Lab Two: TMDB Movies

Datasci 203: An Exploration of the Relationship Between the Budget and Success of a Movie

Ifrah Javed, Ryan Brown, Kodzai Nyakurimwa, Wilford Bradford

13 April, 2022

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Research Question . . . 1

2 Data **1**

2.1 Data Preprocessing and Cleaning . . . 1

2.2 Data Transformations . . . 2

3 Modeling **3**

3.1 Research Design . . . 3

3.2 Model One . . . 3

3.3 Model Two . . . 4

3.4 Model Three . . . 6

3.5 Model Four . . . 7

4 Results **9**

4.1 Model Comparison . . . 9

5 Limitations **10**

5.1 Statistical Limitations: Large-Sample Assumptions . . . 10

5.1.1 Evaluation of Assumption One: I.I.D Data . . . 10

5.1.2 Evaluation of Assumption Two: A Unique-BLP Exists . . . 11

5.2 Structural Limitations: Omitted-Variable Bias . . . 11

5.3 Shared Ancestors . . . 12

5.4 Multicollinearity . . . 12

6 Conclusion **13**

6.1 Summary 13

6.2 Further Study 13

7 Appendix **15**

7.1 Manual Budget and Revenue Corrections 15

1 Introduction

1.1 Motivation

There are many contributing factors that go into creating a "successful" movie. Production companies, directors, and actors work very hard to create the next box office hit in theaters. Some may even work their whole lives without having a movie that is "successful". Budget, particularly, is often presumed to have a positive association with the success of a movie, but evaluating the existence, nature, and magnitude of this relationship may prove useful in tuning budgets to maximize potential value in the future.

1.2 Research Question

Specifically, our research question is as follows,

> *Does a movie's budget have an effect on its success, measured in terms of revenue?*

In this instance, our exposure is budget and our response is revenue. Specifically, budget is defined as the total production budget for the movie, and revenue is defined as the net global box office revenue.

2 Data

Our data source comes from Kaggle and is titled "The Movie Database". This Kaggle dataset is a scrubbed export from themoviedb.org API. This source is entirely user-generated, though there are strict guidelines for contribution, defined colloquially as "The Bible".

The dataset contains 4,083 movies spanning from 1926-2017, and has a total of 20 attributes. Some of the attributes included are: release date, genre, budget, revenue, runtime, and production companies.

2.1 Data Preprocessing and Cleaning

As our data were derived from user inputs, there were a few inconsistencies to tackle. Firstly, we set a hard floor for both budget and revenue. For budget, we required that a given movie have a value greater than $1 million. We felt this was necessary to exclude low-budget films with characteristics far different from the standard movie.

Next, we dropped all rows with reported revenue of $0, as we felt these were most likely representative of missing data.

In this process, we noticed several outliers in which budget or revenue were mere increments of dollars rather than the expected millions. In these cases, we researched the specific movies and made corrections where needed. See the appendix for a detailed view of these cases.

The impact of our filters is shown in the table below. Here "n" represents the resulting number of observations.

There were also several JSON columns that required parsing. These columns supplied lists of given terms for multi-select fields like genre, production companies, and production countries, where multiple options may be selected. For these we converted each list of terms into a matrix of boolean dummy columns.

In the case of production companies, we also added a field for the count of the number of companies involved, as we felt this might influence revenue.

Production countries introduced high cardinality, so for this we used a feature mapping to reduce the feature set to a set of global regions.

Table 1:

Filter	n	impact
Raw	4,803	
Released Only	4,795	-0.2%
Revenue >$0	3,375	-29.6%
Budget >$1M	3,097	-8.2%

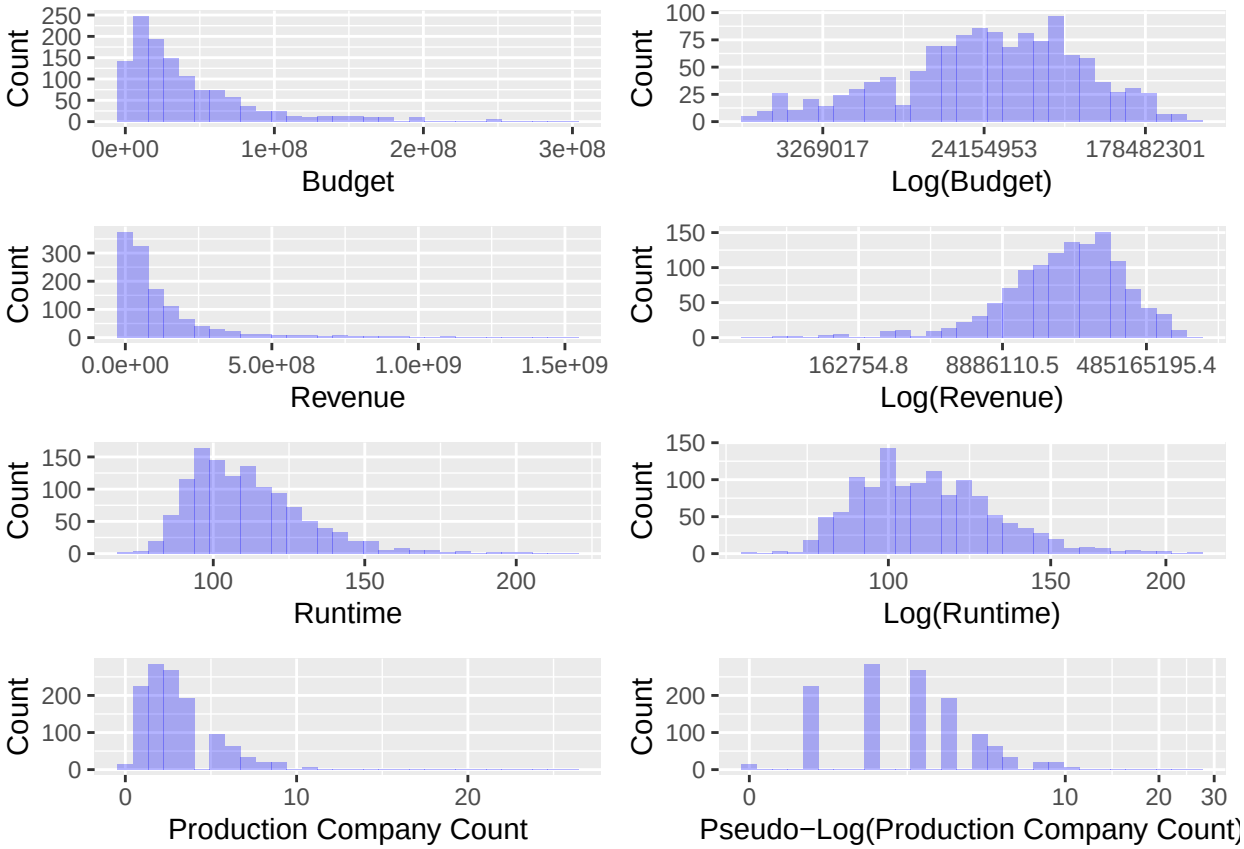
Lastly, to reduce cardinality in original language, we converted it to English vs. non-English, as we again felt this would capture most of the observed variance.

