

Factors Effecting Box Office Success

Andrew Amore

2022-10-14

Abstract

Movie production companies are interested in understanding variables contributing to a successful film, particularly whether film budget and IMBD score are predictive of net profits. Box office data from all 2019 film releases was collected and enhanced with additional features to address this question.

Initially, a linear model was used to explore response associations and establish a baseline. The initial model displayed poor fit according to the residual standard error distribution, warranting a more complicated analysis. Several hierarchical models were used to address heteroscedastic variance concerns and ultimately a mixed effect model was applied to observations based on release month, which enhanced estimation accuracy for most fixed effects.

Final estimates for ‘Budget’ and ‘Critic.Score’ have statistically significant positive effects on profitability at the 95% level. On average, every \$1 increase in budget yields an additional profit of \$2.43 and improving the IMBD score by one point leads to a net profit increase of ~\$45 million. In addition, augmented field ‘Title.Sentiment’, defined as the polarity score of the film’s title, showed suprising significance, indicating a ~\$97 million net profit increase per unit gain of sentiment score.

For production companies, it’s not always feasible to create high budget, critically acclaimed films, but they can control the naming. One actionable recommendation is to title movies positively, as audiences seem more inclined to purchase tickets to these films.

Dataset Overview & Issues

Metadata Information

All fields used in the analysis are shown in Table 1 with metadata descriptions and a sample observation. Fields highlighted in yellow are augmentations.

Table 1: Metadata Information

| Field Name | Description | Sample |
|--------------------|--|---|
| Title | film's title | <i>Isn't It Romance</i> |
| Release.Date | date of movie release | 2019-02-13 |
| Production.Company | name of the movie production company | Warner Bros. Pictures / New Line Cinema / Netflix / Bron Creative |
| Lead.Cast.1 | leading actor/actress in the film | Rebel Wilson |
| Lead.Cast.2 | second leading actor/actress in the film | Liam Hemsworth |
| Lead.Cast.3 | third leading actor/actress in the film | Adam DeVine |
| Director | director of the film | Todd Strauss-Schulson |
| Box.Office | amount (USD) film made at the box office | 48800000 |
| Budget | budget (USD) of film | 31000000 |
| Run.Time | run time (minutes) of film | 89 |
| Critic.Score | IDMB score of film | 5.9 |
| Genre | full genre | Romance/Comedy |
| Genre.1 | main genre of film | Romance |
| Genre.2 | secondary genre of film | Comedy |
| Num.Genres | number of genres of film | 2 |
| Net.Profit | profit (USD) of film | 17800000 |
| Title.Sentiment | polarity score of movie title | -0.289 |
| Day.of.Week | release day | Wednesday |
| Release.Month | release month | February |

Missing Data

Exploratory data analysis (EDA) revealed a large number of missing data across a subset of fields. Figure 1 shows the five fields with missing data. Over a third of movie releases are missing “Net.Profit”...the response variable we want to investigate! Clearly something must be done to address this issue. Bayesian imputation methods were investigated to “draw” missing data from candidate probability distributions derived from available data, but due to the large proportion of missing data in the response variable this strategy was abandoned in favor of dropping affected observations. This avoids adding any statistical bias from an imputation method, however, this increases variability of results. Before proceeding several additional data quality issues were discovered.

