How does a movie's budget affect its box office earnings?

Datasci 203: Lab 2

Heather Rodney, Kolby Devery, Nan Li, Matt Pitz, Ada Guan

April 17, 2022

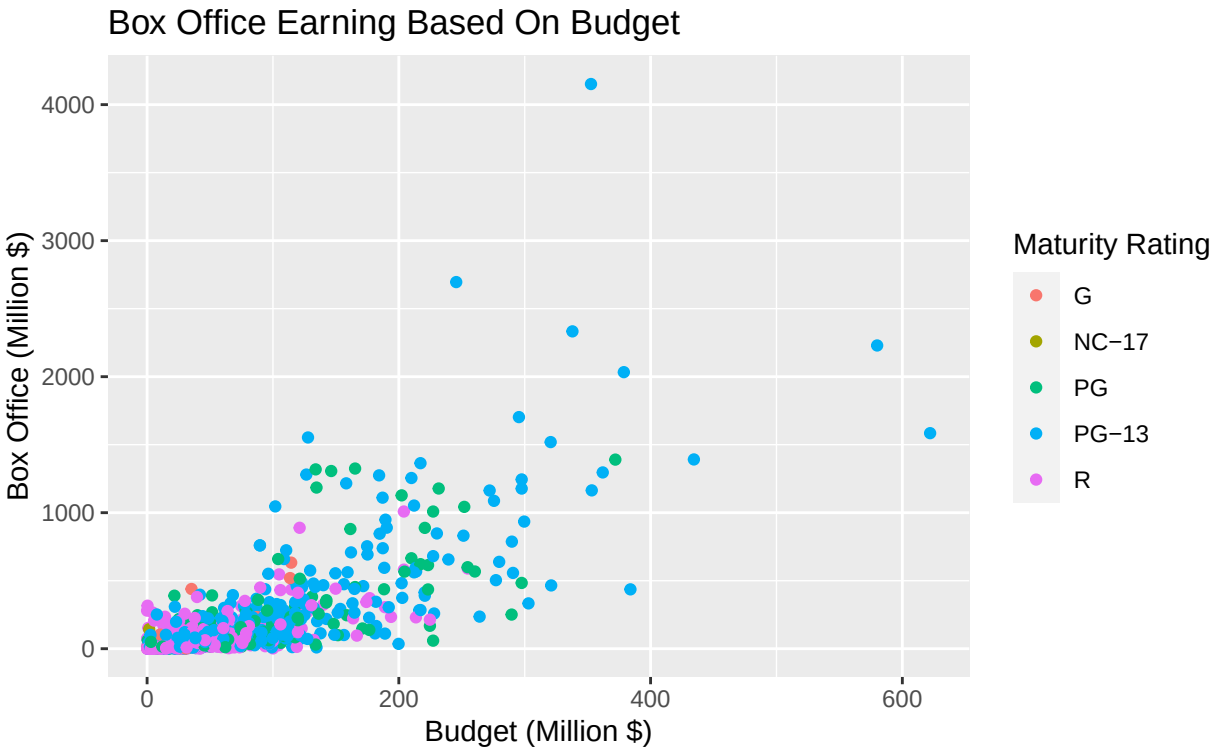
Introduction

Creating movies requires financial support that could range from thousands of dollars to millions of dollars. Until the movie is released, the stakeholders will not be able to recover the financial costs, and if the movie is successful, the stakeholders will be able to make a profit. As such, creating a movie is a financial risk, and for many individuals who fund movies, it would be helpful to know what factors increase and decrease the risks by ensuring the movie is successful and returns a sizable profit.

An article by Carrillat et al (2018)[1] performed an analysis on movies and those that were successful. The results indicated that the actor and their recognition as demonstrated by awards and nominations contributed to the success of a movie. Additionally, movie critic reviews also contributed as they tend to influence the consumer's decision on whether to see the movie. While this may be one article, there are other studies that have shown that recognition and actors likely help the success of the movie, as such those factors likely will affect the budget however that is likely one part of the equation as the genre could contribute if the actor or director does poorly in a movie, and therefore may affect the pay contribution.

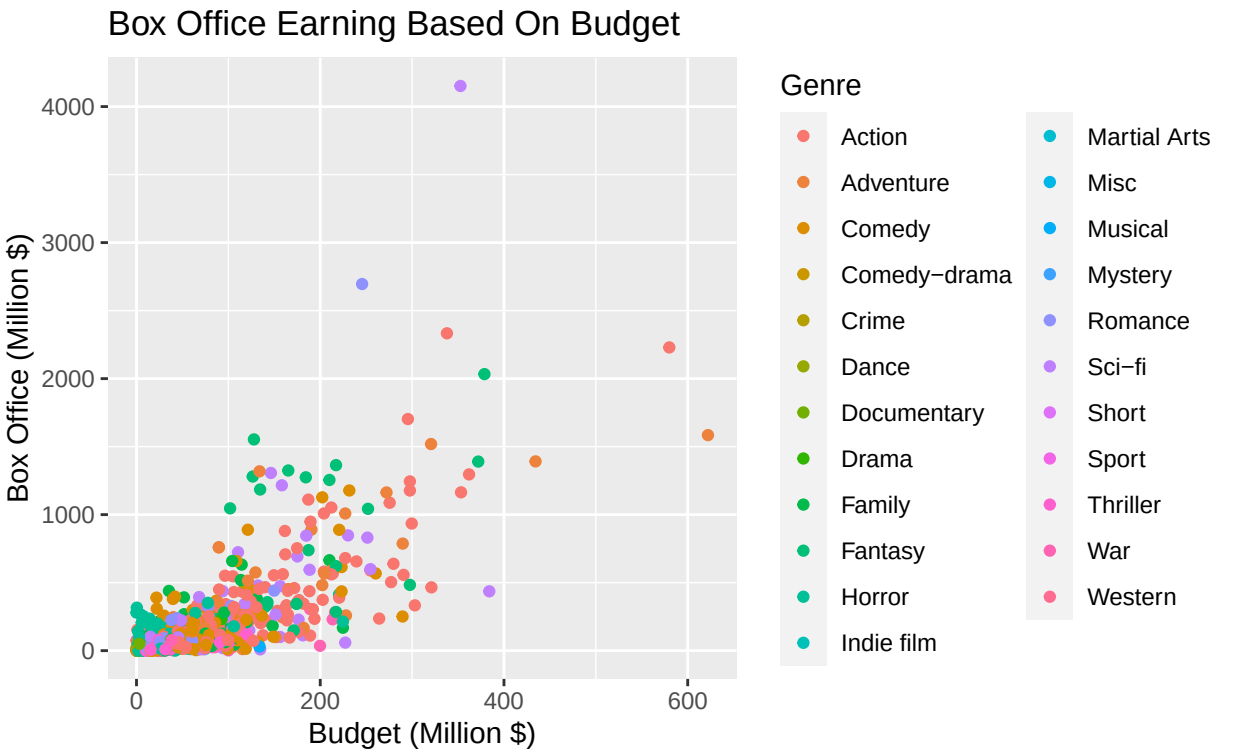
While our data does not include the salary of the actors and their awards prior to production, we do have some information that can help determine the success of the movie and the box office profits. To help understand the data and the patterns, a review of the data was performed. The following information was inferred:

1. Movies with the content rating of PG-13 with higher budgets tended to have higher box office sales compared to those with lower budgets. The movies with G and NC-17 are not fully visible in the scatter plot because they both have a small percentage of data in the dataset.



2. The genre of the movie may have an impact on the budget and the box office sales. As seen in the scatter plot below, action movies with higher budgets had higher box office results, compared to those with lower budgets. For some outliers, romance and sci-fi movies had higher box office results.

[1]Carrillat, F.A., Legoux, R. & Hadida, A.L. Debates and assumptions about motion picture performance: a meta-analysis. J. of the Acad. Mark. Sci. 46, 273–299 (2018), available at: https://doi.org/10.1007/s11747-017-0561-6



3. Overall, it appears that most of the movies tend to go at most $200 million in budget, and still have a reasonable profit.

By using the data from the Google Knowledge Panel on movies scraped from web sources from 1950 to 2020[2], our research will attempt to answer the question: "How does a movie's budget affect its box office earnings ?". Further questions can be answered in association with the research which include:

1. Does genre affect the box office for movies?
2. Does the month of release really matter?
3. What are the effects of duration and maturity rating on box office earnings?

To answer the main and sub-questions, a regression analysis will be performed to determine whether to accept or reject the null hypothesis: Spending has no effect on box office outcome.

Description of Data and Research Design

For this research project, we used movies from 1990 to 2015 that have both budget and box office information.

Description of data

The original dataset, American Movies Data (1950-2020) from Harvard Dataverse (Finberg, 2021) included 17,000 movies. However, due to missing data after 2015 and large amounts of missing box office and budget information, we decided to adjust the data for the analysis, which resulted in 794 movies.

[2]"Google Knowledge Panel Movie Data 1950-2020 - Harvard Dataverse, V1", available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9HDZFU

To answer the research questions we used a variety of numerical and categorical variables. Below is the description of the variables included in the analysis.

Dependent variable
Global Box Office Sales: The amount of money (US Dollars) made after release of movie

Independent variables
1. Budget: The amount of money spent on funding the movie 2. Genre: The primary movie category 3. Metacritic Rating: Movie critique of the movie prior to its release 4. Movie Release Date: The point in time during the year that the movie was released 5. Awards: Binary flag for whether the movie received awards 6. Maturity Rating: Movie content rating 7. Duration: The length of the movie in minutes.

The categorical variables used were genre, maturity rating and movie release. The awards variable is a numerical variable, but transformed to binary to indicate whether the movie received awards.

Descriptive statistics

Most of the movies were rated PG-13 (n=352) followed by R(n=335). For PG-13, 104 were Action, 93 were Comedy and for R rated movies, 88 were Action and 77 were Horror.

Below are additional tables that provide descriptive information on the data prior to the exploratory data analysis and data clean up.

Table 1: Box Office Sales per Rating (USD Millions)

	Ratings	Box.Office.Sales
1	NC-17	143.91
2	G	2,893.70
3	R	26,954.26
4	PG	31,899.59
5	PG-13	95,308.55

Note: Box Office adjusted for Inflation

Table 1 provides the total amount of box office sales in millions for movies in the different maturity ratings. PG-13 movies had the most in box office sales compared to NC-17 that had the least.

Table 2: Number of Movies per Rating

	Ratings	Number.of.Movies
1	NC-17	2
2	G	21
3	PG	154
4	R	335
5	PG-13	352

Table 2 provides the total number of movies that were created after 1990. Based on the information, most of the movies were PG-13, and NC-17 had the least amount of movies in the dataset.

Table 3: Number of Movies per Genre and Rating (Top 10)

	Genre	Rating	Number.of.Movies
1	Action	PG-13	104
2	Comedy	PG-13	93
3	Action	R	88
4	Horror	R	77
5	Comedy	R	74
6	Comedy	PG	54
7	Family	PG	40
8	Sci-fi	PG-13	38
9	Thriller	R	34
10	Romance	PG-13	26

Table 3 provides the top 10 genre and rating with the most amount of movies. PG-13 and R are still part of the top 10 along with some of the most common types of movies such as Comedy and Action.

Research Design

For this study, we are interested in modeling and understanding the relationship between budget and box office revenue of movies in the dataset. As a result of our Exploratory Data Analysis, we decided to create several new variables in our regression model. As the histograms in the next section show, distribution of the box office and the budget amounts is right-skewed and spans multiple orders of magnitude. Based on these characteristics, we added the log of both box office and log of budget to the dataset. We also went through several phases of data cleaning to help operationalize our research question, which will be discussed further in our data cleaning and exploratory data analysis section.

