How to predict a Movie score on IMDB?

Final Poject for Data Analysis 2

Pauline Broussolle

2020-12-20

Link to my github repo https://github.com/Paulinebrsl/Assignment-1.

We are interested in predicting the score of a movie on IMDB. The aim is to identify the variables that are highly correlated to the outcome variable "imdb_score" and to create a regression model.

IMDB is the most important movie database and it is consulted by millions of spectators around the world. This analysis can give some insights to movie producers and distributors.

We are based on the hypothesis that movie notation from IMDB users have a positive association with the movie gross and profit. The movie "gross" refers to gross box office earnings in USD and net profit refers distributor's gross earning minus marketing expenses and distribution costs. Indeed, we can think that if a movie has an important box office score, it means that the public liked it and can give the movie a higher score on IMDB. We will see to what extent this association is true or not. For our analysis, we use the IMDB 5000 movie dataset.

The IMDB 5000 movie dataset comes from Kaggle, we are working. It records data about 5000 movies on the Internet Movie Database (IMBD), from 1916 to 2016.

1. Data Description

The raw dataset is very large, as it contains 26 variables for 5043 observations spanning across 100 years, concerning 66 countries.

I started by cleaning the data. I removed duplicates and I removed missing values from two important variables: budget and gross. Almost 20% of the dataset was concerned. We have left less than 5% of the different rows with missing values, which I consider satisfying.

I eliminated several variables that were not very relevant for prediction of IMDB rating, like aspect ratio and movie link on IMDB. I also noticed that language and color were not important factors, as over 95% movies are in color, which means this variable is nearly constant, and over 95% movies are in English. Thus, I chose to eliminate those two variables. Concerning the country of origin, I noticed that movies mainly come from the USA (almost 80%). Thus, I decided to group the country variable into 2 groups: "USA" and "Others", in order to have less levels. Finally, concerning movie date of release, I noticed that most of the movies in the data are released after 1960. Therefore, I decided to remove movies with a release date before 1960.

In the end we have 3834 observations out of 27 variables.

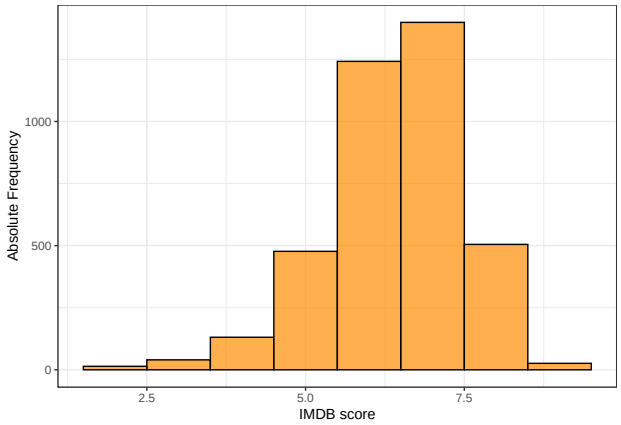
I added two new variables to the dataset: profit, which is equal to gross minus budget, and percentage return on investment, which is the ratio of profit on budget. I believe that those two variables can help us to have a better understanding of IMDB ratings.

The following table shows the descriptive statistics of key variables in the data: imdb score, profit, number of users who voted and movie gross.

Table 1: Descriptive summary of the variables

mean	median	min	max	std	variable
6	6.6	1.60000e+00	9.3	1	imdb score
5559480	800300.0	-1.22133e+10	523505847.0	227664733	movie profit
102454	50402.0	2.20000e+01	1689764.0	150546	number of users who voted
51016740	27996968.0	1.62000e+02	760505847.0	69373669	movie gross

We check the distribution of imdb scores, the distribution limited between 0 an 10. It is skewed with a left tail. Indeed, a large number of scores are between 5.0 and 7.5.



2. Model

Our aim is to identify the factors that are highly correlated to the rating of a movie on IMDB. The outcome variable is "imdb_score".

Thus, we want to regress "imdb_score" on predictive variables of the dataset. Intuitively, we think that the variables "gross", "profit" and "number of voted users" can be significative explanatory variables. Also we would like to add the categorical variable "genre".

