Research Proposal: Analyzing Movie Revenue

Andrew Batmunkh, Khangerel Batzul, Kevin Cheung

2025-10-20

Contributions

Introduction: Khangerel Batzul

Data Description: Kevin Cheung

Ethics Discussion: Khangerel Batzul

Preliminary Results: Andrew Batmunkh

Analysis Plan: Andrew Batmunkh

Introduction

The film industry is one of the most data-rich creative sectors of cultural and economic production. It generates billions in worldwide revenue each year with its success depending on a complex combination of quality, budget, cast and many more. With the availability of large datasets like the TMDB 5000 Movie Dataset, researchers can now study the factors that contribute to a film's success, providing further insight about trends in media production for producers and studios. This project explores the question " How do movie characteristics influence the financial success of films?" Using the aforementioned dataset, we aim to identify which factors best predict box office revenue.

This project applies Multiple Linear Regression (MLR) to investigate the relationship between multiple continuous and categorical predictors (e.g., budget, runtime, genre) and a continuous response variable (e.g., revenue). Using linear regression allows us to estimate average change in success associated with each predictor, using estimated coefficients, confidence intervals, and p-values. Unlike simple correlations that don't account for a multivariate nature or confounding factors, linear regression is ideal for addressing questions of significance and contribution.

Previous research supports the relevance of this question and its variables. Nelson and Glotfelty (2012) found that production budgets, star power, and viewer ratings significantly explain variation in box office return. The primary focus was to quantify the financial value of star power in the film industry. On the other hand, Hsu (2006) examined the role of genre and audience perception in a film's market performance. The findings suggested that while multi-genre films attract wider audiences,they generally receive lower ratings due to the harder categorization, highlighting the significance of genre identity and audience appeal on critical success. Liu (2006) finds that over time, the quantity of discussion or word-of-mouth determines a film's financial success more than whether the conversation is favorable or not. These findings show that financial and critical success depend on multiple interacting factors, and not just one or the other, emphasizing their unique contributions. This supports our decision to include a broader set of predictors to better understand what drives movie performance and how these interactions can be effectively modeled through linear regression.

This analysis could help film studios, producers and marketing teams make data-driven decisions. This study offers a quantifiable way to predict performance and use resources more effectively by showing how different factors affect overall success. Researchers and marketers can also use these predictions to better understand how trends in the entertainment industry are evolving.

Data Description

We use the TMDb 5000 Movies dataset from Kaggle (https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata). The data are from The Movie Database (TMDb). TMDb is a film catalog that gathers movie metadata by user and partner submissions. TMDb helps movie fans find accurate, structured details like title, release date, runtime, genres, budget and revenue.

My response variable is revenue. After filtering to movies with positive revenue, revenue contains 3228 observations with no missing values. The range is 5 to 2787965087 dollars with mean 121279999 and standard deviation 186319835. The first quartile is 17000000, the median is 55191503 and the third quartile is 146343450. This response variable is suitable for use in a linear model. First, revenue is continuous. Also, each movie is mostly treated as an independent observation. However, some movies might still be related. For example, films from the same franchise, made by the same production team. Budget is right-skewed and includes 216 outliers. The variance is large relative to the mean. Popularity and vote_count are also highly right-skewed with 175 and 305 outliers respectively and high variance. Runtime displays a mild right tail with 99 outliers. Vote_average appears near-symmetric and lies within the bounded range from 0 to 10. Release_year covers several decades and contains 203 outliers.

The reason for the choice of interacted predictors is that spending affects each genre differently. Prior research shows genre-specific returns to investment. For example, horror often delivers high ROI (Novoseltsev, 2015), so allowing the budget slope to vary by genre captures those divergent returns. Therefore, we include a log_budget and primary_genre interaction.

Table 1: Summary statistics for numerical predictors in the preliminary model.

