Increasing Box Office Ticket Revenue Report

Hanna Grossman and Sara Kien

Introduction

Background:

The motion picture industry has changed and evolved dramatically throughout the last century, with revenues continuing to grow over time. Although movie sales dipped during the Covid-19 pandemic, going to a movie remains one of the top 10 pastimes. Successful production companies have adapted to changes in technology and consumer interest by implementing different types of modifications to movie features, such as increases in motion and action within a shot. Starting in the mid-2000s, production budgets also began to rise. However, increases in average budget were more likely for some genres (e.g, adventure) than for others (e.g., horror).

Prior analyses have shown that the correlation between budget and revenue is strong, ranging from approximately .70 from 2000 to 2004 to approximately .80 from 2006 to 2017. These analyses also showed that this correlation has grown especially stronger over time for some genres, such as action. Other genres, such as drama, did not show this trend. Other analyses show that overall, revenues from 1995 to 2022 are highest for adventure (64.95 billion), action (50.56 billion), and drama genres (35.68 billion).

Current Project:

The purpose of the current project is to investigate changes in movie production that could be implemented by Acme Movies Inc to increase movie revenue. Specifically, the Data Science team at Acme Movies Inc will examine the influence of budget and genre on movie revenue. Although previous analyses have examined the relationship between budget and revenue as well as the relationship between genre and revenue, very few analyses have directly compared the influence budget and genre have on revenue.

Research Question:

The specific research question examined in the current project is: Is there an effect of production budget and genre on revenue?

Explanatory Models and Operationalizing Variables:

To investigate this research question, the Data Science team at Acme Movies Inc developed explanatory models using OLS regression to examine the influence of budget on revenue while holding genre constant. We also examined the specific influence of genre on revenue, holding budget constant. In a third model we controlled for the potential influence of original language on revenue. Budget is operationalized as US dollars spent on a specific film and revenue is operationalized as US dollars generated by sales of box office tickets for each movie. Budgets and revenues for movies filmed outside of the US were converted to US dollars. Three specific genres were selected for purposes of the present analysis: adventure, action, and drama. These genres were selected because they are associated with higher revenues compared to other genres. In the dataset that was used for the current project (described in more detail below), several movies were classified into multiple genres. For each movie, each of these genres was coded as either present or absent and it was possible for a movie to be classified as more than one genre type. Original language was coded as a binary variable (English vs. other).

Control Variables:

As mentioned above, movie production companies often manipulate multiple features to increase movie revenue. To examine the specific influence of budget and genre above and beyond other features, we considered other variables that might influence revenue. These features need to be "controlled" (i.e., held constant) in the model to examine the specific influence of budget and genre on revenue. One feature that was identified as a potential confound was "original language" of the movie. Prior analyses showed that movies that were filmed in English generated greater revenue in US dollars than movies filmed in other languages, although some movies filmed in other languages are highly profitable. For this reason, original language was included as a control variable in the present analysis.

Predictions:

It was predicted that budget would have an influence on revenue when holding genre constant, and that genre would have an influence on revenue when holding budget constant. It was predicted that these effects would be statistically significant, even when controlling for original language.

Description of the Data and Research Design

As stated above, the research question is: Is there an effect of production budget and genre on revenue? This research question is motivated by an explanatory research design in which we are investigating the causal influence of budget and genre on movie revenue.

The dataset used in this project included a subset of data from a larger Box Office dataset collected from TMDB (kaggle.com) that provides details about 3000 movies released from 1915 to 2017. There were 23 columns in the larger dataset, including information about movie budget, genres, title, popularity, release date, language, countries, production companies, runtime, tagline, cast, crew, and revenue.