It is important to note that our dataset spans many years and potentially puts us in the realm of time series analysis. In order to simplify our data analysis and stay within the scope of this class, we made several decisions to avoid time series analysis while (hopefully) not making major compromises to the validity of our modeling. First we applied inflationary rates to both the budget and revenue data to provide a more reasonable comparison of movies across years (specifics are in the data cleaning section). We also assumed that consumer taste in movies has remained relatively constant from 1990 to 2015. As a final check on this assumption, we applied tests of statistical significance on models that included release year to see if there is evidence of release year playing an important explanatory role in box office revenue.

We also have 2 explanatory variables that are technically generated after the release of the movie. These variables are 'awards received' and 'Metacritic rating'. For the purpose of modeling, we argue that these are both markers of the intrinsic quality innate to the movie before its release, and assume that any increase in box office revenue as a result of these variables is due to the quality of the movie and not necessarily a result of 'buzz' from critics after release. While this point is debatable, we are making this simplifying assumption for the scope of this paper, while acknowledging there is likely room for further study here.

To address our central question, we will be using 4 different linear regression models with a mix of the previously discussed independent variables. Our baseline model includes only one independent variable, which is budget. Our second model includes budget along with movie duration, maturity rating, and primary genre. Our third model has all of the previously mentioned variables while adding Metacritic score. Our fourth model adds release window (summer/holiday) and awards received. Additionally, we have several models that test release month and release year to check the reasonableness of our assumption that release date has little effect on box office revenue. All of these models will be tested for statistical significance through a series of F tests while optimizing residual error and coefficient of determination. Full details of this are in the model and results sections.

Data Cleaning and Exploratory Data Analysis

Data Cleaning

The raw data required a number of cleaning steps prior to beginning our analysis. The genre variable combined each movie's primary and secondary genre into a single field. We split this field into two and relied on the primary genre. There were some genres that had small numbers of observations and were combined with like genres. For example, "Noir" films were combined with "Crime". We also processed the "Movie Release Date" to generate two binary variables: Summer and Thanksxmas, which are binary flags for whether the movie was released during summer months (May, June, July) or winter holiday months (November, December, January), respectively. Next both the budget and box office variables were included in a text field showing either dollars, millions of dollars, or billions of dollars. We cleaned these fields and standardized the values into millions of dollars. Finally, to further standardize the budget and box office variable, we incorporated the Bureau of Labor Statistics measure of inflation. To do this we set the cost inflation level for 1990 for 1 and apply the inflation percentage to each subsequent year. For example, if inflation in 1991 was 4.9%, the factor applied to all dollars in 1991 was 1.049. If the inflation in 1992 was 3.7%, then the updated factor would be (1+4.9%+3.7%) 1.086 for all 1992 dollars, and so on for each subsequent year.

Exploratory Data Analysis

We began by exploring the distribution of numerical variables of interest. The figure below shows that the distributions of box office and budget are highly right-skewed and span multiple orders of magnitude. Therefore, we applied a log transformation to these two variables, which resulted in a more normal distribution. As a result, the original values between 0 and 1 million dollars were dropped after this transformation.

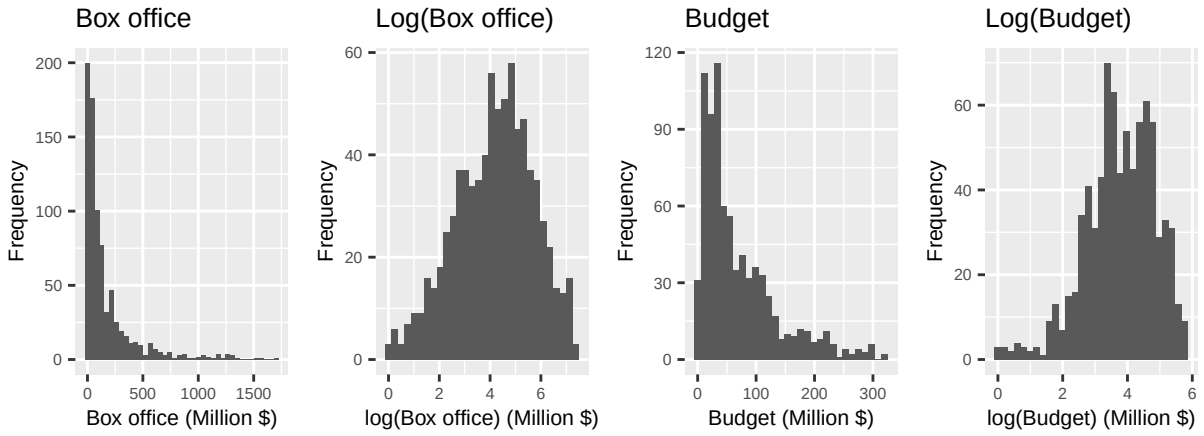


Figure 1: Data Transformations

We removed the extreme outliers (box office >1990 Million $ and budget > 350 Million $) based on observations from the box plots during our EDA. These plots are shown in the Appendix.

We applied a pairwise plot matrix shown in the figure below to quickly identify the visual relationship between these variables. Scatterplots of each pair of numeric variables are displayed on the left part of the plot. Pearson correlation coefficients are displayed on the right. Variable distribution is available on the diagonal. The scatter plots in the first column of this figure show the relationship between the log of

box office and the other independent variables. There is an overall increasing trend at the box office, with increasing the budget, duration, and Metacritic rating. Based on the correlation coefficients that are listed on the right part of this figure, we identify that box office and budget after the log transformation are highly correlated. The other variable pairs are either moderately correlated or slightly correlated. This exploration provided early signals that there is no perfect collinearity between the independent variables.

