Lab 2: What Makes a Movie Successful? Fall 2021

w203: Statistics for Data Science

November 1, 2021

Introduction

What makes a movie "one of the greats"? While that question may elude both data analysts and artists for years to come, we are looking to find some effective indicators of box office success for films, in the hopes of maintaining a thriving movie industry for years to come. This project will focus on several commonly cited reasons for movie fiscal success and analyze how much of an impact each factor has.

The client for this analysis could be any major movie studio or its parent company, e.g. NBCUniversal or Walt Disney Studios. This is a significant issue, because box office sales decreased by 60% in 2020 during COVID-19 pandemic, leaving the movie theaters empty for months. This analysis will provide clues for what a studio can do to speed up the recovery and account for lost revenue.

The main dataset we'll use for this analysis comes from The Movie Database (TMDB - https://www.themov iedb.org/), as well as several assembled lists from online polls (e.g. "Greatest Actors/Actresses of All Time" - Ranker.com).

We're planning on using profitability as the response variable for analysis in this study - using R to clean the empty and irrelevant cells, a "profitability ratio" can be determined for each movie based on its box office sales divided by budget (and controlling for foreign films and currencies).

In order to build our models, we decided to focus on some factors that are usually linked to a movie's financial success, such as budget, time of release, its cast's popularity and director's name brand recognition. We will also take into account some control variables such as runtime and genre.

The products of this analysis will include the code and final datasets, as well as a written summary of the significant results.

Data, Research Question, Underlying Model and Study Design

Data

As mentioned before our main dataset comes from data extracted from The Movie Database (TMDB - https://www.themoviedb.org/). This dataset was compiled by Kaggle for a Machine Learning competition a few years ago.

The original dataset consists of 22 variables for over 7000 films, from the 1920s to 2017. This dataset consists of objective data for each of these movies such as budget, revenue, cast, crew, release date, production company, original language among others.

We also added two more datasets consisting of two lists with a ranking of the "Top 100 movie actors of all time", as well as the top "Top 50 directors". Unlike the previous dataset, these ranks is subjective, but we plan on using these lists to help us create a variable that could stand in for an actor's or director's popularity.

We performed some data cleaning operations on the main dataset in order to have it be ready for our analysis. These included:

- Eliminating any entries that had some vital information missing (such as revenue, budget, cast, crew and genres)

- Parsing certain fields from a stringified JSON.
- Creating the "Actors" column from the "Cast" column
- Extracting the Director from the "Crew" column

After completing this process we still had over 5000 entries.

Research Question

This study intends to find significant relationships between factors known before a movie release and its box office success. We will look at both its profitability ratio (defined as box office divided by its budget) and its box office as ways of measuring a movie's success.

The results of this study could be used by film makers to better inform their investment decisions.

