Keeping it PG-13

Steve Hewitt, Martin Lim, Fidelia Nawar, & Austin Sanders

12/9/2021

1. Introduction

1a. Context

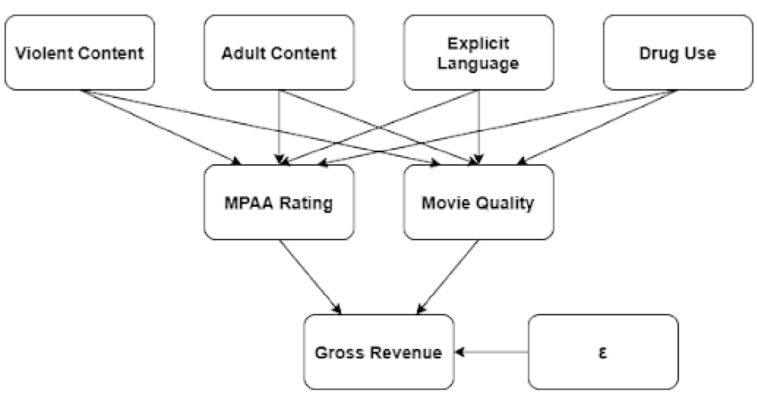
Acme Studios has spent a $50,000,000.00 budget on a superhero movie, and the director insists that the movie should include a scene where the main villain goes on an expletive-laden tirade. We know that including this scene will mean that the movie will be rated R, and cutting the scene will result in the movie being rated PG-13. The director is extremely upset that we want to cut the scene and says we're ruining the film's artistic integrity by trying to make editorial changes after the director's cut. So upset that he went directly to the studio head to complain. Now Acme Studios' executive team has to decide: do they modify the movie for a more family-friendly rating, or do they respect the director's wishes and release it as-is? As data scientists, we would have difficulty quantifying artistic integrity or the value of the relationship between the studio and the director. Still, we feel strongly that we can show the relationship between worldwide revenue for a PG-13 vs. R ratings (holding all other variables constant). The studio head wants to know: How much more money do they expect to make by defying the director's wishes and cutting the movie to make it PG-13?

1b. Research Question

Holding other factors constant, how much more money should a movie studio expect to make on a film that gets a PG-13 rating instead of an R rating from the MPAA?

Our research question intends to measure the impact *MPAA Rating* and *Movie Quality* on the *Gross Revenue* that it will generate. Given other available, quantifiable factors like the *Budget*, *Genre*, and *Runtime*, this study also intends to investigate if they affect the *MPAA Rating* and/or *Movie Quality* which in-turn affect the *Gross Revenue.*

1c. Causal Theory



Our research question seeks to measure the impact of the *MPAA Rating* (more specifically, PG-13 versus R) on the *Gross box office revenue.* A movie typically receives an R rating by the MPAA for some combination of violent content, adult content, explicit language, and drug use. In addition to these factors contributing to the MPAA rating, they also contribute to the quality of the movie. We expect to show that both MPAA rating and movie quality impact the gross revenue of the film. By adjusting the movie's content to secure the desired rating, we may also affect the quality of the movie. Therefore, this study will explore models that include a proxy for movie quality to attempt to minimize omitted variable bias.

2. Research Design and Data

2a. Data Source

The data used for this study is a movie dataset from Kaggle (https://www.kaggle.com/danielgrijalvas/movies). It contains 7512 unique movie titles ranging from the year 1980 to 2020. According to the description of the creator of the dataset, the data was scraped from IMDb.com so the extent of the movie title coverage can go as far as the available information posted on the IMBd website. Below are the important columns that were considered for this study:

- Outcome Variable
  - `gross` : Revenue of the movie in USD

- Explanatory Variables
  - `score` : Average IMDb user rating
  - `rating` : MPAA rating of the movie (R, PG, etc.)
  - `runtime` : Length of movie in minutes
  - `budget` : Budget of the movie in USD
  - `votes` : Number of user votes on the IMBd website

2b. Research Design

Using the data and variables above, this study measured the impact *MPAA Rating* and *Movie Quality* on the *Gross Revenue* that a movie will generate. Causal models were generated using the logic from the causal theory in 1c. `gross` is the main quantified success variable. According to the causal theory, two main explanatory quantities were included in the model. `rating` was used to quantify the MPAA rating of the film while `score` is the main quantified measurement of quality. `budget` and `runtime` both affect the the movie quality, but not the `rating` so they were considered as proxies for quality.

