Screamingly Good Returns of Horror Movie Investments

DATASCI203 Section 1 Team 5

Introduction

Within the film industry, producing a blockbuster film often requires seven, eight or even nine-figure financing. The unpredictability surrounding the financial success of studios' projects creates large uncertainties in the business environment. While action/adventure films dominate the box office, horror movies only account for a small 3.7 percent the market share. Despite their relatively smaller share, the return-on-investment potential for horror films presents a large opportunity for growth.

With the challenge of predicting a movie's financial success, horror historically stands out as a genre that consistently delivers large returns. From this, the inherent question arises: does choosing to make a horror movie, over other genres, cause an increased return on investment for a single film?

This study examines the characteristics of films released between 2000 and 2019. Typically, film budgets are separated into pre-production, production, and post-production budgets. Our dataset is limited to the largest cost category, the production budget. By examining several characteristics of movies, we aim to determine the causal change in return on investment (ROI) in the production budget related to the genre being horror, checking also for potential causal factors like seasonality of release, movie age rating (MPAA rating), and whether a picture was distributed by a high-market-share studio. We believe our results will provide guidance in making better decisions about a company's cinematic investments.

Data and Methodology

Our data was sourced via Kaggle from the site, the-numbers.com, and contains publicly released observational data that tracks key performance metrics of movies released by major studios between 1936 and 2019. The data provides categorical information (genre and distributor), release dates, production budgets, and worldwide and domestic gross revenues for 3,401 films.

To operationalize our analysis, we transformed several dataset variables. We created a binary variable ("genre_indicator") to indicate whether a movie falls into the "horror" genre, which is our primary input variable of interest. ROI, our outcome variable, was computed using two columns, worldwide_gross and production_budget (ROI = (worldwide_gross - production_budget) / production_budget). Using a ratio as our outcome variable allows us to account for year-over-year inflation, as well as assess profitability while managing large differences in film budgets.

We recognize that production_budget does not capture all costs to produce a movie as pre-production and post-production costs are excluded. We also lack data on revenue streams outside of box-office sales, like licensing revenue from streaming and merchandise–these are numbers studios generally do not release. We should also note that ROI does not convey the size of profits, purely their relationship to the initial investment. Looking at Figure 1 below, it is evident that over the 20-year period, horror movies have a greater potential for higher profitability and ROIs compared to other genres.

We applied a log transformation to our outcome variable (ROI) to account for a right skew in our errors, which we discuss with other limitations, below. Because some of the ROI values were negative we added a constant (1) to the ROI variable to avoid the regression dropping undefined values.

Because our database spanned a large period (1936-2019), we omitted films released before 2000 from our analysis, as viewing habits, theater-going, technology, and genre preferences have changed considerably

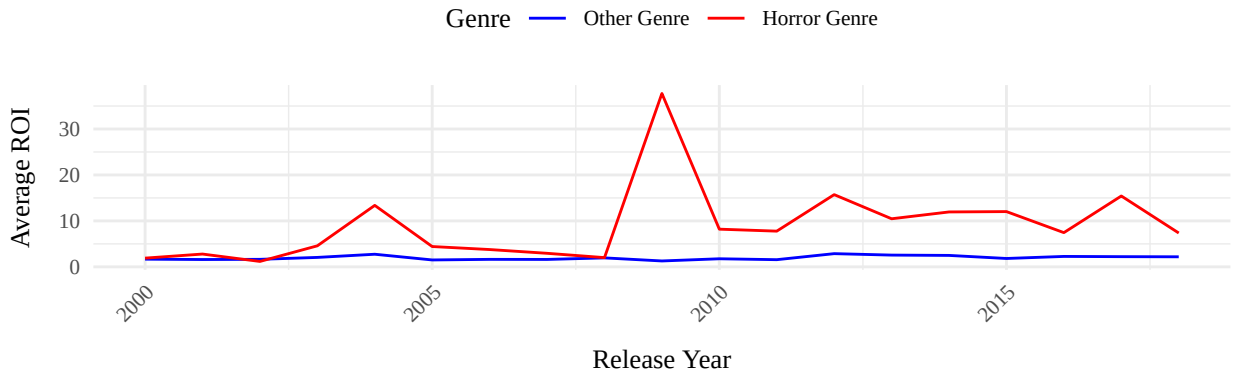

Figure 1: Average ROI by Release Year and Genre

since 1936. Using a large (2,403 samples), but more culturally consistent period allows us to work with more information, while minimizing temporal clustering. To account for what we observed in our data as a gradual increase in ROI during our period of analysis, we created an explanatory time_variable 1-20 mapped to 2000-2019 in our models.