2.2 Data Transformations

In terms of data transformations, we chose to log transform our explanatory variable of "budget" and our dependent variable which was "revenue". As we can see in the figure below, both of these variable were right skewed. This is to be expected with monetary amounts. Prior to transforming the variables, we can see that the data is extremely right skewed. After transforming them, the distribution and spread looks to be more normal. We also transformed the variables "runtime" using a log transform and "production_compaines_n", representative of the count of production companies listed for the movie, using a pseudo-log transform, to handle zeros. By using these transformations we can see a reduction in skew for both variables.

The transformation on count of production companies, while reducing skew, seems less consistent. This transformation, however, distinctly improved our results in modeling linear trend with revenue.

All of our transformed variables were right-skewed with hard floors at 0 and essentially infinite ceilings.



3 Modeling

3.1 Research Design

Our research design is centered on evaluating the relationship/effect that a movie's budget has on its success in terms of revenue. Both of these variables, once log-transformed, are roughly normal in their distributions.

In an attempt to make our model more generalizable to unobserved data, we have split our frame into evaluation and train sets using a 40/60 ratio. We will conduct all of our EDA on the exploration set, then train and evaluate our models on the remaining train set. In the future we may also test our residuals on a third test set.

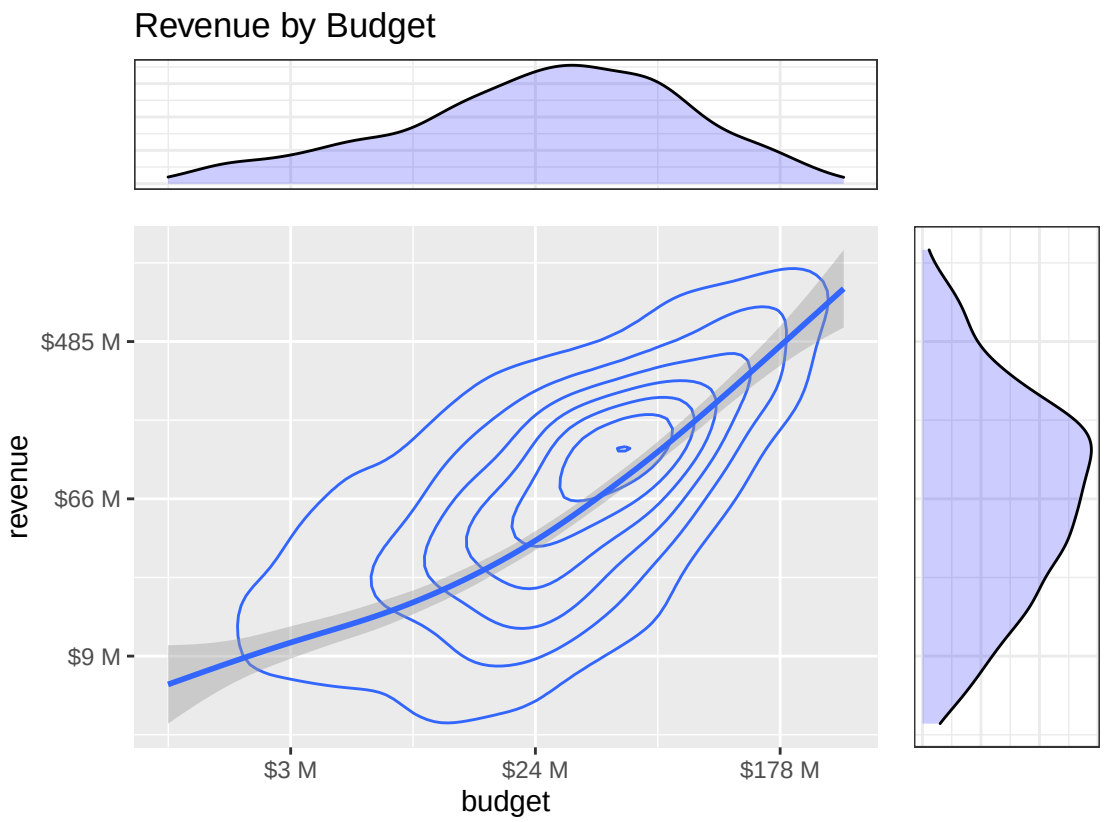
Our train set is comprised of 1,859 rows, which in this case are movies, and as such we will be using large-sample assumptions. This observation count, however, is a tad inflated due to the high cardinality of our data. Some of the low-positive-frequency dummy columns added in latter models may not be as robust as they seem, though our analysis is restricted to evaluating our exposure, budget.