The outcome of interest is `gross`, the revenue of the movie in USD. This is the main outcome variable of all the causal models and is quantified in numeric form as is.

In order to properly quantify `rating` in the model, indicator variables were used. The data was divided into two (2) categories: PG-13, and R. All movies rated PG-13 were given a separate indicator variable (`PG13`) while all movies rated R were treated as the base-case.

Measuring movie quality is primarily measured by the `score` variable. Variables like `budget` and `runtime` also have an effect on quality so they were considered as possible proxies for quality in the causal models.

In order to remove duplicate movie titles, the data point with the larger budget was retained. If both the budgets and titles were equal, the data point with the larger revenue was retained.

Because our dataset covers a period of 50 years, we are applying a CPI-based price adjustment to each monetary variable to account for inflation. The inflation-adjusted valuse (in 2020 dollars) will be represented with the 'adj_gross' and 'adj_budget' variables.

The filtered dataset was used to produce multiple linear models and evaluate them using coefficient tests in R. Stargazer was used to compared the models to determine which model best aligns with the data and our causal theory. Armed with this chosen model, its predictions were evaluated to check how well the model is able to predict the revenue based on the input parameters. Finally, the model was applied to the specific case outlined in the overview section above to predict the revenue for both a PG-13 and R MPAA rating case.

2c. Data Cleaning

Removed entries with budget under 1 or gross revenue under 1 to filter out small-scale releases that do not fit the mold of the type of movie we want to measure.

Removed entries that were not rated PG-13 or R.

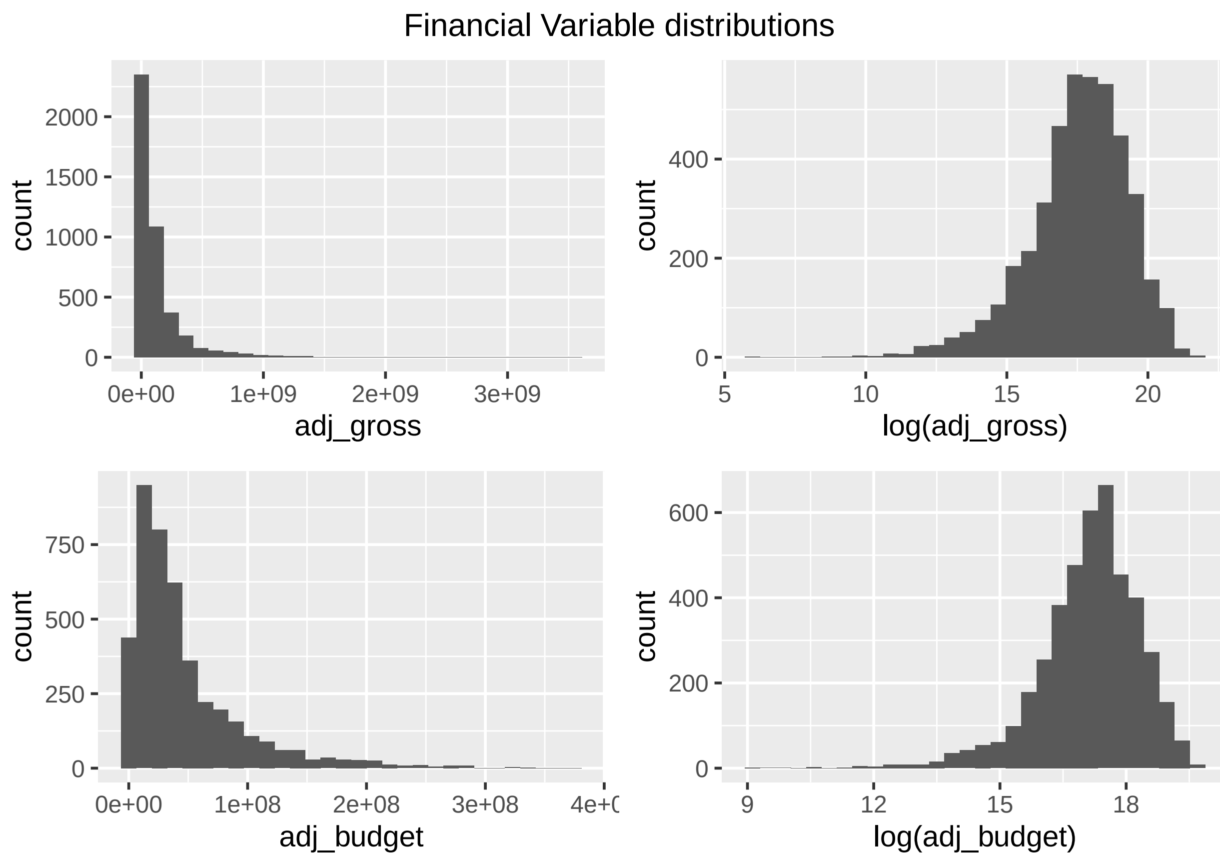
Removed duplicate entries as discussed in the previous section.

