Keeping it PG-13

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

12/9/2021

Contents

1 Introduction

1.1 Context

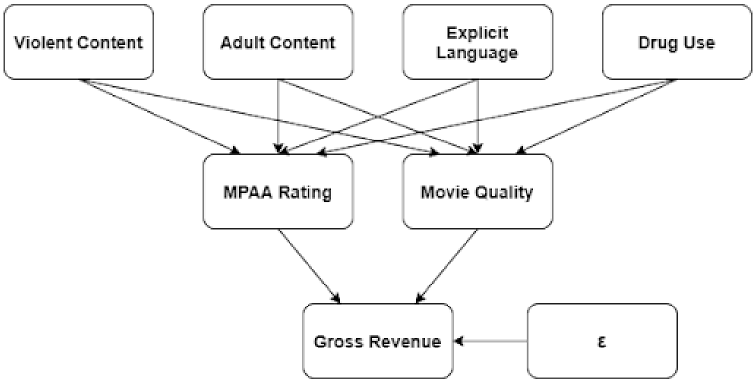
Acme Studios has spent a $50,000,000.00 budget on a superhero movie, and the director insists that the film should include a scene where the main villain goes on an expletive-laden tirade. We know that having this scene will mean that the movie will be rated R, and cutting the scene will result in the 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 making editorial changes after the director's cut. He's 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?

1.2 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 of *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*.

1.3 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 film's gross revenue. 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

2.1 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 IMDb 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 IMDb website

2.2 Research Design

Using the data and variables above, this study measured the impact *MPAA Rating* and *Movie Quality* on the *Gross Revenue* generated by a movie. Causal models were generated using the logic from the causal theory in section 1.3. `gross` is the primary 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 movie quality, but not the `rating` so they were considered as proxies for quality.

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

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 R movies were treated as the base case.

The `score` variable primarily measures movie quality. Variables like `budget` and `runtime` also affect quality, so they were considered as possible proxies for quality in the causal models.

In order to remove duplicated 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 50 years, we apply a CPI-based price adjustment to each monetary variable to account for inflation. The inflation-adjusted values (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 compare 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 can 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.

2.3 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.

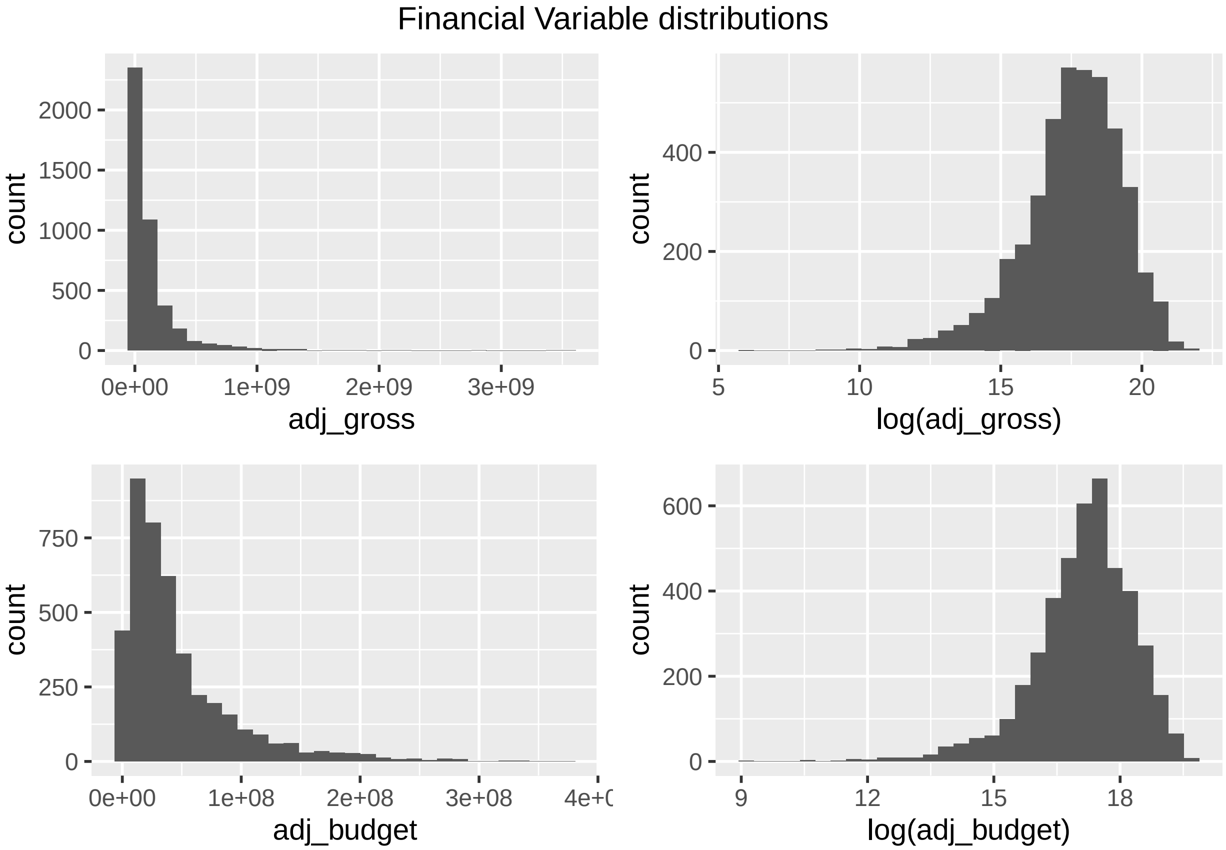
Removed duplicate entries as discussed in the previous section.