Variable	N	Mean	SD	Min	Q1	Median	Q3	Max	Outliers
budget	3228	40666419.51	44398404.53	1.00	1.050e+07	2.500e+07	5.500e+07	3.8000e+08	216
popularity	3228	29.04	36.17	0.02	1.047e+01	2.041e+01	3.734e+01	8.7558e+02	175
release_year	3228	2001.68	13.28	1916.00	1.998e+03	2.005e+03	2.010e+03	2.0160e+03	203
runtime	3228	110.72	20.97	41.00	9.600e+01	1.070e+02	1.210e+02	3.3800e+02	99
vote_average	3228	6.31	0.87	0.00	5.800e+00	6.300e+00	6.900e+00	8.5000e+00	43
vote_count	3228	977.59	1414.42	0.00	1.780e+02	4.710e+02	1.148e+03	1.3752e+04	305

Ethics Discussion

Our dataset, TMDB 5000 Movies, is public and open-access, with its metadata collected through verified users and partner submissions via the TMDB API. With this open-license system, the collection methods can be considered reasonably trustworthy because of its transparency and reproducibility. The easily accessible data also ensures accountability. However, the user-based data has potential risks of inconsistency and incompleteness, especially for older or smaller films. For example, niche films may lack detailed records, making the dataset more reflective of mainstream cinema.

The dataset respects autonomy and informed consent. As previously mentioned, voluntary users contribute to the TMDB data, meaning there's minimal ethical risks related to personal privacy violations or unconsented data extraction. Similarly, the data collects descriptive information about movies only with no mention of confidential details, reinforcing its compliance with ethical principles. Kaggle's distribution of the dataset also maintains data integrity recognizing when and where credit is due.

Despite all this, representational bias is a key limitation in the dataset. Being mainly made up of Hollywood films in English, the dataset leaves international films underrepresented. As a result, models may overrepresent Western characteristics and patterns, limiting the scope of applicability . There may also be bias in TMBD's popularity metric, as its ratings are tailored to its users rather than the general public. WIth this in mind, it is essential to acknowledge such biases when interpreting our model outcomes. From an ethical standpoint, we are using the data for educational and non-commercial purposes, in accordance with TMDB's terms of use and principle of proportionality. The goal of our analysis is to understand trends in success and not to evaluate individuals. To uphold both ethical and statistical integrity, we acknowledge the need to maintain transparency, accurate citations and representational biases.

Preliminary Results

Table 2: Preliminary multiple linear regression results predicting log(revenue). Standard errors are in parentheses and 95% confidence intervals are in brackets. Only significant predictors (p < 0.05) and key continuous variables are shown.

Predictor	Estimate	Std Error	t-value	p-value	CI Lower	CI Upper	Sig.
Intercept	19.0400	2.00700	9.487	<2e-16	15.10600	22.97600	***
log(Budget)	0.8680	0.05400	16.099	<2e-16	0.76200	0.97300	***
Runtime (min)	-0.0020	0.00100	-2.805	0.005	-0.00300	-0.00100	**
Release year	-0.0090	0.00100	-9.370	<2e-16	-0.01100	-0.00700	***
Popularity	0.0020	0.00000	4.437	<0.001	0.00100	0.00300	***
Vote count	0.0001	0.00001	8.879	<2e-16	0.00009	0.00011	***
Vote average	0.1430	0.01600	9.204	<2e-16	0.11200	0.17300	***
English-language	0.0890	0.05700	1.566	0.118	-0.02300	0.20100	
Has homepage	0.0660	0.02400	2.781	0.005	0.02000	0.11300	**
Primary_Genre_Documentary	3.2700	8.23400	3.976	7.16E-05	1.65900	4.88800	***
Primary_genre_Horror	4.1700	0.54700	7.617	<2e-16	3.09600	5.24300	***
Primary_genre_Romance	-2.3300	1.02700	-2.268	0.023	-4.34400	-0.31600	*
Primary_genre_War	-5.2560	2.21300	-2.375	0.018	-9.59400	-0.91800	*
log(Budget):Primary_Genre_Documentary	-0.4890	0.12600	-3.878	0.0001	-0.73700	-0.24200	***
log(Budget):Primary_genre_Horror	-0.5430	0.07600	-7.177	<2e-16	-0.69100	-0.39500	***
log(Budget):Primary_genre_Romance	0.3270	0.14100	2.324	0.02	0.05100	0.60200	*
log(Budget):Primary_genre_War	0.6420	0.29500	2.179	0.029	0.06400	1.22100	*

