

Introduction

In the ever-evolving landscape of the film industry, understanding the dynamics of a movie's financial success is pivotal for producers, investors, and stakeholders alike. This project aims to delve into the intricacies of what factors contribute to a movie's gross revenue, a question of considerable interest in an industry where financial outcomes are as varied as the films themselves. The primary purpose of this analysis is to unravel the relationship between a movie's gross revenue and a range of variables including the number of reviews, IMDb scores, social media popularity of the cast, film duration, budget, and content rating.

The relevance of this study stems from the changing paradigms of film success in the digital age. As streaming services rise and box office dynamics shift, it becomes crucial to reevaluate traditional metrics of success. The background of this topic is rich with studies focusing on limited aspects such as budget, star power, or critical acclaim, but there remains a gap in comprehensive, meta analysis.

Unlike previous research (Ahmed, 2020; Eklund, 2022; Simonoff, 2000), this project employs a linear regression model to quantitatively assess how these variables collectively influence gross revenue. This analysis differs from the background literature in its holistic approach, considering a wide spectrum of variables simultaneously, and providing a more detailed picture of the complex mechanisms behind a movie's financial performance. By bridging these gaps, this study aims to offer new insights into the film industry, contributing to a deeper understanding of what drives a movie's financial success in today's multifaceted entertainment landscape. This project's significance also lies in its potential to inform strategic decisions in film production and marketing, equipping stakeholders, producers, directors, etc. with a more robust framework for predicting and understanding movie revenue in the modern era.

Methods

We start by importing the movie profit dataset into R Studio, reading its summary. After that, we divide the data into two equal parts - 50% for training data and 50% for testing data. During the analysis process, we will use the training data. We then clean the data by removing any "NA" values and deleting incomplete data entries. Our research goal is to investigate the factors that affect a movie's gross revenue. Therefore, we select several variables (including duration, leading actor's Facebook likes, budget, IMDB score, all actors' total Facebook likes, number of critic reviews, and content rating) as predictors. We will use gross as the response variable based on prior literature and our research interest.

When fitting a model, we first check whether the model satisfies conditions. Condition 1 is a conditional mean response; we use a scatter plot of response versus fitted values to see if there is a random diagonal scatter to ensure condition 1 is satisfied. Condition 2 is conditional mean predictors; we use a pairwise scatter plot of predictors to check whether there is any curve or non-linear pattern. If it shows a curve or non-linear pattern, we drop the predictor that has a strong linear correlation with the other predictor.

Next, we check assumptions: uncorrelated error, linearity, constant variance, and normality. We use scatter plots (residuals vs. fitted values and residuals vs. predictors) to check uncorrelated error, linearity, and constant variance. If the plot has large clusters of points, it means that the uncorrelated error assumption is violated, hence the model is not appropriate. Moreover, if the plot has a fanning pattern, then we know that we don't have constant variance, and variance stabilizing transformations will be used. In addition, if the plot exhibits any systematic pattern, especially curves or other functions of predictors, we know that linearity is violated. Lastly, to check normality, we use the QQ plot. If there is a stark deviation/curving/wiggling from the diagonal line, we know that normality is violated. For both violations of linearity and normality, we can use the appropriate Box-Cox transformation. After applying transformation, we recheck the assumptions on the new model.

We then check for multicollinearity by using variance inflation factor (VIF). If predictors have VIF value greater than 5, it means a severe multicollinearity and could lead to incorrect conclusions regarding significance. So, we remove these predictors and recheck the VIF values in the new model to ensure they're below 5.

After checking the multicollinearity, we continue to check the problematic observations. We check for leverage, outliers, and all 3 kinds of influence (Cook's D, DFFITS, DFBETAS). As for cutoffs, we use $2[(p+1)/n]$ for leverage, $[-4, 4]$ for outlier, median of $F(p+1, n-p-1)$ for Cook's D, $2[(p+1)/n]^{1/2}$ for $|DFFITS|$, and $2/(n)^{1/2}$ for $|DFBETAS|$.

Next, we move on to model selection. We first use the ANOVA test to check the overall significance of our model. Then, we perform the t-test on all the predictors. We remove the predictors whose p-values are greater than 0.05. Additionally, We use partial F-test and numerical measures of goodness like adjusted R^2 , AIC, and BIC, to further verify that dropping those predictors is appropriate.

Finally, we validate our model by checking how our model performs in the testing dataset. We compare model components between test dataset and training dataset, including: minimal differences (< 2 s.e.'s) in estimated coefficients, same significant predictors, similar adjusted R^2 , no additional or worsening model violations, similar numbers and types of problematic observations, and similar amount of multicollinearity. If all these characteristics look similar in both datasets, then we can conclude that the model is validated.