We will start with a very basic model and iteratively add features, using partial F-tests to determine whether the additions are worthwhile.

3.2 Model One

The first model we examined was a base model with log transformations applied to both the explanatory and dependent variable. Justification for these transformations is given in Section 2.3: Data Transformations.

In the visualization below, we examine the relationship between these two features. Note the distinctly linear trend observed across the full domain of budget. There is some curvature, but this appears to be most heavily concentrated along the tails of our respective distributions.



This first model takes on the following form:

$$log(revenue) = \beta_0 + \beta_1 log(budget)$$

This first model returned an Adjusted R^2 of 0.3931. The p-value for budget was 1.7e-167. Based on this, we reject the null hypothesis that the coefficient for budget is zero.

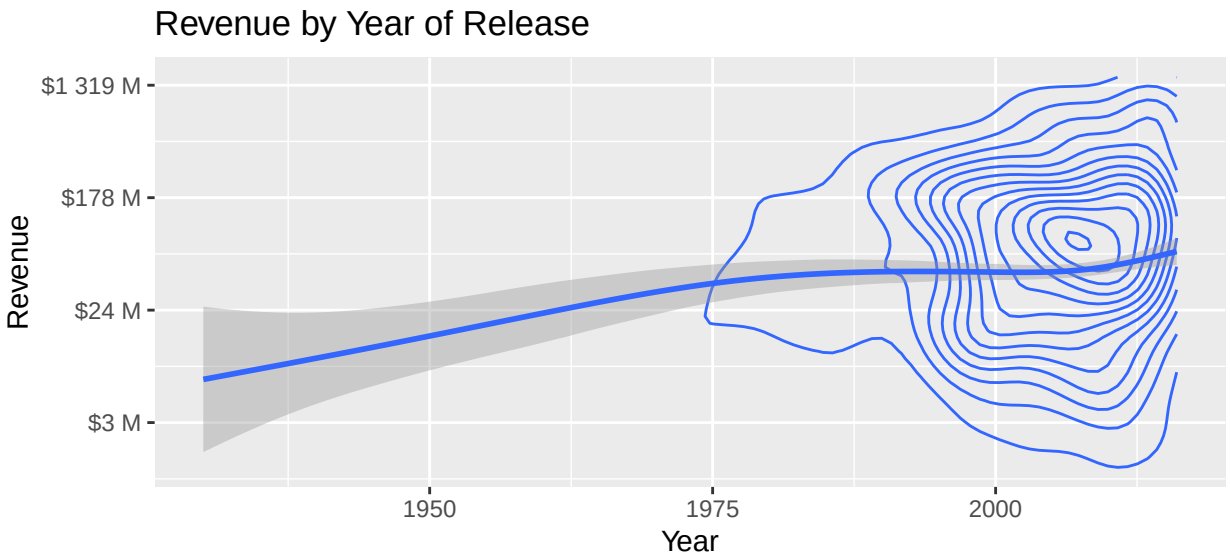
3.3 Model Two

For the second model, we chose to add in some preliminary control variables.

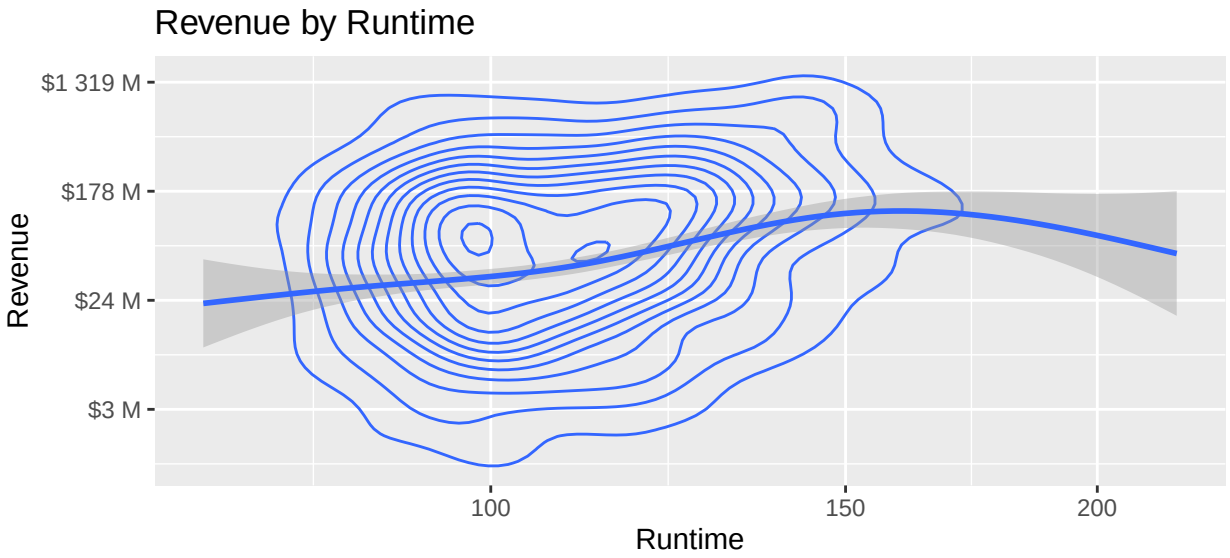
Firstly, we chose to add in year as a way to control for inflation and changes over time.

We then chose to add in runtime, as longer movies may require a higher budget, as more capital and time is needed to produce them. Longer movies may also attract more interest and impact revenue.

We examine the distributions for each field below.



