W203 Lab 02 Report - Team 02

Ayoade Israel, Divya Menghani, Hamsini Sankaran, Sivakumar Thiyagarajan

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 its Revenue. Movie production firms have control over the budget and are highly motivated to enhance their movie revenue. The global movies and entertainment market generated an estimated revenue of USD 90.92 billion in 2021. According to projections, the market is expected to grow at a compound annual growth rate (CAGR) of 7.2% between 2022 and 2030. By 2030, the market is anticipated to reach a total revenue of USD 169.68 billion[1].

Although numerous studies have been conducted on predicting movie revenue, there have been instances in which movies with large budgets and compelling plotlines failed to perform at the box office, such as Mulan (Budget: $200 million, Revenue: $69 million). In contrast, movies with relatively lower budgets succeeded, such as Super Size Me (Budget: $65K, $22 million)[2]. It raises questions about what factors influence a movie's success.

Our research aims to gain insights into the factors that affect movie revenue and provide recommendations for filmmakers to optimize their marketing and production strategies, using a comprehensive dataset. One of the drawbacks of existing research in predicting movie revenue is the difficulty in accurately capturing the impact of other complex factors that influence a movie's success apart from budget. In this study, by applying a set of regression models, we estimate a movie's revenue based on budget along with several other factors.

2 Data and Methodology

We gathered the dataset from Kaggle. The dataset comprises metadata of more than 700,000 movies listed in the TMDB Dataset. It has 722986 rows and 20 columns representing various factors influencing the movie's success. 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 movies, with each movie being distinct from the others and identified by an id column in the dataset.

We performed exploratory studies and model building on a 30% subsample of the data. The remaining 70%, totaling 9832 rows, is used for testing the model and evaluating the results in this report. To meet the large sample assumptions, we are evaluating the assumptions of I.I.D and Unique BLP. From the exploratory data analysis, although the budget and revenue distribution was roughly identical, we observed that our dataset is not strictly I.I.D (further discussed in the Limitations section), and it does not have finite variance from the histogram plots, eventually violating the unique BLP assumption. We performed a log transformation of the predictor variables (budget and title length) to overcome this. Because each movie is not subjected to multiple samplings, the dataset is regarded as cross-sectional.

We cleaned data by ignoring the rows with missing values in all dependent and independent variables. From the exploratory plots, we observed few outliers for budget, and we are considering the movies with a budget greater than $999 and excluding movies with a budget greater than $799 million. After filtering the data, we are left with a sample size of 9832 observations.

To best fit the relationship observed in exploratory plots, we operationalize the variables Budget(X) and Revenue(Y) in terms of amount (in dollars, $). Figure 2 shows a plot of Revenue as a function of budget,

[1] Grand View Research, "Movies And Entertainment Market Size, Share & Trends Analysis Report, By Region, And Segment Forecasts, 2022 - 2030".

[2] BuzzFeed, "10 Films That Flopped At The Box Office Despite Mammoth Budgets, And 9 Low-Budget Movies That Made Hefty Profits", 2023.

depicting a linear relationship between the two variables. Figure 1 shows the "pearson" correlation between the dependent and all independent variables. Based on the correlation score, we chose the following covariates that affect the dependent variable (Revenue). The color shades are used to represent correlation, spanning from dark red to dark blue. In this scheme, dark shades of red (+1.0) indicate a strong positive correlation between variables, whereas dark shades of blue indicate a strong negative correlation (-1.0).. By including them in our analysis, we will better understand the fundamental factors that influence the movie revenue.

- vote_count (count): An indicator of the level of public engagement a movie has received.
- runtime (minutes): An essential aspect of the movie-watching experience that has a correlation with Revenue. It can impact the number of screenings and profitability.