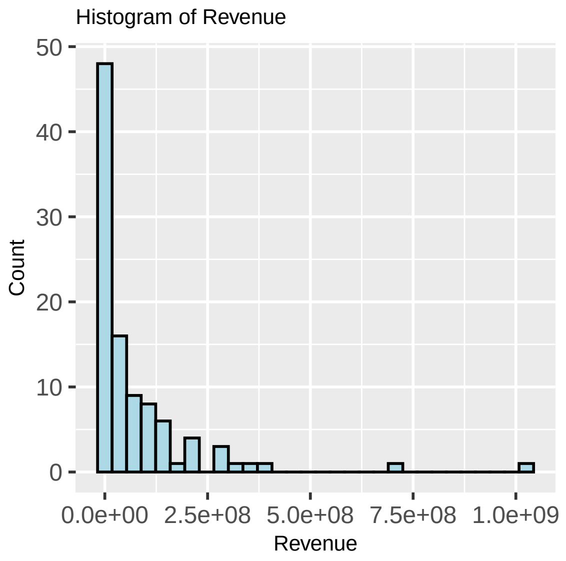
Movies with values of 0 for budget were excluded from the dataset. Several movies that were listed as having a budget of $0 included large-budget movies like Muppet Treasure Island, indicating that this was an inputting error in the original dataset. To avoid time-series dependencies within the data, the final dataset was restricted to movies that were released in 2010. This specific year was selected because the sample size was the largest for this year. There were 26 movies in the 2010 dataset that included the budget error described above (i.e., budget = $0). After removing those movies, the final dataset for 2010 consisted of 100 rows (n = 100).

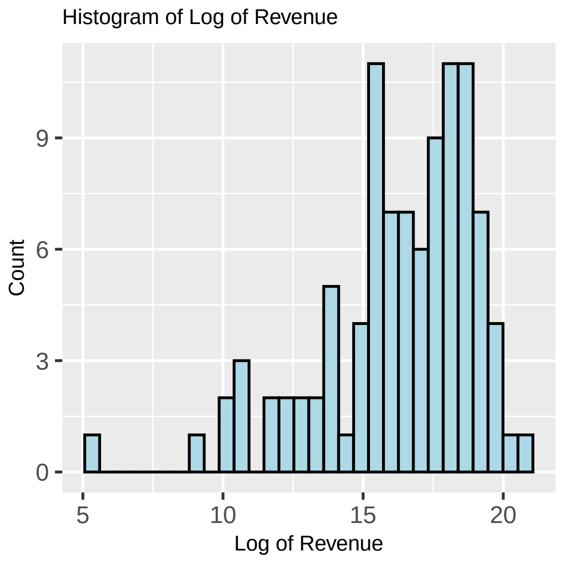
To investigate the research question, we conducted OLS regression analyses that compared three regression models. The outcome variable for all three models is movie revenue. The first regression model served as the reduced model and only included the movie budget as an input variable. In the second regression model, we included genre as a second input variable, specifically adventure (binary variable), action (binary variable), and drama (binary variable). We conducted an F test to compare the second model to the first model and to examine whether adding genre significantly reduces the mean squared residual (MSR). In the third model, we added "original language" as a control variable to examine whether English vs. other languages affect movie revenue. An F test was conducted to compare the third model to the second model. Popularity was excluded from the regression models based on the expectation that other input variables, such as budget, genre, and language, have a direct effect on popularity, indicating that popularity is actually an outcome variable.

Given the predictions associated with the three regression models outlined above, the final dataset was restricted to the following variables: Budget, Genre, Original Language, and Revenue.