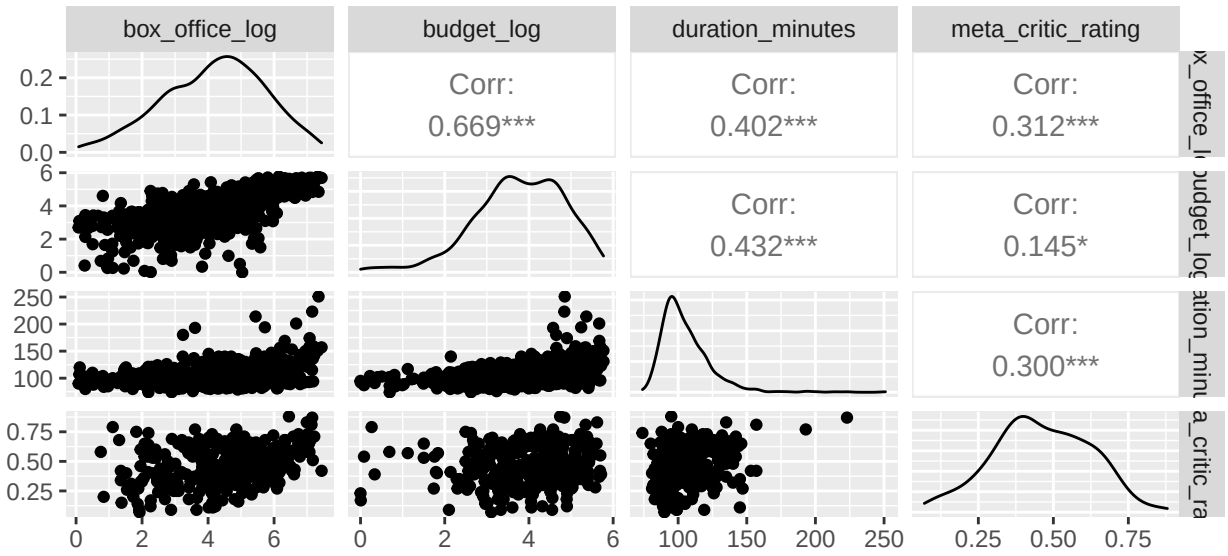


Figure 2: Correlation plots

Modeling

Our goal is to measure how the budget of a movie affects its box office earnings, while holding all other characteristics of the movie constant. We will consider duration, maturity rating, primary genre, Metacritic rating, release month and received award as our covariates.

Base Model
The baseline model only include the primary independent variable of interest, log transformed budget variable:

$$\log(\text{Box office}) = \beta_0 + \beta_1 * \log(\text{Budget})$$

The budget shows a strong positive correlation to box office and is statistically significant. The coefficient estimate tells us that for every 1% increase in the budget, the box office increases by about 0.94%.

Second Model
In the second model, we include additional independent variables of interest, duration, maturity rating, and primary genre:

$$\log(\text{Box office}) = \beta_0 + \beta_1 * \log(\text{Budget}) + \beta_2 * \text{duration} + \beta_3 * \text{maturity rating} + \beta_4 * \text{primary genre}$$

The null hypothesis is that the first and second model have equal residual sum of squares. We conducted a F test between the first and second model, since the p-value is less than 0.05, we can reject the null hypothesis. The additional covariates included makes the second model more fully representative of the true population.

Third Model
In the third model, we include independent variable Metacritic rating:

$$\log(\text{Box office}) = \beta_0 + \beta_1 * \log(\text{Budget}) + \beta_2 * \text{duration} + \beta_3 * \text{maturity rating} + \beta_4 * \text{primary genre} + \beta_5 * \text{Metacritic rating}$$

Based on the F test between the second and third model we observe that the covariate Metacritic rating has additional explanatory power (able to explain the variance in the dependent variable better than the first and second model) and is improving the model performance.

Fourth Model
In the fourth model, we include binary variables thanksxmas, summer and award:

log(Box office) = β_0+β_1*log(Budget)+β_2*duration+β_3*maturity rating+β_4*primary genre+β_5*Metacritic rating+β_6*thank

From the results of the F test we failed to reject null hypothesis since the p-value is greater than 0.05. The three additional covariates in the fourth model do not have additional explanatory power and do not improve the performance of the model.

Fifth - Seventh Model
We included three additional models to assess the time of release and its effect on box office. We did not see statistical significance or additional explanatory power in covariates; release month, summer (created from release month) and release year compared to the third model. The lack of significance in release year suggests that treating the models as a snapshot in time rather than a time-series can be justified.

Even though the fifth model has the highest R^2, the lack of statistical significance of time of release makes the third model the most appropriate model.

Assessment of Classic Linear Model Assumption

Although our dataset is large enough to allow us to rely on the large sample assumptions. We nonetheless check all assumptions required for the classical linear model to better understand where our model may be limited.

IID Data
As mentioned earlier, the movie dataset from the Google knowledge panel is web scraped and sampled from multiple sources. We can assume IID is satisfied based on the method used for sampling, however we believe movies that are sequels and franchises may raise IID concerns.

No Perfect Collinearity
In our early stages of EDA, there were indications of no perfect collinearity between our independent variables of interest. To further support this claim we computed the variance inflation factor. The VIF values were less than 2 across our independent variables (the largest value was 1.4) and it did not suggest that redundancy is present between our predictor variables.

Normally Distributed Errors
We can assess the normality of the error distribution by utilizing Q-Q plot (theoretical quantities on the x-axis and residuals on the y-axis). The points in the Q-Q plot should form a relatively straight line, indicating that the quantiles of the dataset match what the quantiles of the dataset would theoretically be.The histogram of our residuals, shown in the appendix, appears normal however the tails of the Q-Q plot suggests that the model is not fitted well for movies with very low or high budget. The Shapiro-Wilk test was used to further investigate the distribution of the residuals. With a p-value < 0.05 we can reject the null hypothesis and conclude that the errors are not normally distributed. However, the areas where the residuals are off appear to be largely around the tails of the distributions. This suggests that our model may not be suited for indie films and extremely large blockbusters, but everything else in between may be fine.