Year appears to be somehwat linearly associated with revenue, though there is a lot of skew towards the edges of each respective distribution. These could be indicative of outliers, or lesser observation counts.



Runtime appears to exhibit a weak, albeit present, linear association with revenue, though this again presents some skew towards the extremities of the domain.

The first phase of this second model, model 2.a, takes on the following form:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year$$

After adding release year, our Adjusted R^2 improves to 0.4009. The p-value for release year in a model just adding it to the previous was 3.0e-04. In this single addition model, our partial F-test, relative to the fist model, returns the same p-value. Thus, we can reject the null hypothesis that the coefficient for release year is zero.

The second phase of this second model, model 2.b, takes on the following form:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_3 log(runtime)$$

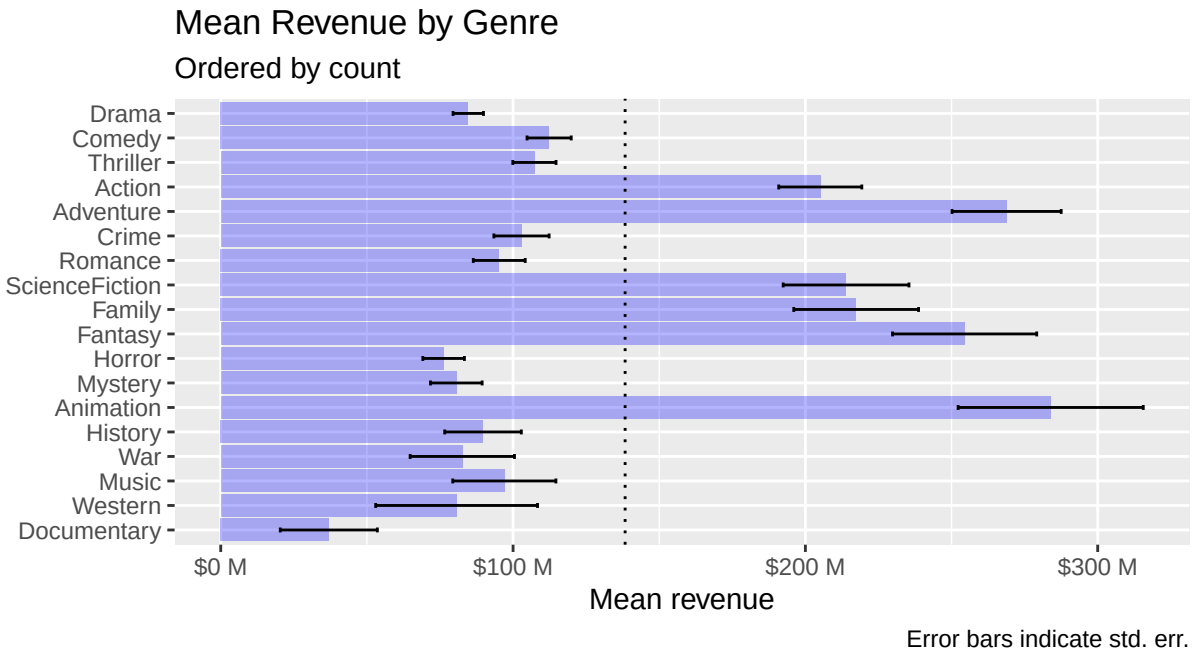
In repeating the same procedure to add runtime, our Adjusted R^2 only improved to 0.4015. This time, our partial F-test p-value, relative to the last significant model (adding year), was 0.12. Thus, we fail to reject the null hypothesis that the coefficient for runtime is zero.

3.4 Model Three

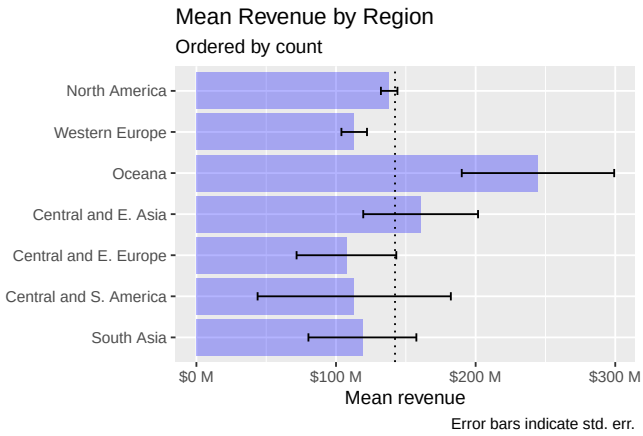
In the third phase, we decided to take a look at additional control variables. These variables were genres, production regions, and whether English was the original language. Each of these features has a theoretical association with revenue. Action movies, for instance, may produce greater revenue on average, and also may require increased financing for special effects. Regions and language are more regional. We aim to add these to control for regional and cultural differences in both supply and demand.

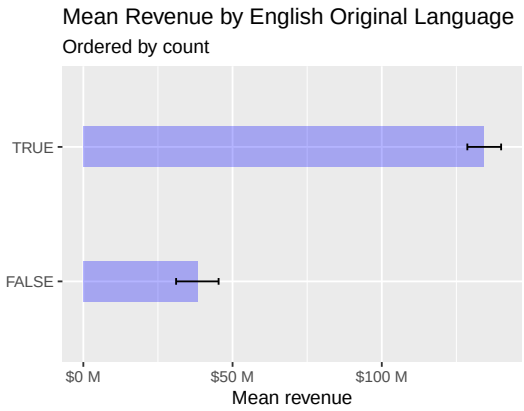
In order to do this we created three smaller models and ran partial F-tests to compare which controls we should keep.

We examine the distributions for each of these categorical fields below.



There are some obvious trends here with genres like Animation, Adventure, Fantasy, and others presenting averages distinctly above that of others. Others like Documentaries seem to lag behind.





