Does a film's rating depend of the choice of genre?

Karola Takács

02/01/2021

1. Executive Summary

2. Introduction

The primary interest of this analysis was whether there is a causal connection between the imdb rating of a film and it's genre. I was simply interested if certain genres are more 'likeable' by viewers or more successful from a director's point of view. My original (broad) research question was: Is there a favorable genre to achieve higher movie rating? And also What other factors might influence the rating? However after the data cleaning part, I managed to narrow down the research question to: Does the genre choice of a director yield to certain imdb ratings for movies older than 1920? (more precisely: for movies with more than 25000 reviews). With this information at a director's hands it would be a direction what are the seemingly "more successful" genres.

3. Data

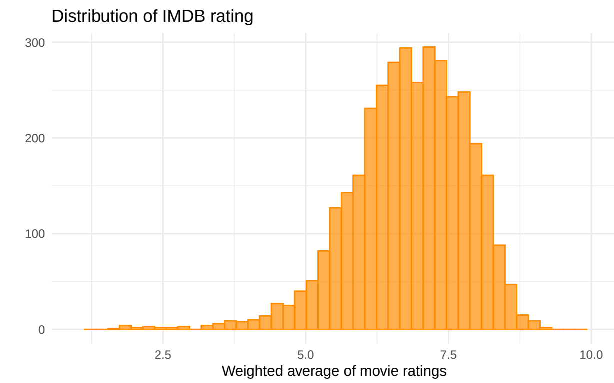
What is the used data, data quality issues, descriptive statistics and graphs and variable transformation

The main variables of the joined dataset to be analyzed are the movies' title, year of making, their genre along with ratings from two different sources: imdb and Metacritic (Metacritic aggregates movie reviews from the leading critics) and also number of reviews by imdb users and critics and finally the worldwide gross income in millions. The left-hand sided variable is rating and the right-hand side variable is genre.

To begin with I had two tables about movies both downloaded from the Kaggle webpage. They differed in record count, but after joining them based on the imdb movie ID field, I had 43605 observations. The year variable ranges from 1874 to 1995.
During the cleaning process I excluded all the observations with missing values for the most important variables, imdb rating and genre, but these accounted only for 2031 observations. Let me explain these two variables in more detail.

- IMDB rating: all registered members of IMDB can cast their votes (good quality). For one movie a user can vote several times but then the last value is going to be updated, so basically there is 1 user for 1 film at a single time.
- Genre: this is a problematic variable since it not only is categorical but for a single movie there can be more genres attached. For simplification my assumption was: the firstly listed genre is the most relevant or primary genre. Even if this is true, it really makes sense to have more categories, because there are few main categories, like action or drama and it is easy to put most of the movies into one of these which results in that these categories will be overrepresented without other genres to balance out.

Descriptive statistics show that IMDB rating values are close to normal distribution, bit skewed to the left, signaling that he mean is well above 5, it is around 7. Similar with metacritic score, which has a better looking bell-shape, the distribution of values are wider, values are more spread on the scale of 100. IMDB Votes would make sense to transform for the analysis part as it shows right skewness, but since it is already incorporated into the IMDB rating variable it would be a bad conditioning variable in the visual analysis, so I decided not to use it.



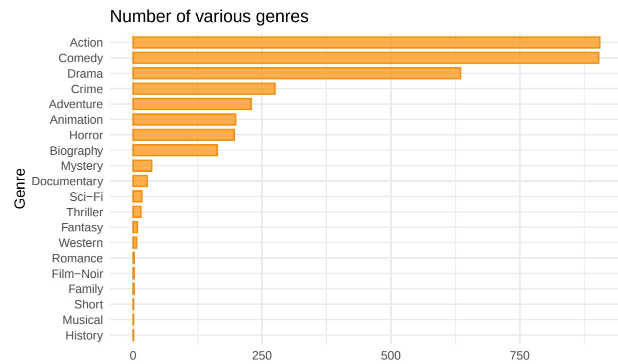


Table 1: Summary stat of IMDB ratings

Variable	mean	median	std	min	max	skew	numObs
IMDB ratings	6.8	6.9	1	1.6	9.3	-0.82	3624

The distributions (see in Appendix - Histograms) of votes, gross income and user review suggest possible log transformations. After the transformations we can see an approximately normal distribution for gross income and number of reviews, however number of votes is still not close to normal distribution. In this case it is not a concern since I argued earlier why I am not intended to use this variable.