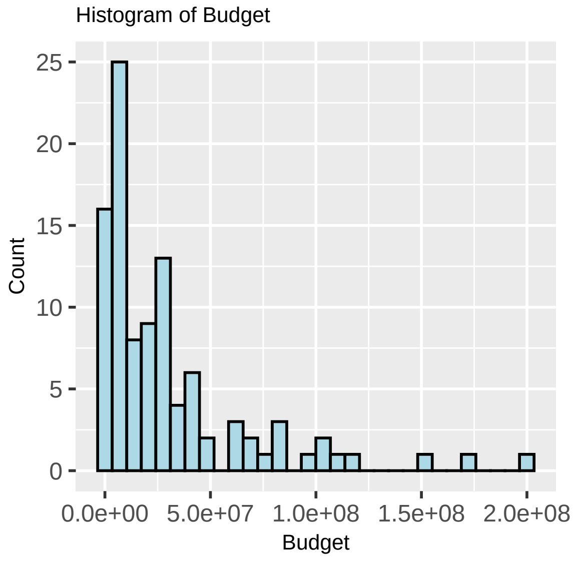
Model Building Process

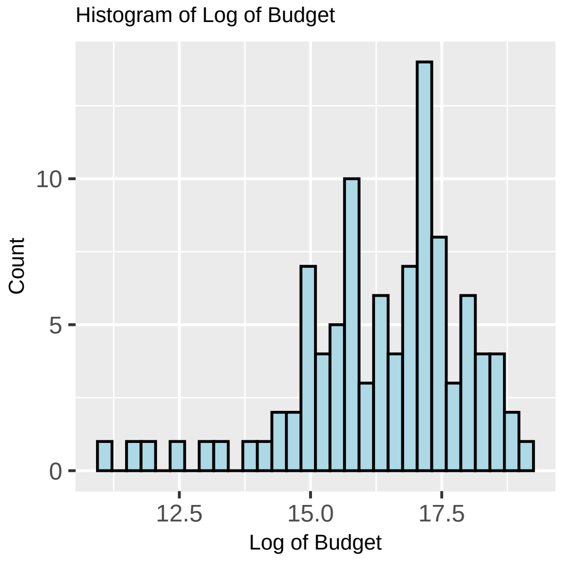
After identifying our key variables and three models of interest above, we began exploratory data analysis to gain a better understanding of our key variables. We see below that the histogram of revenue is strongly skewed right, and therefore we will used a log transformation of this variable in our three models.



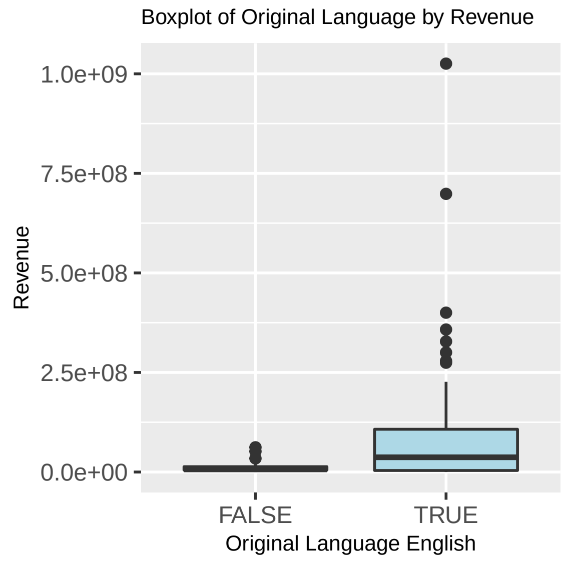


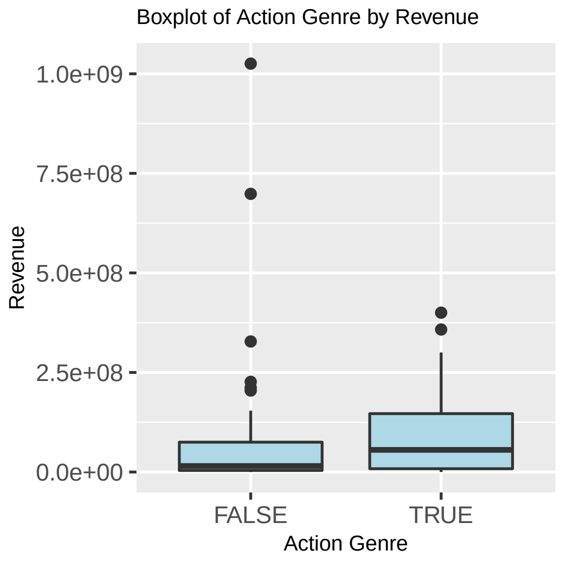
From there, we examined our x-variables, budget, genre, and original language. Budget is skewed right, but when we take the log of budget, the histogram becomes slightly skewed left. As this transformation does not seem to make a large difference, budget is an input variable, and one of our goals is to interpret our models, we chose to not transform budget in our models.