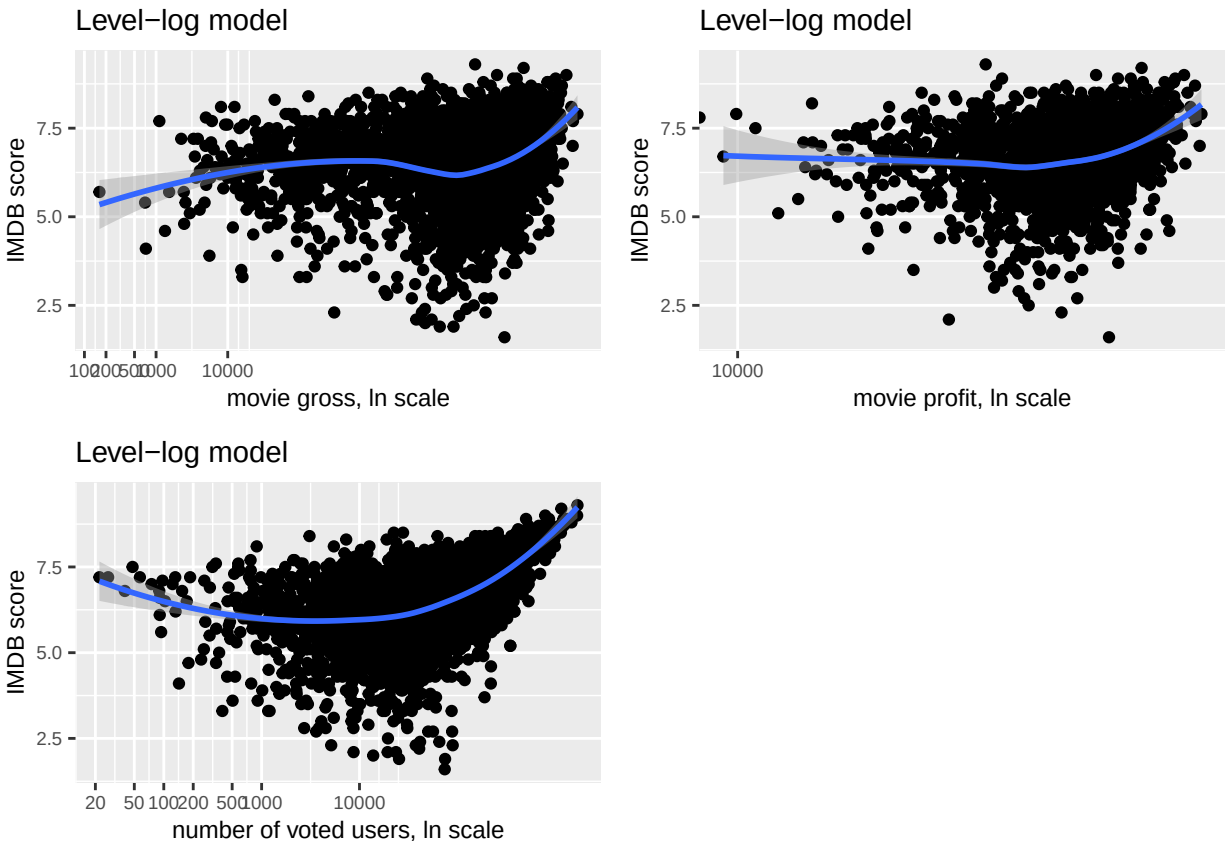
2. a) Pattern of association

First we check the pattern of association between y and each key x variables, with non-parametric regression, by plotting different scatterplots with lowess. We check possible different ln transformation for the variables: gross", "profit", "num_critics_for_review" and "num_voted_users".

For the simple model without scaling, the pattern is non-linear for all the variables. Most of observations are either concentrated on the left or the right of the plot. the bottom and there are some extreme observations.

Instead, the model with the level-log transformation creates a more linear association for all four variables. We see a linear upward trend on the right of the plot. We see that there are some outliers on low scores below 5.0.

We create new variables for log of the four variables: ln_profit, ln_gross, ln_num_voted_users and ln_num_critic_for_reviews. It is easier to interpret and it gives a better approximation to the average slope of the pattern.