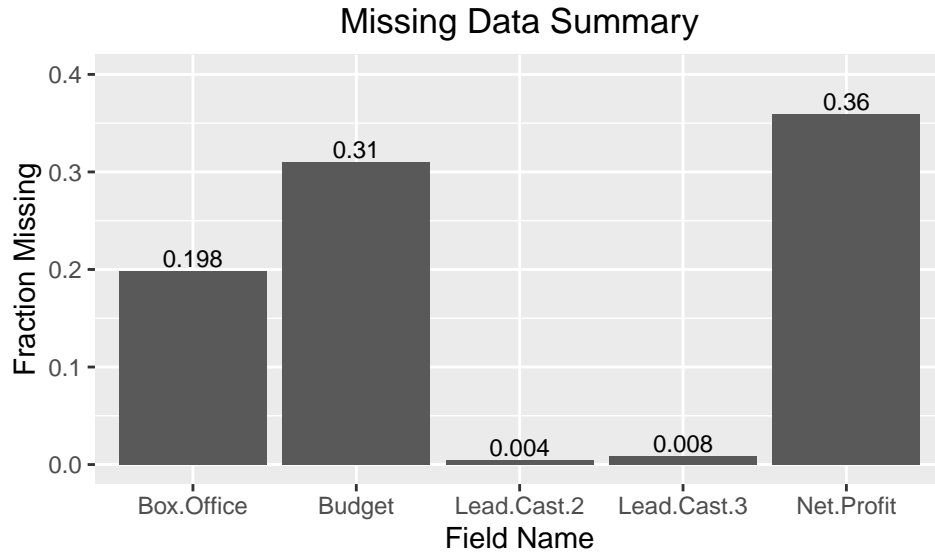


Figure 1: Missing Data Summary

Limited Releases: There are some additional concerns with the data. First, we have no information about how box office metrics were collected, such as release duration or the number of screening theaters. Films debuting as limited releases have fewer opportunities to generate box office profits and have skewed revenue numbers. A known example of a limited release is “The Irishman”, shown below. At first glance the film appears to have lost over \$100 million, but additional research indicates the film only released at eight theaters nationally. Because of this concern, all Netflix films were dropped from analysis except “Isn’t it Romance” as secondary research confirmed a full theatrical release.

Table 2: Outlier Sample

| Title | Production.Company | Box.Office | Budget | Net.Profit | Removal.Reason |
|-------------------|--|---------------|-------------|---------------|-----------------|
| Avengers: Endgame | Marvel Studios | 2,798,000,000 | 365,000,000 | 2,433,000,000 | Outlier |
| The Lion King | Walt Disney Pictures | 1,663,000,000 | 250,000,000 | 1,413,000,000 | Outlier |
| The Irishman | Netflix / TriBeCa Productions | 8,000,000 | 159,000,000 | -151,000,000 | Limited Release |
| Frozen II | Walt Disney Pictures / Walt Disney Animation Studios | 1,450,000,000 | 150,000,000 | 1,300,000,000 | Outlier |

Outliers: The dataset also contains a number of outliers which can lead to large leverage points and/or violate modeling assumptions. One such example is “Avengers: Endgame”, which released as the final installment to a series of films from the past decade to much fanfare and the box office numbers are almost twice as large as the next closest film. It’s reasonable to assume that many people were motivated by a different set of factors to watch this film and as a result it should be included in the analysis. Other “sequels” with extraneous factors were also removed (Lion King & Frozen II) for the same reason.

Variable Selection

The main objective of this analysis is to investigate factors influencing ‘Net.Profit’. Before beginning to build a model the covariance structure of the dataset was investigated for films with no missing values. Table 3 displays the correlation coefficients for numeric fields and indicates mostly weak relationships (correlation < 0.5) with the response.

Unsurprisingly, ‘Box.Office’ and ‘Budget’ show the strongest relationship. Figure 2 shows the relationship between the response and the two main covariates of interest: budget and critical score.