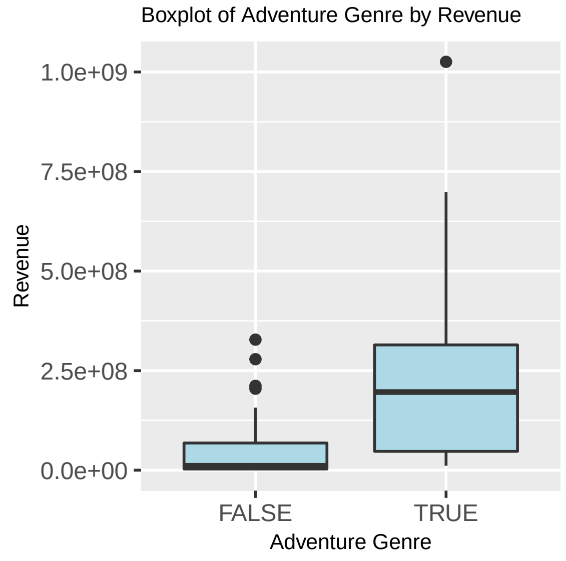


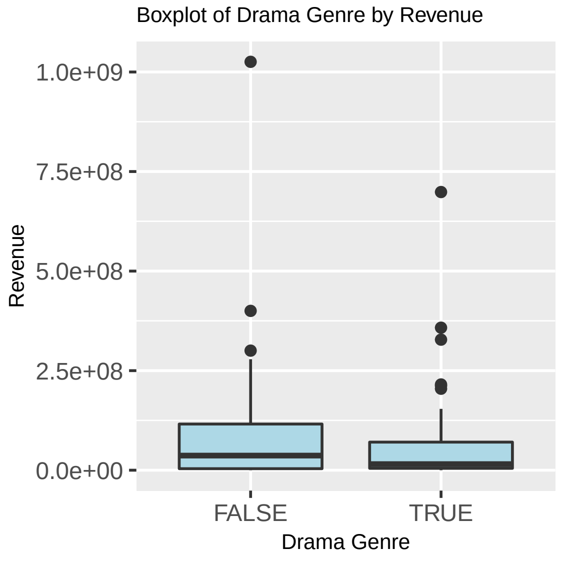


We coded the genre variable into three variables, action, adventure, and drama, with each having a value of true or false for each row. We similarly coded the original language variable as being true for English and false otherwise. As these are all binary variables, no transformations are needed. Please see the distributions for these four variables in the boxplots below.


Boxplot of Original Language by Revenue


Boxplot of Action Genre by Revenue


Boxplot of Adventure Genre by Revenue


Boxplot of Drama Genre by Revenue

From there, we ran our three models and will discuss the results in the next section:
Model 1: log(revenue) = β_0 + β_1 budget
Model 2: log(revenue) = β_0 + β_1 budget + β_2 action + β_3 adventure + β_4 drama
Model 3: log(revenue) = β_0 + β_1 budget + β_2 action + β_3 adventure + β_4 drama + β_5 original langugage english

Results

Please see the results of our three models in the table below. Also note that we used robust standard errors. We looked at the residual vs fitted plots for each of our three models, and the width of the band of data was not consistent from left to right on these plots, signifying that our data are not homoscedastic. This led us to use robust standard errors.

```
##
## =========================================================================================
##                                              Dependent variable:
##                      --------------------------------------------------------------------
##                                                  log(revenue)
##                               (1)                    (2)                    (3)
## -----------------------------------------------------------------------------------------
## budget                     0.00000***             0.00000***             0.00000***
##                             (0.000)                (0.000)                (0.000)
##
## genres_action                                       -0.153                 -0.162
##                                                    (0.731)                (0.706)
##
## genres_adventure                                    0.695                  0.682
##                                                    (0.582)                (0.648)
##
## genres_drama                                        -0.029                 -0.034
##                                                    (0.546)                (0.538)
##
## original_language_en                                                       -0.058
##                                                                           (0.616)
##
## Constant                   14.973***              15.022***              15.069***
##                             (0.341)                (0.504)                (0.543)
##
## -----------------------------------------------------------------------------------------
## Observations                  100                    100                    100
## R2                           0.358                  0.363                  0.363
## Adjusted R2                  0.351                  0.337                  0.330
## Residual Std. Error     2.232 (df = 98)        2.258 (df = 95)        2.270 (df = 94)
## F Statistic          54.644*** (df = 1; 98) 13.558*** (df = 4; 95) 10.735*** (df = 5; 94)
## =========================================================================================
## Note:                                                         *p<0.1; **p<0.05; ***p<0.01
```