We see a similar trend with regions, as Oceana appears to have higher average revenue than other regions. It is important to note, however, this dataset is imbalanced, and most movies are produced in North America. English as the original language also shows a distinct trend, but again, the data are highly imbalanced.

Our first model of the third phase add genres, and is defined as follows:

Model 3.a:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n}$$

After adding genres, our Adjusted R^2 improves to 0.4163. Our partial F-test relative to the model adding release year return a p-value of 3.1e-07, so we can reject the null hypothesis that the coefficients for genres are all zero.

Next, we move to the next set of variables, production regions:

Model 3.b:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_{4...n} production.regions_{1...n}$$

In attempting to add production regions, we observe a partial F-test p-value of 0.65, relative to the previous model adding genres, so we fail to reject the null hypothesis that the coefficients for regions are also non-zero.

Lastly, we add whether the original language was English, defined as follows,

Model 3.c:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_{4...n} original.language.english_{1...n}$$

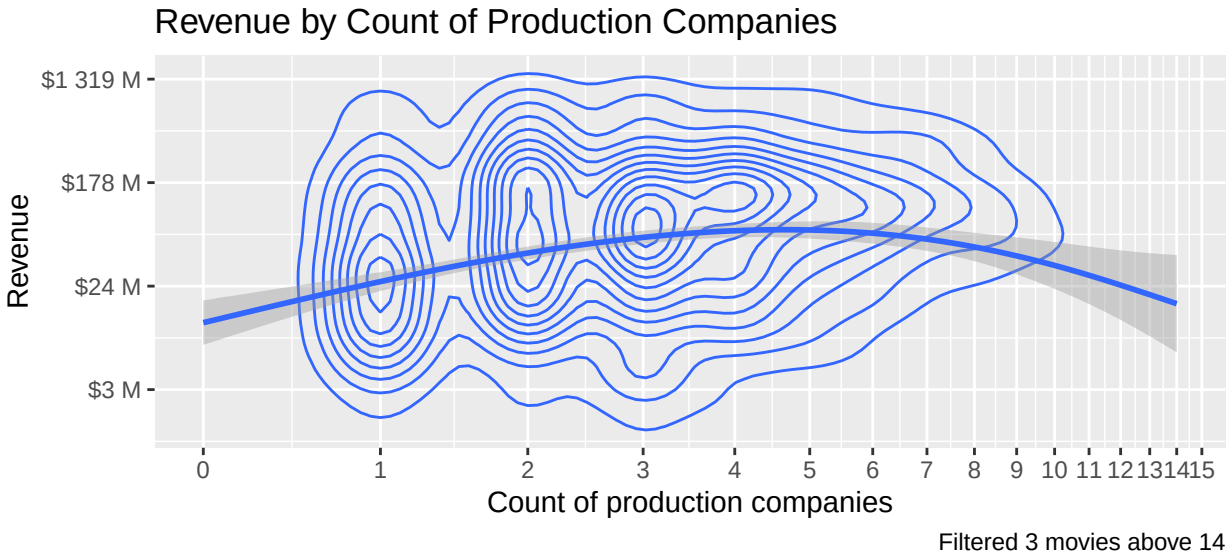
For 3.c, our partial F-test relative to the last confirmed model adding genres, 3.a, returns a p-value of 0.97. Thus, we again fail to reject the null hypothesis that this coefficient is zero.

3.5 Model Four

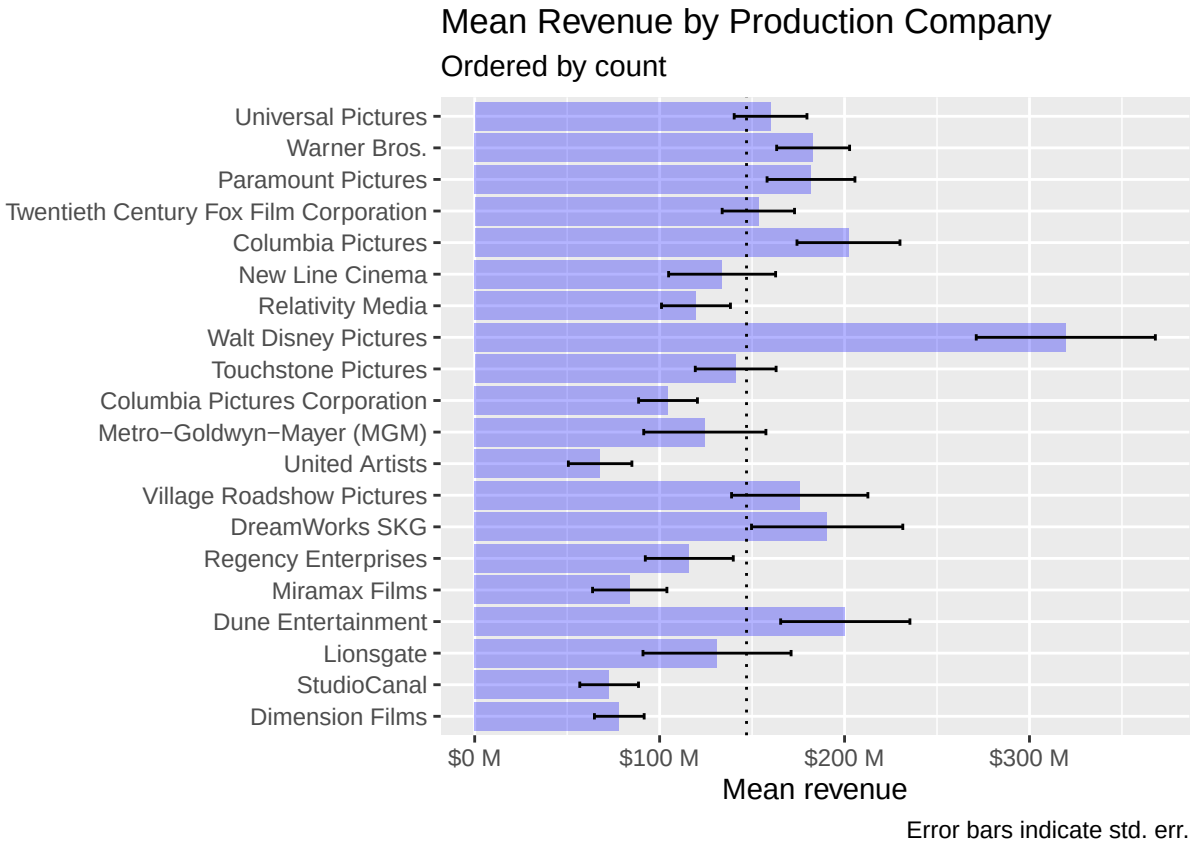
The fourth phase added in a few more controls. Firstly, we will test count of production companies, then try specific, high frequency companies with at least 20 movies produced. We feel this threshold is neccesary as without reduction, we would add thousands of features to the existing model.

