# 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/Final_Project_Pauline_Broussolle.

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. 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 and in English, which means these variables are nearly constant. 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 categories. 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.

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.

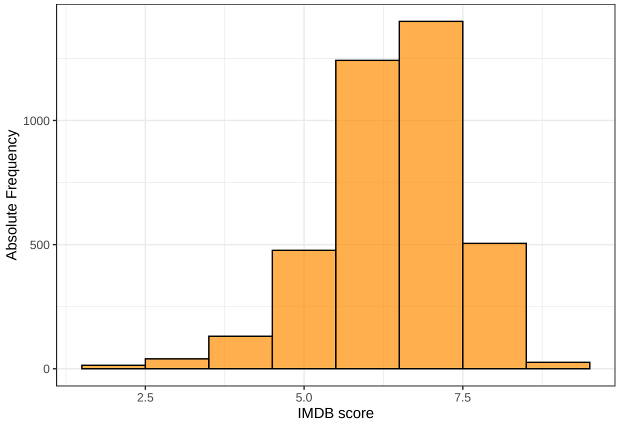
In the end we have 3834 observations out of 27 variables.

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 located between 5.0 and 7.5 as the mean is 6.0.



# 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”.

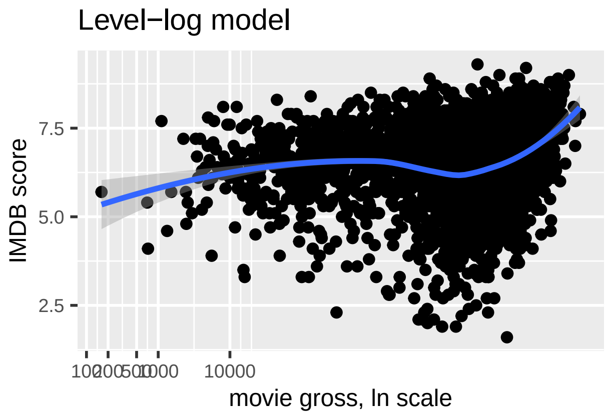
## 2. a) Pattern of association

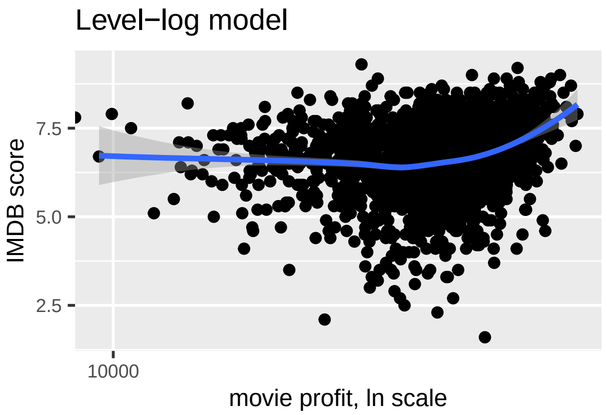
First we check the pattern of association between y and each key x variables, with non-parametric regression, by plotting different scatterplots with lowess. We check possible different ln transformation for the variables: “gross”, “profit”, “num_critics_for_review” and “num_voted_users”.

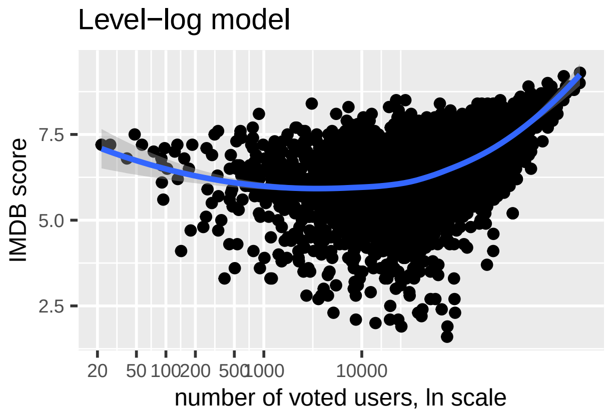
For the simple model without scaling, the pattern is non-linear for all the variables. Most of observations are either concentrated on the left or the right of the plot. There are a lot of extreme values.

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. The patterns are not totally linear, we could use a Piecewise Linear Spline model for a better fit.

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.