Our first model included solely the genre_indicator and time_variable input variables.

Our second model added a "seasonality" variable we derived from each movie's release_date (Winter, Jan-Mar; Spring, Apr-June; Summer July-Sept; Fall, Oct-Dec). Since the summer and Thanksgiving-Christmas holiday release seasons are thought to be high-grossing, we expect to see an increase in ROI over the summer and fall seasonality variables.

Our third model includes age-restricted rating (mpaa_rating). Since G-rated and PG-rated movies are appropriate/available for a wider audience and tend to include a large number of animated movies, which can be cheaper to produce, we might expect to see a lower age rating increase ROI.

Our fourth model includes a categorical variable for movies distributed by "major_studios," the four distributors with highest historical market shares (Walt Disney, Warner Bros, Sony Pictures, and Universal). Any film released by a non-major studio is labeled "other," the default in our model. Since major studios have big marketing departments, relationships with movie stars, relationships with theater chains, we might expect to see that advantage for a film pay off as a causal increase in ROI.

To account for autocorrelation in our residuals (Ljung-Box test, p-value = 2.2e^-12), we adjusted our models with Newey-West autocorrelation-robust estimates for coefficients and standard errors.

Results

The regression table below shows the outcomes of our four models. Across all models, the intercept and the coefficient for our primary input variable (genre_indicator) are statistically significant, with the coefficient ranging from .76 to .87. What that means in the practical sense, is that for example, in model 1 holding all other variables constant, a horror movie could expect ROI to be 1.5 [calculation:exp(.19 + .71) -1] compared to 0.2 [calculation: exp(.19)-1] for a non-horror movie, 1.3 more than non-genre.

The age-restricted rating (MPAA rating) of a movie also showed a statistically significant and non-trivial downward effect on ROI for non-G movies. The vast majority of horror movies in our sample were rated PG-13 or R (203 out of 207). However, the downward impact for horror movies was tempered in model3 by a higher coefficient for the genre_indicator, suggesting that MPAA rating may be an outcome of genre, rather than an independent variable.

A summer or winter release date also improved ROI to a statistically significant degree. A horror movie released in summer or winter, for instance, could expect a roughly 1.8 ROI compared to 1.1 in fall. Distribution by a major studio also improved ROI to a statistically significant degree for all four major studios.

A horror movie released by Sony Pictures, for instance, could expect a 2.2 ROI compared to .5 for a horror movie released by a non-major studio.

Overall, our models suggest that investments in horror movies yield higher returns compared to other genres. Releasing movies in winter or summer, and having the backing of a major distributor could further improve ROI significantly. Further research needs to be conducted to determine the size of the expected returns compared to individual movie genres, as other genres may yield the studio overall greater profits. But with the information available, in situations where projects have a limited production budget, we recommend studios to opt for horror over other genres to maximize ROI.

	Output Variable: Return on Investment				
	log(ROI + 1)				
	(1)	(2)	(3)	(4)	(5)
Genre Indicator	0.71***	0.72***	0.86***	0.75***	0.87***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
Time Variable	0.01	0.01	0.01*	0.01*	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Seasonality: Spring		0.01			−0.03
		(0.10)			(0.10)
Seasonality: Summer		0.30***			0.23**
		(0.08)			(0.08)
Seasonality: Winter		0.28**			0.23*
		(0.09)			(0.09)
MPAA Rating: PG			−0.54***		−0.41**
			(0.14)		(0.14)
MPAA Rating: PG-13			−0.69***		−0.57***
			(0.13)		(0.13)
MPAA Rating: R			−1.05***		−0.87***
			(0.14)		(0.15)
Major Studio: Sony Pictures				0.46***	0.43***
				(0.09)	(0.09)
Major Studio: Universal				0.74***	0.68***
				(0.11)	(0.10)
Major Studio: Walt Disney				0.70***	0.46***
				(0.11)	(0.10)
Major Studio: Warner Bros				0.51***	0.47***
				(0.10)	(0.10)
Constant	0.19**	0.04	0.97***	−0.05	0.52***
	(0.07)	(0.09)	(0.12)	(0.09)	(0.15)
Observations	2,436	2,436	2,430	2,436	2,430
R^2	0.02	0.02	0.04	0.05	0.07
Residual Std. Error	1.60 (df = 2433)	1.59 (df = 2430)	1.58 (df = 2424)	1.57 (df = 2429)	1.55 (df = 2417)
Note:					*p<0.05; **p<0.01; ***p<0.001
					Newey-West Adjusted Standard Errors
					in parentheses