Underlying Model

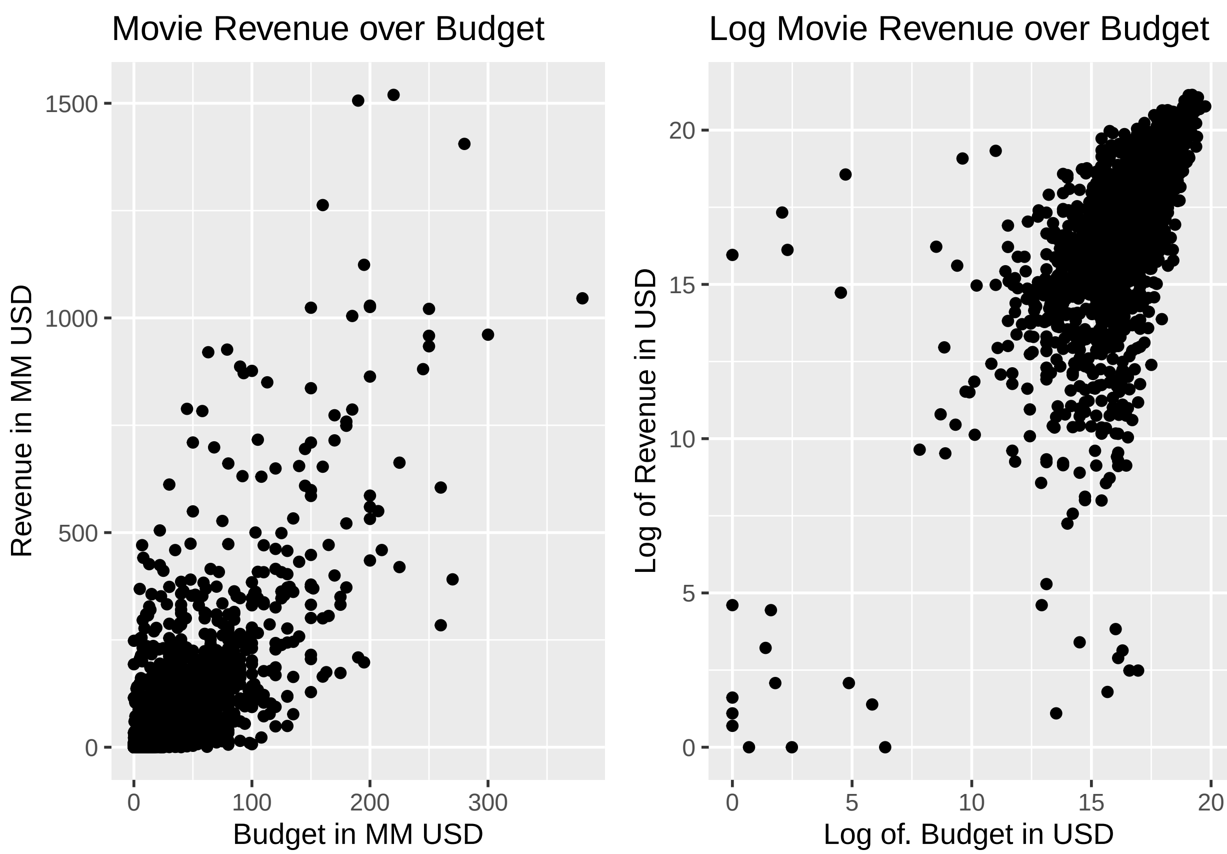
As mentioned in the introduction, we plan on using commonly cited reasons for movie's financial success to create our models, starting with a simplified model with just the key variables, and then add more variables we hypothesize could have an effect on a movie's financial success to the following models

Simplified model The most simple and straight forward way of predicting a movie's box office would be its budget. The relationship is pretty clear. The more money that is put into a project the better it is expected to do. One would also expect the increase budget would go into improving the quality of the movie that in turn should increase its box office performance.

```r
p1 <- ggplot(final_table, aes(x=budget/1000000, y=revenue/1000000)) + geom_point() + labs(
  title = 'Movie Revenue over Budget',
  x = "Budget in MM USD",
  y = "Revenue in MM USD"
)

p2 <- ggplot(final_table, aes(x=log(budget), y=log(revenue))) + geom_point() + labs(
  title = 'Log Movie Revenue over Budget',
  x = "Log of. Budget in USD",
  y = "Log of Revenue in USD"
)

grid.arrange(p1, p2, nrow = 1)
```



As seen in the previous charts, there is a clear positive relationship between budget and revenue, both for with and without a logarithmic transformation. This contributes to our hypothesis than a higher budget leads to a better box office.

Seasonality Just like for a lot of other industries, seasonality in the movie industry is a well documented phenomena. [1]

Certain times of the year tend to attract more moviegoers, such as the summer, the holiday season and certain weekends that coincide with important holidays (Christmas, 4th of July weekend, Memorial Day weekend, among others). This increase demand would lead to a better box office performance.

On the other hand, supply is also affected by seasonality. Big budget movies are more likely to be released during a time of the year that could help them maximize their potential revenue.

This self-selection on the supply side is likely to have an effect on the demand as well, as a greater number of quality movies during certain times would attract more movie goers.

Our model will control for this effect by adding the month of release to the model. We would expect the summer months, as well as December to have a significant positive effect on revenue.

The Cast and Director's Name Recognition It shouldn't be a surprise to anyone that famous actors attract crowds, which in turn increase the movie's box office performance. Audiences want to see their favorite actors in their movies, for a variety of reasons.

A recent example that comes to mind is the 2021 movie adaptation of Dune. The actress Zendaya was a big part of the movie's marketing, despite the fact that sheonly briefly appears in the movie. A lot of movie

[1] Einav, L. (2007). Seasonality in the U.S. Motion Picture Industry. The RAND Journal of Economics, 38(1), 127–145. http://www.jstor.org/stable/25046296

goers were upset after seeing the movie because they expected her to have a bigger role.[2].