After running our three models, we then conducted F-tests to compare model 1 with model 2, and model 2 with model 3. For our first F-test, we received a p-value of 0.847. We therefore failed to reject our null hypothesis that model 1 is the correct population model. Model 2, with the additional genre terms, is therefore not more appropriate than model 1. From there, we conducted an F-test to compare model 2 and model 3. We received a p-value of 0.9253, causing us to fail to reject our null hypothesis that model 2 is the correct population model. Model 3 with the additional original language english variable is therefore not more appropriate than model 2. From these tests, we conclude that model 1, with just budget included as an input variable, is the most appropriate choice to model revenue.

Diving into model 1, we see that the intercept coefficient, β_0, is 14.97. Because our output variable, revenue, is log transformed in our model, we can take eˆ14.97 to better interpret this variable. This means that when the budget is set to zero, our revenue is about $3,172,403. This is of course not a meaningful interpretation, as our budget cannot be zero for a movie, but does give an idea of what our model looks like. From there, we see that the coefficient for budget, β_1, is 0.000000044426, or approximately zero. This means that revenue will increase by about 0.0000044426% for every one-unit increase in budget. Although revenue is increasing when we increase budget, and this coefficient is statistically significant, we argue this increase is not practically significant. This change is so small that we would not recommend for Acme Movies Inc to increase budget with the goal of increasing revenue at a high rate. We instead argue that there are variables other than budget, which were not included in this dataset, that have a stronger impact on revenue. We recommend a followup study to analyze these additional variables and their relationship with revenue.

In models 2 and 3, we see that the original language and genre variables were not statistically significant. We therefore have no recommendations to make on these variables, as they are not significantly related to revenue. We do notice that for each of the three models, the coefficient for budget remains somewhat constant, keeping the same sign for each model and varying only slightly. The coefficient for budget for model 1 is 0.000000044426, for model 2 is 0.00000004102, and for model 3 is 0.00000004129. Budget also remains statistically significant in all three models. This shows that there is a strong, consistent relationship between budget and revenue. However, it is not practically significant, as the coefficient is so small that we do not advise making decisions to increase revenue using budget.

Statistical Limitations

The main large sample model assumption that may pose a problem is I.I.D. data. Our data is identically distributed, as it is all from the same time period, 2010, and all collected from the same population of movies. However, our data is likely not independent. There is likely clustering by company, as how a company decides to invest in one movie will likely tell us something about how that same company decides to invest on their other movies. However, most companies are only listed once in our dataset and therefore this violation is small. There also may be clustering by country, as we will likely find patterns in films all produced in the same country. However, most movies in our dataset are produced in the U.S. and therefore this violation is minimal. Violating independence should not bias our estimates. However, it may produce incorrect standard errors on our estimates. If standard errors are underestimated, this may lead to incorrect conclusions being reached. To combat this, we used robust standard errors. We also could combat this issue in the future by focusing on just the U.S. to avoid the issue of clustering by country. We also could focus on one production company to avoid the issue of clustering by company. Finally, we should meet the BLP assumption without any issues.

Structural Limitations

There are several potential variables that might serve as omitted variables in this project. Omitted variables are variables that are related to other independent variables that are included in the model and to the dependent variable. Exclusion of omitted variables has the potential to introduce bias in the model, such that the effects of the independent variables are either overestimated or underestimated (depending on the nature of the effect of the omitted variables).

Although there may be several omitted variables in this project, three prominent variables include marketing via social media, the number of theaters in which a movie is playing during opening week, and whether the movie can be streamed at home.