Created indicator variable for PG-13 rating.

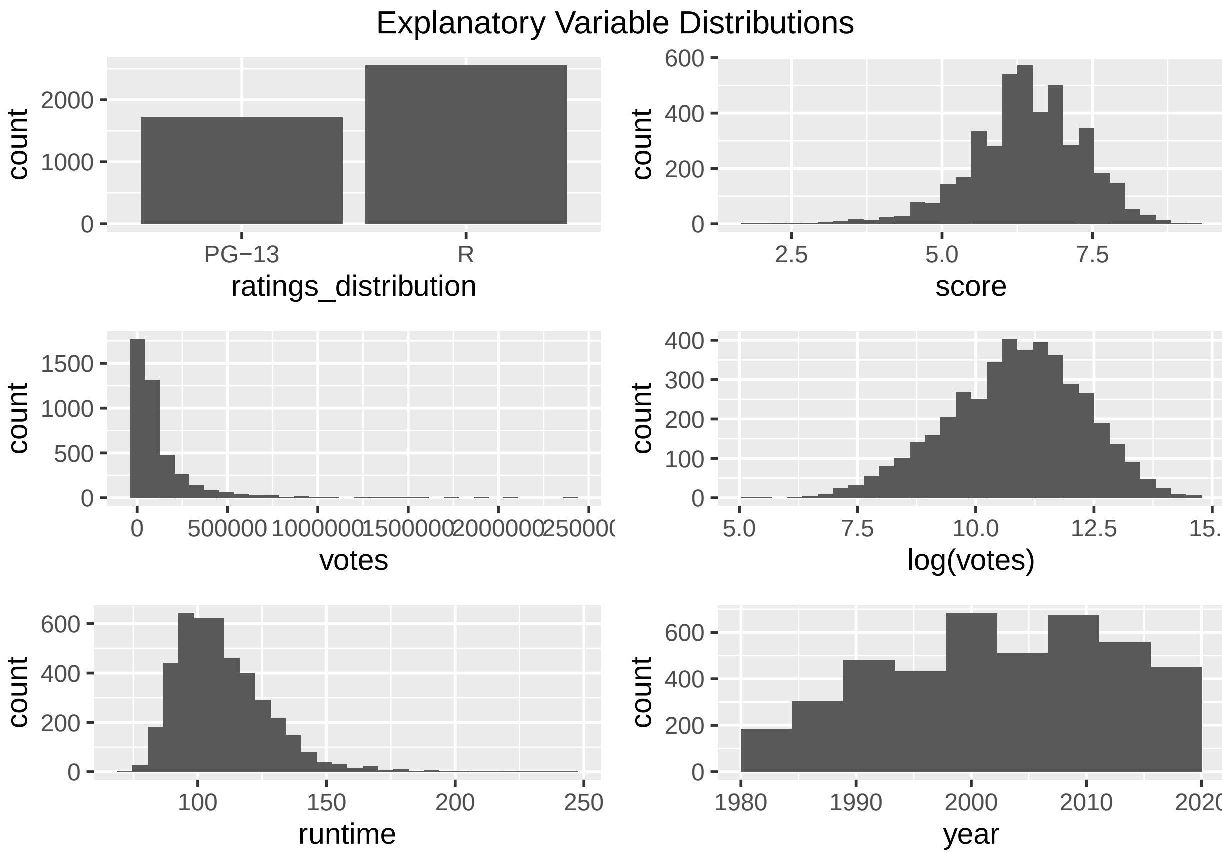
Created CPI-adjusted variables for gross revenue and budget.