Result

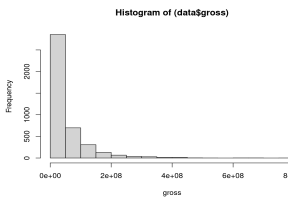
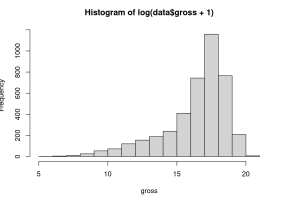
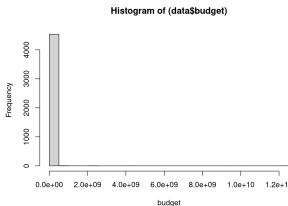
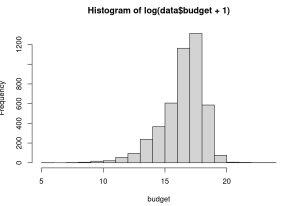
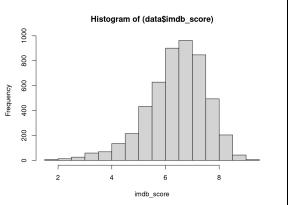
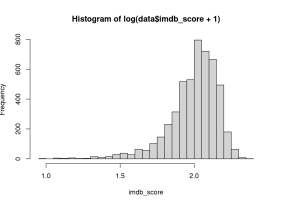
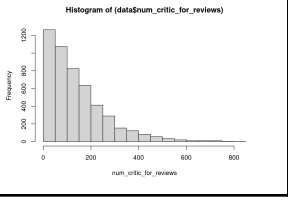
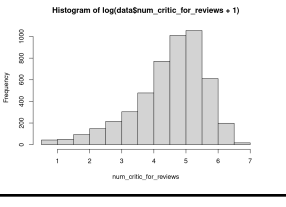
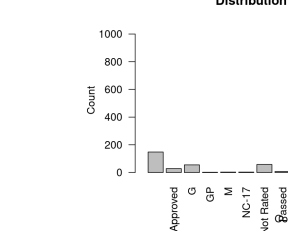
Data Summary

After our comprehensive analysis, there were four predictors in the final model, and their relationship can be explained by the following equation.

$$Gross^{1/4} = -115.67107 + 0.06460 \times imdb_score^{2.6} + 8.56970 \times \log(budget) + 5.72736 \times num_critic_for_reviews^{1/3} + 14.45511 \times content_ratingPG + 8.37339 \times content_ratingPG - 13 - 0.57652 \times content_ratingR$$

The following table summarizes the four predictors' name, type, their distribution, and why they should be included. The distribution of the data after log transformation is also included.

Table 1: Variable summary table

Variable Name	Type	Distribution	Distribution of the log of variable	Justification to use
Gross (dollars)	numerical			Investigate what factors might affect the gross.
budget (dollars)	numerical			Large budget of a movie is a predictor for success, and a way to prevent loss. Budget and gross are positively related.
imdb_score	numerical			IMDB scores are highly recognized by the public, representing the quality of movies, and audience's satisfaction.
num_critic_for_reviews	numerical			Critics have significant influence. Many audiences depend on critics to gauge the nature of the film.
Content Rating	categorical			The content rating (MPAA) influences box office gross. Well-targeted films fit certain market segments.

Model Selection

In the preliminary selection, we selected some predictors over others based on contextual implications. For instance, *genres* would be something meaningless to study, as each can refer to a wide variety of movies showing limited validity.

We used Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) to compare different models. By comparing the AIC and BIC of all five models, we find the latest one highest in both dimensions.

In the ANOVA test, the overall p-value is much smaller than 0.05, while some predictors showed much larger p-values in their separate t-tests, including *duration* and *actor_1_facebook_likes*. We then excluded the variables in our following models. Comparing the model with and without *duration* by partial F test, we maintained the decision as the predictor contributes little to the regression model.

In the Box-Cox transformation, the hypothesis showed that we should conduct boxcox (log) transformations for some variables but not all of them. Based on the range of power, we conducted a log transformation for the variable *budget*, as its range of power is closest to 0.