Marketing via Social Media:

In 2010, researchers at Hewlett-Packard found that the number of Twitter posts about a movie can predict movie revenues (Asur & Huberman, 2010). It is possible that a marketing campaign focused on advertising via social media can create the kind of buzz needed to increase movie watching and revenue. The predicted relationship between social media posts and budget would be positive because the budget would increase to fund marketing via social media posts, and the predicted relationship between social media and revenue would also be positive because advertising would increase ticket sales. In this case, the direction of bias would be away from zero. This means that the effect of budget on revenue would be overestimated when social media is excluded from the model. The data that would need to be collected to resolve this omitted variable bias would be the number of social media posts about a movie prior to its opening.

Number of Theaters:

Asur and Huberman (2010) also found that the number of theaters in which a movie is playing during opening week accounted for a significant proportion of the variance in opening week revenue. Number of theaters is another feature that could be manipulated by production companies. The predicted relationship between number of theaters and budget would be positive because budget would increase as number of theaters increases, and the predicted relationship between number of theaters and revenue would also be positive (based on findings from previous research). In this case, the direction of bias would be away from zero. This means that the effect of budget on revenue would be overestimated when the number of theaters is excluded from the model. The data that would need to be collected to resolve this omitted variable bias would be the number of theaters in which a movie is showing during opening week.

Streaming:

In 2021, worldwide subscriptions to streaming services reached 1.3 billion. Streaming services impact box office ticket sales and affect revenue. The predicted relationship between streaming and budget would be negative because it may cost less to offer movies via streaming services than in theaters, so budget would decrease as streaming increases. The predicted relationship between streaming and revenue would also be negative because consumers pay less for streaming than for box office tickets. In this case, the direction of bias would be away from zero. This means that the effect of budget on revenue would be overestimated when streaming is excluded from the model. The data that would need to be collected to resolve this omitted variable bias would be the number of consumers who are watching a movie in a theater vs. at home via streaming services.

Conclusion

The results of the present research partially supported predictions. While the models showed that budget does affect revenue, they also revealed that genre does not affect revenue. The relationship between budget and revenue was positive, indicating that increases in budget cause increases in revenue. This finding supports previous findings showing that budget is positively associated with revenue. Importantly, the effect of budget on revenue remained significant even when holding genre and original language constant. In contrast, the prediction regarding the effect of genre on revenue was not supported. When holding budget and original language constant, none of the genres examined in the explanatory models (including adventure, action, and drama) had a statistically significant effect on revenue. This finding conflicts with previous findings showing that these genres are associated with greater revenue than other genres.

It is important to note that the overall impact of budget on revenue, while statistically significant, was potentially negligible with respect to profit. The coefficient was very small, indicating that increases in budget lead to similar increases in revenue. The interpretation of these results is that the unique influence of budget on revenue may only allow a production company to “break even.” This finding provides insight regarding the importance of budget. A production company should not rely on budgets alone when seeking profitable revenues. While it may be assumed that movies with large budgets typically tend to be more profitable, these findings suggest that budget alone is insufficient.

It is also important to note that there was overlap among the binary coding for adventure, action, and drama genres. Specifically, one movie might be associated with multiple genres. For this reason, the unique effects of each genre may be difficult to assess. Future research could consider alternative strategies for collecting data in ways that classify genre as mutually exclusive variables. In addition, other genres were excluded from the model, and it is possible that other genres (such as comedy) could impact revenue.

Recommendation:

Overall, based on the present findings, the Data Science team at Acme Movies Inc concludes that budget alone is insufficient to increase profitability of the company. We recommend that Acme Movies Inc continue to sponsor this project so that we can further examine the potential influence of omitted variables and other features on revenue that the company can manipulate to increase profits.

References

https://www.kaggle.com/competitions/tmdb-box-office-prediction/data?select=train.csv
https://www.the-numbers.com/market/
https://www.thetoptens.com/top-10-pastimes/
https://www.nyfa.edu/student-resources/the-evolution-of-film-over-time-a-brief-history/
https://stephenfollows.com/how-has-the-cost-of-making-a-movie-changed-over-the-past-twenty-years/
https://www.statista.com/statistics/188658/movie-genres-in-north-america-by-box-office-revenue-since-1995/
https://www.screendaily.com/box-office/box-office-analysis-foreign-language-films-in-the-us/5096804.article
https://arxiv.org/pdf/1003.5699.pdf
https://deadline.com/2022/03/streaming-services-mpa-1234977814/