For our model we will give each movie an “Actor Score” which will be just a counter of how many well known actors are part of their cast.

We have also added a “Director Score” which will reflect a movie’s director ranking in the following manner:

Top 10 - 5 points Top 20 - 4 points Top 30 - 3 points Top 40 - 2 points Top 50 - 1 point Not on the ranking - 0 points

Genre There are a lot of different movie genres, and a movie can be considered to have multiple of those. We do not have very specific expectations regarding which genres may have a significant effect on revenue.

Is it part of a franchise/collection? There have always been movies that have been part of something bigger. Successful movies tend to spawn sequels, or even entire series. Just being associated with something greater may help a movie do better in the box office.

Omitted Variables

While we have covered a lot of possible variables for our model, it’s almost certain that there are more variables that could explain a movie’s financial success. Here we will cover the ones we consider to be the main omitted variables.

The Quality of the Movie So far we have discussed more objective measurements, such as the budget, release date, cast and crew. However, these are not the only factors that can affect a movie’s success.

Word of mouth has been studied as an important factor in a movie’s success [3].

Average movies tend not to create a lot of word of mouth, so they are unlikely to benefit from that effect, while excellent movies will get talked about and recommended.

We do run into a problem when trying to include it in our model. While we could use rankings and ratings as proxy variables, this research focuses on predicting a movie’s revenue, and these kinds of rankings and ratings tend to come out after a movie is released, making including them on the research pointless.

We have decided that the best way to capture the effect of the quality of a movie would be through the Director Score variable. A movie by a great director is probably the safest bet someone can make when it comes to the movie’s quality.

We would also expect some of its effect to be captured by the Actors score, and its budget, both are things are associated with a movie’s quality.

Since the relationship is positive, and we assume that so is the Actor’s score and the budget, then it would be safe to assume that omitting this variable is pushing these coefficients away from zero, thus overestimating their effect on a movie.

The Marketing strategy and budget Another possible omitted variable is the movie’s marketing. If nobody hears about a movie, they are much less likely to go see it. A good marketing strategy can get people to the movie theatres, and it can also maximize the effect of an actor’s appeal, as was the case with the Zendaya example.

Quantifying a “good” marketing strategy is complicated, but we could use the movie’s marketing budget as a proxy variable. However, the marketing budget is not part of our dataset’s budget column, as the marketing and distribution budget are not usually part of the movie’s reported budget, which tends to focus on just production costs.

[2] https://nypost.com/2021/10/26/fans-outraged-that-zendaya-is-in-dune-for-only-7-minutes/

[3] Liu, Yong. “Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue.” Journal of Marketing, vol. 70, no. 3, American Marketing Association, 2006, pp. 74–89, http://www.jstor.org/stable/30162102.

Having said that, it's very likely that a marketing budget will be heavily correlated with a movie's production budget. So by including the budget, we are including a proxy variable for a movie's marketing strategy.

Just like with the quality, we would expect a good marketing strategy to have a positive effect on the revenue, as well as a positive effect on related variables such as the Actors Score.