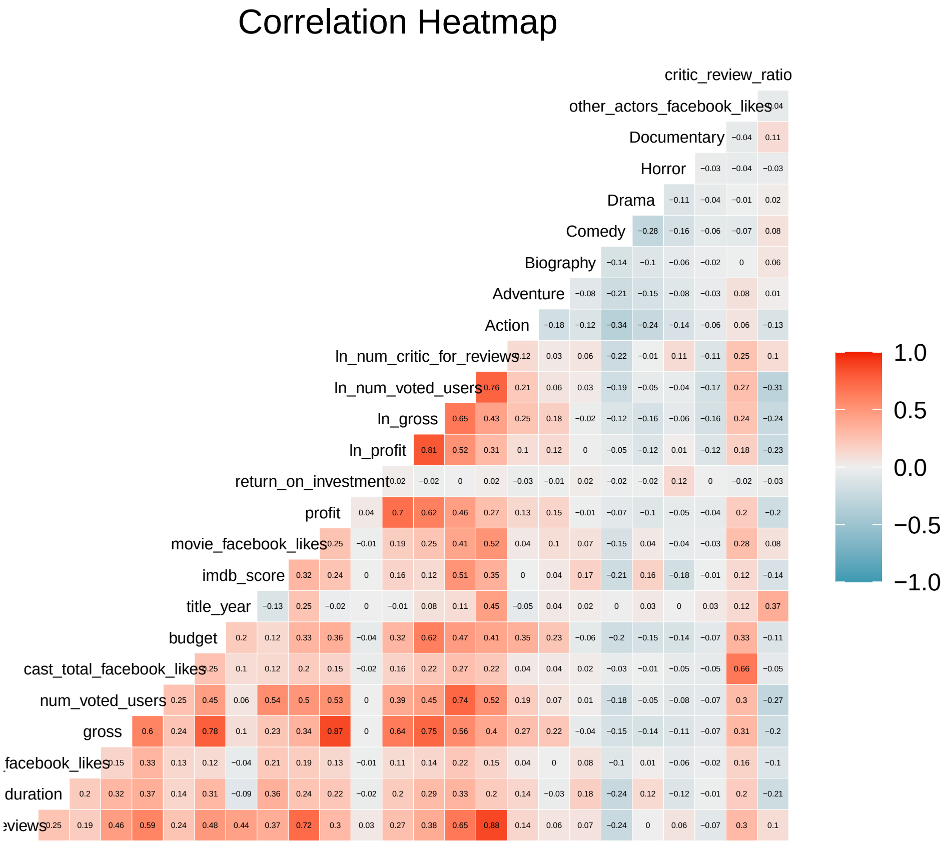
## 2. b) Compare explanatory variables

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. As we saw from the patterns of association above, we choose tu use Piecewise Linear Spline with one knot for gross, profit and number of users who voted. We use following models:

**First regression model with no controls:**

imdb_score = alpha + Beta* lspline( log(num_voted_users), log(12000) )

**Second regression model with controls:**

imdb_score = B0 + B1* lspline( log(num_voted_users), log(12000) ) + B2* lspline( log(gross), log(8000000) ) + B3* lspline( log(profit), log(8000000) ) + B4* log(num_critic_for_review) + B5* duration

**Third regression model, extended model:**

imdb_score = B0 + B1* lspline( log(num_voted_users), log(12000)) + B2* lspline( log(gross), log(8000000)) + B3* lspline( log(profit), log(8000000)) + 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: Dim, jan 03, 2021 - 15:55:08

According to the first regression without controlling, using Piecewise Linear Spline, when comparing movies with less than log(12000) voted users, imdb_score is 0,229 units lower on average, for movies with ten percent more voted users. When comparing movies with more than log(12000) voted users, imdb_score is 0.527 units higher on average, for movies with ten percent more voted users. This model captures the flattening of the regression line at the start. The R-squared is 0.32.