Table 3: Correlation Matrix

| | Box.Office | Budget | Run.Time | Critic.Score | Num.Genres | Net.Profit | Title.Sentiment |
|-----------------|------------|--------|----------|--------------|------------|------------|-----------------|
| Box.Office | 1 | 0.77 | 0.27 | 0.3 | −0.04 | 0.99 | 0.06 |
| Budget | 0.77 | 1 | 0.28 | 0.14 | −0.05 | 0.65 | −0.09 |
| Run.Time | 0.27 | 0.28 | 1 | 0.47 | −0.06 | 0.25 | 0.07 |
| Critic.Score | 0.3 | 0.14 | 0.47 | 1 | 0.03 | 0.32 | 0.18 |
| Num.Genres | −0.04 | −0.05 | −0.06 | 0.03 | 1 | −0.03 | −0.04 |
| Net.Profit | 0.99 | 0.65 | 0.25 | 0.32 | −0.03 | 1 | 0.1 |
| Title.Sentiment | 0.06 | −0.09 | 0.07 | 0.18 | −0.04 | 0.1 | 1 |

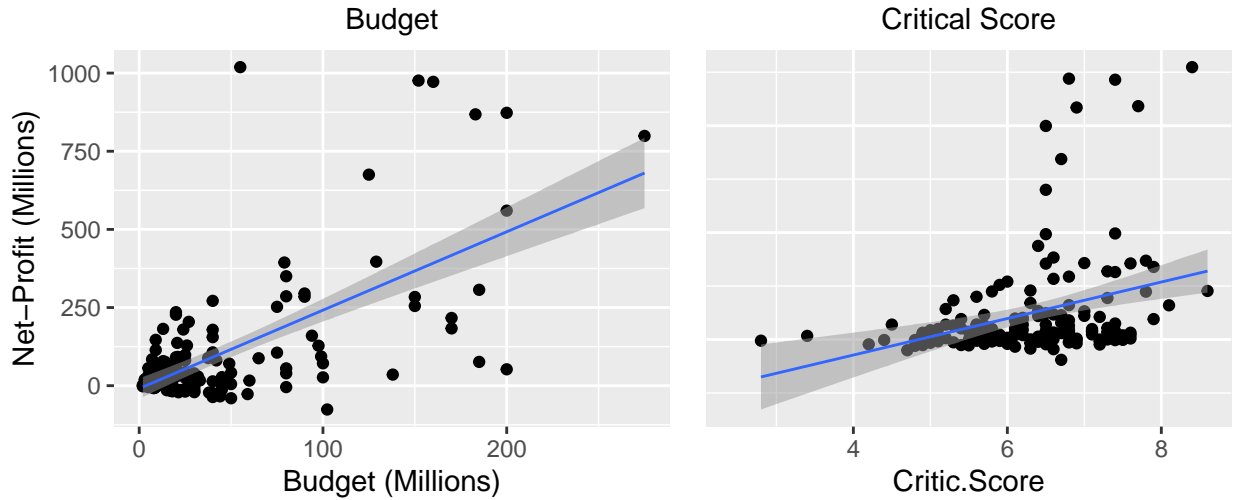


Figure 2: Relationship with Net-Profit

Model Building

Standard Regression

Before more complex modeling, a baseline linear regression model was fit on the data. Let $y \in \mathbb{R}^n$ be a vector of movie net profits, $X \in \mathbb{R}^{n \times 2}$ be a matrix of covariates (‘Budget’ & ‘Critic.Score’), and β_0, β be regression coefficients and intercept respectively. The model can be formalized as follows:

$$y = \beta_0 + X\beta + \epsilon, \text{ where } \epsilon_{iid} \sim N(0, \sigma^2 I)$$

The results of the model fit including coefficient estimates, uncertainty quantification via 95% confidence interval and p-values are shown in Table 4. All coefficients, including the intercept are shown to be statistically significant, but standard errors (SE) are large for Critic.Score and intercept.