Research Design

Statistical Model

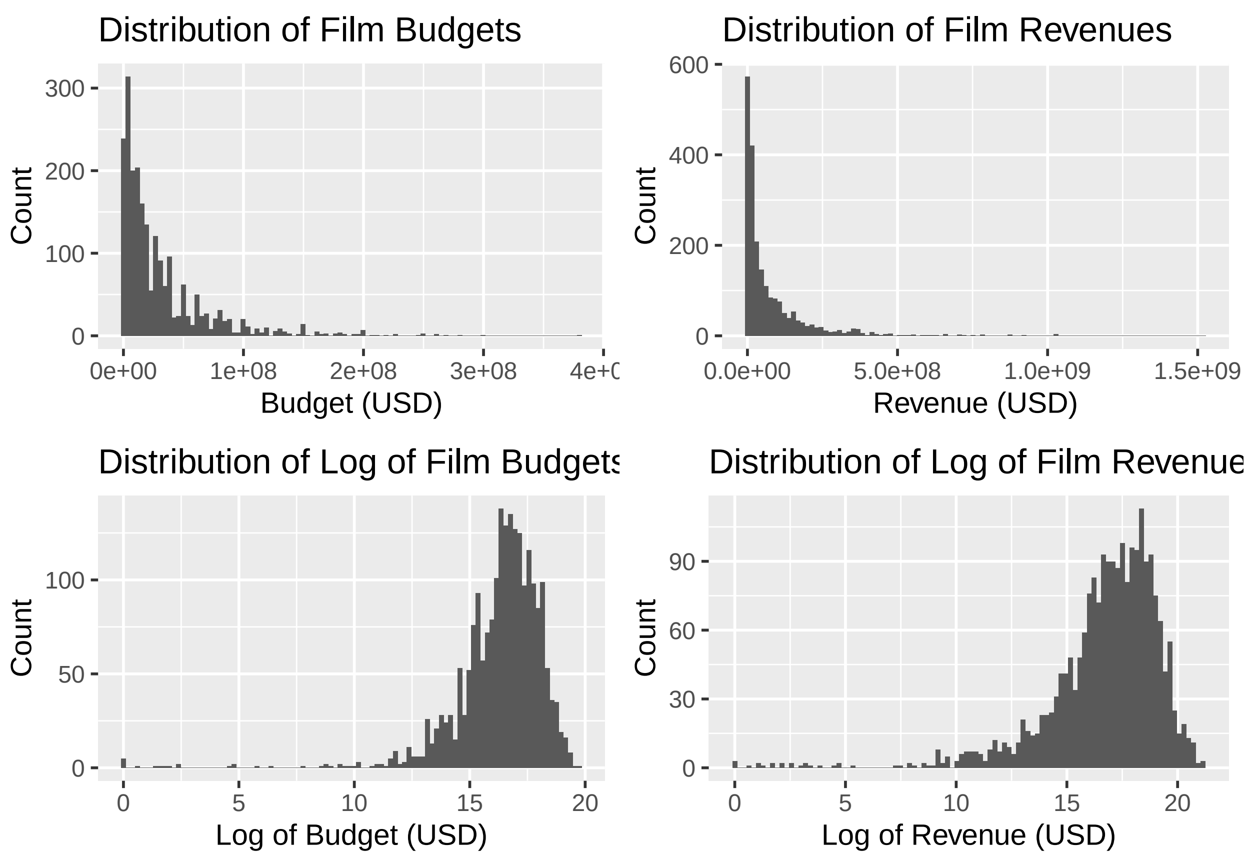
```
p1 <- ggplot(final_table, aes(x=budget)) + geom_histogram(bins = 100) + ggtitle("Distribution of Film Bu

p2 <- ggplot(final_table, aes(x=revenue)) + geom_histogram(bins = 100)  + ggtitle("Distribution of Film


# The logarithm of both budget and revenue appear more homogenously distributed
p3 <- ggplot(final_table, aes(x=log(budget))) + geom_histogram(bins = 100) + ggtitle("Distribution of Lo

p4 <- ggplot(final_table, aes(x=log(revenue))) + geom_histogram(bins = 100) + ggtitle("Distribution of L

grid.arrange(p1, p2, p3, p4, nrow = 2)
```