Created indicator variable for PG-13 rating.

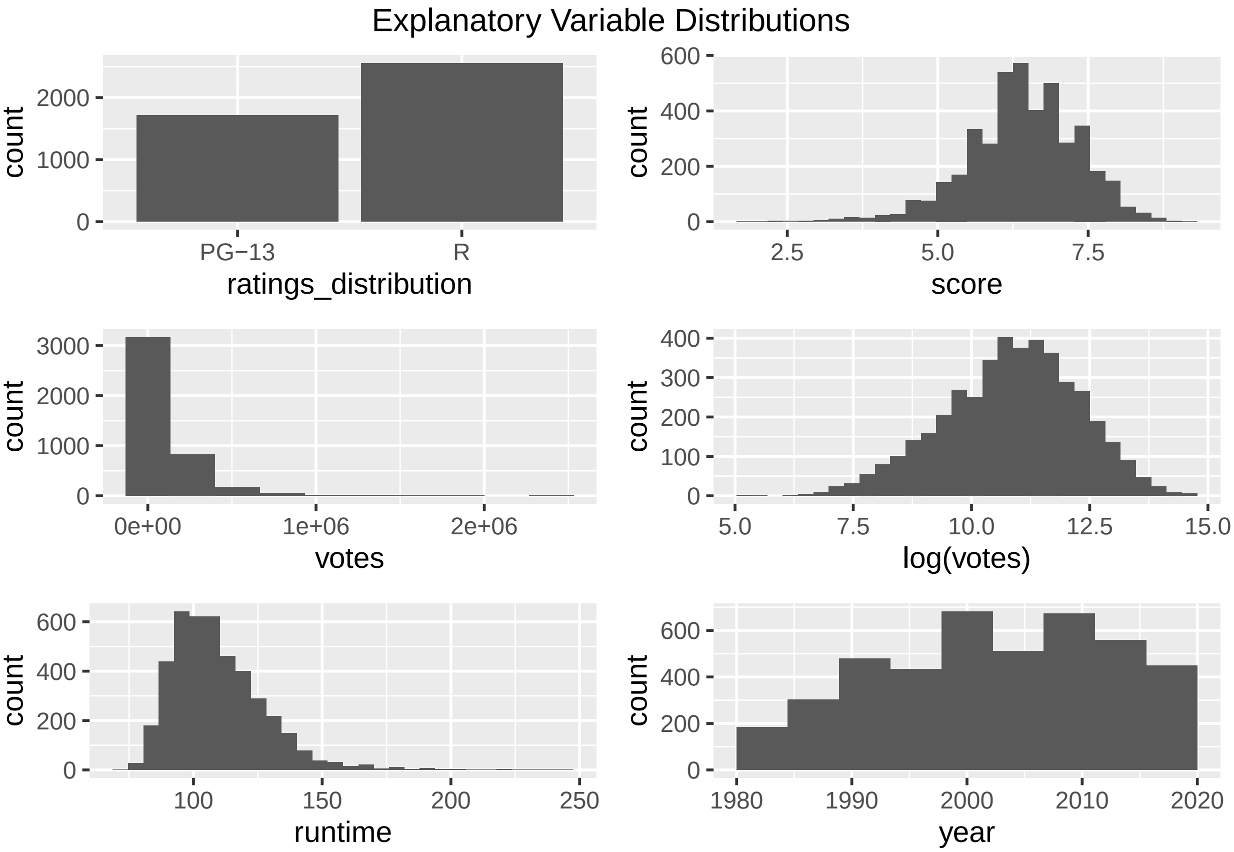
Created CPI-adjusted variables for gross revenue and budget.

