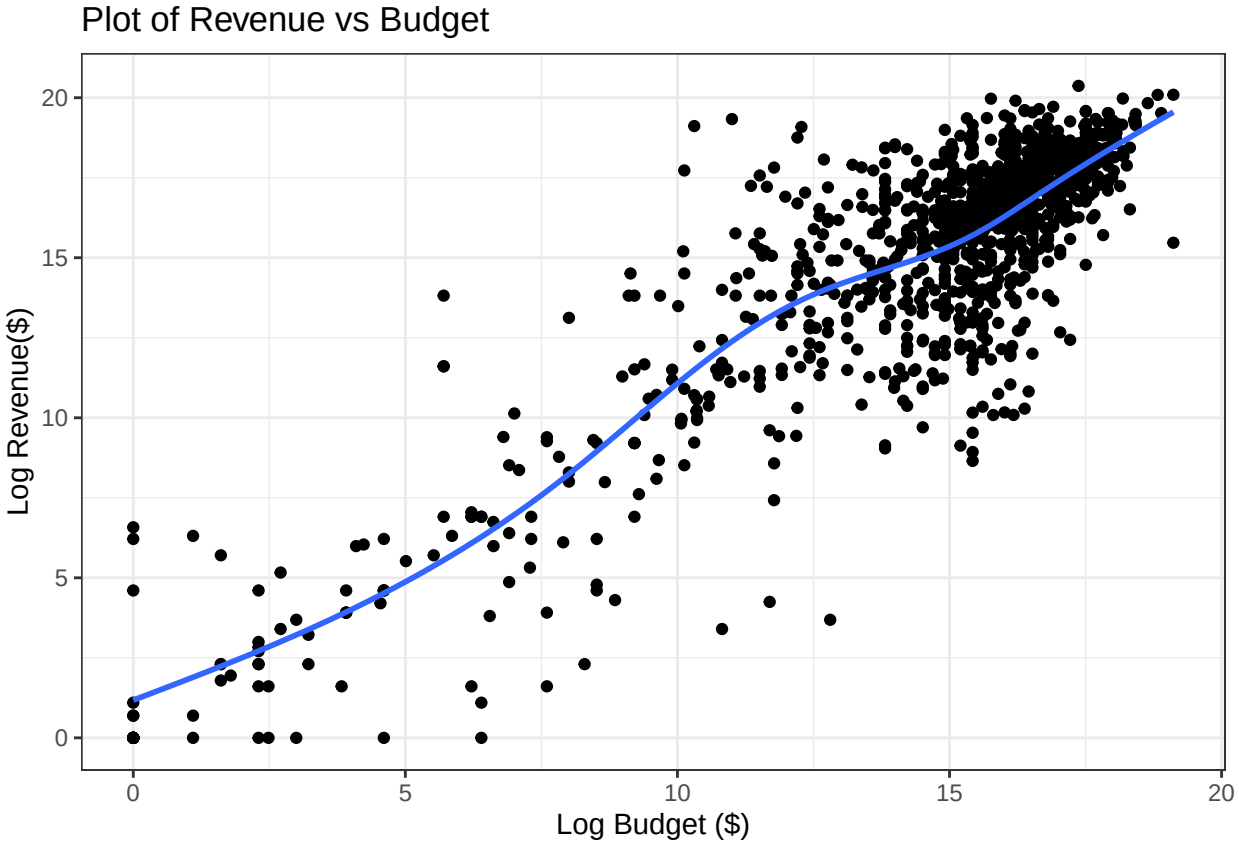
Our dataset consists of horror movies dating back to the 1950's. The dataset was extracted from *The Movie Database* using the TMDB API. There are approximately 35,000 movie records in the dataset, where each record represents a unique movie ID. Thus, the data is observational.

We filtered and cleaned the dataset for budget, revenue, and other variables of interest we plan to include in the second and third models. We first filtered the dataset to include only movies that were already released and had both non-zero and non-negative values for revenue and budget. This ensures that our model would not include movies that are currently in production and have finalized a set budget without reaping in any revenue. We then transform some variables to better suit a regression model. For example, we created a new column that represents the difference between the year a movie was released and the year of the oldest movie in our dataset, where it is possible that there might be changes in budget and revenue for horror movies over time. We also decided to transform the "language" column to an indicator variable that shows if a movie was in english or not. The "collection" column was also converted to an indicator variable, where movies that were part of well-known collections such as "The Conjuring", "Freddy", and "Ghostbusters" were labeled as 1

[1] Odell, Tracy. "High-Budget vs Low-Budget Horror Movies: Budgets and Box Office Stats." FinanceBuzz, FinanceBuzz, 20 Nov. 2022, https://financebuzz.com/high-budget-vs-low-budget-horror-movies.

and 0 otherwise. In the process of building and testing our models, we also decided to remove the popularity variable since we are unsure how popularity is calculated and what it accounts for.

According to the scale-location plot and the Breusch-Pagan test when checking for homoscedasticity between our output variable revenue and our primary input variable budget, since our p-value is 8.39e-13, which is much smaller than significance level 0.05, we reject the null hypothesis that the residuals are homoscedastic or evenly-spread. In an attempt to dampen this effect, we decided to apply a natural log transformation to budget and revenue, with a new p-value of 1.181e-08.