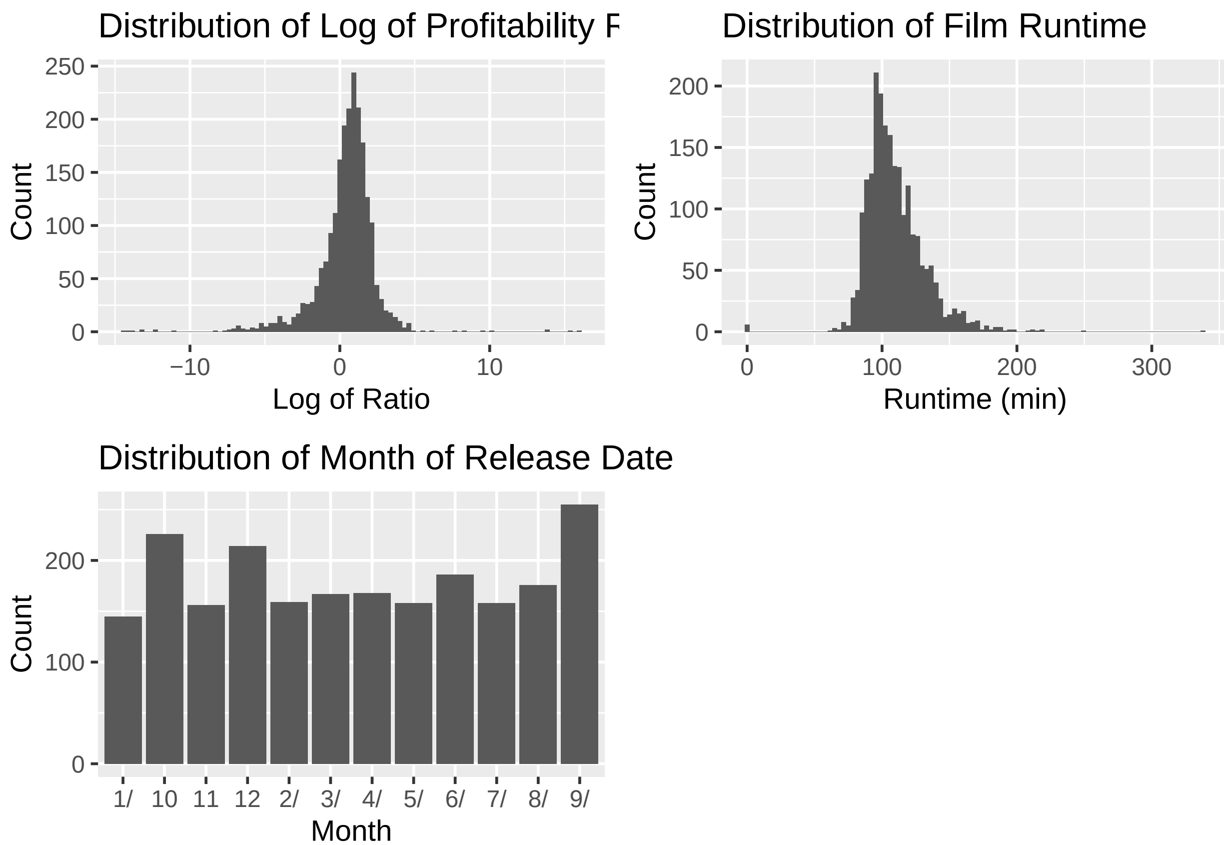

```
p1 <- ggplot(final_table, aes(x=log(profitability_ratio))) + geom_histogram(bins = 100) + ggtitle("Dist
#Looking at the shape and distribution of profitability ratios, the log of the graph looks pretty well

p2 <- ggplot(final_table, aes(x=runtime)) + geom_histogram(bins = 100) + ggtitle("Distribution of Film
#runtime appears to be relatively normally distributed overall, except for the narrow group on the lowe
```

```
p3 <- ggplot(final_table, aes(x=substr(release_date, start = 1, stop = 2))) + geom_bar() + ggtitle("Dis
#Release date overall looks fairly well distributed over time - the number indicates month, with 1/ - 9

grid.arrange(p1, p2, p3, nrow = 2)
```