2.4 Exploratory Data Analysis

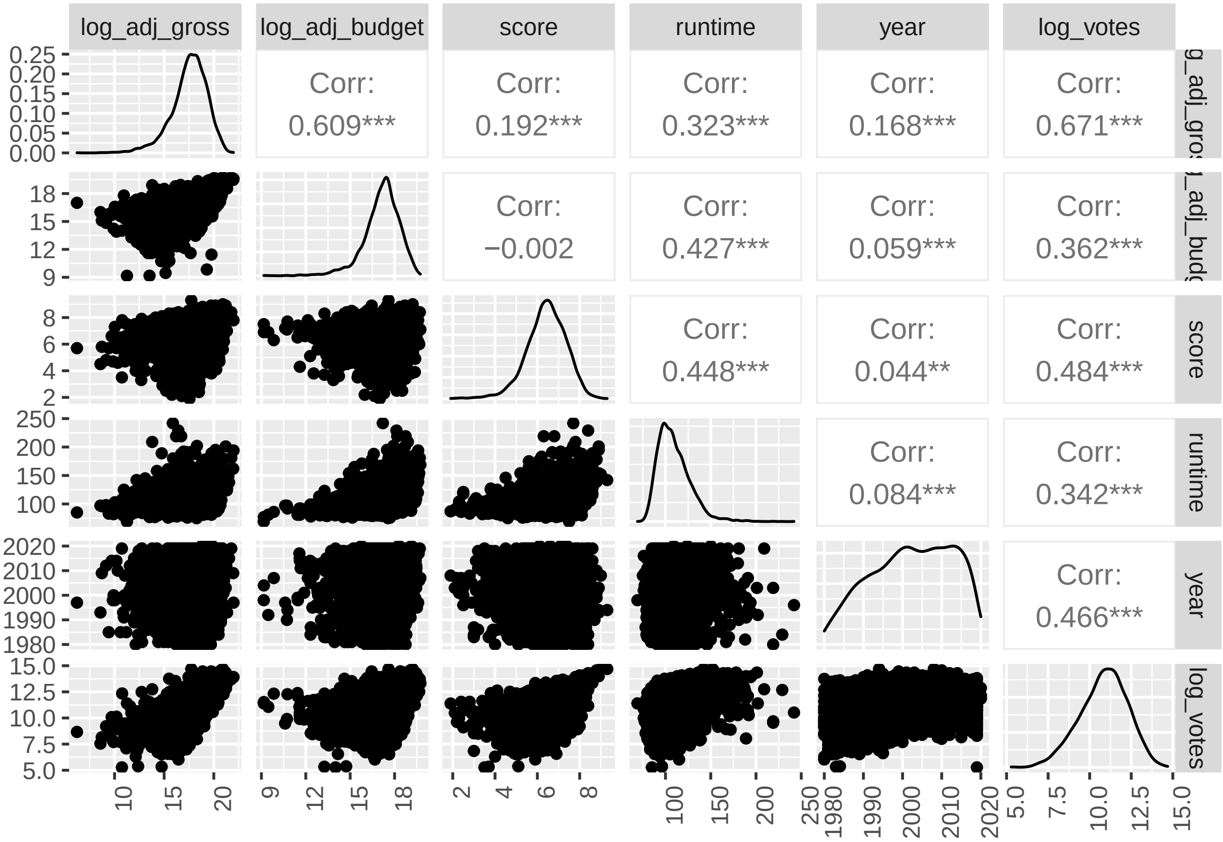
There are 4269 unique titles considered for this study.