There are some **extreme values** for user_review, namely 8232, 6938 and 5392. But these are for films in the top 15, so it is possible, they might have a huge fan base and users left that many reviews. There is no need to drop these values. Gross income is strongly left skewed and there will be a need for log transformation in the modelling part.

The number of records drop significantly after I decided to filter based on the number of votes. The baseline was 25 thousand number of votes and my rationale behind it was the fact that when creating the top 250 list, IMDB considers only movies with minimum this vote number. I think this makes the rating more credible so I decided to apply the same. In the very end I had 3624 observations to work with.

When thinking about **representativeness** I would say that IMDB ratings are only reliable for the IMDB users who rated the movie, but not necessarily reliable indicators of what general movie audiences thought; it depends a lot on personal taste and preferences. Old movies are not that much represented since I would argue lot less registered users have seen those (even less if we consider the whole population). I also thought of the geographical coverage: only those can vote who have internet access (can register on the IMDB page) and also language might be an issue. We also cannot be 100% sure that all voters have seen the film and their vote is the true assessment of their liking. All in all this dataset captures information for registered users' preferences and the findings would indicate the favorite genre of these users only.

4. Model

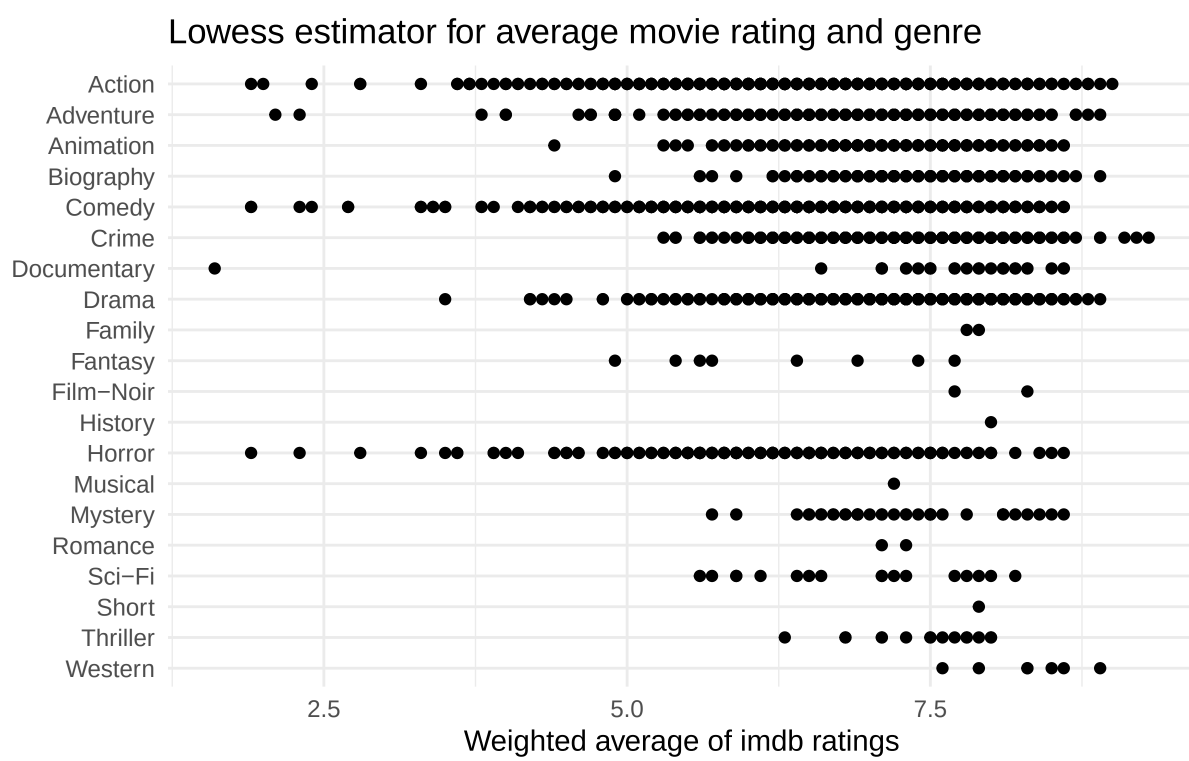
4.1 Setup

As mentioned earlier, the outcome variable is rating, the weighted (on number of votes) average user rating on IMDB and the parameter of interest is genre, for which I created a new column - their number of occurrence in the dataset - for possible weighting. Before checking for potential confounder in the existing dataset, I noted several other factors which possibly can influence my outcome variable:

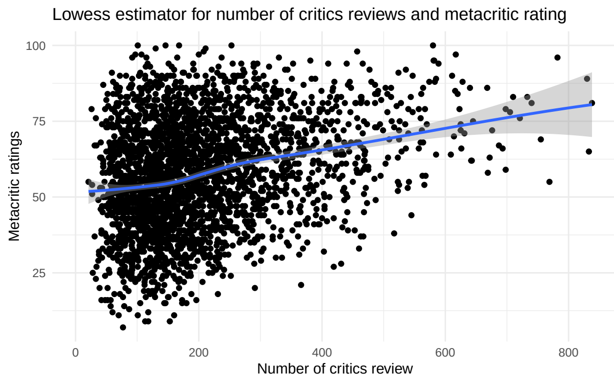
- availability of dubbing
- screening in how many countries (how worldwide the screening was) - this might relates to worldwide gross income
- lot harder to acquire and watch old films since they are not digitized or simply rare
- motion picture content rating system (whether a movie is rated only above 18 years of audience)
- leading actor/actress in the movie is famous/well known

Since I want to regress genres on ratings as the base idea, after checking these with lowess smoother (and after transforming genre into a factor), I could detect a few things:

- the most common genres are Action, Comedy, Adventure, Drama and Horror.
- Their ratings move on a wide range, while for instance Western has an average rating only above 7.5, similarly with Family or Film-Noir.



I wanted to check for a potential confounder effect between metacritic score and number of critics review since metacritic score is also based on critics reviews and scores, but there seems no pattern between the two.



In this phase I also checked correlations among my X variables.The highest correlation (where correlation is above 0.7) is between user_review and votes. This is not surprising since I filtered for observations only with at least 25 thousand votes and these movies tend to have more user reviews. ucratio is also obviously highly correlated with user_review since it is a calculated variable based on user - and critics reviews. For this reason it is enough to use either the ratio or the two other variables but not both. The moderately correlated pairs are number of votes - gross income and votes - ucratio (I would guess since it incorporates user rating which is highly correlated with ucratio).

```
## [1] 3
```

Table 2: Mid - Highly correlated variable pairs

Var1	Var2	corr_val
rating	metacritic	0.7443394
metacritic	rating	0.7443394
worldwide_gross_income	votes	0.5607189
user_review	votes	0.7984952
votes	worldwide_gross_income	0.5607189
votes	user_review	0.7984952

For an example on interaction analysis see Appendix.