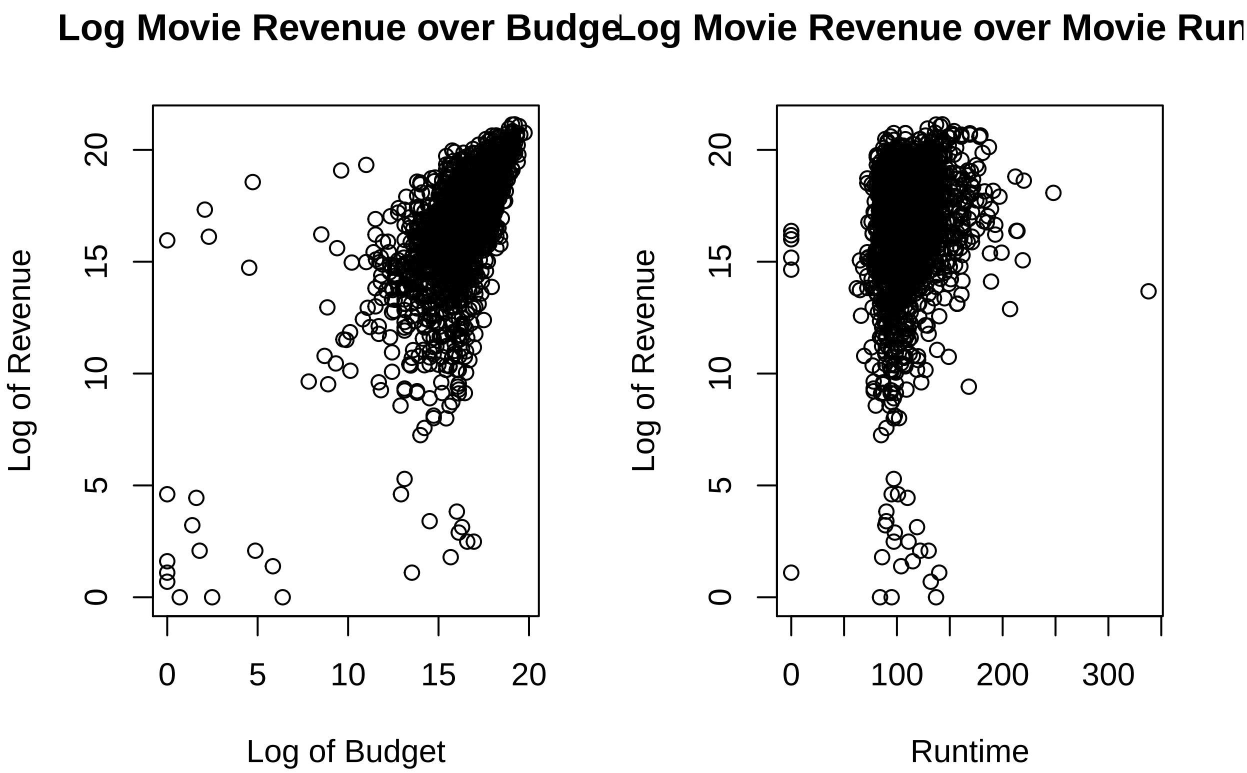
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




```
par(mfrow=c(1,2))

plot(log(final_table$budget), log(final_table$revenue), main = 'Log Movie Revenue over Budget', xlab =
# The shape of the relationship looks to be positively correlated, and seems to suggest that a higher b


plot(final_table$runtime, log(final_table$revenue), main = 'Log Movie Revenue over Movie Runtime', xlab
```



We're looking to see which model and variables will be a good approximation for future films' natural logarithm of revenue(in U.S. dollars). We will be building 4 predictive models to help assess our research question of which variables play a crucial role in box office revenues to help the industry recover from the decline in the last year. We start with a simple model with just the `budget` variable. The EDA done earlier in the report shows a relationship between budget and film success measured in revenue. We add covariates to the 2 following models to control for additional variables that we also expect to be significant. And the final 4th model will include only the variables that proved to be statistically and practically significant.

Variable that we're most interested in is budget β_1, which measures a percent increase or decrease in revenue for every 1% increase in budget. We applied a natural logarithm transformation to both variables because the distribution was skewed towards lower variables. The transformation spread out the values, and the results show that the original data has a more log-normal distribution, making this transformation a good choice to improve our model.

Some issues with the covariates could be

1) Model 1: Natural logarithm of revenue on a constant and natural logarithm of budget

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget})$$

2) Model 2: Natural logarithm of revenue on a constant, natural logarithm of budget and genre

$$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget}) + \beta_2 \text{genre}$$

3) Model 3: Natural logarithm of revenue on a constant, natural logarithm of budget, genre, runtime, actor_score, director_score,release month, indicator variable of native english film and belongs to a collection(if the film is part of a sequel)

$log(\text{revenue}) = \beta_0 + \beta_1 log(\text{budget}) + \beta_2 \text{genre} + \beta_3 \text{runtime} + \beta_4 \text{actor score} + \beta_5 \text{director score} + \beta_6 \text{native english film} + \beta_7 \text{release mo}$

Results

We're presenting a regression table below to compare results and find variables that prove to be significant.

```
stargazer(
  model_1,
  model_2,
  model_3,
  #model_4,
  type = 'latex', header = FALSE
  #se = list(get_robust_se(model))
)
```

All models are overall significant when looking at F-statistic which means that our linear regression model provides a better fit to the data than a model that contains no independent variables. As we're adding more predictors into the model we want to also evaluate the Adjusted R2 since that metric will account for predictors that improve the overall model more than expected by chance. Across the model we're seeing an increase in Adjusted R2 from most simple model to the most complex, which means that independent variables are capturing the variation in the dependent variable quite well and the addition of covariates is improving the model.