2d. Exploratory Data Analysis - Fidelia

There are 4269 unique titles considered for this study.

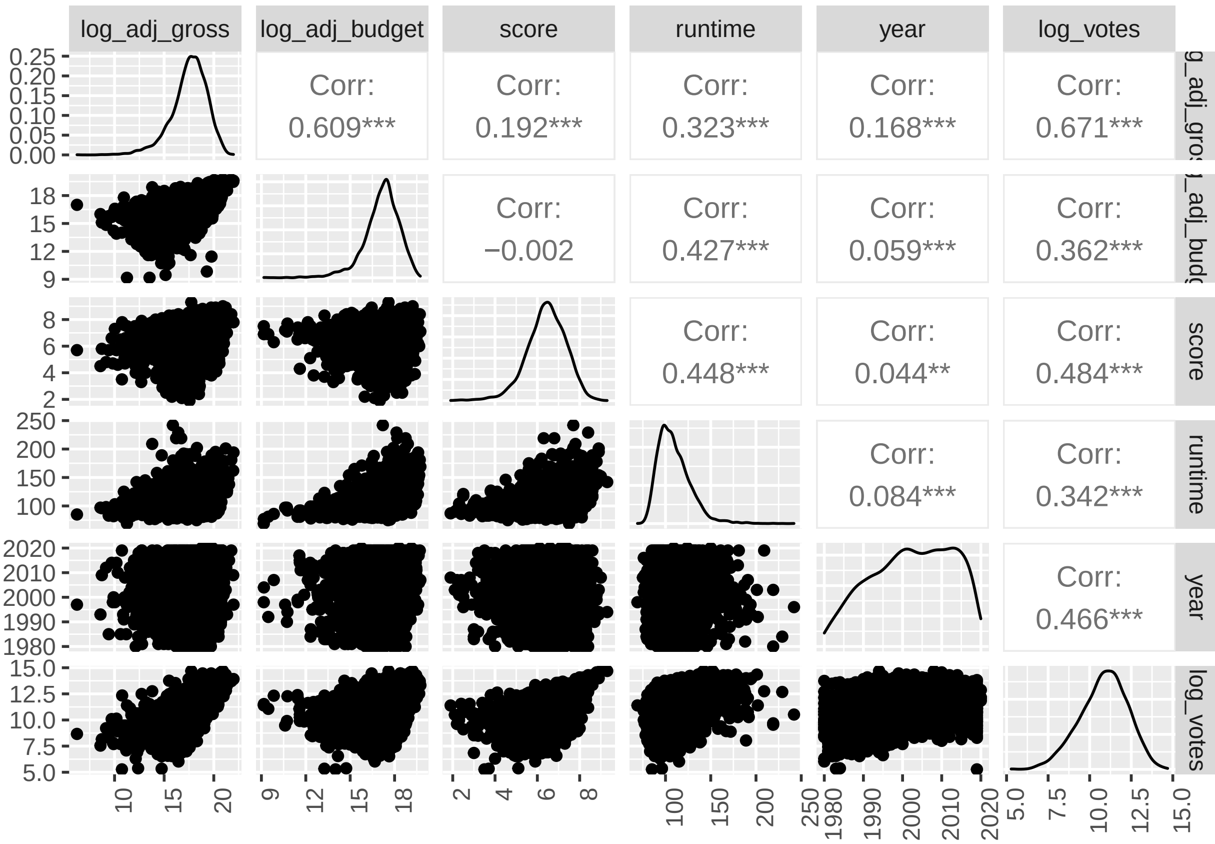


Analysis on each of the budget and gross variables.



Analysis on each of the six graphs.

Based on our EDA, log transformations of both the gross revenue (`gross`) and `budget` variables seemed helpful to produce a better-fitting linear model. This is because of the skewed nature of the data where both the financial variables have most data points toward zero with some extreme outliers (movies that generated a very large revenue). Other numeric variables did not seem to require any transformations because the distribution seemed symmetrical enough.



According to the correlation plots above, the explanatory variables do not seem to have any strong correlation to each other. The log-transformed data for `adj_budget` and `votes` may have decent correlation with the log-transformed data for `adj_gross` (the outcome variable), but this should not produce any problems for finding an BLP for this data.

3. Statistical Model - rearranged, move some parts to results

3a. A Model Building Process

Write up on progression models and justification

Proposed Models:

(1)

$$ln(Gross) = \beta_0 + \beta_1 * PG13$$

(2)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score$$

(3)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * PG13 * Score$$

(4)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * ln(Budget)$$

(5)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * PG13 * Score + \beta_4 * ln(Budget) + \beta_5 * PG13 * ln(Budget)$$

In order to answer the research question, according to our causal theory, the two main factors that affect a movie's revenue that will be measured are *Quality* and *MPP Rating*. Other variables would either contribute to, or interact with, the two main factors.

The first model (1) measures only the effect of changing the MPAA rating (PG-13 or R) on the gross revenue of the movie. β_0 is the estimated revenue that an R-rated movie is estimated to generate while β_1 is the estimated gain or in revenue if the movie is rated PG-13.

(1)

$$ln(Gross) = \beta_0 + \beta_1 * PG13$$