Figure 1. Correlation Plot

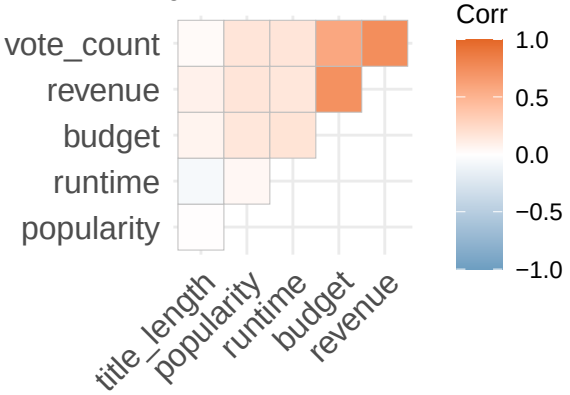
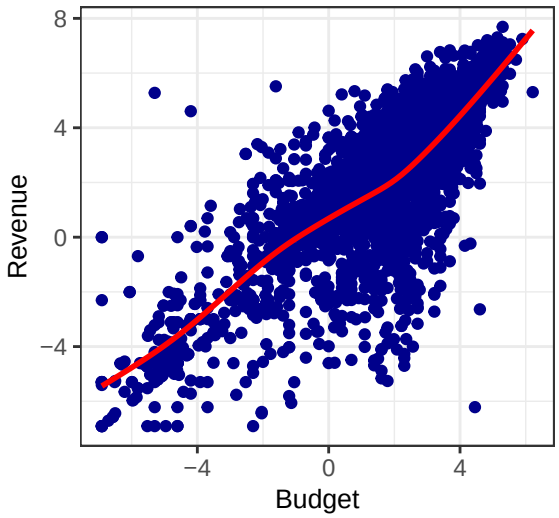


Figure 2. Budget vs Revenue



- popularity (score): A comprehensive metric that considers various factors such as promotions and social media mentions.
- title_length (count): A factor that can impact audience interest, with shorter titles being easier to remember and share. This column is derived from the 'title' column in the dataset.
- release_season (binary): Many film studios release their most significant and most anticipated movies during Christmas and New Year to take advantage of the increased audience turnout and box office revenue potential. Here we considered two prominent holiday months (December & January) as one category and the rest as a different category.
- release_language (binary): Movies released in widely spoken languages such as English have a higher potential audience and, therefore may generate higher Revenue. The model can capture this effect by including release language as a covariate and make more accurate revenue predictions. Here we took English as a primary category and all other languages as a non-English category.

Before running the regression model, we checked the multi-collinearity between all predictors by running a variation inflation factor to reduce unstable and unreliable estimates of the regression analysis. We observed there is no evidence of multi-collinearity between the predictor variables.

We created three regression models on the 30% dataset to understand better to understand the factors contributing to the Revenue of a movie. The first model studied the correlation between budget and Revenue. In the second model, we gradually added covariates such as 'Vote Count' and 'Run time'. Finally, in the third model, we added more covariates such as Popularity, Movie Title Length, Release season, and Release Language.

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

where β_0 is the constant in the model, β_1 represents the budget coefficient, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

Additionally, we evaluated homoscedasticity to check if the variance of the errors or residuals in the model is constant across all the predictor variables, but the results showed heteroscedastic behavior. Since our dataset is large, we will continue to build our model with robust standard errors to handle the heteroscedastic behavior.

3 Results

Table 1 shows the results of three regression models with 70% dataset.

Table 1: Estimated Regressions

	Output Variable: Revenue of the movie		
	(1)	(2)	(3)
Budget	0.909^{***}	0.780^{***}	0.772^{***}
	(0.008)	(0.009)	(0.010)
Vote Count		0.0002^{***}	0.0002^{***}
		(0.00001)	(0.00001)
Run time		0.007^{***}	0.008^{***}
		(0.001)	(0.001)
Popularity			0.001^{**}
			(0.0003)
Movie Title Length			0.107^{**}
			(0.036)
Release season			0.181^{***}
			(0.049)
Release Language			0.156^{**}
			(0.051)
Constant	0.514^{***}	-0.325^{***}	-0.800^{***}
	(0.026)	(0.072)	(0.128)
Observations	6,880	6,880	6,880
R^2	0.625	0.671	0.673
Adjusted R^2	0.625	0.671	0.673
Residual Std. Error	1.734 (df = 6878)	1.624 (df = 6876)	1.620 (df = 6872)
Note:	*p<0.1; **p<0.05; ***p<0.01. Release season is Dec/Jan and rest of the months are considered as non-holiday season. Release Language Category is either English or Non-English movies.		