We found that quite a few variables in our consideration proved to be statistically significant. Budget, our main variable in question in the first model is significant because the p-value is below 0.05 threshold and is showing practical significance with a high coefficient. Since both variables were transformed using natural logarithms we can conveniently interpret the results in percents. For every percent increase in budget will have 0.81% increase in revenue. Since this could mean that for every additional dollar in the budget the return will be 80 cents. Which is why we want to interpret additional variables to understand what is the differentiating factor between profitable and not profitable movies.

In the second model when we include genre we have to interpret the statistical significance differently to avoid getting a significant genre by chance since we're evaluating a large number of genres. To make the significance stricter we will multiply the p-value fro genre by 19, since we have 19 coefficients for genre in the model. We conclude that drama may only be significant by chance because we ran multiple t-tests. And we can say that other variables, adventure, horror and foreign genres are significant predictors of revenue.

We will evaluate the release month in the same way to make sure we're not proving a variable to be significant by chance. We'll multiply each p-value by 11 because there are 11 month variables in the model. Which leaves us with july as a significant variables, where June and Decmber don't meet that level of siginificance and only July is proving to be significant.

Some additional variables that proved to be a good predictor of revenue are Director Score, Actor Score and Native English Film, Belongs to Collection as well as Runtime with a caveat that runtime is not practically significant with a very low coefficients which means that there will be insignificant increase to revenue with every additional minute of runtime. On the other hand we want to pay attention to variables that are statistically significant and have a high practical significance, such as belonging to a collection of movies, having a higher budget.

Our suggested final model will be model 4 that will have practically and statistically significant variables.

Limitations

Statistical limitations

We will evaluate this model against 2 large model assumptions since we have a sample of well above 100. The first is IID and presence perfectly coliniear variables.

Table 1:

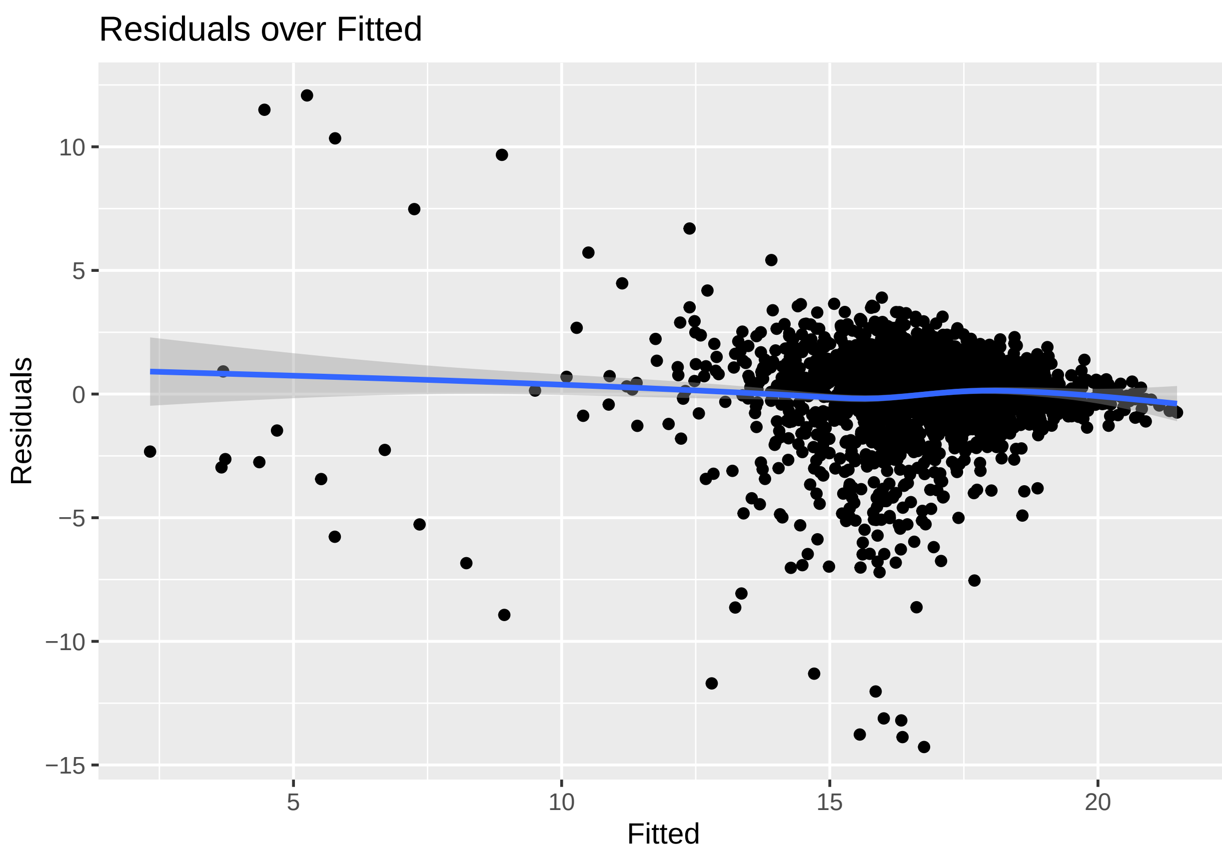
	Dependent variable:		
	log(revenue)		
	(1)	(2)	(3)
log(budget)	0.853***	0.819***	0.749***
	(0.021)	(0.022)	(0.023)
comedy		−0.020	0.140
		(0.108)	(0.106)
drama		−0.281***	−0.173*
		(0.103)	(0.102)
family		0.318*	0.268
		(0.183)	(0.176)
romance		0.115	0.166
		(0.118)	(0.113)
thriller		−0.169	−0.088
		(0.114)	(0.109)
animation		−0.058	0.237
		(0.231)	(0.227)
adventure		0.498***	0.177
		(0.132)	(0.128)
horror		0.499***	0.418***
		(0.153)	(0.150)
music		0.276	0.255
		(0.252)	(0.241)
crime		−0.044	−0.109
		(0.125)	(0.121)
sci_fi		−0.236	−0.241*
		(0.145)	(0.139)
action		0.160	0.115
		(0.114)	(0.110)
war		−0.097	−0.157
		(0.233)	(0.223)
western		−0.420	−0.340
		(0.361)	(0.347)
fantasy		−0.042	−0.068
		(0.162)	(0.155)
foreign		−2.281***	−2.185***
		(0.606)	(0.583)
mystery		0.149	0.153
		(0.162)	(0.156)