Looking at the adj_budget vs. count and adj__gross vs. count graphs, we can see how there is a significant skew towards the left. 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).



Looking at the ratings_distribution graph after EDA, we can see how the dataset contains at least 1,000 more rated R movies than PG-13 movies. This could potentially affect our training/testing sets and results because we have more R data movies to work with rather than PG-13 but should still be appropriate to answer our research question. For the second graph of score vs. count, we see a relatively normal distribution that is slightly off-centered to the right with a peak around 6.25. This bell curve is expected because the ratings are scored from 0-10 and it is expected that the peak will be at the average rating (at around 5). The votes vs. count graph has a left skew with a heavy tail near 0. This is due to most of the movies in the dataset having no more than a couple thousand or tens of thousands of reviews. This plot is not as visually intuitive, so by taking the log(votes), we can see a much cleaner normal distrbution with an off-center to the right once again. The runtime vs. count graph also has a normal distribution but centered to the left at about 100 minutes, meaning most of the movies in the dataset don't run longer than an hour and a half to two hours. For the final graph, we see a relatively even distribution of the number of movies released each year, except during 1980-1990 where fewer movies were released in the dataset than in 2000 or 2010. Overall, the visualizations of the histograms from the data seem to follow our expectations. The other numeric variables (except for gross revenue (`gross`) and `budget`) 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

3.1 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 categorical variable with most of the 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 factors that determine *MPAA rating* and *Movie Quality*, which also influence *gross revenue*. Because we don't have an exact measurement of *Movie Quality*, we use 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.

3.2 Proposed Models:

(1)

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

We started with the most simplistic possible model to see if a measurable difference existed between the *gross revenue* of an *R* or *PG-13* rated movie. 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 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 Results