Table 4: Initial Results

| Coefficient | Estimate (USD) | SE (USD) | P-Value | CI Lower (2.5%) | CI Upper (97.5%) |
|--------------|----------------|------------|----------|-----------------|------------------|
| (Intercept) | −315,071,394 | 82,493,418 | 0.000197 | −478,106,904 | −152,035,885 |
| Budget | 2.38 | 0.233 | 7.71e−19 | 1.92 | 2.84 |
| Critic.Score | 49,072,204 | 12,985,672 | 0.000229 | 23,408,027 | 74,736,381 |

| Observations | Residual Std. Error | R ² | Adjusted R ² |
|--------------|---------------------|----------------|-------------------------|
| 149 | 147,033,657 | 0.476 | 0.469 |

Based on the high SEs, the model fit was assessed. The scale location plot is the most interesting as it addresses the large differences in film net profits which were left as raw dollar amounts. Clearly, as the fitted value increases the standardized residual increases which is a violation of the linear model homoscedastic variance assumption. Before moving to a more complex modeling technique additional covariates were added to try and improve the baseline regression.

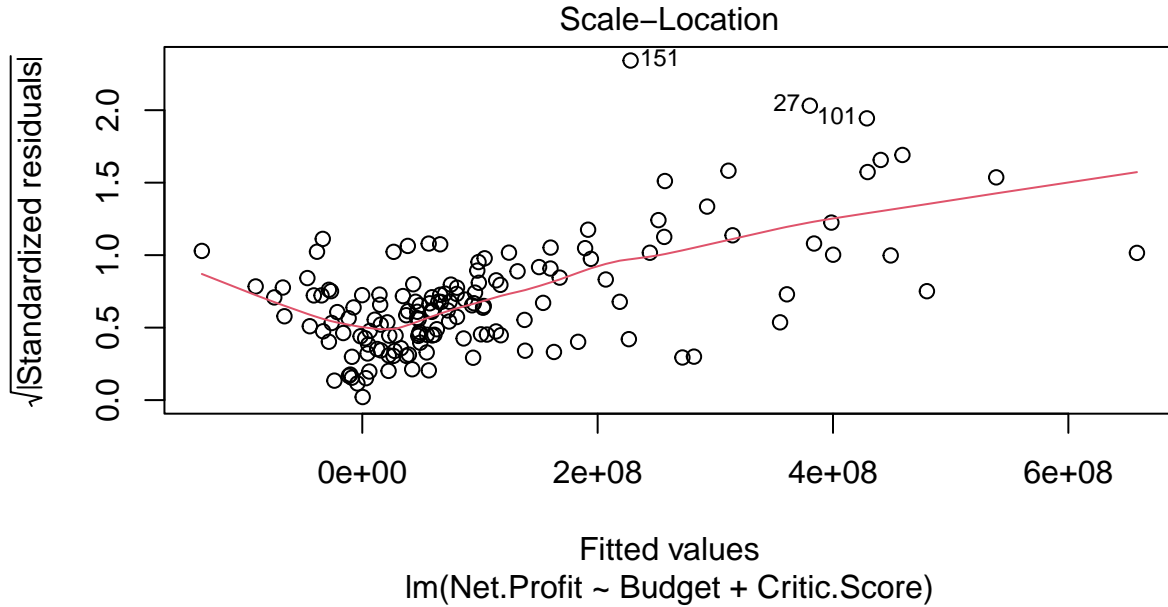


Figure 3: Standardized Residual Plot

Adding Additional Covariates to the Standard Regression

The correlation matrix in Table 3 suggests ‘Run.Time’ and ‘Title.Sentiment’ may be good predictors to add to the model. Models with each additional covariate were compared to the reference model using an F-test to determine inclusion criteria. In addition to the two numeric covariates, ‘Day.of.Week’ and ‘Genre.1’ were also investigated. F-test summaries are provided below and based on these results only ‘Title.Sentiment’ was added to the model as closer inspection of ‘Day.of.Week’ revealed a high variance estimate from Tuesday (only has three observations).