Limitation

(8a) Statistical Limitations

Although our sample size satisfies the criteria for a large data sample, to validate the use of the linear regression model conservatively we evaluated each of the assumptions of the Classical Linear Model.

Assumption 1: I.I.D. Data: Our data is not fully Independent and Identically Distributed (I.I.D.), as clustering is present. Small-budget films are not included in our dataset, as the smallest production budget included is $250,000, and most production budgets are in the multi-million-dollar range. Moreover, as the data spans across multiple years (2000-2019) there is temporal clustering as movie-going may be impacted by economic booms and downturns, which we addressed with the inclusion of our time_variable. Moreover, movie budgets, genres, and gross revenues may not be independent of one another. A high-performing horror or action movie may inspire sequels. Multi-picture contracts with directors or actors may impact production budgets and genre choices.

Assumption 2: No Perfect Collinearity: Since the model runs without dropping any features, it shows that there exists no perfect collinearity.

Assumption 3: Linear Conditional Expectation: After applying a log transformation to the ROI with log(ROI +1) in the model, the Prediction vs. Residuals plot in Figure 2 suggests that we do not have a reason to conclude that there is no linear relationship between the genre of a movie and its ROI. The residuals for model predictions are centered around zero, so the linear conditional expectation assumption is not violated.

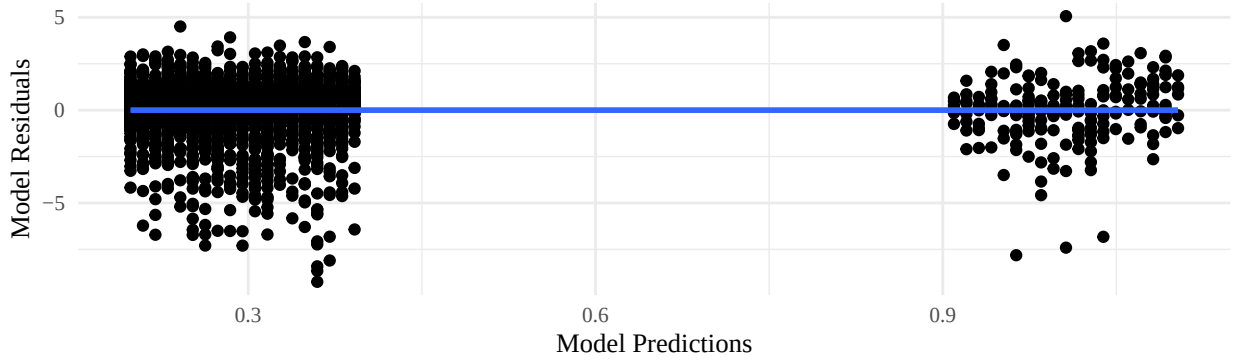

Figure 2: Model Residuals v.s. Model Predictions

Assumption 4: Homoscedasticity: We employ the Breusch-Pagan test to evaluate the homoscedasticity assumption. The null hypothesis is that the variance of the errors in the regression model is constant (homoscedasticity), while the alternative hypothesis is that the variance of the errors is not constant (heteroscedasticity). In our model, the p-value that is very close to zero (0.0074) which suggests strong evidence against the null hypothesis. Therefore, the presence of heteroscedasticity indicates that the variance of the errors is not constant.

Assumption 5: Normally Distributed Errors: This assumption specifies that the residuals of our model should follow a normal distribution. However, our finding reveals a skewness value of 27.75219, indicating a significant right skew in the distribution of the errors. The QQ Plot also shows an exponential shape with points deviating from the quantile line as x grows larger. We addressed the concern by applying a log transformation to ROI + 1, evident in the result seen in Figure 3.