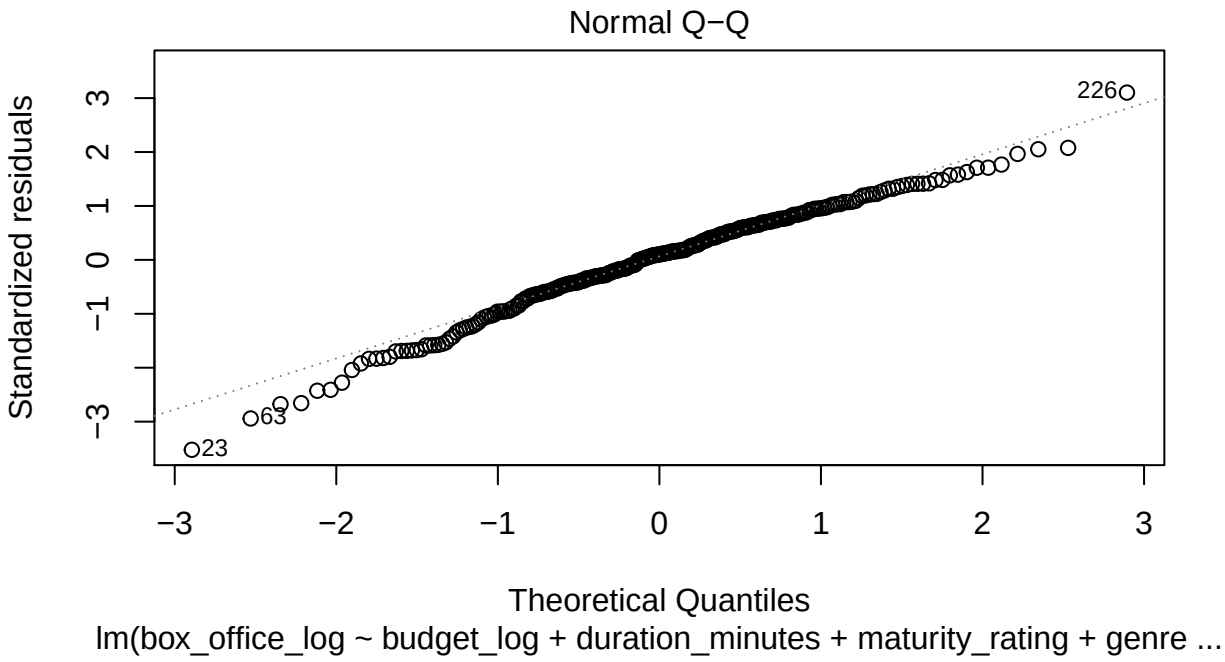


Figure 3: Residual QQ plot

Linear Conditional Expectation

To assess the linear conditional expectation we can plot the predicted number of views on x-axis and model residuals on the y-axis. Look for residuals that are units above the predicted value at that point. The residuals appear to be linear across the range of x. Based on the residuals and fitted graph, it does not seem like there are any points where the residuals are significantly different from 0. The model satisfies the assumption of a linear conditional expectation.

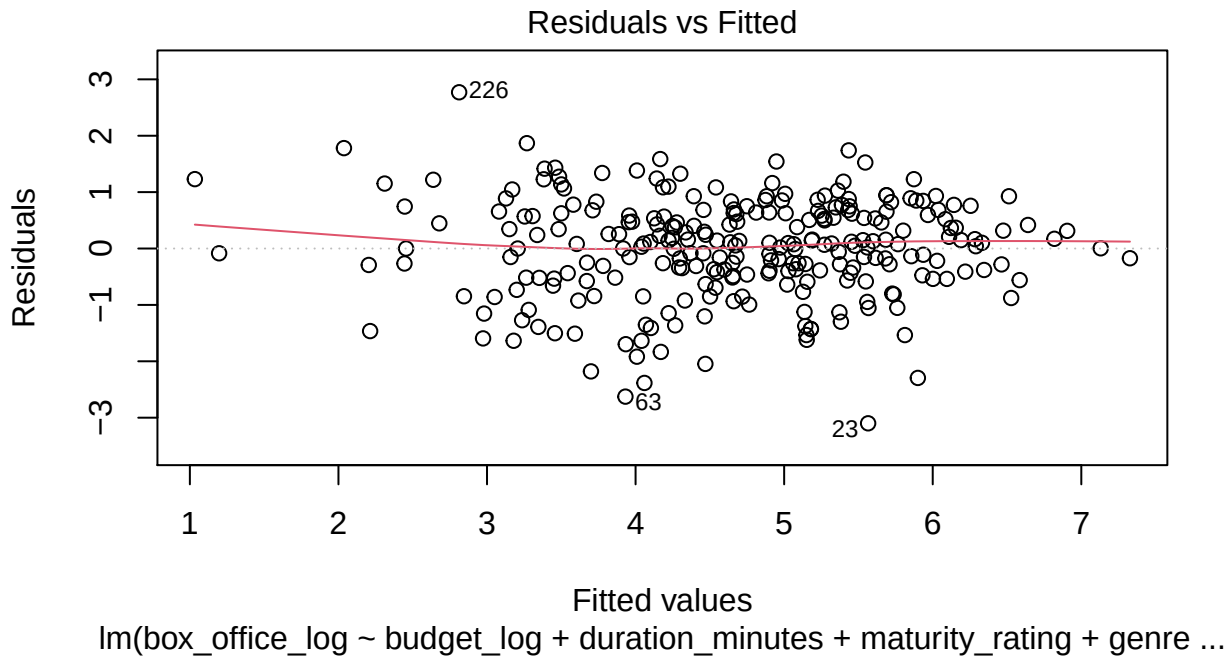


Figure 4: Residuals vs Fitted

Model Results and Intrepretation

Across our model sets, we had a range of both statistically significant and statistically insignificant coefficients.

The base model indicates a log(budget) of .94 with an adjusted coefficient of determination of .489 and a residual error of .995. Due to the log transformation on both variables, we can interpret from this model that every 1% increase in budgets may result in a .94% increase in box office revenue. However, there are likely other explanatory variables that capture this effect.