Table 5: F-Test Results

| Added Covariate | Df | RSS | RSS Reduction from Ref. | Test-Statistic (F) | P-Value |
|-----------------|-----|----------|-------------------------|--------------------|---------|
| Reference Model | 146 | 3.16e+18 | NA | NA | NA |
| Title.Sentiment | 145 | 3.08e+18 | 7.21e+16 | 3.390 | 0.0676 |
| Day.of.Week | 143 | 2.99e+18 | 1.68e+17 | 2.683 | 0.0491 |
| Run.Time | 145 | 3.15e+18 | 1.08e+16 | 0.496 | 0.4824 |
| Genre.1 | 117 | 2.31e+18 | 8.50e+17 | 1.487 | 0.0725 |

The final “baseline” linear regression model results are shown in Table 6.

Table 6: Final Baseline Reference Results

| Coefficient | Estimate (USD) | SE (USD) | P-Value | CI Lower (2.5%) | CI Upper (97.5%) |
|-----------------|----------------|------------|----------|-----------------|------------------|
| (Intercept) | -282,789,925 | 83,683,651 | 0.000934 | -448,187,277 | -117,392,573 |
| Budget | 2.43 | 0.232 | 2.09e-19 | 1.97 | 2.89 |
| Critic.Score | 44,256,400 | 13,143,548 | 0.000973 | 18,278,709 | 70,234,091 |
| Title.Sentiment | 93,913,069 | 51,006,166 | 0.0676 | -6,898,556 | 194,724,694 |

| Observations | Residual Std. Error | R ² | Adjusted R ² |
|--------------|---------------------|----------------|-------------------------|
| 149 | 145,844,747 | 0.488 | 0.478 |

This final model improves the residual standard error estimate from the reference model, but is still a relatively poor fit (high coefficient SEs). To try and reduce these values further hierarchical modeling is explored.

Hierarchical Modeling

To perform an adequate hierarchical analysis a “clustering” category is needed. At first glance, ‘Genre’ shows promise, but many films are recorded with multiple or unique even entries with limited overlap. Table 7 displays the number of films with multiple genres. The median film has more than one genre, making grouping more challenging.

Table 7: Genre Clustering

| Number of Genres by Film | Number of Observations |
|--------------------------|------------------------|
| 1 | 81 |
| 2 | 110 |
| 3 | 17 |
| 4 | 1 |

One could circumvent this by using the first listed entry as the primary genre and proceed with analysis, but there are still quite a few clusters with few observational data points shown in Figure 4. A proposed solution is to categorize similar genres into a hierarchy for analysis. For example, maybe “War” and “Action” involve enough similarity to warrant combining into one super-category, however, extensive domain knowledge may be required to infer less obvious groupings. This procedure might introduce statistical bias that could alter results.