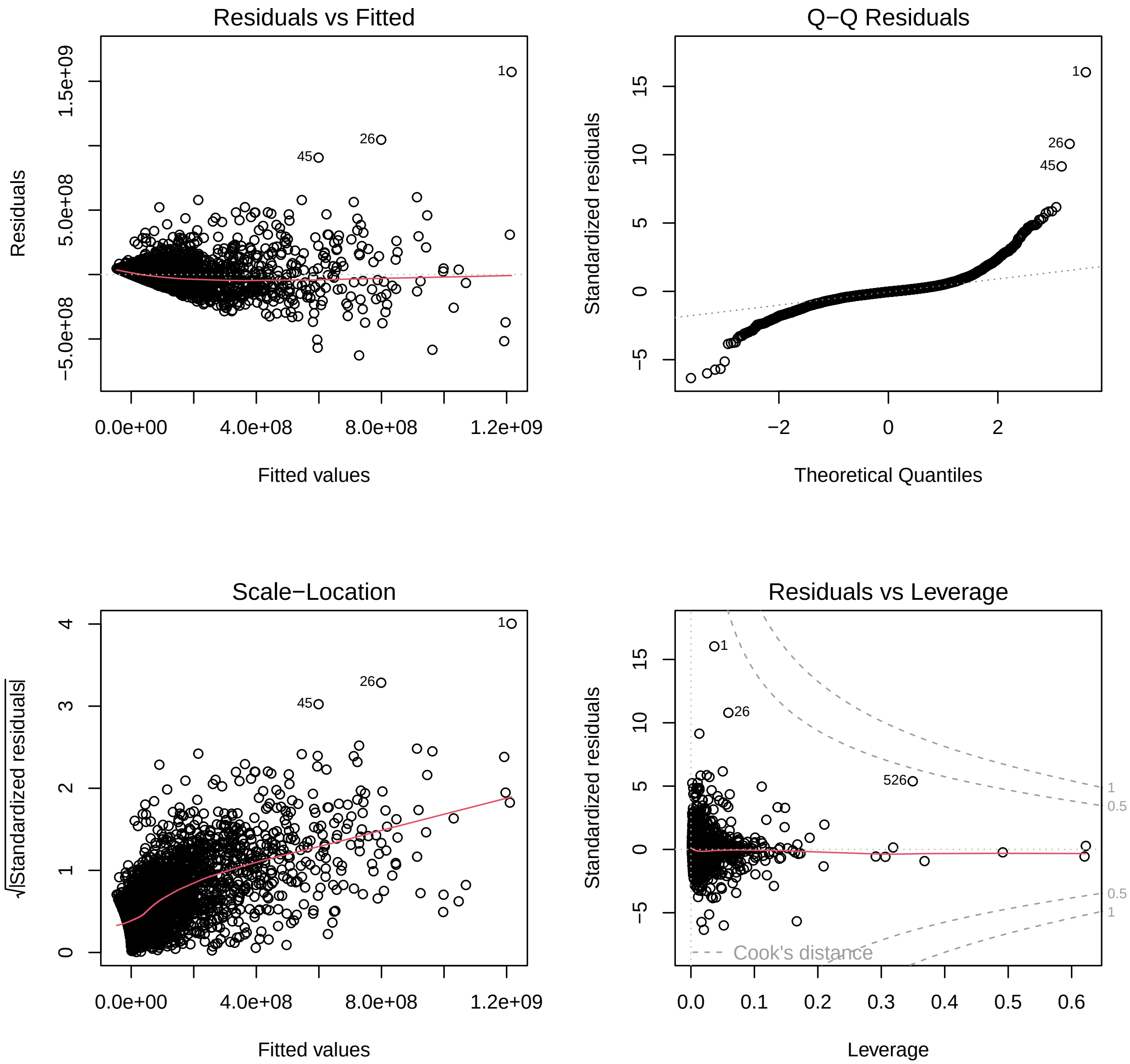

Figure 1: Residual diagnostics for the preliminary multiple linear regression predicting log(revenue) using 10 predictors: log(budget), runtime, release_year, popularity, vote_count, vote_average, is_english, has_homepage, primary_genre, and log(budget)×primary_genre. Model based on n = 3,228 observations with Adjusted $R^2 = 0.565$, F-statistic = 98.6, p < 2×10^{1} . Figure shows 4 separate plots: (A) Residuals vs Fitted, (B) Normal Q-Q, (C) Scale-Location, and (D) Residuals vs Leverage with Cook's distance contours.