The second model includes several additional explanatory variables. These include the quantitative variable of duration (in minutes), the categorical variable of maturity rating, and the categorical variable of genre. After performing an F test between this model and the base model, the performance of the second model was statistically significant and reduced residual errors while increasing the adjusted coefficient of determination (R^2: .549, RSE: .935). As expected, the coefficient of log(budget) was reduced to .815 while explanatory effects were captured in other variables. The movie duration coefficient was found to be .011 and statistically significant. It is also worth noting that the interpretation of our explanatory variable coefficients (that have not been log-transformed) would need to be interpreted through the following equation: (exp(coefficient) – 1) * 100[3]. Of the maturity rating categories, PG and R were mildly significant, with coefficients of -.808 and -1.18 respectively. The genres that pulled box office revenue in the positive direction were Documentary, Fantasy and potentially Horror, while the genres that pulled revenue in the negative direction were Sport, Mystery, and potentially Comedy and Drama.

[3]"Interpreting Log Transformations in a Linear Model", available at: https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

Table 4: Regression Results

	Dependent variable:			
	box_office_log			
	(1)	(2)	(3)	(4)
budget_log	0.940***	0.815***	0.823***	0.799***
	(0.074)	(0.084)	(0.084)	(0.085)
duration_minutes		0.011***	0.008*	0.007
		(0.004)	(0.005)	(0.004)
maturity_ratingPG		−0.808*	−0.504*	−0.647*
		(0.412)	(0.294)	(0.340)
maturity_ratingPG-13		−0.742	−0.387	−0.521
		(0.451)	(0.350)	(0.395)
maturity_ratingR		−1.180**	−0.837**	−0.939**
		(0.466)	(0.372)	(0.411)
genre_primaryAdventure		0.214	0.159	0.146
		(0.328)	(0.311)	(0.297)
genre_primaryComedy		−0.316*	−0.342*	−0.365**
		(0.176)	(0.180)	(0.182)
genre_primaryCrime		−0.245	−0.429	−0.415
		(0.655)	(0.661)	(0.713)
genre_primaryDocumentary		1.448***	1.405***	1.227***
		(0.375)	(0.375)	(0.379)
genre_primaryDrama		−0.848*	−0.837**	−0.740*
		(0.485)	(0.407)	(0.413)
genre_primaryFamily		0.030	0.059	0.031
		(0.316)	(0.314)	(0.321)
genre_primaryFantasy		0.459**	0.384*	0.490**
		(0.206)	(0.201)	(0.207)
genre_primaryHorror		0.419*	0.502**	0.515**
		(0.245)	(0.239)	(0.241)
genre_primaryMystery		−1.794***	−1.659***	−1.611***
		(0.148)	(0.150)	(0.156)
genre_primaryRomance		−0.018	0.0004	0.020
		(0.204)	(0.210)	(0.206)
genre_primarySci-fi		−0.302	−0.331*	−0.330*
		(0.201)	(0.194)	(0.192)
genre_primarySport		−0.977***	−1.224***	−1.111***
		(0.187)	(0.202)	(0.215)
genre_primaryThriller		−0.008	−0.109	−0.108
		(0.293)	(0.296)	(0.302)
meta_critic_rating			1.320***	1.234***
			(0.411)	(0.414)
thanksxmas				0.022
				(0.163)
summer				0.239
				(0.147)
received_award				
Constant	0.817**	1.024	0.393	0.671
	(0.326)	(0.681)	(0.649)	(0.703)
Observations	266	266	266	266
R^2	0.491	0.579	0.597	0.602
Adjusted R^2	0.489	0.549	0.566	0.568
Residual Std. Error	0.995 (df = 264)	0.935 (df = 247)	0.917 (df = 246)	0.915 (df = 244)
F Statistic	254.531*** (df = 1; 264)	18.910*** (df = 18; 247)	19.209*** (df = 19; 246)	17.610*** (df = 21; 244)

Note: *p<0.1; **p<0.05; ***p<0.01

The third model includes all of the explanatory variables in model 2 while also adding Metacritic ratings. After performing F tests of this model against both models 1 and 2, it was determined that Metacritic ratings did have a statistically significant impact on explanatory power. This model has an adjusted coefficient of determination of .566 and residual error of .917, which is the best performance of a model that passed tests of statistical significance. The coefficient of log(budget) remained relatively constant at .823, while the statistical significance and coefficients of some of the genres changed. This may be due to the introduction of Metacritic ratings, that capture some of the explanatory power of genre (some genres may inherently perform better with critics and sell more tickets). Additionally, the statistical significance of movie duration decreased significantly, meaning the explanatory power of movie length may be captured elsewhere.

The fourth model includes all of the explanatory variables of the third model, while adding the binary coefficients for summer releases, holiday releases, and whether the movie received any awards. Like mentioned in the model section, this did not pass any F test for statistical significance, so will not be used. Additionally, we created a series of models (5-7) to test the explanatory power of time of release on box office sales. As discussed, we treated our data as a single snapshot in time to avoid time series analysis. These models were used to test against our assumption that we can reasonably analyze this data without time series assumptions as long as we account for inflation. This assumption is supported by the lack of statistical significance that we have for models that include release month, year, or release window (summer/holiday).

Limitations of the Model

Data Issues

Since the goal of our report is to determine the impact of additional spending on a movie's box office, we need both variables along with the other RHS variables considered to be present in our data for the analysis. While the dataset captures information through 2020, these inputs are not available in the data after 2015, so all of those movies are dropped. Additionally, the majority of films prior to 2015 do not have both pieces of information. Overall this reduces the number of observations to 266.

In an ideal scenario, we would be able to break the spending into its various components like actor salaries, script writers, director, and production expenses like CGI. This would allow us to determine the impact of specific types of spending. However, this information is almost exclusively non-public information. There are instances where individual actor's salaries have become public, but after researching for additional data, that appears to be the exception rather than the norm.

We make a strong assumption that when we restrict the data to the time period between 1990 and 2015 and control for inflation, we can treat the data as a panel dataset. We believe that customer tastes and movie going trends have remained relatively constant during this time and support this assertion with the general observation of the continued prevalence of movie franchises, sequels, and reboots. However, we checked our assumption by including the release year as a time trend variable as a modeling sensitivity. The coefficient on release year and the F-test between model 3 and this new model both showed that including release year did not improve the fit of our model.