We are specifically interested in seeing whether we can predict revenue from the budget variable. As part of our exploratory data analysis, we plotted the linear relationship between budget (log) and revenue (log) as shown above, and the visual suggests a roughly linear relationship between both variables. Regressions of the following form are fitted:

$$\widehat{log(revenue)} = \beta_0 + \beta_1 \cdot log(budget) + \mathbf{Z}\gamma$$

where β_0 represents the y-intercept, β_1 represents the increase in revenue (log) for every one dollar increase in budget (log), $\mathbf{Z}$ represents the vector of all additional covariates we plan to test other than our base model with only budget (log) and revenue (log), and γ represents coefficients as a column vector.

Our second and third models have a more flexible structure, where we added additional covariates including movie vote count, vote average, (year released - oldest release year), whether or not the movie was in English, and whether it was part of a famous collection. We selected features as we thought they may also contribute in some way when predicting revenue (log). We did consider using other variables in our second and third models, but noticed issues with outcome variables on the right-hand-side which may affect the precision of our model. For example, we noticed that budget may affect movie runtime, so we decided to leave the variable out of our three regression models since we want to avoid having outcome variables on the right-hand-side.

Results

Table 1: Estimated Regressions

	Output Variable: Log Revenue in Dollars		
	(1)	(2)	(3)
Log Budget	0.97***	0.90***	0.86***
	(0.02)	(0.02)	(0.02)
Vote Count		0.0004***	0.0004***
		(0.0000)	(0.0000)
Vote Average			0.10*
			(0.04)
(year released - oldest release year)			−0.03***
			(0.004)
English			−0.17
			(0.19)
Collection			1.17***
			(0.11)
Constant	0.90**	1.38***	2.62***
	(0.29)	(0.29)	(0.44)
Observations	1,098	1,098	1,098
R^2	0.75	0.78	0.82
Residual Std. Error	2.09 (df = 1096)	1.96 (df = 1095)	1.81 (df = 1091)
Note:	HC_1 robust standard errors in parentheses. Additional features are vote count, vote average, year diff, english, collection		

Our regression results for all three models are summarized above in Table 1. The key coefficient on budget (log) was found to be statistically significant, where point estimates range from 0.86 to 0.97, which suggests that as budget (log) increases, revenue (log) also increases. If we look at our covariate, votecount, we notice it has coefficient estimates around 0.0004 for models 2 and 3, which means that for every one unit increase in vote count and holding budget (log) constant, revenue (log) increases by 0.0004. Our base model's r-squared score increases, which means more variability is explained in models 2 and 3, as more covariates are added.

The following hypothetical scenario summarizes how model 3 works: If, for instance, we have a horror movie with budget of $5 million, has a vote count of 300, vote average of 4.8, year release of 2022, made in English, and is not part of a collection, then our prediction for revenue would be around $143 billion, after converting log(revenue) back to revenue. Here, model 3 also suggests that boosts in revenue fade slightly over time and if the movie is made in the English language. However, we also want to note that a negative prediction of revenue can be unrealistic.

Limitations

In the process of running our regressions, there are several statistical and structural limitations we would like to note, which includes the following: assumptions for large-sample models, omitted variables, and outcome variables on the right-hand-side.

There are no dependencies between individual movie records based on vote counts, whether or not the movie was in English etc. However, because our data contains all horror movies over time, we note there may be some temporal dependence between the values, so we cannot assume independence entirely. For example, the time of the year may affect the talent involved in filming, where they could either be able or unable to recruit the cast of their choice or find the right people to invest, which will affect movie budgeting. More horror movies may be produced during certain times of the year based on past knowledge on peak popularity periods (i.e. Halloween).

Moreover, since the histogram of our main X variable, budget (log), is left-skewed, where all the points are concentrated towards the tail, this implies that variance might be infinite since points may not converge to a finite value. Thus, we note that a unique BLP may not exist even though no perfect collinearity exists.

Other than the limitations with large-sample assumptions listed above, there are also a few omitted variables that may bias our estimates, which are not included in the models. An example of an omitted variable is how much investors/shareholders contribute. Since how much investors/shareholders contribute has a positive correlation with the key variable, budget, which is positive in our model(s), the bias is away from zero. Thus, our budget variable is higher than it should be when considering this omitted variable.

We also identified issues withoutcome variables, where movie runtime could be dependent on budget, which we also removed from all three models. This was supported since the correlation between runtime and budget was high, around 0.72. As the budget increases, so does movie runtime. If we were to remove movie runtime from our regression, we would expect the coefficient for budget to increase since the coefficient of budget and movie runtime are both positive, removing movie runtime would transfer the positive effect to budget.

Conclusion

This study estimated the revenue of a horror movie using data that dates back to the 1950s. According to all three of our models, every one unit increase in budget (log), we can expect an increase in revenue (log) between 0.86 and 0.97. We also found that increases in revenue are expected to fade slightly over time and decrease further if the movie is made in the English language. The variables we were able to include in our model were limited based on the available features in our dataset and other limitations including violations of large-sample assumptions, outcome variables on the right-hand-side, and omitted variables, as discussed earlier in the *Limitations* section. For future research, if there is a better-structured dataset with more variables, like the type of film studio or how much investors/shareholders contributed, that could affect revenue prediction or a full description of what the popularity variable means, the precision of our regression model can be improved. We could also conduct a similar exploration where we focus on other types of movies genres like thriller, drama, or adventure. Ultimately, our goal is to help provide the tools and methods for production budget managers in the film industry to allocate and plan their budget accordingly.