We can also observe from Table 1 that model 3 has the highest adjusted R-squared value of 0.67, indicating that it is the best-fitting model out of the three. Additionally, the results of anova test of the three models also reflects the same as the residual sum of squares is low for model 3 as compared to other models.

The coefficient of the X-concept variable "Budget" is having higher statistical significance in all three models. Point estimates range from 0.91 to 0.77, showing that the variable budget has a positive relationship between the other covariates and the Revenue. It is also evident from the table that all the covariates are statistically significant, having a p-value less than 0.05 (alpha level). The co-variates vote count, runtime, and release season show more statistical significance than other covariates. We also conducted a "Wald" test to study

the statistical significance of the two categorical covariates (Release Season and Release Language), and the results reassured us that they are also statistically significant variables in the model.

To better understand the practical significance of the results, let's consider a hypothetical use case with 180 minutes movie being produced with a million-dollar budget with a shorter title length of 10, the Model 3 shows that the Revenue of the movie could increase by 77.2%. Similarly, the model also shows that for every movie which is released in the English language, the revenue of the movie is expected to increase by approximately 15.6% keeping all the other covariates constant. Likewise, on average, the model predicts if a movie gets released during December/January, the revenue of the movie is expected to increase by 18.1%.

These results emphasize that budget is a crucial factor in determining the Revenue of a movie. It also indicates that along with the budget, the other covariates are also having significance in determining the revenue of a movie. But considering the statistical significance of popularity, movie production firms should not spend more money in promotions. Having all these into considerations, we believe that Model 3 can assist movie production companies in making informed decisions about the various factors involved in the overall planning process, which is essential for the success of the movie.

4 Limitations

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 the dataset cannot be representative of all movies that have been produced globally. The data may also be biased toward movies that are more popular than others.

The I.I.D assumption becomes questionable after closely examining the movie dataset because two movies may share the same cast & crew, and production company and have identical release dates.

We observed that the dataset has some omitted variables. For example, it does not mention any information about Motion Picture Association of America(MPAA) Ratings, which have the possibility of influencing the Revenue. Since MPAA rating are categorical, different MPAA ratings may generate different revenue. Here budget impacts the content of a movie which in-turn affects the rating of a movie, hence it is difficult to determine the direction of omitted variable bias accurately. Likewise, actors, a categorical variable, may influence both budget & Revenue of the movie where we may not be able to determine the direction of omitted variable bias.

The dataset may not represent a true picture of the current movie industry/trends since it contains information on movies which has a release date before 2013. Moreover, there is no consideration of inflation when concluding the statistical analysis between covariates such as budget & Revenue. More inflation leads to more budget & revenue and the direction of the bias is away from zero.

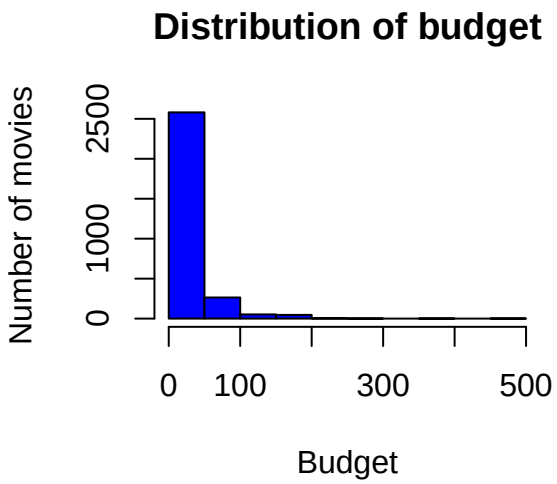
Also, in the model, we have an outcome variable, "Vote count" on the RHS. This variable can be an outcome of the predictor "Popularity". The reason could be because of promotions, more people can engage with the movie, which could lead to an increase in the number of votes.