Structural Limitations

Movies are complex productions and there are some potentially important omitted variables from our models. In addition to the components of spending, we would ideally have information on the specific actor, director, or studio effects. The employment of a specific headliner could lead directly to increased box office returns. Our use of the Metacritic rating variable attempts to apply a proxy for this. We believe that the Metacritic rating indicates the movie's intrinsic quality that exists prior to release in theaters.

Since our model uses data through 2015, the relationships described here may no longer hold in a post-covid movie going world. Throughout 2020 and 2021 we observed the movie industry struggles and saw many

production studios shift some major releases to streaming services. Although by the end of last year, there were strong signs of a return to normalcy and consumer's continued desire to see movies in person.[4] Taking these concerns into consideration, we believe that our findings will remain directionally consistent.

Finally, a regression based analysis may not be the most appropriate method to use when analyzing box office returns. A model that allows for non-linear relationships, a broader application of categorical variables, and controls for missing data, like a gradient boosting model.[5]

Statistical Limitations

One potential criticism of whether the data is IID is movie sequels and franchises. Consumers may already have direct experience with the quality of these movies and that may in part drive the box office returns. Studios may flood the market with sequels as studios seek to cash in on what they believe is a sure thing.[6] However, the findings on actual returns are less clear.[7] When a movie studio invests in releasing a sequel, we believe that it is little different than investing in a movie's intrinsic quality, which we capture through the use of rating. They may also fail to sufficiently invest in that quality, as there have been numerous failed sequels and reboots.

Much of the data that we drop relates to movies with small budgets or small box office returns, where one or both pieces of information is missing. These may represent the world of indie films, which implies our model may not be useful to describe the impact of spending when spending will be extremely limited.

There may be interdependence between some of our right hand side variables. We argue that Metacritic rating, although observed after the movie has been released, are instead indicative of intrinsic quality. However, if there is a strong interaction between this quality variable and spending itself, our beta may be unreliable. If we believe that to be an issue, we can instead rely on our most simplistic model, model 1, that only compares spendings to box office investment. This more simplistic model also finds that increases in spending are related to increased box office performance.

Finally, we have run multiple regressions and multiple tests on the regression coefficients in preparing this analysis. As there are a number of coefficients, especially in the genre category, that are significant at the 90% or 95% level, this suggests at least some of the significance findings on the coefficients may be a type 1 error (reject the null that the beta is 0).

Conclusion

Returning to the original research question, "how does a movie's budget affect its box office earnings ?", we find there to be a significant relationship between spending and earnings. Specifically, we find that an increase in spending is associated with box office success, so if a studio wants a larger return, they may feel safe investing money. However, the data relied on for this study is survey data, so our findings do not suggest a causal relationship, but merely a strong relationship between increased spending and box office returns. This effect shows up in each of the various model versions run, indicating that the relationship is robust.

Our analysis of the sub questions suggest that some industry norms may be correct. R ratings appear to have a negative impact and some genres are better money makers than others. Longer movies also appear to generate larger returns. However, seasonality, peaks around holidays or summer blockbusters may be less of a function of the specific month than the studio schedules themselves. If the studio believes that they have a successful movie on their hands, or even a mediocre one they are attempting to find a release timing for, they do not need to put as much weight on the release calendar.

[4] "Box Office Rebound: Wall Street Analysts – The Hollywood Reporter", available at: https://www.hollywoodreporter.com/business/business-news/box-office-rebound-wall-street-analysts-1235024902/

[5] See "Gradient boosting - Wikipedia"

[6] "Nearing the endgame: is Hollywood's lust for sequels destroying cinema? | Movies | The Guardian" available at: https://www.theguardian.com/film/2019/may/16/hollywood-sequels-cinema-avengers-endgame

[7] "The sequels that are bombing badly at the box office this year", available at https://www.businessinsider.com/2016-sequels-bomb-at-box-office-2016-6

Future considerations

Should we or other researchers return to this topic, there are a number of areas they could look into to improve upon this study. First, future researchers could partner directly with a studio to gain access to more detailed financial information. This would allow them to assess the impact of spending in specific areas. Next, they may also consider performing a true time series analysis. While we have manipulated the data in order to remove time varying factors, there may be important industry or macro trends that could be included with a time series analysis. Finally, researchers could look for ways to directly address a broader range of categorical variables like the impacts of specific talents, sequels, and whether basing a movie on pre-existing materials, like a successful novel.