1. For the IID assumption we want to make sure that our samples are Independent and Identically Distributed. It is hard to find a sample that is perfectly independent. Our data source is coming from The Movie Database (TMDB) which is a community built movie and TV database including international movies. It's one of the largest community managed database of movies, which should provide a good collection of movies including international releases. And we used a sample of approximately 2000 randomly selected movies that were compiled for a Kaggle competition. The randomized sample should help with the assumption of indepedence.

Although multiple factors could cause the movies to be dependent on each other. For example presence of the same crew and cast can produce movie samples that may not be independent. Movies that are produced as part of the sequel do dependent on the revenue and success of the previous releases. There could be multiple other scenarios that can influence other movies in the sample, and we're seeing some pattern in the Residuals vs. Fitted plot below also may indicate error terms not to follow a constant variance, but given random sample that is drawn from a large database we will consider this to be IID.

```
modf_2 <- fortify(model_4)
ggplot(modf_2, aes(x = .fitted, y = .resid)) + geom_point() + geom_smooth() + labs(
  title = "Residuals over Fitted",
  x = "Fitted",
  y = "Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



2. The second model assumption is that there is no perfect colliniarity between variables which is proven by the fact that none of the variables were dropped from the model when we were fitting the model. In the presence of multicolliniariry the `lm` function in R would drop a variable and not display a coeffcient value or significance.

We are meeting the 2 large model assumptions that are necessary to prove the model is reflecting a true information about the world and that we are reporting a results and coming to a conclusion that will be useful in the industry.

Structural limitations

Although we've considered a multititude of variables in our model there are some limitations, our actor and director scores are measured only for top 100 actors, actresses and directors at a time when in a movie industry and through social media the popularity of a cast and crew could evolve.

We're also not controlling for inflation when comparing movie revenues throughout the years. A revenue 10 or 20 years ago will be higher if we account for inflation and compare it to movies that were released more recently.

Some of the structural limitations of the model are omitted variables. If there is presence of another significant predictor that we don't measure or include in the model that influences the revenue. We suppose some of those variables are quality and promotion budget or strategy. It's harder to measure the true quality of a movie and in our model we use more of a proxy for quality through evaluating actor and director score as well as budget, which may not capture the whole measurement of quality. Another omitted variable is marketing budget since we only have the production budget in our dataset. Promotional strategy and budget could be a significant influence on revenue, since more popular movies have more box office revenues.

Conclusion and Discussion

- significant variables and their contribution to answering the question