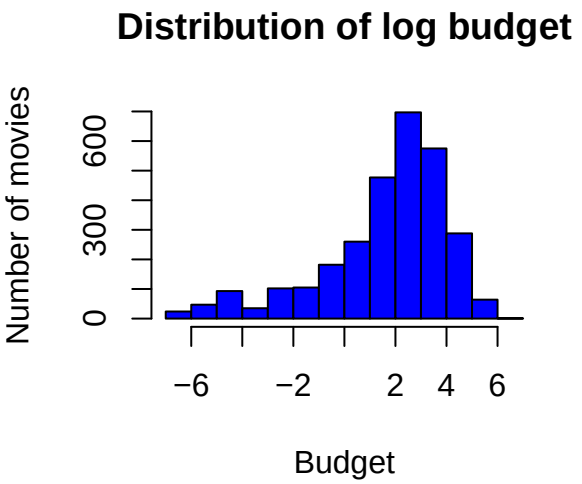
5 Conclusion

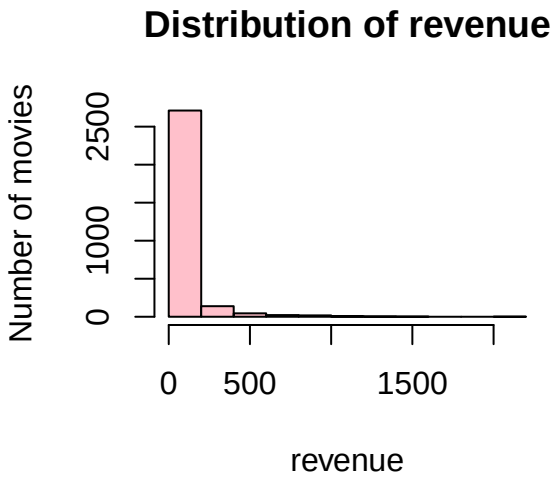
This study estimated the economic value of movie production, specifically examining the relationship between movie budget and Revenue. We also found that several covariates, such as vote_count, runtime, popularity, title length, release season, and release language, have a significant impact on movie revenue. Additionally, we identified the holiday season and language as other important categorical factors that can impact the success of a movie. We hope 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 work may explore the correlation of Revenue with additional data like MPAA ratings, cast/crew information, and sequel information on previous releases for identifying the latest trends in the movie industry.

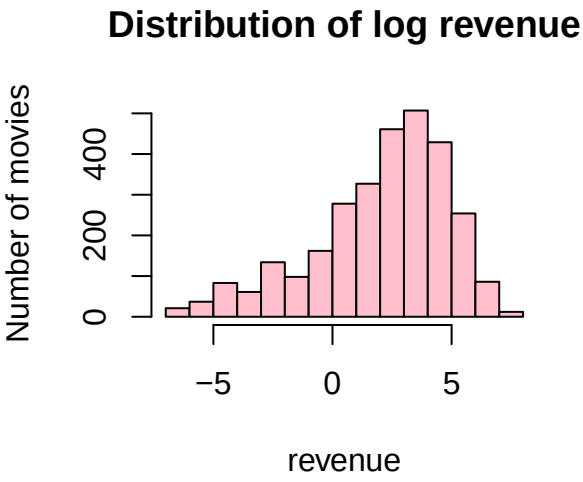
6 Appendix

6.1 Exploratory Analysis









6.2 Homoscedasticity validation

From below plot, we see that the red line on the plot is not horizontal and also the spread of the residuals is not equal at all the fitted values, highlighting evidences for Heteroscedasticity.

We have also used Breusch-Pagan test to validate homoscedasticity. The null hypothesis(H0) of the Breusch-Pagan test is that the residuals are Homoscedastic.The alternate hypothesis(Ha) is that the residuals are heteroscedastic. If the p-value from the test is less than the chosen significance level (e.g., 0.05), we can