For our prediction of log(revenue), we fitted a multiple linear regression model with ten predictors: seven numerical (log_budget, runtime, release_year, popularity, vote_count, vote_average, is_english) and three categorical (has_homepage, primary_genre, and log_budget × primary_genre). These variables predict how production scale, audience engagement, and genre-specific factors influence financial performance. The model explains 57% of the variance in log(revenue) ($R^2 = 0.571$, Adjusted $R^2 = 0.565$) with an RSE of 0.596 (df = 3,184) and shows strong overall significance (F(43, 3184) = 98.6, $p < 2×10^{1}$).

Table 1 shows that several numerical and categorical predictors significantly impact revenue. Log(budget) ($\beta = 0.87$), popularity ($\beta = 0.002$), vote count ($\beta = 0.0001$), and vote average ($\beta = 0.143$) all positively correlate with revenue, meaning higher budgets, audience attention, and ratings lead to increased earnings, consistent with Liu (2006) on visibility and word-of-mouth. The negative coefficient for release year ($= -0.009$) suggests older, high-earning films are common in the dataset, skewing newer movie revenue after log transformation. Runtime has a small negative effect, possibly due to fewer screenings for longer films.

Categorical predictors show clear genre effects. In Table 1, Documentary and Horror films have strong, significant positive coefficients, indicating higher revenue, while Romance and War films show negative effects, supporting prior findings that emotional or historical genres earn less commercially (Hsu, 2006; Novoseltsev, 2015). The interaction between budget and genre suggests that higher budgets yield smaller gains for Horror and Documentary films but greater benefits for Romance and War genres.

Figure 1 shows that model assumptions are mostly met. Residuals appear randomly scattered around zero with mild heteroskedasticity and slightly heavy tails, however still suggesting linearity and independence are reasonably satisfied.The Q–Q plot indicates near-normal residuals, and leverage plots reveal a few high-influence outliers, such as major blockbusters, as noted by the R message "not plotting observations with leverage one: 3199." These minor deviations suggest the log transformation was effective, though small variance issues remain. Overall, the model shows strong explanatory power consistent with past research. Future refinements will address mild heteroskedasticity and test if further transformations improve generalizability.

Analysis Plan

Continuing on our preliminary model, the next part of this project will focus on refining and validating the multiple linear regression predicting log(revenue). Our current cleaned model allows basic linear assumptions but shows mild heteroskedasticity and potential redundancy among correlated predictors. With the concepts covered up to Module 6, we will aim to test, and re-evaluate the model to improve accuracy and interpretability.

We will begin by refining the preliminary model, reviewing overall model fit and diagnostic results from the residual plots provided in Figure 1. This includes examining the Residuals vs. Fitted plot for signs of non-linearity or non-constant variance and the Q–Q plot to evaluate the normality of residuals. Any noticeable curvature or fan-shaped patterns would indicate mild heteroskedasticity. To improve model performance, we will compare several refined models using an improved Adjusted R^2 and AIC/BIC as selection criteria. Predictors with weak statistical significance or minimal contribution to explanatory power will be considered for removal to improve better prediction , while also ensuring that our key predictors of log(budget), primary_genre, and popularity will remain in all versions, as these variables consistently explain a large portion of revenue variation in both our results and prior research (Liu, 2006; Hsu, 2006; Nelson & Glotfelty, 2012).

Model assumptions will be checked after every major adjustment using residual and Q–Q plots. Because the residuals showed mild heteroskedasticity, we will test different transformations such as Box–Cox and check the lambda for the most sufficient transformation and fix for skewed predictors like vote_count or popularity. We may also explore additional interaction terms (e.g., popularity × vote_average) to capture audience behavior more accurately.

Bibliography

Hsu, G. (2006). Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly*, *51*(3), 420–450. https://doi.org/10.2189/asqu.51.3.420

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89. https://doi.org/10.1509/jmkg.70.3.74

Nelson, R. A., & Glotfelty, R. (2012). Movie stars and box office revenues: An empirical analysis. *Journal of Cultural Economics*, *36*(2), 141–166. https://doi.org/10.1007/s10824-012-9159-5

Novoseltsev, M. (2015, October 6). *Cinema: Analysis of genres and plot texts and their impact on "box office" performance* [Project work, DAS Data Science, School of Engineering]. Zurich University of Applied Sciences (ZHAW). https://www.zhaw.ch/storage/engineering/institute-zentren/cai/DAS15_Movie_Performance_Novoseltsev.pdf