Keeping it PG-13

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

12/9/2021

1. Introduction

1a. Context

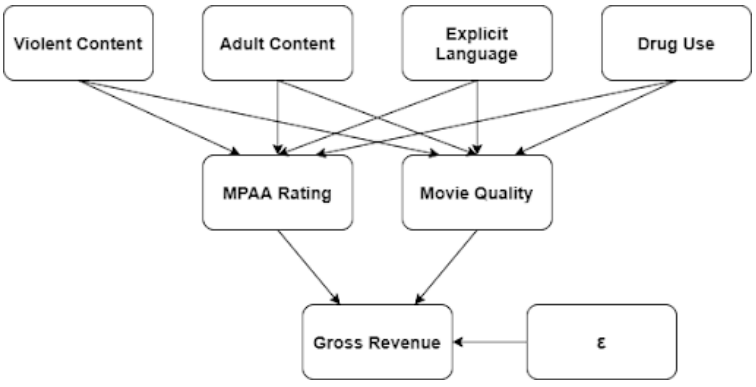
Acme Studios has spent a $50,000,000.00 budget on a superhero movie, and the director insists that the movie should include a scene where the main villain goes on an expletive-laden tirade. We know that including this scene will mean that the movie will be rated R, and cutting the scene will result in 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.

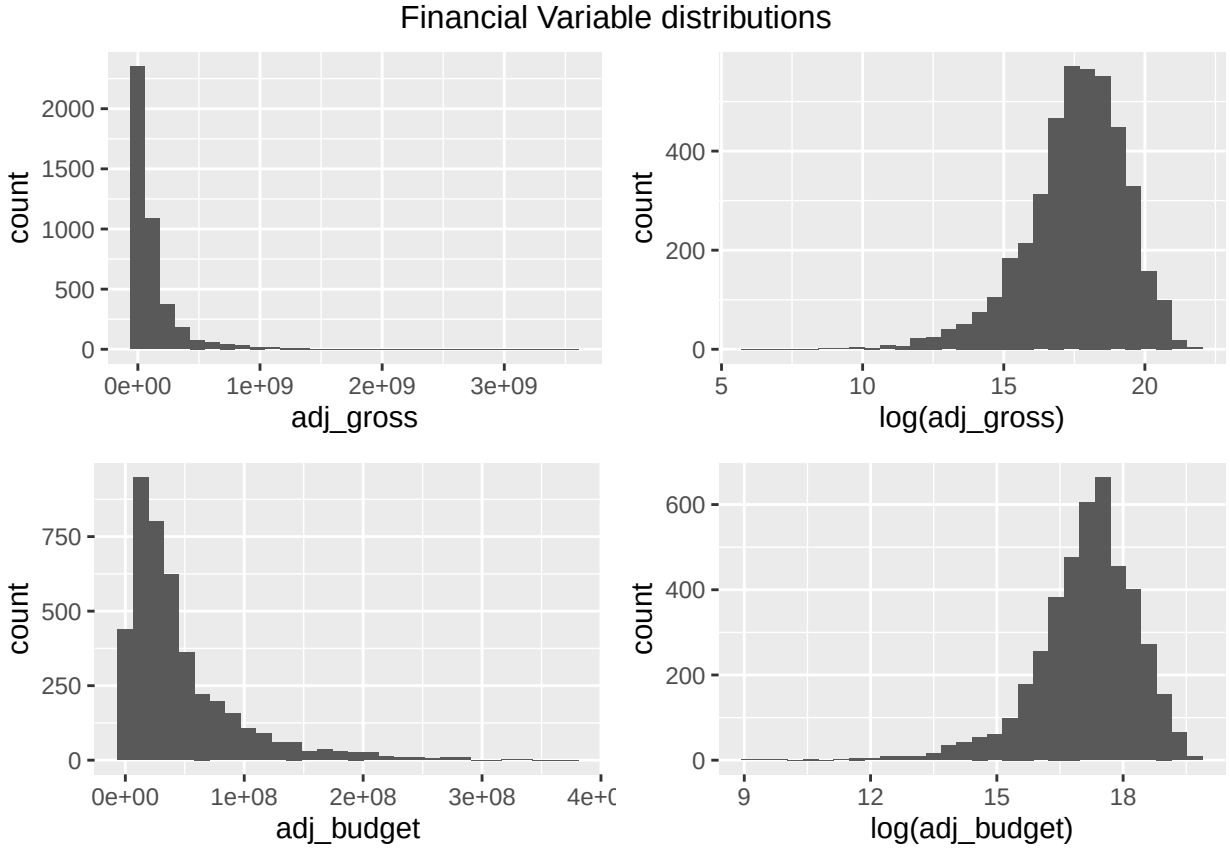
Removed duplicate entries as discussed in the previous section.