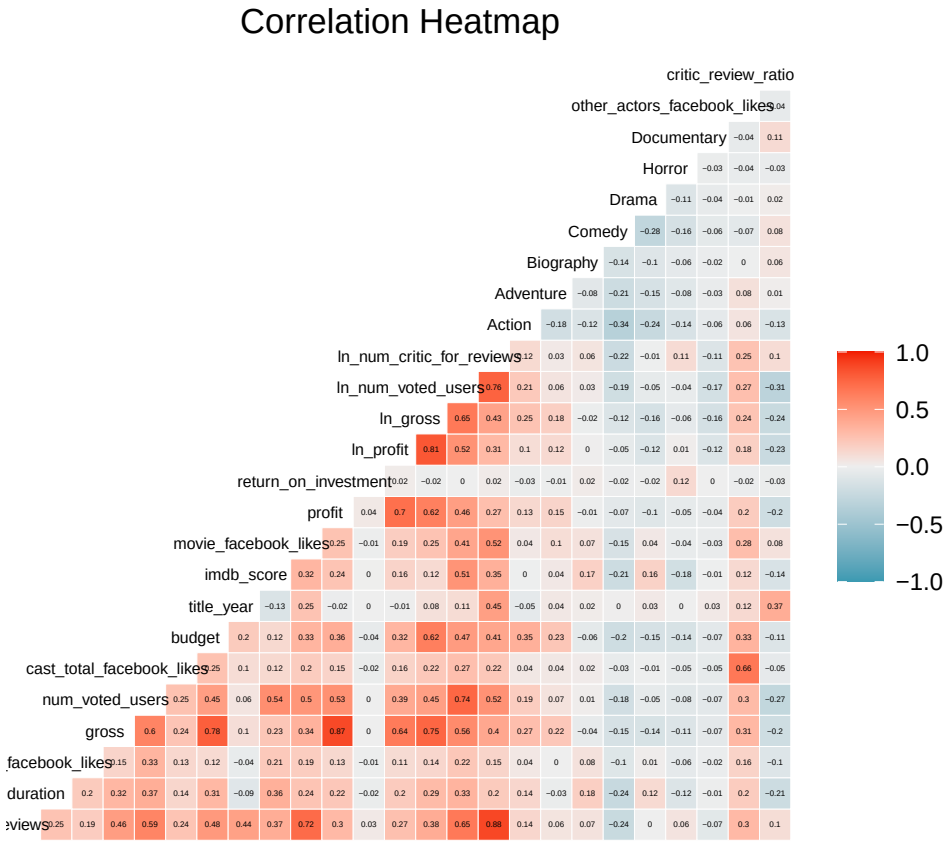
2. b) Compare explanatory variables

We check correlation between variables by creating a correlation heatmap. For creating the correlation heatmap, we do some modifications on the dataset: we remove variables that we won't use for our analysis like actors names and plot keywords.

We created a binary variable for title year: the variable is 1 if the movie is released after 2000. The variable is 0 otherwise. Same for duration, we create a binary variable: 1 if duration is superior to 120 min, 0 otherwise.

We created dummy variables for genre: Action, Adventure, Biography, Comedy, Drama, Horror, Documentary.

We can identify which variables are most correlated with imdb score, which are log of number of critic for review, log of number of voted voted users, duration, and gross.



2. c) Model choice

We estimate three different regression models: from least to most extended model. We use following models:

First regression model with no controls:

imdb_score = alpha + Beta* log(num_voted_users)

Second regression model with controls:

imdb_score = B0 + B1* log(num_voted_users) + B2* log(gross) + B3* log(profit) + B4* log(num_critic_for_review) + B5* duration

Third regression model, extended model:

imdb_score = B0 + B1* log(gross) + B2* log(num_voted_users) + B3* log(profit) + B4* log(num_critic_for_review) + B5* duration+ B6* Action + B7* Adventure + B8* Drama + B9* Comedy + B10* Horror

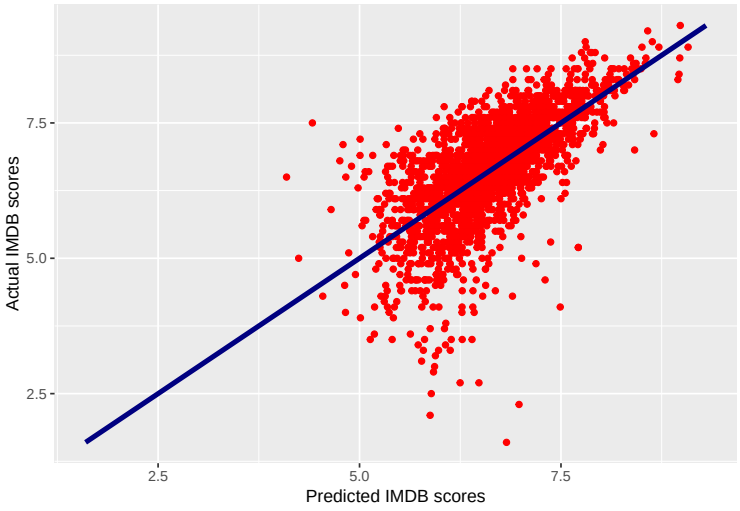
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Ven, jan 01, 2021 - 22:56:09

The results suggest that without controlling for any other variable, IMDB score is 0.04 units higher, on average, for movies with ten percent more users who voted. Then we would like to extend our model to make it more precise. We compare movie that have the same gross, profit, number of critics with reviews and duration less than 120 min – but that differ in terms of number of users who voted. We find that movies with ten percent more voted users have a 0.06 units higher IMDB score, on average. We extend this model even more by adding dummy variables for movie genres. When we compare movies with the same gross, profit, number of critics with reviews, duration and genre, we find that movies with ten percent more voted users have a 0.06 units higher IMDB score, on average. From our first to our third model, the R-squared