Production companies in general have a logical association with both revenue and budget. Companies with more financing may have higher budgets on average and also produce more revenue from the goodwill of their brands.

Below, we examine the associated distributions.



Here, there is somewhat of a positive association between count of production companies and revenue, at least visually. There is a distinct curvature to this relationship, however. This skew is most pronounced towards the higher end of our data, perhaps indicating that high positive values are either rare or different from standard movies.



For production companies, the most pronounced trend is with Disney, as their mean revenue is substantially higher than other companies. This tracks with our own a priori assumptions about Disney in general, as they operate under the wing of an expansive franchise. We have chosen to only display those with at least 20 observations, as estimates for averages with less observations present a lot of skew.

Model 4.a: The first model of our fourth phase, adding count of production companies is defined as follows:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_4 pseudo.log(production.companies.count)$$

After adding the count of production companies, our Adjusted R^2 improves to 0.4194. Our partial F-test p-value, relative to the previous model adding genres, was 0.0032, so we can reject the null hypothesis that the coefficient for count of production companies is zero.

Model 4.b: Our second step, adding production companies with at least 20 observations is defined as follows:

$$\begin{aligned} log(revenue) = \beta_0 &+ \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_4 pseudo.log(production.companies.count) \\ &+ \beta_{5...n} production.companies_{1...n} \end{aligned}$$

After adding specific production companies with at least 20 observations, our Adjusted R^2 again improves to 0.4352. With this addition, our F-test p-value, relative to 4.a, is 1.6e-10, so we can reject the null hypothesis that the coefficient for all of the production companies is collectively zero. We also note that our previously tested coefficient, count of production companies, loses its significance. This may suggest multicollinearity between these fields.

With this in mind, we will test a final model retaining production companies and dropping count, defined as such:

Model 4.c:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_{4...n} production.companies_{1...n}$$

After dropping count, our Adjusted R^2 reduces to 0.4345. With this change, our F-test p-value, relative to the previously retained model, is 1.1e-12, so we can again reject the null hypothesis that the company coefficients are zero.

We also tested this last model against the previous model including count of production companies. Here the F-test p-value was 0.1047, and we fail to reject the hypothesis that the coefficient for count is zero. Again, this suggests multicollinearity, so we will take the more granular features in companies and remove count.

4 Results

4.1 Model Comparison

As discussed above, we iteratively added features, or ranges of dummy features in some cases, and used partial F-tests to determine whether we could reject the null hypothesis that the coefficients were zero. The one backwards revision was when we first added count of production companies, then the companies themselves, and noticed the p-value for the former coefficient dropped to insignificant. This encouraged us to retain only the companies and drop the count of companies.

Our final model is as follows:

$$log(revenue) = \beta_0 + \beta_1 log(budget) + \beta_2 release.year + \beta_{3...n} genre_{1...n} + \beta_{4...n} production.companies_{1...n}$$

It is important to note that we used robust standard errors in all of our tests to control for heteroscedasticity.

In our final model, and really all of our models, our coefficient for budget remained signifcant with a very low p-value, far under 5%. Our large data-set could influence this value.

Our practical significance is clear; for every 1% increase in budget, we can expect revenue to increase by 0.927%.

Table 2:

	Dependent variable:			
	log(revenue)			
	1	2a	3a	4c
	(1)	(2)	(3)	(4)
log(budget)	0.977***	1.030***	1.022***	0.927***
	(0.032)	(0.038)	(0.044)	(0.047)
release_year		−0.014***	−0.015***	−0.009**
		(0.004)	(0.004)	(0.004)
Constant	1.004*	27.338***	29.780***	20.396***
	(0.558)	(7.159)	(7.412)	(7.786)
Control genres	N	N	Y	Y
Control companies	N	N	N	Y
Partial F-test		13.13***	3.57***	5.13***
Partial base		1	2.a	3.a
Observations	1,859	1,859	1,859	1,859
Adjusted R^2	0.393	0.401	0.416	0.434
Residual Std. Error	1.371 (df = 1857)	1.362 (df = 1856)	1.344 (df = 1837)	1.323 (df = 1817)
Note:				*p<0.1; **p<0.05; ***p<0.01

This seems to be roughly uniform. If you increase budget, you can practically expect revenue to increase by almost the same amount, at least in holding the given covariates constant.

Our final adjusted R^2 was 43.4%, meaning our model explains roughly 43% of the total variance in revenue, penalizing feature additions.