Created indicator variable for PG-13 rating.

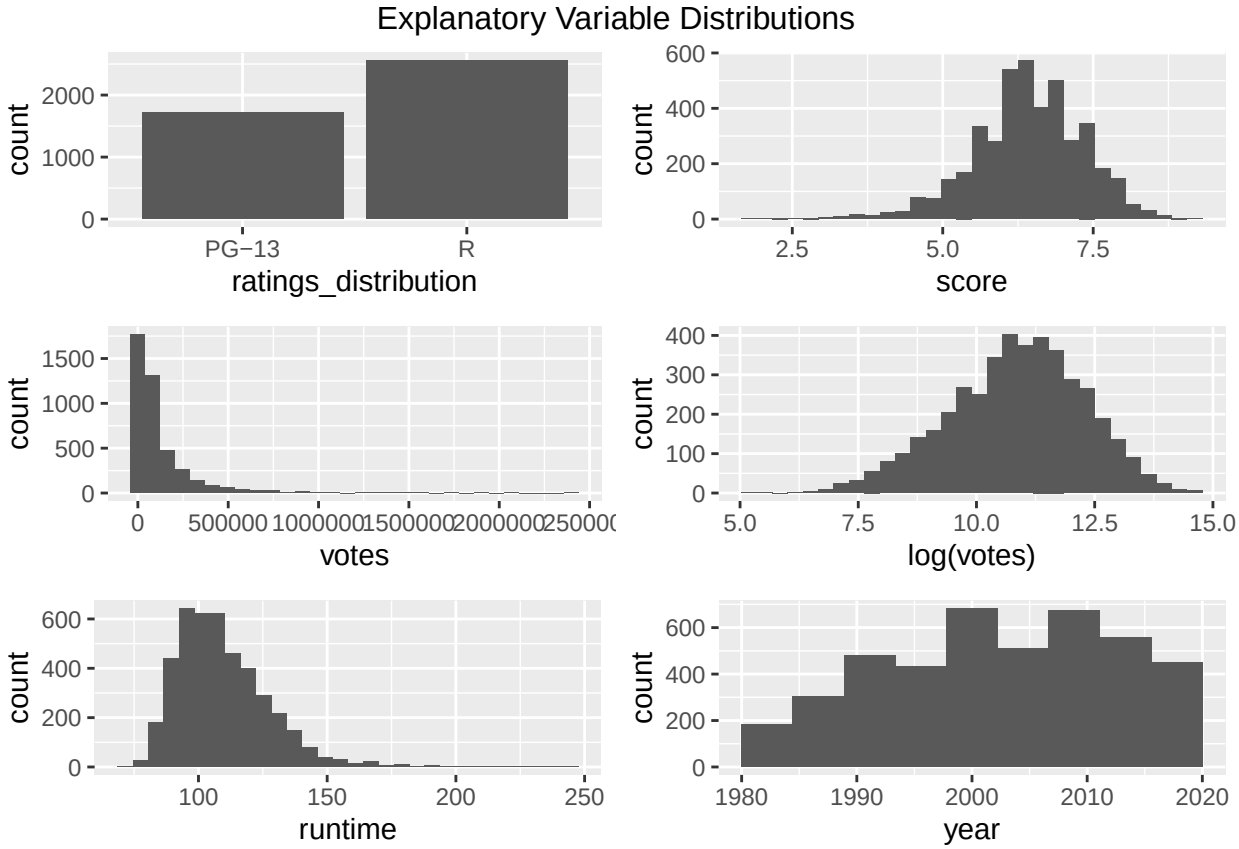
Created CPI-adjusted variables for gross revenue and budget.