4.1 Model Comparison

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. Model (2) adhered most strictly to the causal theory we hoped to advance, but the vastly superior performance of model (5) could not be ignored. We performed an F-test on these models and it reinforced our decision to reject the null (simple) model (2) in favor of the fuller model (5).

```
## Analysis of Variance Table
##
```

```
## Model 1: log(adj_gross) ~ PG13 + score
## Model 2: log(adj_gross) ~ PG13 + score + PG13 * score + log(adj_budget) +
##     log(adj_budget) * PG13
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   4266 11855.5
## 2   4263  7825.9  3    4029.6 731.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2 Model Interpretation

After running our linear regressions and calculating the beta coefficients, our final model (5) is set:

$$ln(Gross) = 1.309 - 2.435 * PG13 + 0.419 * Score - 0.046 * PG13 * Score$$

$$+0.778 * ln(Budget) + 0.186 * PG13 * ln(Budget)$$

A closer look at each of the model coefficients and how they can be interpreted:

$\beta_0 = 1.309$: As a baseline, movies rated *R* with a *budget* of 0 and an *IMDb score* of 0 are expected to generate #3.703 in gross revenue.

$\beta_1 = -2.435$: The baseline for *PG-13* movies with a *budget* of 0 and an *IMDb score* of 0 is lower than movies with an *R* rating. This puts the baseline prediction for *gross revenue* for a PG-13 movie at 0.324 Although both β_0 and $beta_1$ are statistically significant, they are of little practical significance. We are not interested in movies with a *budget* of 0 or an *IMDb score* of 0. The other coefficients tell a more important story.

$\beta_2 = 0.419$: For every one point increase in *IMDb Score* our *gross revenue* prediction increases by 41.9%. This is both statistically and practically significant. It supports our prior assumption that *movie quality* as measured by *IMDb score* is an important factor in the success of a movie as measured by *gross revenue*.

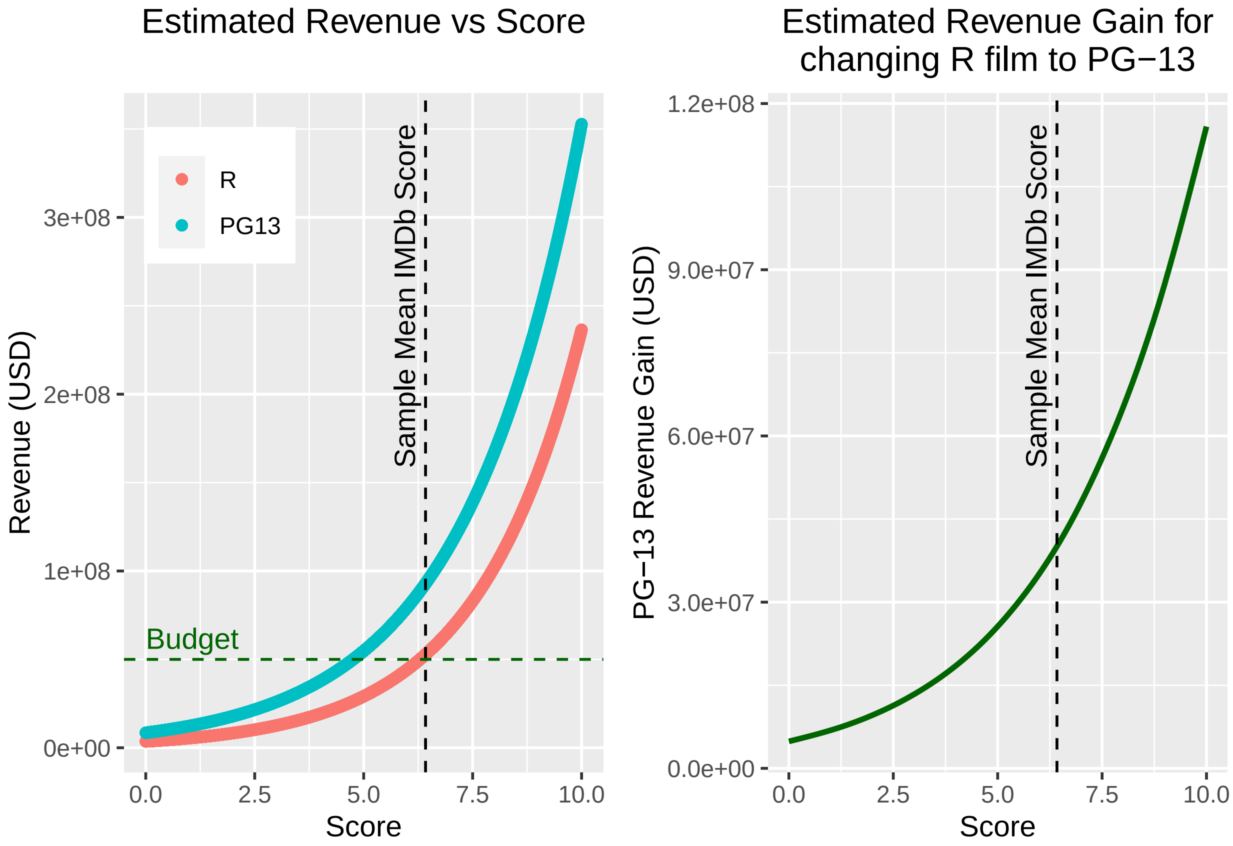
$\beta_3 = -0.046$: This is the first interaction term in our model, and it serves to illustrate the difference in impact that increasing *IMDb score* has on *PG-13* movies versus the *R* movie base case. For a PG-13 movie we combine the effects of β_2 and β_3 so for each one point increase in *IMDb* rating we expect to see a 37.3% increase in *gross revenue*. This is the only coeeficient in our model that is not statistically significant, so it is not particularly valuable in drawing practical conclusions. It suggests that *PG-13* movies benefit less per point increase in *IMDb score*, but the relationship is small and the standard error overlaps zero.

$beta_4 = 0.778$: Because *budget* proved to be a strong predictor of *gross revenue*, this term is especially important. Due to both variables having undergone a log transformation, the interpretation is less straightforward. For each 10% increase in *budget* we expect to see a 0.778% increase in *gross revenue*. This term is both statistically and practically signficant. It shows that increasing the *budget* is not always profitable in-and-of-itself. The excess spending may only be worthwhile if it also increases the *movie quality* in a meaningful way. This is a question that could warrant future research.

$beta_5 = 0.186$: The final coefficient to evaluate is another interaction term. It serves to illustrate the difference in impact that increasing *budget* has on *PG-13* movies versus the *R* movie base case. This is another case where log transformations have been applied to both *gross revenue* and *budget*, making the coefficient less straightforward to interpret. For each 10% increase in *budget* we expect to see a 0.186% increase in *gross revenue*. This term is both statistically and practically signficant. It suggests that increasing the *budget* of a *PG-13* movie has a greater positive effect on *gross revenue* than it does for an *R* movie. This helps support our hypothesis that the movie will perform better if we change from an *R* rating to a *PG-13* rating.

4.3 Model Predictions