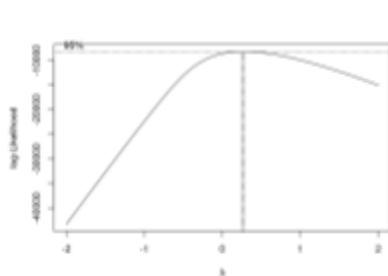


Figure 1. Log-Likelihood Profile of λ with 95% Confidence Interval

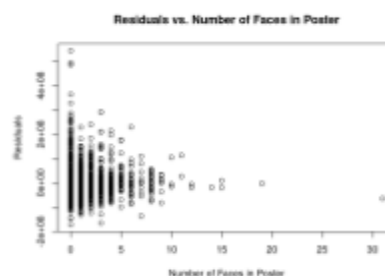


Figure 2. Scatterplot of Residual vs number of faces

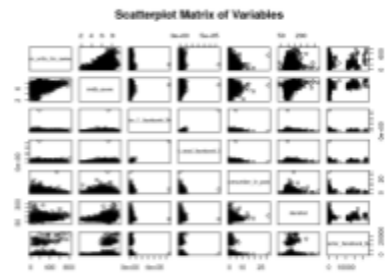


Figure 3. Scatterplot matrix of predictor variables

Model Assessment

We checked the conditions and assumptions a total of three times, two times before the model selection and validation procedure, and once after. In the first round, as a violation of the constant variance assumptions was found for the predictor *faces_in_poster*, we removed the predictor from the model (figure 2).

When checking the conditions of collinearity, we found the linearity between *num_critic_for_reviews* and *imdb_score* are (figure 3). Yet, when checking multicollinearity with VIF, neither of the variables is deviant. GVIF of both *actor_1_facebook_likes* and *cast_total_facebook_likes* is abnormally high. We eventually kept *actor_1_facebook_likes* and removed *cast_total_facebook_likes* as the former one predicts the movie gross better (figure 6).

When conducting problematic observation checks, we looked up the leverage, outlier, and influential points. We found 6 outlier observations, but as they do not impact the regression or hold contextual meaning, we did nothing but acknowledge it. For the leverage points, we found 141 observations matching the criteria. But considering the skewed distribution of movie gross, it's still understandable. The same is also true for influential points.

Model Validation

In the final validation process, we validated our model by checking how our model performs in the testing dataset. All the estimated coefficients exhibit minimal differences (< 2 s.e.'s). They have the same significant predictors (despite the categorical variable *content_rating* being slightly different). we also didn't observe any additional or worsening model violations. In addition, adjusted R^2 , multicollinearity, and numbers and types of problematic observations are all similar between training and testing models. Since all characteristics look similar in both datasets, we conclude that our model is validated (Figure 5).

Discussion

Final model equation:

$$Gross^{1/4} = -115.67107 + 0.06460 \times imdb_score^{2.6} + 8.56970 \times \log(budget) + 5.72736 \times num_critic_for_reviews^{1/3} + 14.45511 \times content_ratingPG + 8.37339 \times content_ratingPG - 13 - 0.57652 \times content_ratingR$$

Referring back to the research question: “How do the number of reviews, IMDb scores, leading actors’ Facebook likes, total casts’ Facebook likes, duration, budget, and content rating of a movie relate to its gross revenue? ”, it is obvious that the content rating as PG and PG-13, budget, numbers of movie critics, and IMDb scores, have a highly positive correlation with the movie gross. Compared to the result of the study from Eklund, our model shows consistency in the influence of budget and content rating in the movie gross and examines the correlation between celebrity popularity’s impacts on the gross in the process of data analysis just like Simonoff’s study demonstrated(Eklund,2022; Simonoff, 2000). We have also included the movie critics in our model which is a future analysis application mentioned in the Eklund study. To interpret this model, taking budget as an example, for each one unit increase of the log budget while maintaining other predictors as 0, the gross with fourth root decrease will increase by -107.10137. The model provides an overview of the components that contributed to the gross revenue and offers insight for the movie worker to produce movies with higher gross by considering the predictors, especially the content-rating mentioned in the final model.

Many limitations still exist in the final model. First, the final model still violates the assumption by observing the fanning pattern from the residual vs. fitted values scatterplot and clusters in the residuals vs. log(budget), implying non-constant variance and uncorrelated errors. These violations may be the result of the influential points. The categorical variables are difficult to transform as the transformation might interfere with the validity of the model.

Second, the data on actor popularity is based on Facebook likes, which is more-or-less outdated. Further studies can utilize data from more on-trend social media like Instagram and TikTok, which might be more reliable in assessing the relationship between the actor’s popularity and movie gross.