Based on model (1) movies rated R are expected to generate \$2.6853165 $\times 10^7$ in gross revenue (with no other factors).

Based on the same model, movies rated PG-13 are likely to generate 185.95% more gross revenue than a movie rated R.

| The second model (2) introduced the effect of a film's IMDb score on the gross revenue alongside the basic coefficients. β_2 indicates the estimated gain in revenue per increase in IMDb score. This model includes the variable for quality (`score`).

(2)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score$$

Based on model (2) movies rated R are expected to generate \$1.5771292 $\times 10^6$ in gross revenue (with no other factors).

Movies rated PG-13 are likely to generate 212.72% more gross revenue than a movie rated R.

Movies are expected to generate 54.65% more gross revenue per 1 point increase in score.

| The third model (3) introduced the possible effect of the interaction of a film's IMDb score and rating on the gross revenue alongside the basic coefficients. β_3 indicates the estimated extra gain in revenue per increase in IMDb score if the movie is rated PG-13. The effect of β_3 would capture any interaction between the two main quantities and its effect on the movie's gorss revenue.

(3)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * PG13 * Score$$

Based on model (3) movies rated R are expected to generate \$1.5368591 $\times 10^6$ in gross revenue (with no other factors).

Movies rated PG-13 are likely to generate 231.64% more gross revenue than a movie rated R.

Movies are expected to generate 55.27% more gross revenue per 1 point increase in score.

Movies rated R are expected to generate -0.92% more gross revenue per 1 point increase in score than movies rated PG-13.

| The fourth model (4) introduced the possible effect of a film's budget on the gross revenue alongside the basic coefficients for rating and score. β_3 indicates the estimated extra gain in revenue per increase in IMDb score if the movie is rated PG-13. Since it is a well-known fact that the budget is a big contributing factor in the quality and revenue of a movie, it was included to measure its effect on the gross revenue of a movie without any interactions.