4.2 Modelling

The simple linear regression (1) was the following: Rating E =

$$\beta_0$$

+

$$\beta_1$$

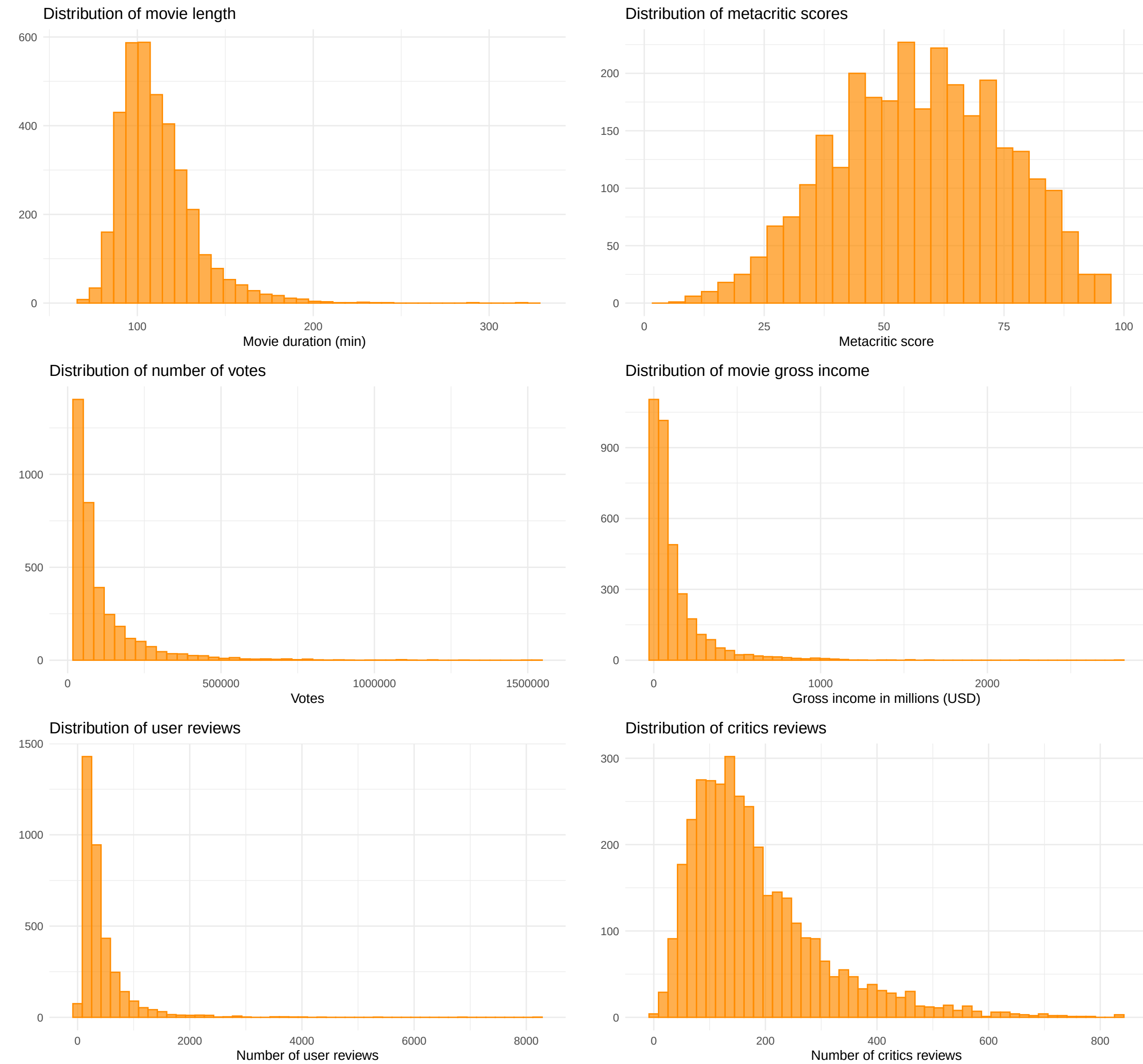
* genre. At first glance on the model output, there are two catgeories - Horror and Fantasy - which have negative coefficients meaning these genres affect the ratings adversely. I don't intend to go into more details with this model's interpretation, since its R square is really low: 0.10, similarly with models (2) and (3).They