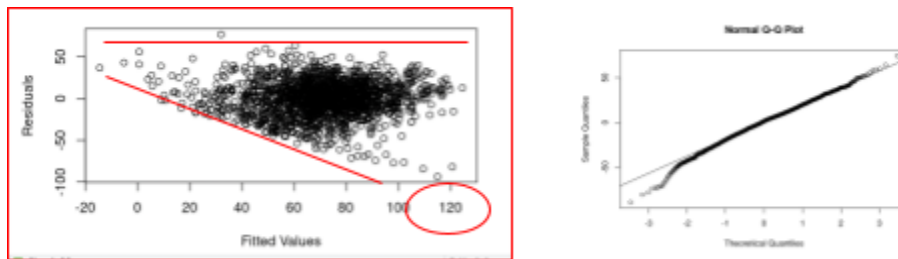


Figure 4. Assumption and condition checking for final model

Ethics Discussion

In constructing our model, we chose manual selection tools over automated methods. This decision was primarily due to the size and complexity of our dataset, which contains a large number of variables. Manual selection allows us to efficiently focus on variables of particular interest and better understand the relationships between specific predictors and the response variable.

Manual selection offers the advantage of incorporating contextual decision-making into the model construction process. It facilitates a more nuanced consideration of various aspects than automated tools. However, it also brings with it the risk of personal biases and subjectivity, potentially leading to oversights and imprudent decisions.

Conversely, automated selection tools can carry inherent biases in their algorithms. The design of these systems often lacks transparency, complicating the understanding of relationships between certain predictors and the response variable. This lack of clarity, combined with an over-reliance on these tools, can heighten the risk of negligence and errors.

It's important to recognize that both methods have their strengths and weaknesses. Manual selection, while more susceptible to human error, provides a level of flexibility and adaptability that automated systems lack. Automated tools, although efficient in handling large datasets, might miss out on nuanced relationships that manual inspection could reveal.

In conclusion, regardless of the selection method employed, it's imperative to engage in critical thinking and be aware of potential biases.

References

Ahmed, U., Waqas, H. & Afzal, M.T. Pre-production box-office success quotient forecasting. *Soft Comput* 24, 6635–6653 (2020). <https://doi.org/10.1007/s00500-019-04303-w>

Eklund, J.; Kim, J.-M. (2022). Examining Factors That Affect Movie Gross Using Gaussian Copula Marginal Regression. *Forecasting*, 4, 685–698.
<https://doi.org/10.3390/forecast4030037>

Simonoff, J. S., & Sparrow, I. R. (2000). Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. In *Chance* (New York) (Vol. 13, Issue 3, pp. 15–24). Taylor & Francis Group. <https://doi.org/10.1080/09332480.2000.10542216>

Appendix

Call:
lm(formula = t_gross ~ log(budget) + t_numcritic + t_imdb_score +
content_rating, data = train_2)

Residuals:
Min 1Q Median 3Q Max
-95.303 -12.652 0.629 13.494 85.517

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -115.67187 5.67774 -20.373 < 2e-16 ***
log(budget) 8.56970 0.35514 24.131 < 2e-16 ***
t_numcritic 5.72736 0.40992 13.972 < 2e-16 ***
t_imdb_score 0.06460 0.01028 6.285 4.05e-10 ***
content_ratingPG 14.45511 2.31148 6.254 4.92e-10 ***
content_ratingPG-13 8.37339 2.14294 3.907 9.65e-05 ***
content_ratingR -0.57652 2.02642 -0.285 0.776

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 1935 degrees of freedom
Multiple R-squared: 0.4903, Adjusted R-squared: 0.4887
F-statistic: 310.2 on 6 and 1935 DF, p-value: < 2.2e-16

Final model(trained)

Call:
lm(formula = t_gross ~ t_budget + t_numcritic + t_imdb_score +
content_rating, data = test)

Residuals:
Min 1Q Median 3Q Max
-93.671 -12.614 1.787 13.171 76.806

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -99.12723 6.12396 -16.187 < 2e-16 ***
t_budget 7.86335 0.37812 20.796 < 2e-16 ***
t_numcritic 6.55978 0.42783 15.333 < 2e-16 ***
t_imdb_score 0.04542 0.01033 4.398 1.15e-05 ***
content_ratingPG 10.12756 2.49378 4.061 5.08e-05 ***
content_ratingPG-13 1.97114 2.33465 0.844 0.39861
content_ratingR -7.64517 2.23834 -3.416 0.00065 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.77 on 1882 degrees of freedom
Multiple R-squared: 0.4586, Adjusted R-squared: 0.4569
F-statistic: 265.7 on 6 and 1882 DF, p-value: < 2.2e-16

Final model(tested)

Figure 5. Final model in trained and tested data sets

```
##              GVIF Df GVIF^(1/(2*Df))
## log(budget)    1.581145 1      1.257436
## t_numcritic    1.535617 1      1.239200
## t_duration     1.465433 1      1.210551
## t_imdb_score   1.551513 1      1.245597
## actor_1_facebook_likes 11.316400 1      3.363986
## cast_total_facebook_likes 11.785148 1      3.432950
## content_rating  1.310009 3      1.046034
```

Figure 6. VIF values