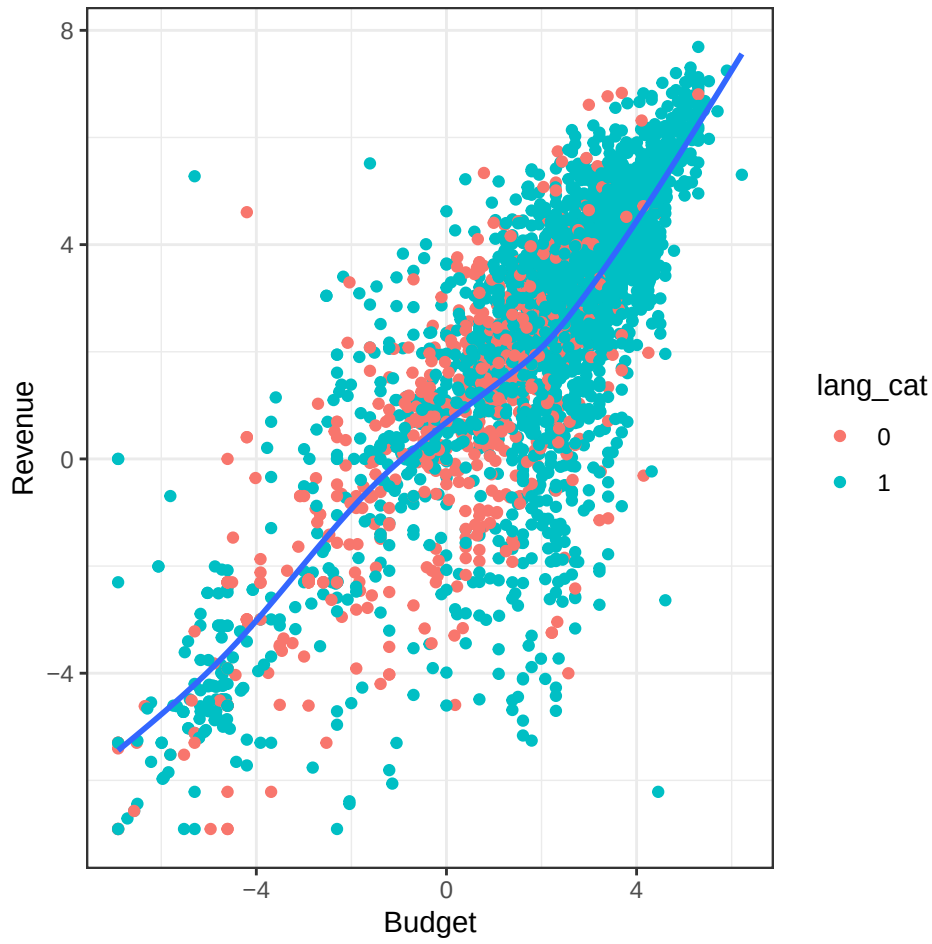


Figure 1: Revenue as a function of Budget with language category.