5. Generalization, external validity and causality

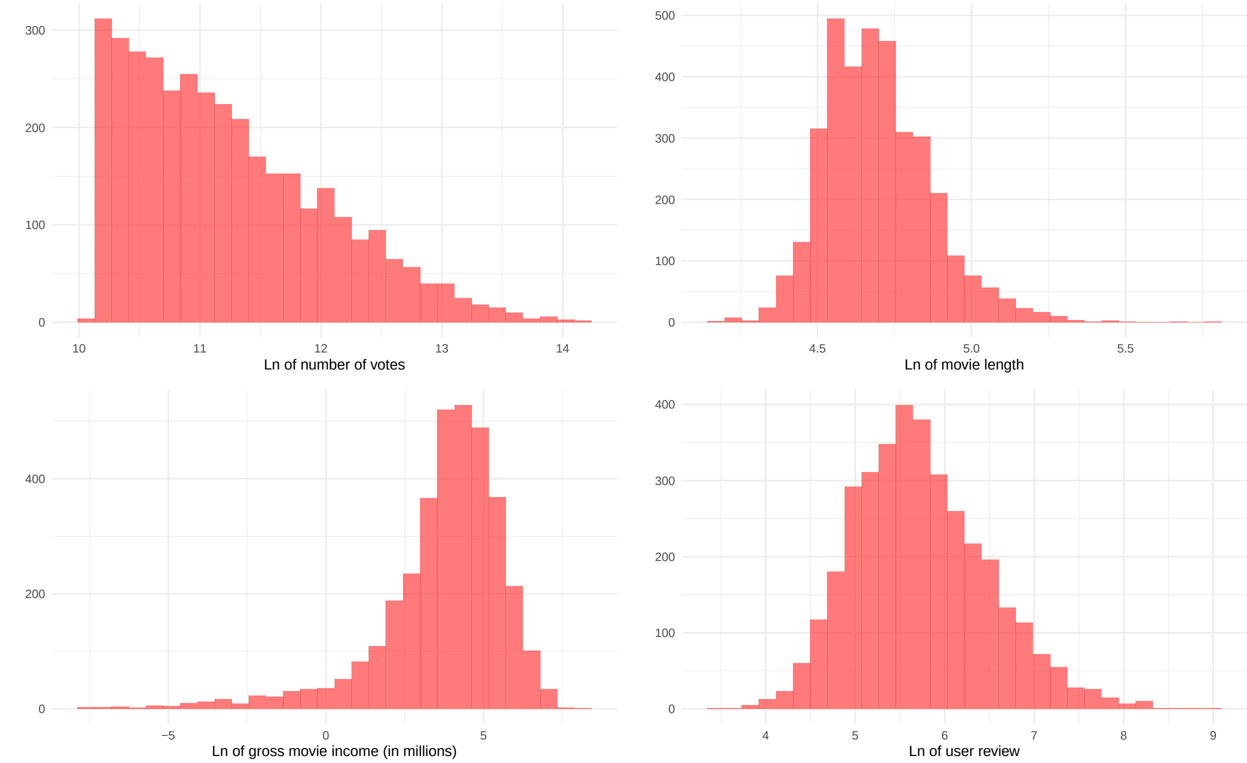
6. Summary

Appendices

Histograms

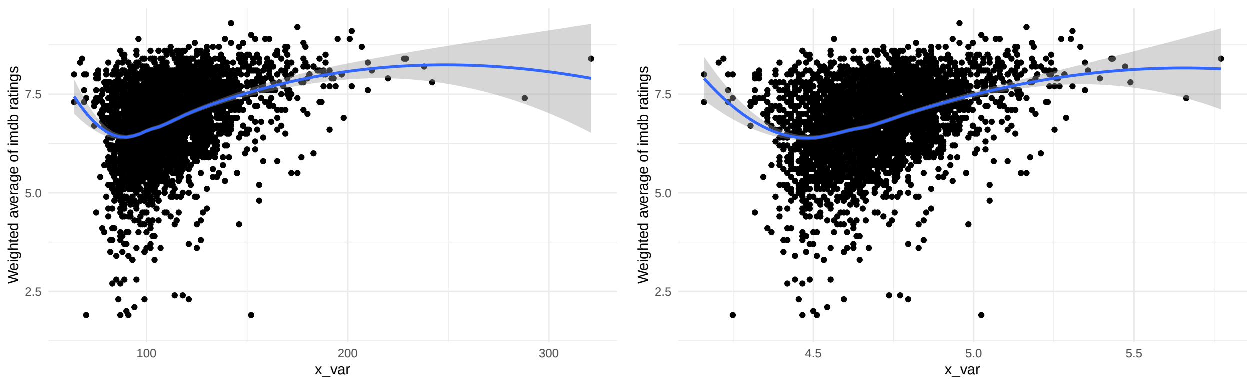


Possible log transformations

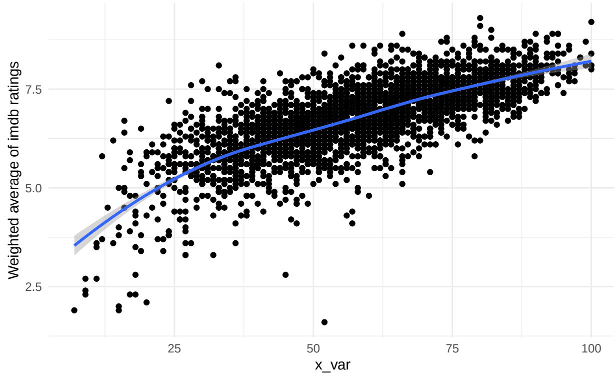


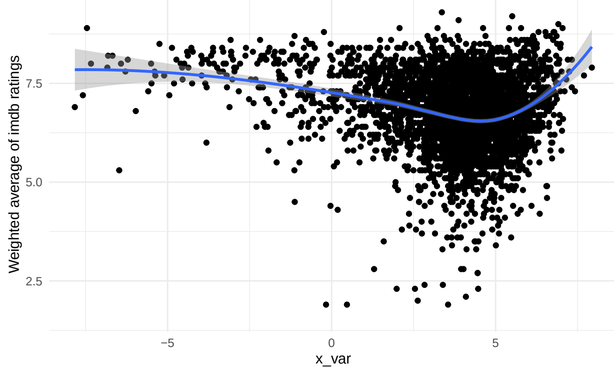
Transformations for x variables

Duration Log seems better

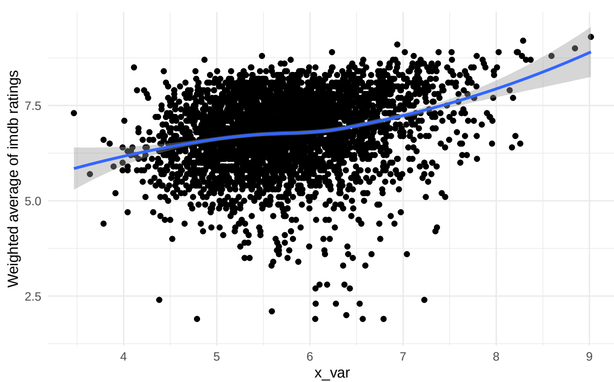


Metacritic The only variable that shows a linear pattern with IMDB rating thus this is going to be an important variable in the regression model



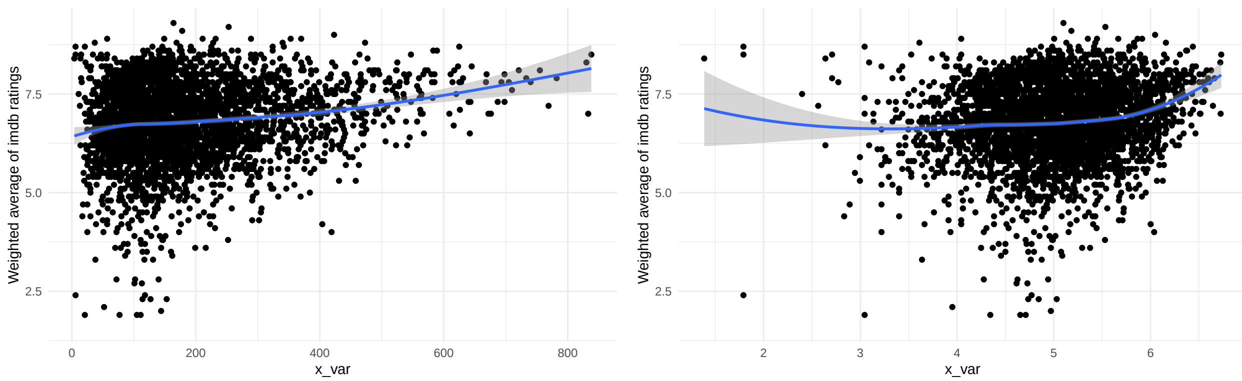


Gross income



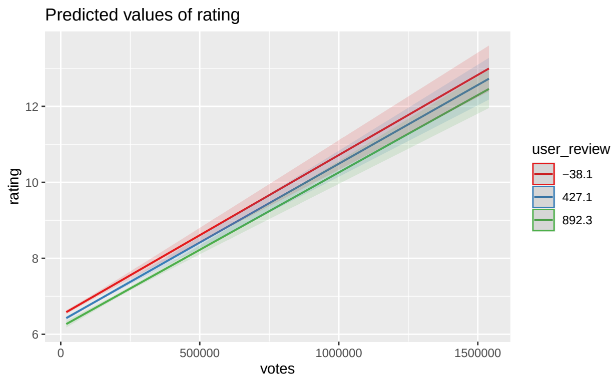
User_review

Critics review Taking the log does not really help.



Interaction

The lines between the variables votes - gross income are parallel, so no interaction occures between them, while for ucratio - votes we can see the following graph:



The plot shows that when user review raises from low to medium there is a slight change (drop) in the output rating. The interaction term is significant at 1%, still these slope differences are not that extreme, almost parallel.

Model summary