Given the constraints of the movie as input parameters, model (5) was used to estimate the possible revenue of the Acme Studios superhero movie. The movie's budget is set at $50,000,000.00 and the MPAA rating can be either R or PG-13. Assuming that the IMDb score (or perceived quality) may vary, the estimated gross revenue predicted by the model also varies. The figure below shows the estimated gross revenue for different IMDb scores:



According to the model, a PG-13 movie with this budget is expected to generate more revenue than an R movie. As the IMDb score (perceived quality) increases, the estimated gain in revenue increases at an exponential rate. If this model properly predicts the revenue, then it would suggest that there is much more to gain by allowing the movie to be rated PG-13. The horizontal dotted line on the left plot represents our movie's financial break-even point and shows that a PG-13 movie is expected to break-even at a lower IMDb score. Using the sample mean (6.419) as an estimate for IMDb score (represented by the vertical dotted line on both plots), and holding all other variables constant, our model estimates the gross revenue for the movie to be $52,781,381 if rated R versus $92,884,941 if rated PG-13 (an improvement of $40,103,560). The predicted revenue for an R rating is dangerously close to our break-even point, whereas the gross revenue prediction for a PG-13 rating suggests a sizable profit margin.

5 Limitations of your Model

5.1 Statistical limitations of your model

Inflation

Initially, we were concerned that our data was not identically distributed because the movies range over 40 years. The cumulative price increase from 1980 to 2020 was 214%. This change undoubtedly affects our outcome variable, gross revenue. We used the consumer price index (CPI) to calculate an inflation adjustment for gross revenue and the budget. Accounting for inflation allows us to level the playing field for movies produced across the time span.

Common Themes

Sequels and movies with common themes are not independent of each other. The success of one Star Wars movie directly impacts the success of another one. Specific movie themes and genres can also generate more attention at times. We did not take action to adjust for this. In future modeling, we can look at clustering by genre, year, and movie theme.

Best Linear Predictor

We ensured that there were no concerns with co-linearity and non-finite variance within our data modeling. Do we need to elaborate on this? According to guidelines, we only need to highlight assumptions that pose problems for analysis.

5.2 Structural limitations of your model

While our model accounts for many important variables of interest, other factors could impact our final model. We will discuss how omitted variables might affect each of our explanatory variables and what steps could resolve any omitted variable bias.

Production Company

Different movie production companies have their niche within the Hollywood community. Certain companies will draw a strict line on violent, sexual, and other explicit content in their films. Those factors play a large role in MPAA rating. For example, Pixar Animation Studios, a subsidiary of Walt Disney Studios, consistently produces high-quality movies with a G or PG rating. While initially aimed for an adolescent population, the movies have gained success among all age groups and generated revenue reaching 1 billion dollars. Well-known production companies with specific standards for content will likely still make movies with high gross revenue. Additionally, movie critics might be partial to certain companies. Therefore, we expect the production company to bias the MPAA rating and quality score away from zero. We would need to find quality data on production companies to resolve this issue.

Popular Actors

There are a lot of similarities between the argument for the production company and famous actors. Actors can have the same standards for content. Additionally, actors gain cult-like followings among the public. People will insist on watching all Matthew McConaughey or Julia Roberts movies regardless of the quality or content. Both known factors will influence the bias for MPAA rating and quality score away from zero. We would need to investigate actor popularity to resolve this bias.

Movie Genre