5 Limitations

5.1 Statistical Limitations: Large-Sample Assumptions

There are two different large-sample assumptions to evaluate for our research design which are: (1) Independent and Identically Distributed (I.I.D.) Data (2) A Unique BLP Exists

5.1.1 Evaluation of Assumption One: I.I.D Data

The final cleaned dataset is not considered I.I.D data. There are multiple dependencies throughout our dataset. The first one is that the there are many movies in our dataset that are made by the same production companies. This creates dependency between movies with regards to release schedules and allocation budget.

In addition, many movies are dependent on each other with respect to release date. For example, in order to maximize revenue, a highly anticipated movie would not be released the same day as another highly anticipated move.

The dataset also has movies that are sequels, such as the "Harry Potter" franchise. These movies' release dates, past success/budget are all dependent on each other.

Essentially the population in question is constrained in many ways and as such this dataset cannot be considered I.I.D.

5.1.2 Evaluation of Assumption Two: A Unique-BLP Exists

This second assumption requires no perfect collinearity between the variables in the final model. Every variable in our final model has an associated coefficient as we can see in the table in the Results section. If any of the variables from the model had been dropped by R, this would indicate perfect collinearity, as otherwise, the model would not converge. As such we can conclude that there is a unique BLP. In addition to this, we have transformed our variables to control for "fat tails", which we can see in the Data Transormation section. This proves that a BLP does indeed exist.

5.2 Structural Limitations: Omitted-Variable Bias

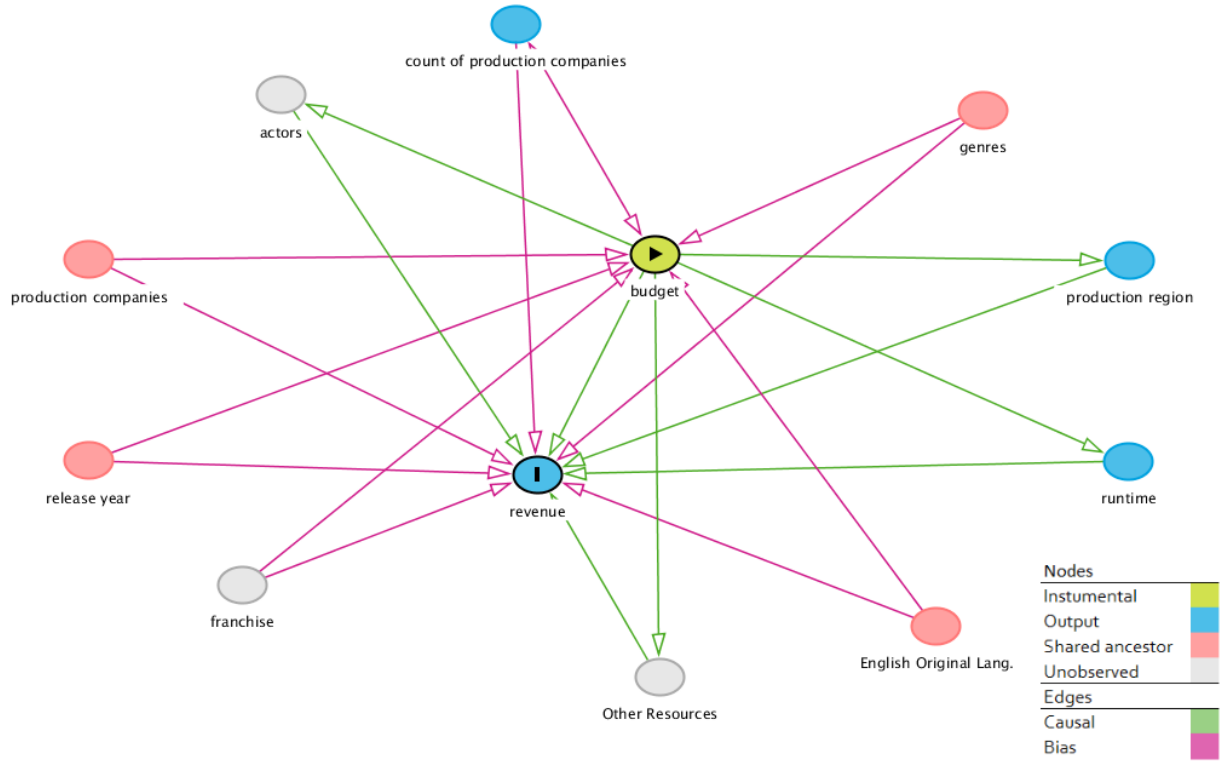
There are two primary omitted variables that we would have wanted to include in our model.

One of the omitted variables in our model are the actors in each movie. If a movie has a cast with high-profile celebrities it would most likely have a higher budget (well known celebrities have higher salaries) and it would increase revenue as people tend to want to watch movies with recognizable names. By adding in a variable representing the popularity of actors in a movie the bias would move away from zero.

A third omitted variable is whether or not a movie is based on something that is already popular (specifically book series, biographies and high profile criminal cases). For example, if the movie is based off of a very popular book series, there would be a positive relationship with budget (as the production companies would need to buy the rights to use the content) and there would be a positive relationship with revenue as fans of the book series expand the target audience. Including this variable would move the bias away from zero.