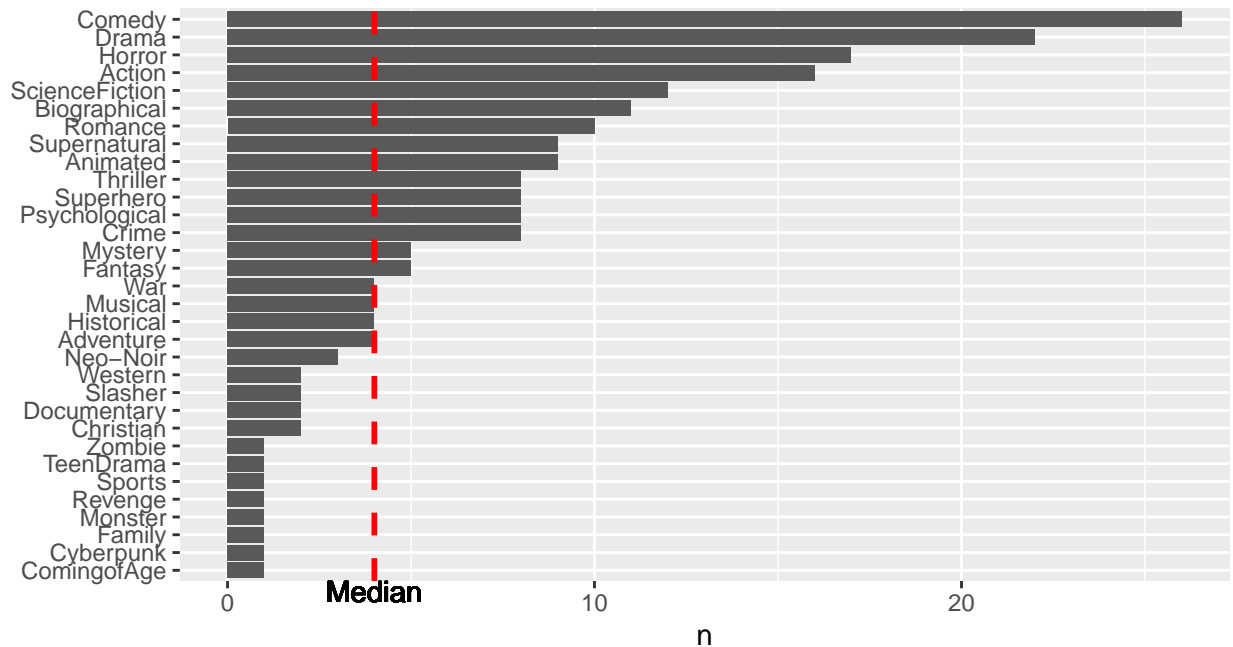


Figure 4: Count of Films by Genre.1

A more objective hierarchy is needed and one hypothesis is that audiences may be sensitive to the seasonality of a film’s release. For example, during the winter months individuals might be more likely to attend a new film because there are less alternative outdoor activities compared to other seasons. Production companies might alter a film’s budget accordingly. Figure 5 displays linear regression fits contrasting the group effect. Clearly, there are differences in Net.Profit trends across groups, warranting the hierarchical model.

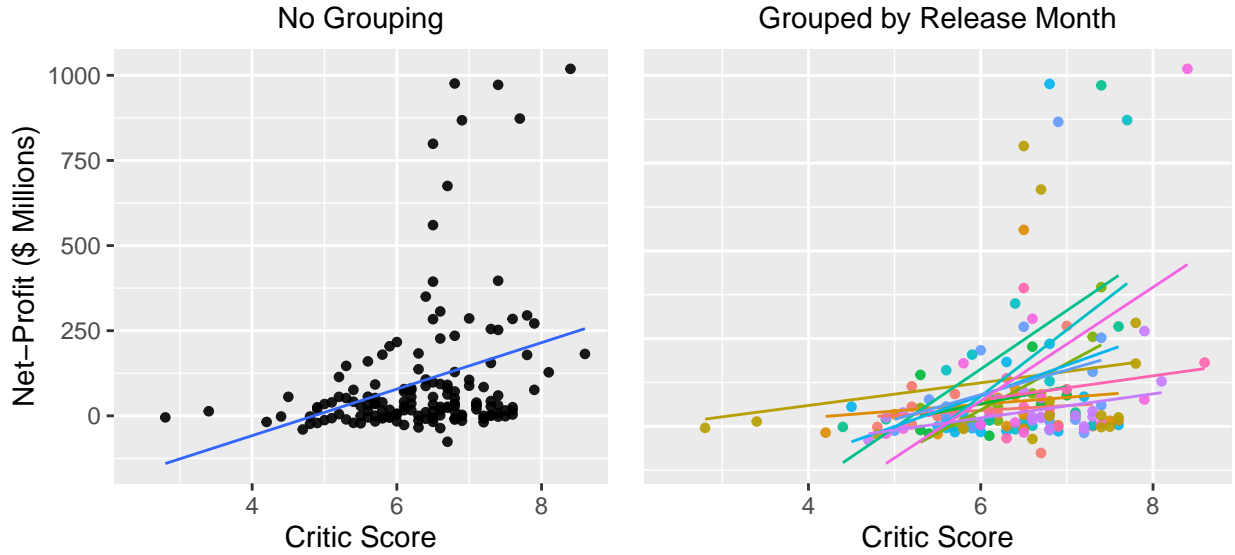


Figure 5: Motivating the Hierarchical Model

Several hierarchical models were tried with the data. First, a random intercept on release month was added. The model can be described using the same notation as the final reference linear model with new parameter $\alpha_j \in \mathbb{R}^{12 \times 1}$ denoting the random effect for release month j .