The genre of the movie often affects the MPAA rating. An action movie will have more violence, and a comedy will likely have more explicit language. The genre will impact movie quality and viewership because people gravitate to certain genres. The bias impact is up for debate and not necessarily clear. While certain genres will drive up the MPAA rating, others will drive it down. We need more information on the popularity of genres to see their impact on movie scores.

6 Conclusion

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

For this research study, we expected to show that both MPAA rating and movie quality impact the film's gross revenue. In particular, we wanted to focus on how much more revenue a movie studio expects to make on a film that gets a PG-13 rating instead of an R rating, holding other factors constant. We did this by creating a model that does the following: measures the effect of changing the MPAA rating on the gross revenue of the movie, the effect of a film's IMDb score on gross revenue, the effect of the interaction of score and rating on gross revenue, and the effect of on gross revenue. The combination of all these effects proved to be the most effective based on the adjusted R-squared score and because of its best-quality predictions.

According to this model (5), we can see how a PG-13 film with a budget of $0 and a score of 0 is expected to make 91.24% *less* revenue than a rated R-movie. However, when looking at the coefficient of the interaction term for PG13 * Budget, we can see how β_5 is positive and statistically significant, which suggests that as the budget increases, the PG-13 movies will generate more revenue. The negative coefficient might have mitigated this for PG13 * Score, where, as the score increases, PG-13 movies are expected to yield 4.51% less revenue. But it seems statistically insignificant because the p-value is not within the cut-off for *and* its 95% confidence interval (+/- 2 standard deviations) contains zero.

Based on the findings of the results table above, we can see how our primary variable of interest, rating of PG-13, is expected to generate less revenue than movies rated R for every 1 point increase in the score (which is our proxy for movie quality). It's also interesting to note that even if there is an increase in the rating of a PG13 movie, it is still expected to generate less revenue than a rated-R movie with a similar rise in rating. In general, we see how an increase in a film's score drastically increases how much gross revenue a movie generates (52.02%). There is also a positive relationship between increasing budget to generate more revenue but only by 7.69%. Therefore, we can conclude that improving the quality of the movie (score) increases the gross revenue generated, as expected. Compared to our original prediction, though, PG-13 movies are not gaining more revenue in every case (only when the budget is high).

Based on these results, our model seems to favor R movies with a lower budget but favors PG-13 movies with a higher budget with an R-Squared value of 0.428.

As a result, our study seems fairly conclusive about the causal relationship between rating and revenue, but only depending on the budget (meaning without holding other variables, aka budget, constant as we stated in our problem statement). We do not fail to detect a significant causal effect between these variables because, depending on the budget, we see a trend in the increase in revenue for the respective rating.

Given the limitations in our data, including the fact that it is not entirely IID and the subjective interpretation of what makes a movie "high" quality based on a good score, it is interesting to see our estimation models detect an effect for our variables of interest (gross revenue). Our theoretical causal model accurately captured the existing causal paths using OLS regression under classical linear model assumptions to find an effect. For future refinement, perhaps we could work with a larger sample size (with movies before 1980) and compare their models and R-squared values to see if there is a similar effect.

If we stated earlier that we're using runtime as a proxy for movie quality, is there a reason it's not included in any of the models? We do not fail to detect any significant causal effect between these variables