Table 2: Linear regression results

	Dependent variable:		
	imdb_score		
	(1)	(2)	(3)
ln_gross		−0.482*** (0.027)	−0.463*** (0.027)
ln_profit		0.140*** (0.019)	0.146*** (0.019)
ln_num_voted_users	0.384*** (0.014)	0.580*** (0.024)	0.561*** (0.023)
ln_num_critic_for_reviews		−0.114*** (0.033)	−0.090*** (0.032)
duration		0.589*** (0.043)	0.468*** (0.043)
Action			−0.390*** (0.056)
Adventure			−0.042 (0.069)
Drama			−0.005 (0.058)
Comedy			−0.423*** (0.054)
Horror			−0.901*** (0.082)
Constant	2.355*** (0.162)	6.735*** (0.255)	6.696*** (0.257)
Observations	2,014	2,014	2,014
R^2	0.262	0.415	0.471
Adjusted R^2	0.261	0.413	0.469
Residual Std. Error	0.880 (df = 2012)	0.785 (df = 2008)	0.747 (df = 2003)
Note:			*p<0.1; **p<0.05; ***p<0.01

increased from 26% to 47%. Thus we keep the third model, which is a better fit. Below we plot our predicted IMDB scores on actual IMDB scores to visualize the fit of our model. Based on our chosen model, most x variables are significant at 1%, except dummy variables "Adventure" and "Drama". For exemple, we can see that when we compare movies with the same gross, profit, number of voted users, critics with reviews and duration, but differ in genre, we find that horror movies have a 0.9 units lower IMDB score, on average. We can state with 95% confidence the score of a horror movie is between 0.98-0.819 units lower.




```
##        fit      lwr      upr
## 1 6.861436 5.392584 8.330289
```

```
##       fit      lwr      upr
## 1 7.30596 5.835484 8.776435
```

2. d) Residual analysis

We analyse residuals: we check for our highest and lowest residuals. We can see that negative errors are more important than positive errors. From our visualization above, we can infer that there are more outliers on movies with low scores (from 2.7 to 1.6). Thus our model is less fitted for low-rated movies. The

Table 3: Movies with largest negative error

movie_title	imdb_score	reg3_y_pred	reg3_res
Meet the Spartans	2.7	6.482357	-3.782357
Date Movie	2.7	6.248481	-3.548481
Epic Movie	2.3	6.979753	-4.679753
Justin Bieber: Never Say Never	1.6	6.824063	-5.224063
Crossover	2.1	5.876426	-3.776426

Table 4: Movies with largest positive error

movie_title	imdb_score	reg3_y_pred	reg3_res
Dolphins and Whales 3D: Tribes of the Ocean	6.5	4.094197	2.405803
Marilyn Hotchkiss' Ballroom Dancing and Charm School	7.1	4.795176	2.304824
Sholem Aleichem: Laughing in the Darkness	6.8	4.755904	2.044096
Short Cut to Nirvana: Kumbh Mela	7.2	5.007467	2.192533
Call + Response	7.5	4.414972	3.085028

2. e) Robustness check

Table 5: Movies with largest negative error

movie_title	imdb_score	reg3_y_pred	reg3_res
Meet the Spartans	2.7	6.482357	-3.782357
Date Movie	2.7	6.248481	-3.548481
Epic Movie	2.3	6.979753	-4.679753
Justin Bieber: Never Say Never	1.6	6.824063	-5.224063
Crossover	2.1	5.876426	-3.776426

Table 6: Movies with largest positive error

movie_title	imdb_score	reg3_y_pred	reg3_res
Dolphins and Whales 3D: Tribes of the Ocean	6.5	4.094197	2.405803
Marilyn Hotchkiss' Ballroom Dancing and Charm School	7.1	4.795176	2.304824
Sholem Aleichem: Laughing in the Darkness	6.8	4.755904	2.044096
Short Cut to Nirvana: Kumbh Mela	7.2	5.007467	2.192533
Call + Response	7.5	4.414972	3.085028

We investigated registered death due to covid and number of confirmed covid cases pattern of association. We used the model of Weighted linear regression, using population as weights. The slope for our chosen model is 0.953. This shows that countries with 10% more recorded covid cases have on average 9.5% more death

due to covid. This means that the number of registered death due to covid and the number of registered covid cases are positively correlated. The analysis helps us to identify countries with unexpectedly low and high mortality number due to covid.