$$y = \beta_0 + \alpha_j + X\beta + \epsilon, \text{ where } \epsilon_{iid} \sim N(0, \sigma^2 I)$$

This model detects very little differences with the addition of the random effect and all parameter estimates are almost identical to the reference linear model.

Table 8: Mixed Model 1 Results

| Coefficient | Estimate (USD) | SE (USD) | P-Value | CI Lower (2.5%) | CI Upper (97.5%) |
|-----------------|----------------|------------|----------|-----------------|------------------|
| (Intercept) | -282,789,925 | 83,683,651 | 0.000727 | -445,638,767 | -1.2e+08 |
| Budget | 2.43 | 0.232 | 1.45e-25 | 1.98 | 2.88 |
| Critic.Score | 44,256,400 | 13,143,548 | 0.000759 | 18,678,984 | 69,833,816 |
| Title.Sentiment | 93,913,069 | 51,006,166 | 0.0656 | -5,345,209 | 193,171,348 |

| Coefficient | CI Lower (2.5%) | CI Upper (97.5%) |
|------------------------------|-----------------|------------------|
| sd_(Intercept) Release.Month | 0 | Inf |
| sigma | 128,969,424 | 161,893,468 |

Clearly the model doesn't detect any differences under the first model and ideally, we would like to add random slope terms for all covariates, but unfortunately, the small sample size can lead to boundary fits and singularity for too many parameters.

The final hierarchical model is defined with similar notation as follows. To avoid boundary conditions, the random intercept is removed and instead estimate the random slope for 'Critic.Score', $\gamma_{j[i]}$ the random effect for critic score of film c_i , in addition to all other parameters from the baseline reference linear model.

$$y_{ij} = \beta_0 + X\beta + \gamma_{j[i]}c_i + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2)$$

Accounting for this new parameter alters the parameter estimates compared to the baseline reference model, but does not dramatically change the coefficient SE estimates... a disappointment. Comparing Table 9 to Table 6 the estimate for 'Budget' is identical, but 'Critic.Score' and 'Title.Sentiment' see an increase with lower p-values that almost nudge 'Title.Sentiment' into a statistically significant effect at the 95% level.

Table 9: Mixed Model 2 Results

| Coefficient | Estimate (USD) | SE (USD) | P-Value | CI Lower (2.5%) | CI Upper (97.5%) |
|-----------------|----------------|------------|----------|-----------------|------------------|
| (Intercept) | -287,034,469 | 83,429,515 | 0.000584 | -448,679,088 | -121,223,595 |
| Budget | 2.43 | 0.232 | 1.23e-25 | 1.98 | 2.88 |
| Critic.Score | 45,088,054 | 13,131,958 | 0.000605 | 18,880,324 | 70,552,765 |
| Title.Sentiment | 96,888,305 | 50,903,239 | 0.0571 | -4,563,154 | 195,533,850 |

| Coefficient | CI Lower (2.5%) | CI Upper (97.5%) |
|-------------------------------|-----------------|------------------|
| sd_Critic.Score Release.Month | 0 | 8,369,437 |
| sigma | 127,953,194 | 161,739,097 |