(4)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * ln(Budget)$$

Based on model (4) movies rated R are expected to generate a in \$1.54 in gross revenue (with no other factors).

Movies rated PG-13 are likely to generate 63.96% more gross revenue than a movie rated R.

Movies are expected to generate 49.38% more gross revenue per 1 point increase in score.

Movies are expected to generate 8.3% more gross revenue if the budget is increased by 10%.

| The fifth model (5) introduced the possible effect of the interaction of a film's IMDb score and rating as well as the interaction between MPAA rating and budget on the gross revenue alongside the basic coefficients. β_3 indicates the estimated extra gain in revenue per increase in IMDb score if the movie is rated PG-13. β_4 indicates the estimated extra gain in revenue per increase in IMDb score if the movie is rated PG-13. β_5 indicates the estimated extra gain in revenue per increase in budget if the movie is rated PG-13. In addition to testing for any significant effect of the interaction of quality (`score`) and rating, model (5) also tests to check if the effect if the changing rating changes the effect of budget on the film's revenue as well.

(5)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * Score + \beta_3 * PG13 * Score + \beta_4 * ln(Budget) + \beta_5 * PG13 * ln(Budget)$$

Based on model (5) movies rated R are expected to generate \$3.7 in gross revenue (with no other factors).

With a score of 0 and a budget of 0, movies rated PG-13 are likely to generate 91.24% less gross revenue than a movie rated R.

Movies are expected to generate 52.02% more gross revenue per 1 point increase in score.

Movies are expected to generate 7.69% more gross revenue if the budget is increased by 10%.

Movies rated PG-13 are expected to generate -4.51% more gross revenue than movies rated R per 1 point increase in score.

Movies rated PG-13 are expected to generate 1.79% more gross revenue than movies rated R if the budget is increased by 10%.

4. A Results Section - Steve

You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table.

In your text, comment on both *statistical significance and practical significance*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

5. Limitations of your Model - Austin

5a. Statistical limitations of your model

IID Concerns: Movie data stretches over a 40 year period. We have to account for the impacts of inflation. We adjusted our models to account for inflation. Sequels and movies with common themes are not independent of each other. We did not take action to adjust for this. Clustering by genre. Clustering by year.

Unique BLP Concerns: No perfect co-linearity

Non-infinite variance: No concern

As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant

Table 1:

	Dependent variable:				
	log(adj_gross)				
	(1)	(2)	(3)	(4)	(5)
PG131	1.051***	1.140***	1.199***	0.494***	−2.435***
	(0.054)	(0.052)	(0.361)	(0.045)	(0.719)
score		0.436***	0.440***	0.401***	0.419***
		(0.028)	(0.037)	(0.023)	(0.030)
PG131:score			−0.009		−0.046
			(0.056)		(0.045)
PG131:log(adj_budget)					0.186***
					(0.039)
log(adj_budget)				0.836***	0.778***
				(0.018)	(0.022)
Constant	17.106***	14.271***	14.245***	0.435	1.309**
	(0.034)	(0.183)	(0.241)	(0.333)	(0.411)
Observations	4,269	4,269	4,269	4,269	4,269
R^2	0.083	0.133	0.133	0.425	0.428
Adjusted R^2	0.083	0.133	0.133	0.424	0.427
Note:			*p<0.05; **p<0.01; ***p<0.001		

problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the large sample model.

5b. Structural limitations of your model

Production company - positive bias away from zero for production companies. Certain companies won't make movies with explicit content (Pixar) Key actors and actresses - positive bias away from zero for popular actors Genre - certain genres will have more viewers.

Collect data on quality of production companies, actors/actresses, most popular genres etc.

What are the most important *omitted variables* that you were not able to measure and include in your analysis? For each variable you name, you should *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting. What data could you collect that would resolve any omitted variables bias?

6. Conclusion - Fidelia

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.