5.3 Shared Ancestors

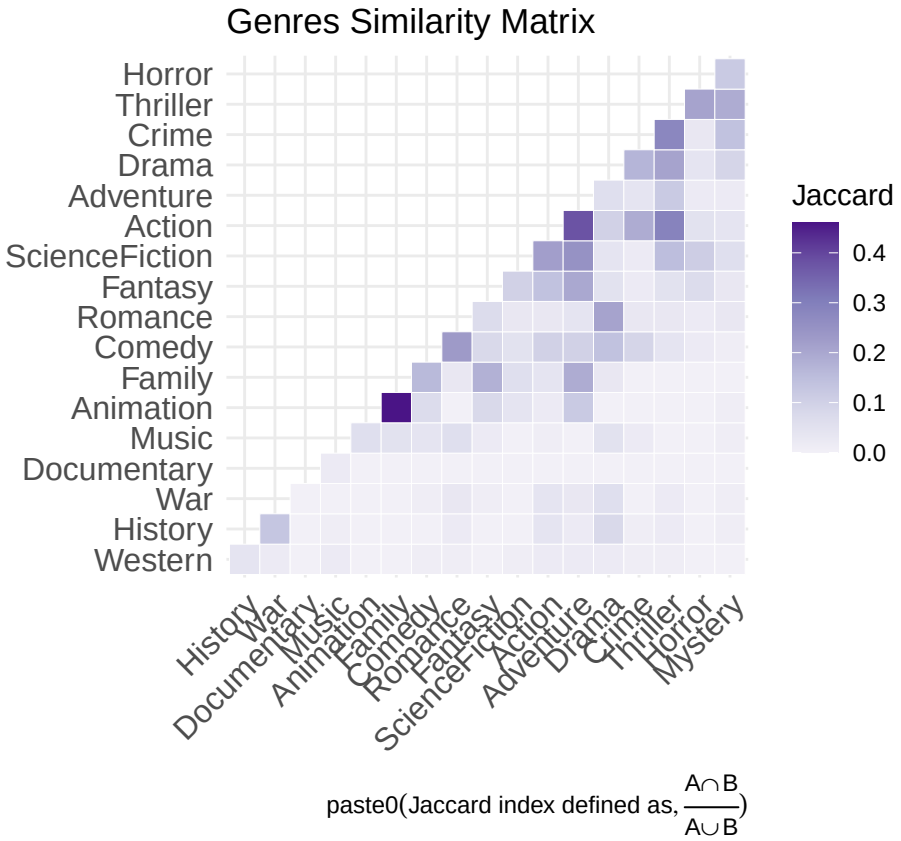
Our original causal graph, including controls whose H_0s we failed to reject in the current analysis is as follows:



In this image, it is clear there are many, *many* interactions within our feature set. As this is not a perfect world with perfect experiments, and we are relying instead on observational data, this is oftern the case. The main problem here is that there may be shared ancestors between our exposure and response. This could bias our coefficient estimate and invalidate the causality of our analysis.

5.4 Multicollinearity

Along with our interactions, many of our dummy fields are associated with one another. The similarity matrix below demonstrates this for genres in particular.



This multicollinearity could skew coefficients for genres in particular. If these interactions extend to other movie characteristics, this could, in turn, extend to budget. That said our overall fit should hold. In the future, we may want to explore MCA as a means of reducing our feature set.

6 Conclusion

6.1 Summary

In our analysis, we parsed user-inputs into neat orthogonal sets with which to evaluate the impact of budget on revenue. Each step in our process was intentional and methodical, and allowed us to reject the null hypothesis that there is no association between budget and revenue. Given the aforementioned shared ancestors and extensive interactions betwee our exposure and covariates, we cannot definitively say there is a *causal* relationship between budget and revenue.

That said, when holding each provided covariate constant, we could expect a 0.927% increase in revenue for every 1% increase in budget, a relationship of nearly 1-to-1. We achieve an adjusted R^2 value of 43.4%.

This data is not necessarily actionable, however, as this model is more explanatory than prescriptive or prescriptive. This is something we would want to expand on in the future.

6.2 Further Study

For future study there are more variables that can be added that were not available in this dataset. For example, one of them would be the MPAA movie rating. Movies that are rated PG-13 movies have a much

larger pertinent audience than movies that are rated R, since only a subset of the population are watching them.

We also would want to include data on actors in movies and their popularity. We could create a binary variable that identifies whether or not a movie had a popular celebrity in their cast. A popular celebrity could be a Top 10 actor in the year the movie was released.

Conducting additional analysis where we stratify our samples by things such as genre and production company could also be useful in an iterative manner.

Lastly, our dataset is open source, so in the future, we may want to standardize fields or pull from multiple sources. In line with this, we could also try to conduct experiments with given producers to map financial trends against fixed controls.

7 Appendix

7.1 Manual Budget and Revenue Corrections

id	title	release date	original revenue	original budget	corrected revenue	corrected bu
16340	Rugrats in Paris: The Movie	2000-09-14	$103	$30	$103,000,000	$30,000,000
14844	Chasing Liberty	2004-01-09	$12	$23,000,000	$12,000,000	$23,000,000
1613	The 51st State	2001-12-07	$14	$28	$14,000,000	$28,000,000
10397	Angela's Ashes	1999-12-25	$13	$25	$13,000,000	$25,000,000
2196	Death at a Funeral	2007-02-09	$46	$9,000,000	$46,000,000	$9,000,000
18475	The Cookout	2004-09-03	$12	$16,000,000	$12,000,000	$16,000,000
10944	In the Cut	2003-09-09	$23	$12,000,000	$23,000,000	$12,000,000
28932	F.I.S.T.	1978-04-26	$11	$11	$20,300,000	$8,000,000
217708	Of Horses and Men	2013-08-30	$11	$10	$239,969	$10,000,000
78383	Nurse 3-D	2013-09-28	$10,000,000	$10	$80,231	$10,000,000
13006	Split Second	1992-05-01	$5	$7	$5,000,000	$7,000,000
38415	Bran Nue Dae	2009-08-09	$7	$7	$7,000,000	$7,000,000
108346	Dreaming of Joseph Lees	1999-10-29	$7	$2,000,000	$12,044	$3,250,000
3082	Modern Times	1936-02-05	$8,500,000	$1	$1,800,000	$1,500,000
11980	The Prophecy	1995-09-01	$16	$8	$16,000,000	$8,000,000
51942	I Married a Strange Person!	1998-08-28	$203	$250	$203,000,000	$250,000,00