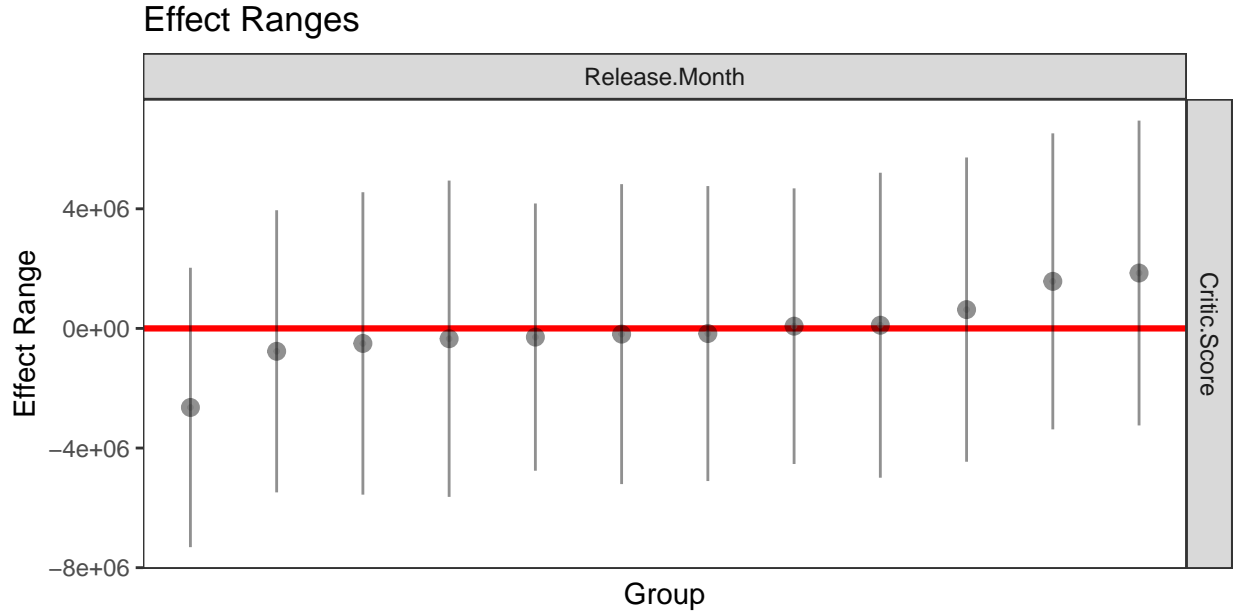


Figure 6: Interval Estimates for the Random Effect

Conclusions, Limitations & Enhancements

In this analysis factors affecting box office success were evaluated. Before modeling, covariates were cleaned, analyzed and enhanced to better understand dependence. Roughly 100 observations were dropped from the analysis related to data quality. Standard linear regression and mixed effect models were evaluated to assess the statistical significance of covariates on profitability. ‘Budget’ and ‘Critic.Score’ were found to have significant positive effects at the 95% confidence level, while ‘Title.Sentiment’ was found to be significant at the 90% confidence level, adding a surprising positive effect. However, it’s important to note that more analysis is needed to determine the causal relationship between ‘Title.Sentiment’ and net profitability as only associativity is assessed. In addition, there are several limitations of this analysis and the main concerns are addressed.

First, the sample contains a large fraction of missing data, is relatively small, and only contains a single year of observations which might not be representative of overall film releases. Second, no knowledge of the data collection procedure is provided and raises concerns about the use of overall profit as a metric which may be misleading. Outlier detection revealed several films with excessive losses that released limitedly, but more lower budget limited releases may be present. To combat this, one could code profitability as a binary response indicator, $\{0, 1\}$, and fit a logistic regression on the new dependent variable, minimizing the influence of any outliers and potentially make the analysis more robust. However, the ability to quantify direct covariate effect on overall profit would be lost, which feels relevant to any profit evaluation. Third, there is no obvious clustering field for a hierarchical model. Spending more time to manually group like genres may yield higher standard error reduction and more accurate coefficient estimates. Lastly, more analysis is needed to determine the causal relationship between ‘Title.Sentiment’ and net profitability as the current report analyzes association.