We also observed that there are several outliers (ROI>400%) that might have an impact on prediction accuracy and they could potentially contribute to the violation of CLM assumptions mentioned above. However, we decided not to remove them as they're valid data points in real-world scenarios and can contain valuable information relevant to our research question. Serial autocorrelation issues are addressed above.

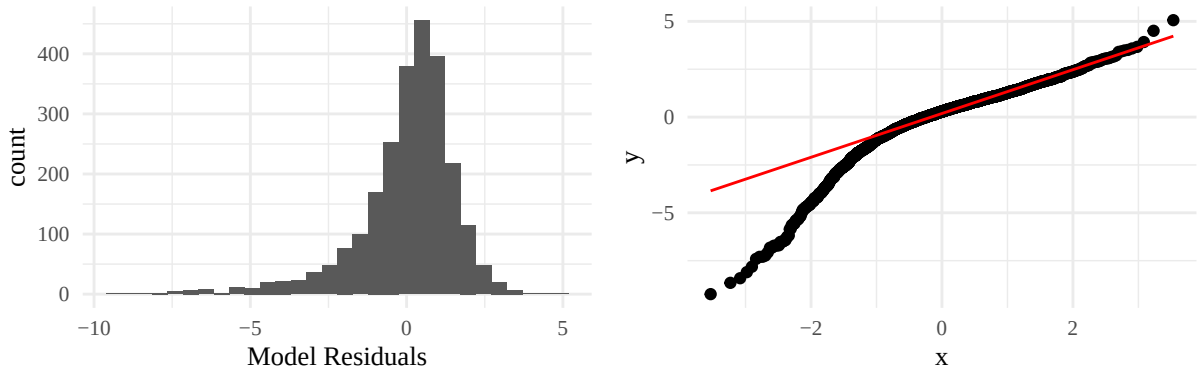

Figure 3: Histogram of Residuals and QQ Plot After Log Transformation

(8b) Structural Limitations

Due to the availability of cost data, calculations of ROI were limited to just the production budget. Excluding other cost categories results in an incomplete representation of the total expenses, potentially resulting in overestimates of returns. The pre-production and post-production costs vary largely among films, potentially producing largely different ROIs across genres. Consequently, the true model would produce different coefficients comparing horror and other genres, potentially yielding differing insights.

As our analysis is limited to our dataset, several omitted variables could have an impact on a film's ROI, such as having a star actor/actress feature. We expect the relationship between popular talent and ROI to be positive, as popular stars attract more viewers. Typically, horror movies are associated with smaller budgets. In our data, the median production budget for horror was $13M, the smallest among all genres, 4.5 times less than $58M the median action-genre film budget, the genre with the highest budget. With horror movies having smaller budgets, we expect a negative correlation with casting popular talent and the horror genre indicator variable. The omitted variable bias would be negative, and the bias would be driven toward zero, understating our results. Similar reasoning could be applied to the omitted variable of reputable directors causing more negative omitted variable bias and driving the bias further towards zero.

Furthermore, the ROI of a movie is heavily dependent on effective promotion to capture the public's attention. With a minimal production budget, we estimate that the marketing budget would be limited as well, resulting in a positive correlation between the production budget and the marketing budget. The positive correlation between ROI and marketing budget, as well as production budget and marketing budget would cause our omitted variable bias to be positive. From this, the bias would be driven away from zero, overestimating our results.

Conclusion

Our investigation reveals a statistically significant positive association between the horror genre and ROI. The difference in ROI between horror and non-horror movies in our most basic model is 1.3. Thus, our analysis presents solid evidence that characteristics that come along with movies in the horror genre result in a greater ROI compared to other genres. This conclusion remains stable even as we introduce other factors such as time, season, and MPAA rating.

There are, however, limitations in our dataset and our model–most notably, our inability, from given data, to estimate all of a movie's costs and revenue streams. A more extensive investigation, encompassing these missing pieces of information , would increase the precision and application of our results. As the film industry evolves, constant research efforts are vital for staying informed of emerging trends and guaranteeing the applicability of predictive models.