2d. Exploratory Data Analysis - Fidelia

There are 4269 unique titles considered for this study.

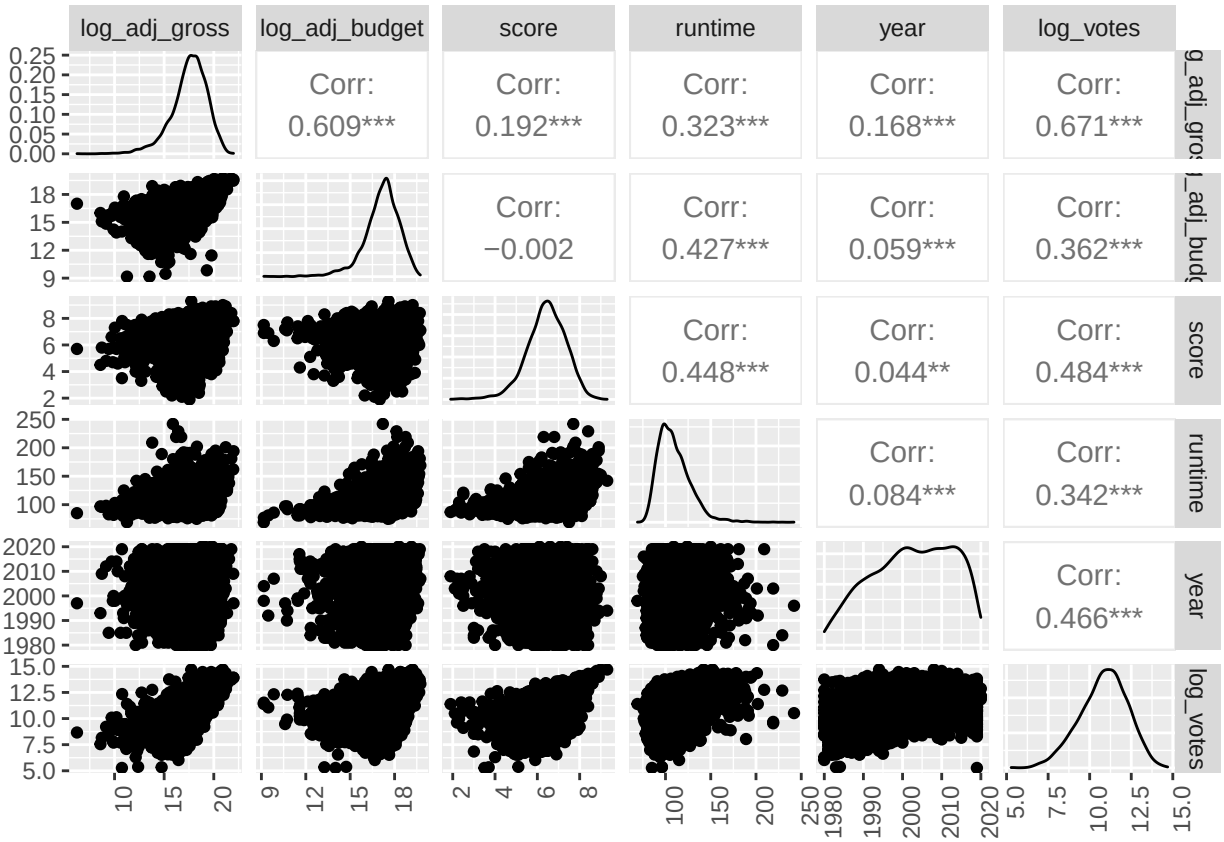


Analysis on each of the budget and gross variables.



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

Our primary input variable of interest is the *MPAA rating*, with an *R* rating as the base case and a *PG-13* rating as the alternative case. *MPAA rating* is a categoric variable with most of our dataset falling under the *R* rating and a sizable minority falling under the *PG-13* rating.

Our causal theory has identified *Movie Quality* as another key input variable to measure because there is a relationship between the factos that determine *MPAA rating* and *Movie Quality*, and both also influence *gross revenue*. Because we don't have an exact measurement of *Movie Quality* we are using the IMDb Rating (*score*) as a proxy to help mitigate omitted variable bias. The distribution of this variable seems roughly normally distributed so no transformations were applied.