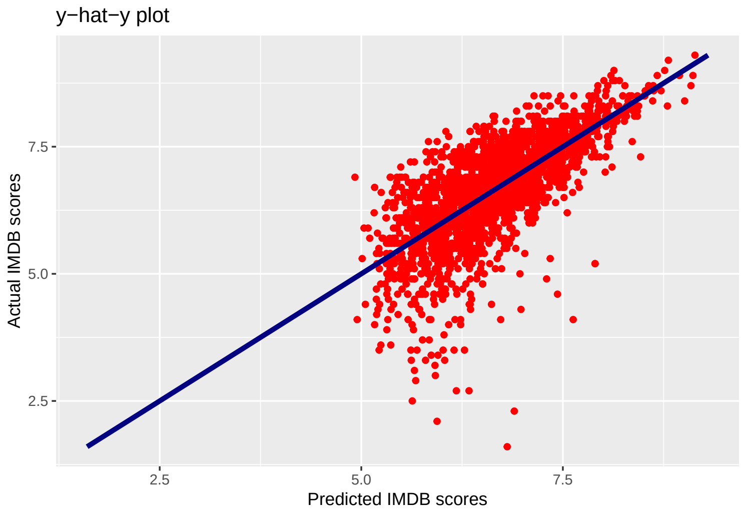
Table 2: Linear regression results

| | | *Dependent variable:* | |
|---|---|---|---|
| | | imdb_score | |
| | (1) | (2) | (3) |
| lspline(ln_profit, log(8e+06))1 | | 0.020 (0.030) | 0.026 (0.028) |
| lspline(ln_profit, log(8e+06))2 | | 0.251$^{***}$ (0.031) | 0.259$^{***}$ (0.031) |
| lspline(ln_gross, log(8e+06))1 | | −0.268$^{***}$ (0.055) | −0.226$^{***}$ (0.053) |
| lspline(ln_gross, log(8e+06))2 | | −0.540$^{***}$ (0.035) | −0.530$^{***}$ (0.036) |
| lspline(ln_num_voted_users, log(12000))1 | −0.229$^{***}$ (0.049) | 0.100$^{*}$ (0.054) | 0.101$^{*}$ (0.052) |
| lspline(ln_num_voted_users, log(12000))2 | 0.527$^{***}$ (0.018) | 0.658$^{***}$ (0.025) | 0.637$^{***}$ (0.025) |
| ln_num_critic_for_reviews | | −0.076$^{**}$ (0.032) | −0.055$^{*}$ (0.032) |
| duration | | 0.542$^{***}$ (0.043) | 0.428$^{***}$ (0.042) |
| Action | | | −0.378$^{***}$ (0.055) |
| Adventure | | | −0.051 (0.068) |
| Drama | | | 0.023 (0.057) |
| Comedy | | | −0.398$^{***}$ (0.052) |
| Horror | | | −0.864$^{***}$ (0.080) |
| Constant | 7.787$^{***}$ (0.443) | 9.343$^{***}$ (0.635) | 8.768$^{***}$ (0.608) |
| Observations | 2,014 | 2,014 | 2,014 |
| R$^2$ | 0.320 | 0.451 | 0.504 |
| Adjusted R$^2$ | 0.319 | 0.448 | 0.501 |
| Residual Std. Error | 0.845 (df = 2011) | 0.761 (df = 2005) | 0.724 (df = 2000) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Then we would like to extend our model to make it more precise. We compare movies that have the same gross, profit, number of critics with reviews and duration – but that differ in terms of number of users who voted. We find that correlation is now positive between imdb scores and log number of voted users, both less and more than ln(12000) voted users. The R-squared is 0.45.

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, genre, and more than log(12000) voted users, imdb_score is 0.637 units higher on average, for movies with ten percent more voted users.

From our first to our third model, the R-squared increased from 32% to 50%. Thus we keep the third model, which is a better fit. We plot below our predicted IMDB scores on actual scores to visualize the fit of our model. Based on our chosen model, most explanatory variables are significant at 1%, except log of profit below log(8000000) and dummy variables "Adventure" and "Drama". For exemple, we can see that when we compare movies with the same gross, profit, number of voted users, number of critics with reviews and duration, but that differ in genre, we find that horror movies have a 0.86 units lower IMDB score, on average. We can state with 95% confidence the score of a horror movie is between 0.78 and 0.94 units lower. For this third model, we see that correlation is positive between imdb scores and log(profit), but it is negative between imdb scores and log(gross).