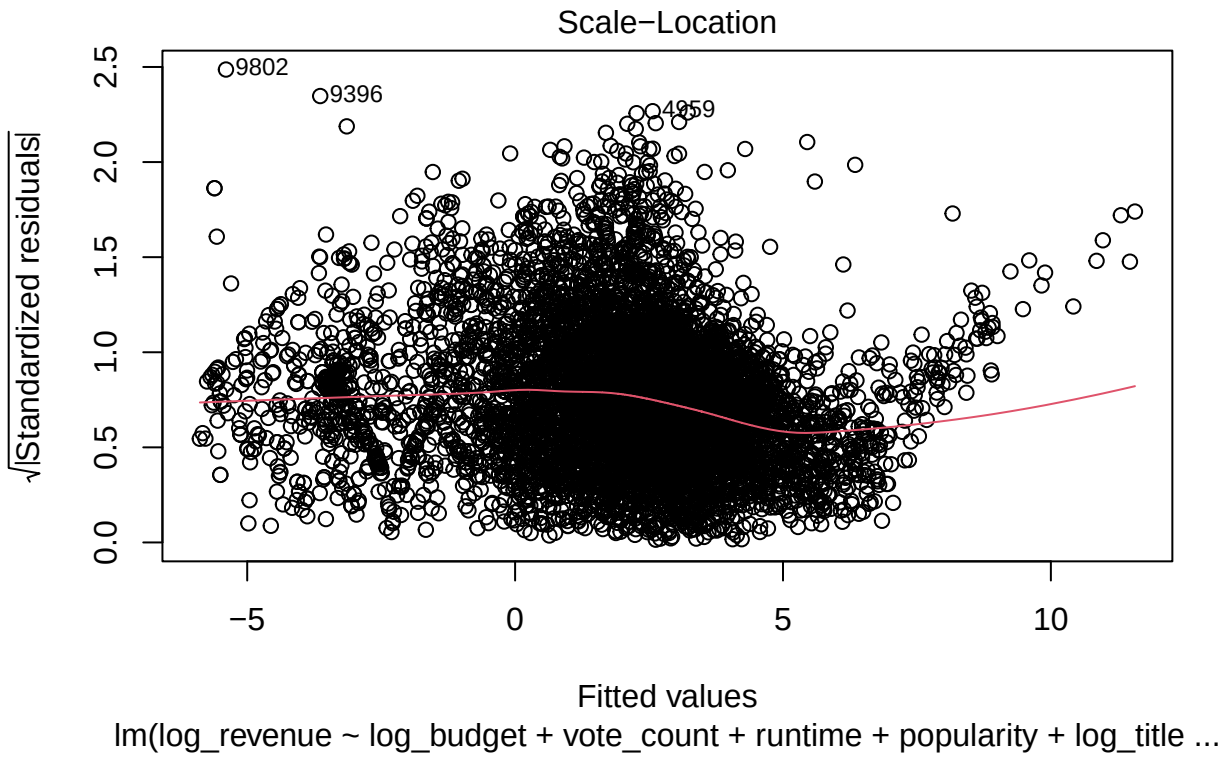


Figure 2: Scale-Location plot for Homoscedasticity validation

reject the null hypothesis of homoscedasticity. From the bptest results, it is seen that the p-value (3.422e-12) is less than 0.05 and we can reject the null hypothesis and conclude that there is strong evidence of heteroscedasticity in the linear regression model.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 68.193, df = 7, p-value = 3.422e-12
```

6.3 t-test

Below is the results of t-test highlighting the statistical significance of each and every predictor in the selected model (Model 3 from stargazer table). Since the model is of heteroscedastic, robust standard errors are used as an effective solution to handle it.

```
##
## t test of coefficients:
##
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)       -8.0015e-01  1.2792e-01 -6.2553 4.205e-10 ***
## log_budget         7.7181e-01  9.9935e-03 77.2314 < 2.2e-16 ***
## vote_count         2.0712e-04  9.1937e-06 22.5287 < 2.2e-16 ***
## runtime            7.5076e-03  6.7953e-04 11.0483 < 2.2e-16 ***
## popularity         7.5792e-04  2.9229e-04  2.5930 0.0095344 **
## log_title_length   1.0723e-01  3.5811e-02  2.9943 0.0027610 **
## release_date_cat1  1.8139e-01  4.9100e-02  3.6943 0.0002222 ***
## lang_cat1          1.5601e-01  5.1479e-02  3.0306 0.0024501 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.4 Anova test

We have conducted anova test to find out best fitting model. Below results of anova test of the three models shows that model 3 is fitting better than other two models as it has significant F-statistic value along with lower residual sum of squares, low residual degrees of freedom and low p-value.

```
## Analysis of Variance Table
##
## Model 1: log_revenue ~ log_budget
## Model 2: log_revenue ~ log_budget + vote_count + runtime
## Model 3: log_revenue ~ log_budget + vote_count + runtime + popularity +
##     log_title_length + release_date_cat + lang_cat
##   Res.Df   RSS Df Sum of Sq       F    Pr(>F)
## 1   6878 20682
## 2   6876 18142  2    2540.2 483.998 < 2.2e-16 ***
## 3   6872 18033  4     108.5  10.337 2.408e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.5 Wald test

```
## Wald test
##
## Model 1: log_revenue ~ log_budget + vote_count + popularity + runtime +
##     log_title_length + release_date_cat
## Model 2: log_revenue ~ log_budget + vote_count + runtime + popularity +
##     log_title_length + release_date_cat + lang_cat
##   Res.Df Df      F  Pr(>F)
## 1   6873
## 2   6872  1 9.1843 0.00245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Wald test
##
## Model 1: log_revenue ~ log_budget + vote_count + popularity + runtime +
##     log_title_length + lang_cat
## Model 2: log_revenue ~ log_budget + vote_count + runtime + popularity +
##     log_title_length + release_date_cat + lang_cat
##   Res.Df Df      F    Pr(>F)
## 1   6873
## 2   6872  1 13.648 0.0002222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```