Appendix

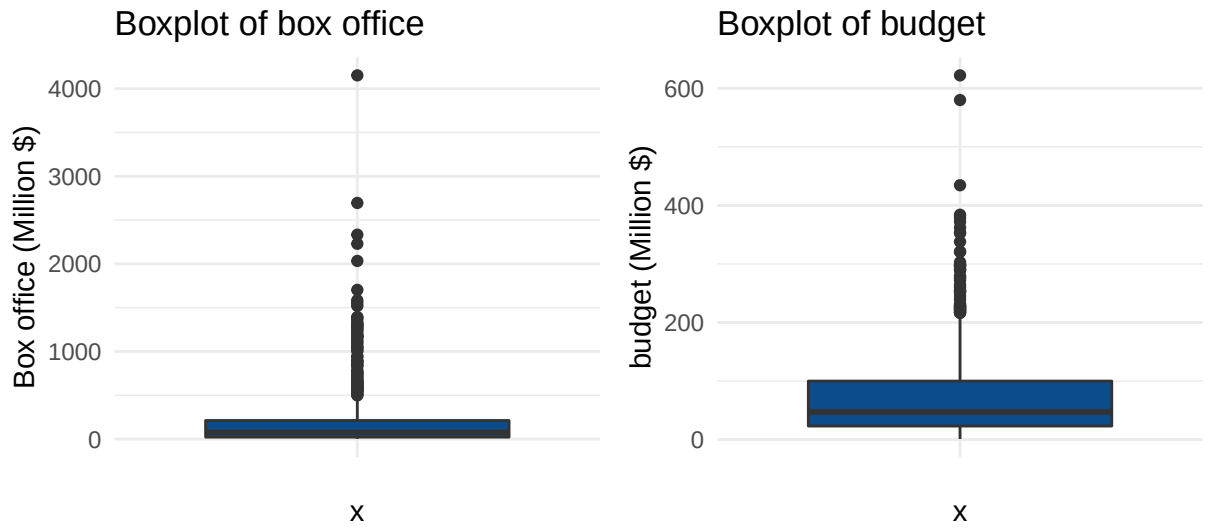


Figure 5: Outlier box plots

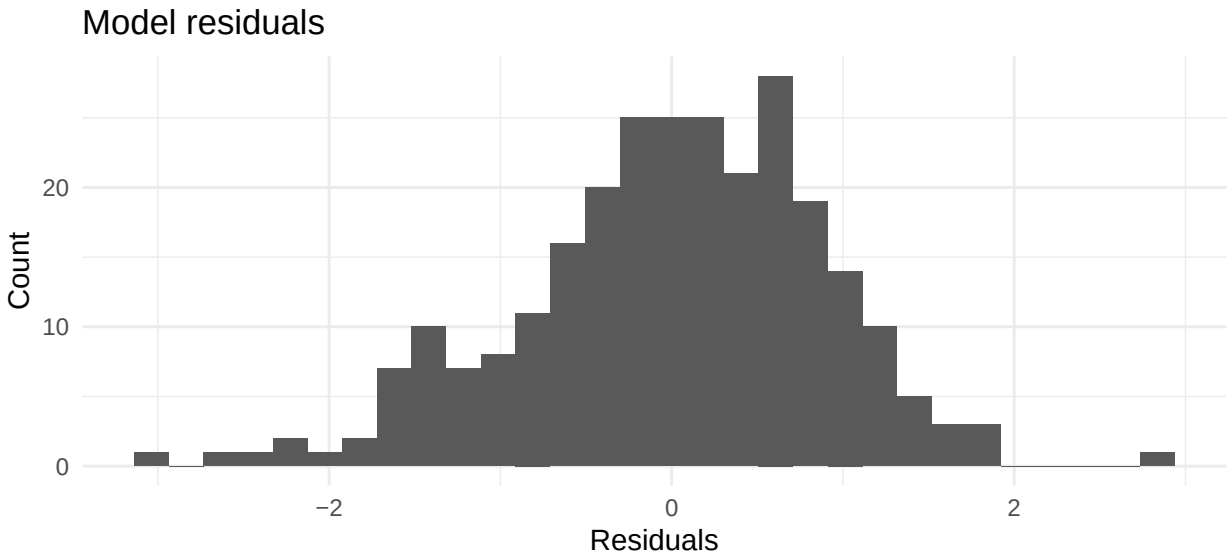


Figure 6: Residual Histogram

Table 5: Sensitivity Regression Results

	Dependent variable:		
	box_office_log		
	(5)	(6)	(7)
budget_log	0.807***	0.800***	0.825***
	(0.084)	(0.085)	(0.092)
duration_minutes	0.008*	0.007	0.008*
	(0.005)	(0.004)	(0.005)
maturity_ratingPG	−0.576	−0.658**	−0.502*
	(0.393)	(0.330)	(0.295)
maturity_ratingPG-13	−0.421	−0.534	−0.385
	(0.429)	(0.373)	(0.351)
maturity_ratingR	−0.834*	−0.953**	−0.836**
	(0.442)	(0.391)	(0.374)
genre_primaryAdventure	0.154	0.147	0.158
	(0.299)	(0.298)	(0.309)
genre_primaryComedy	−0.345*	−0.366**	−0.342*
	(0.181)	(0.181)	(0.181)
genre_primaryCrime	−0.383	−0.413	−0.427
	(0.809)	(0.713)	(0.662)
genre_primaryDocumentary	1.400***	1.229***	1.417***
	(0.393)	(0.380)	(0.447)
genre_primaryDrama	−0.815*	−0.742*	−0.839**
	(0.452)	(0.411)	(0.408)
genre_primaryFamily	0.078	0.033	0.059
	(0.332)	(0.323)	(0.314)
genre_primaryFantasy	0.495**	0.495**	0.387*
	(0.242)	(0.204)	(0.207)
genre_primaryHorror	0.550**	0.517**	0.504**
	(0.251)	(0.239)	(0.237)
genre_primaryMystery	−1.855***	−1.616***	−1.665***
	(0.327)	(0.151)	(0.191)
genre_primaryRomance	0.070	0.022	0.003
	(0.228)	(0.208)	(0.217)
genre_primarySci-fi	−0.326	−0.327*	−0.332*
	(0.200)	(0.190)	(0.193)
genre_primarySport	−1.358***	−1.117***	−1.217***
	(0.374)	(0.207)	(0.246)
genre_primaryThriller	−0.104	−0.105	−0.105
	(0.316)	(0.301)	(0.296)
meta_critic_rating	1.224***	1.235***	1.318***
	(0.415)	(0.413)	(0.420)
release_monthAugust	0.140		
	(0.340)		
release_monthDecember	0.179		
	(0.375)		
release_monthFebruary	0.319		
	(0.314)		
release_monthJanuary	0.374		
	(0.371)		
release_monthJuly	0.403		
	(0.333)		
release_monthJune	0.536*		
	(0.305)		
release_monthMarch	0.423		
	(0.355)		
release_monthMay	0.595*		
	(0.348)		
release_monthNovember	0.367		
	(0.339)		
release_monthOctober	0.537		
	(0.410)		
release_monthSeptember	0.476		
	(0.378)		
summer		0.231*	
		(0.134)	
release_year			−0.001
			(0.011)
Constant	0.177	0.685	1.961
	(0.777)	(0.690)	(22.494)
Observations	266	266	266
R^2	0.612	0.602	0.597
Adjusted R^2	0.563	0.570	0.565
Residual Std. Error	0.921 (df = 235)	0.913 (df = 245)	0.919 (df = 245)
F Statistic	12.364*** (df = 30; 235)	18.564*** (df = 20; 245)	18.175*** (df = 20; 245)
Note:			*p<0.1; **p<0.05; ***p<0.01