## 2. d) Residuals 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.

Table 3: Movies with largest negative error

| movie_title | imdb_score | reg3_y_pred | reg3_res |
|---|---:|---:|---:|
| Fifty Shades of Grey | 4.1 | 7.628119 | -3.528119 |
| Meet the Spartans | 2.7 | 6.336408 | -3.636408 |
| Epic Movie | 2.3 | 6.897544 | -4.597544 |
| Justin Bieber: Never Say Never | 1.6 | 6.809846 | -5.209846 |
| Crossover | 2.1 | 5.939797 | -3.839797 |

Table 4: Movies with largest positive error

| movie_title | imdb_score | reg3_y_pred | reg3_res |
|---|---:|---:|---:|
| The Conjuring 2 | 7.8 | 6.049233 | 1.750767 |
| Secondhand Lions | 7.6 | 5.940776 | 1.659224 |
| The Muppet Christmas Carol | 7.7 | 6.084359 | 1.615641 |
| Lights Out | 6.9 | 4.921403 | 1.978597 |
| Instructions Not Included | 7.6 | 5.832710 | 1.767290 |

## 2. e) Robustness check

We check the robustness of our model. Since we have many observations, we use a test sample where we re-run our regression and prediction, to check if the results are true. Therefore we use an alternative sample which represents roughly 10% of our original dataset. For creating this new sample, we choose to:

- Keep only movies from "Others" countries, thus remove movies from USA
- Include movies released before 1960
- Remove movies released after 2010

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Dim, jan 03, 2021 - 15:55:10

Table 5: Model Robustness analysis

| | *Dependent variable:* | |
|---|---|---|
| | imdb_score | |
| | Other countries | Other countries |
| | (1) | (2) |
| lspline(ln_profit, log(8e+06))1 | 0.078 (0.063) | 0.077 (0.058) |
| lspline(ln_profit, log(8e+06))2 | $0.498^{***}$ (0.084) | $0.427^{***}$ (0.079) |
| lspline(ln_gross, log(8e+06))1 | $-0.278^{***}$ (0.106) | $-0.191^{*}$ (0.098) |
| lspline(ln_gross, log(8e+06))2 | $-0.798^{***}$ (0.081) | $-0.701^{***}$ (0.079) |
| lspline(ln_num_voted_users, log(12000))1 | 0.080 (0.134) | 0.041 (0.123) |
| lspline(ln_num_voted_users, log(12000))2 | $0.525^{***}$ (0.069) | $0.499^{***}$ (0.063) |
| ln_num_critic_for_reviews | 0.001 (0.106) | 0.019 (0.099) |
| duration | $0.590^{***}$ (0.105) | $0.471^{***}$ (0.100) |
| Action | | $-0.501^{***}$ (0.136) |
| Adventure | | −0.209 (0.172) |
| Drama | | −0.045 (0.139) |
| Comedy | | $-0.467^{***}$ (0.146) |
| Horror | | $-1.172^{***}$ (0.205) |
| Constant | $8.979^{***}$ (1.388) | $8.243^{***}$ (1.280) |
| Observations | 218 | 218 |
| $R^2$ | 0.532 | 0.621 |
| Adjusted $R^2$ | 0.514 | 0.597 |
| Residual Std. Error | 0.649 (df = 209) | 0.591 (df = 204) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

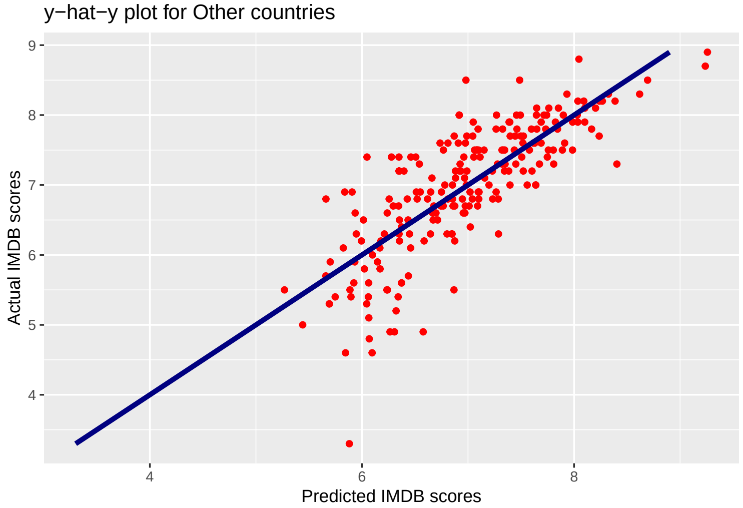


Table 6: Movies with largest negative error

| movie_title | imdb_score | reg3_y_pred | reg3_res |
|---|---:|---:|---:|
| Superman III | 4.9 | 6.577406 | -1.677406 |
| Seed of Chucky | 4.9 | 6.306672 | -1.406672 |
| Spice World | 3.3 | 5.880957 | -2.580957 |
| Confessions of a Teenage Drama Queen | 4.6 | 6.095792 | -1.495792 |
| White Noise | 5.5 | 6.867553 | -1.367552 |