Lastly, we are also prepared to consider a model that includes *budget* as a parameter because it has a strong correlation to *gross revenue*, and including it has the potential to make the predictions from our model much more accurate even if it is not directly named in our causal theory. Including *budget* has the potential to absorb some of the effect we are measuring, but there is also the potential of value in examining the interaction between *budget* and *MPAA rating*; we may find that a certain rating provides better return as the budget increases, or conversely that a certain rating does well with even very small budgets. This variable has a lopsided distribution with a heavy tail near zero and quickly becomes sparse moving to the right. Applying a log transformation makes this variable appear more evenly distributed.

We took an iterative approach to model building by starting with something very simple and adding terms as we went along until we reached something that satisfied our need for something that was both significant and sufficiently predictive.

Proposed Models:

(1)

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

We started off with the most simplistic possible model to see if a measurable difference between the *gross revenue* of an *R* or *PG-13* rated movie existed. This 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 *gross revenue* that an *R*-rated movie is estimated to generate while β_1 is the estimated gain in *gross revenue* if the movie is instead rated *PG-13*. This model is too simple to be our final choice, but if it had shown no significant difference in *gross revenue* it would have been a major red flag.

(2)

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

The second model (2) introduced the effect of a film's *IMDb score* on the *gross revenue* on top of the components of model (1). β_2 indicates the estimated percentage increase in *gross revenue* per one point increase in *IMDb score*. Including *score* was considered a bare minimum requirement for our final model, because we are using it as a proxy for *Movie Quality* from our causal theory.

(3)

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

The third model (3) introduced an interaction term between a film's *IMDb score* and *rating*, alongside the components of model (2). β_3 indicates the estimated percentage increase in revenue per one point increase in *IMDb score* if the movie is rated *PG-13*. Like in the previous model β_2 indicates the estimated percentage increase in *gross revenue* per one point increase in *IMDb score*, but having an interaction term allows *score* to impact predictions for *PG-13* and *R* differently. The magnitude of β_3 shows if the change in *gross revenue* per point of *score* is different for a *PG-13* movie than it is for an *R* movie.

(4)

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

The fourth model (4) was built on top of model (2), with the addition of a term for *budget*. Unlike model (3), an interaction term between *MPAA rating* and *score* was not included. The purpose of this model was to offer something with more predictive power than the previous models by including *budget*. We know that *gross revenue* and *budget* are strongly correlated and we expected to see this model perform better at predicting results (as measured by adjusted R-squared) than the previous models. The term β_3 in this model represents the relationship between percentage increases in *budget* and percentage increases in *gross revenue*; for every 1% increase in *budget* we expect a β_3% increase in *gross revenue*.

(5)

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

The fifth and final model (5) builds on all of the models before it. The terms β_0, β_1, β_2, and β_3 have identical meanings to model (3). β_4 now represents the relationship between percentage increases in *budget* and percentage increases in *gross revenue*. The new term β_5 is an interaction term between *MPAA rating* and *budget*. Having an interaction term allows *budget* to impact predictions for *PG-13* and *R* differently. The magnitude of β_5 shows if the percentage change in *gross revenue* per percentage increase in *budget* is different for a *PG-13* movie than it is for an *R* movie.

4. A Results Section

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

Of all models considered, our team determined that model (5) was the most effective. It produces the best-quality predictions of all models, as shown by the adjusted R-squared score. It also captures the all of the effects that any of the models flagged as being significant. The one term that is not significant in this model is the interaction term between PG-13 and score, with a standard error that overlaps zero. A closer look at each of the model coefficients and how they can be interpreted:

β_0 - As a baseline, movies rated *R* with a *budget* of 0 and an *IMDb score* of 0 are expected to generate $3.70 in gross revenue.

β_1 - The baseline for *PG-13* movies with a *budget* of 0 and an *IMDb score* of 0 is even lower, as they are expected to generate Movies rated PG-13 are likely to generate -91.24% more 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%.

5. Limitations of your Model

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 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.