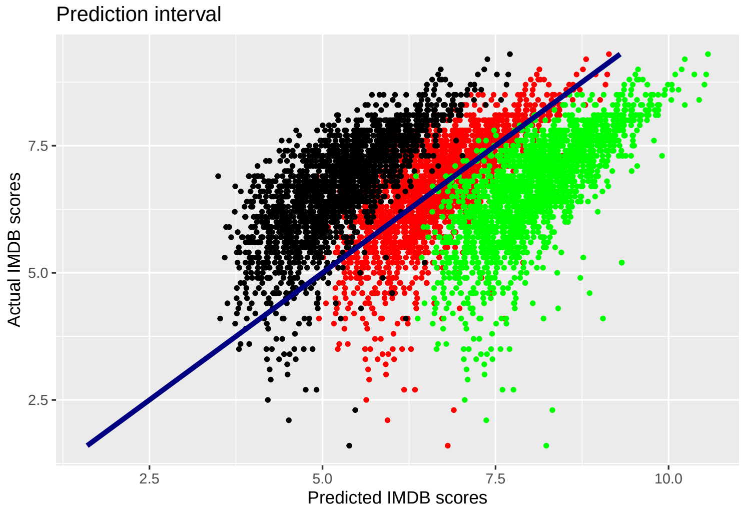
Table 7: Movies with largest positive error

| movie_title | imdb_score | reg3_y_pred | reg3_res |
|---|---:|---:|---:|
| Casino Royale | 8.0 | 6.917241 | 1.082759 |
| Shadowlands | 7.4 | 6.047278 | 1.352722 |
| Quigley Down Under | 6.8 | 5.660255 | 1.139745 |
| Alien | 8.5 | 6.980187 | 1.519813 |
| Monsoon Wedding | 7.4 | 6.278089 | 1.121911 |

According to our table of estimation results above, we see that R-squared has increased for both regressions, thus the model is a better fit with this sample. The value of each parameter has increased, but they do not show a significant change. We can see from our y-hat-y visualization that there are still extreme values, especially for low-rated movies on bottom part of the plot. However we see from our analysis of residuals that positive and negative errors have reduced, which can be due to the fact that we reduced a lot the size of the sample and we may have removed some outliers. Finally, we can say that our model is still significant with this sample, and it has the same weaknesses.

## 3. Prediction and uncertainty

We created a visualization of the prediction interval for IMDB scores. The green dots are the upper parts of prediction intervals and black dots are the lower parts. We can be 95% confident that IMDB scores are between these values. We also plotted the actual IMDB scores, to see if they are actually within the prediction interval. We see that our prediction interval contains the actual IMDB scores, so we can validate our model. However the weaknesses of our model are: there are some outliers, especially for low-rated movies ; some of the upper parts and lower parts of the prediction interval overlap with actual IMDB scores. The model do not fit completely, there could be other factors at play. Our predictions are true within the prediction interval, but we would need more explanatory variables to have much more exact predictions.



## Summary

We created a prediction model for movie scores on IMDB. We investigated potential explanatory variables correlated to IMDB scores. We used a log transformation for movie gross, profit, number of users who voted and number of critics for reviews. We also added the binary variable duration (1 if duration is superior to 120 min and 0 otherwise) and dummy variables for genres to our multiple regression. Using Piecewise Linear Spline for gross, profit and number of users who voted, we obtained a R-squared of 0.5. We checked for robustness, restricting our attention to movies from countries other than USA and released between 1922 and 2010, which did not have a significant change on our predictions.

We analyzed the differences between our predicted scores and actual IMDB scores. We found that actual IMDB scores are within our prediction interval. Howerver our model is not precise, we would need more explanatory variables to have much more exact predictions. Furthermore, our model is less fitted for movies with low IMDB scores, we should uncover specific factors why people can dislike a movie. This underlines that some factors that make a good movie are difficult to quantify, like artistic talent of the director and actors.