THE COUNTRY-OF-ORIGIN (COO) EFFECT ON MOVIE PREFERENCES: A COMPARISON BETWEEN CHINESE AND AMERICAN AUDIENCES USING MOVIE RATINGS DATA FROM DOUBAN AND IMDB

Xiaotong Chen

May 2019

A thesis submitted to the department of Mathematical Methods in the Social Sciences (MMSS)

Advised by Professor Robert Porter

Northwestern University
Contents

Abstract iv

Acknowledgements v

1 Introduction 1

1.1 Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2.1 Insights into the Preferences Using Ratings Data . . 3

1.2.2 Prediction v.s. interpretation . . . . . . . . . . . . . . 4

1.2.3 Larger Data Sets . . . . . . . . . . . . . . . . . . . . . 4
2 Data

2.1 Overview of the Data ........................................... 6

2.1.1 Data Collection Methods .............................. 7

2.1.2 Limitations to the Data and the Solutions ............. 7

2.2 Descriptive Statistics of the Data Sets ................. 9

2.2.1 A Description of the Data Sets ......................... 9

2.2.2 Advantages of Using the Subset of Raw Data ........ 11

3 Methods ........................................................ 15

3.1 Hypotheses .................................................. 16

3.2 Models ....................................................... 19

3.2.1 "Naive" Multiple Linear Regressions (MLR) ......... 19

3.2.2 Adding Two-Way Interaction Terms ................. 21

3.2.3 Detecting and Fixing the Multicollinearity Issue .... 22

3.2.4 Models for Extreme Ratings Analysis ............. 25

3.2.5 Three-Way Interaction Models .................. 25

4 Results ..................................................... 27
Abstract

This paper aims to compare the preferences of Chinese and American audiences influenced by a set of factors using web-scraped Douban and IMDb data. With two-way, three-way interaction models and extreme rating models, this paper explores how variables such as budget, genres, countries of production, and length impact film ratings differently across audience groups. Such variations are found to widely exist across genres. Extreme rating behaviors are found to be influenced differently across audience groups by a film’s country of production, while average ratings are not. American audiences are found to be more likely to give extreme low ratings to foreign films than to domestic ones, while Chinese audiences are not. Both audience groups are found to give more extreme high ratings to domestic films. No evidence show that budget or length of film influence film ratings in either audience group.
Acknowledgements

First and foremost, I would like to thank my thesis advisor, Professor Robert Porter, for helping me formulate my research ideas, advising me on statistical modelling, answering my endless stream of questions during meetings or over emails, and always encouraging me with his kindness and patience. I would also like to thank Professor Joseph Ferrie for hosting MMSS Thesis Seminars, and thank Nicole Schneider for her constant help and caring throughout the past four years. Finally, I would like to thank my family and friends for their love. I am so grateful that they are always there to support my decisions, and to witness my growth.
1. Introduction

1.1 Background

Movie Ratings on platforms such as IMDb, Rotten Tomatoes, and Douban not only serve as a reliable means to predict the quality of movies, but also show the aggregated sentiments and inherited preferences of the respective user groups and audiences. Prior research using ratings data often focuses on either the prediction of box office [Yang and Yecies, 2016, Moon et al., 2010, Hennig-Thurau et al., 2007] or the inference of movie quality [Feng, 2018, Moon et al., 2010, Cui, 2018]. Some studies also discuss the differences in prediction algorithms or inference methods applicable to different movie markets and platforms [Feng, 2018, Han and Ji-Fan, 2017]. Surprisingly, there is little satisfying research that focuses on the systematic differences between film ratings across different audience groups and the factors that influence their respective preferences of films. This paper will use scraped data from movie rating websites (Douban and IMDb) to help fill in this gap.
Due to the platforms' wide coverage and the high volume of data, Douban and IMDb (U.S. users) are ideal proxies of the Chinese and American audiences under inspection. IMDb (Internet Movie Database) is an online database (established in 1990) that provides fan reviews and ratings on films and other media. Users are from all over the world, although a significant portion are American. The platform provides breakdowns of ratings given by U.S. and non-U.S. users, which enables this study to isolate ratings and the number of raters corresponding to U.S. audiences only. People rate films on a scale of 1 to 10, and the final ratings are weighted averages of the scores. Douban (established in 2005) can be thought as the "IMDb" in China. Its language and primary users are Chinese. Audiences rate films on a scale of 1 to 5, and the website gives a weighted average rating.

This paper will focus on the interpretation of regression results of ratings on other features, treating the audience group (i.e. their country of origin) as a dummy variable. Other features include number of rating votes, year of production, genre, budget, production country, language, length, etc. The paper will summarize descriptive statistics on the data and use multiple linear regression models with two-way and three-way interaction terms. It will also fix the multicollinearity issue via different selections of variables. Finally, it will further discuss the regression results to understand the relationship between ratings and the audience groups from different angles.

The rest of this chapter is a review of previous literature on relevant topics. The second chapter of the research will explain the data sets used in the paper, as well as the web-scraping and cleansing procedures. The third chapter will discuss the methods used in the study and the theories in detail. The fourth chapter will focus on the presentation and analysis of regression results. The final chapter is the discussion of the contributions and
the limitations of the study.

1.2 Literature Review

This section will discuss three major limitations of previous research, and explain how this paper fills in the gaps. As readers will see, previous literature that share the goals and approaches with this paper is limited. Research on the preferences of audience groups using large-enough data, which focuses on interpretation instead of prediction, is currently missing in the field.

1.2.1 Insights into the Preferences Using Ratings Data

Most previous research using Douban or IMDb ratings data does not aim to provide comprehensive insights into the preferences of the audience groups. Instead, they focus on the prediction of variables other than the preference, such as the box office and the true quality of movies.

For example, Jie Yang and Brian Yecies’s study [Yang and Yecies, 2016] provides an improved programming and algorithm on analyzing Douban movie reviews. The study focuses on developing more efficient machine learning models to capture the audience’s sentiments towards a specific movie. However, it does not further analyze the pattern of these sentiments among a set of movies. Another example is Mingyu Cui’s paper on rerating the Douban scores using word embedding [Cui, 2018]. The paper focuses on adjusting for the density of data to generate adjusted Douban ratings, improving the quality of Douban as a rating system, instead of studying Douban’s audience group.
Many studies either use Natural Language Processing (NLP) methods to analyze text data, rather than ratings and other numeric features, or focus on specific aspects of audience’s preferences, without providing insights into the whole picture. For instance, Yang et al.’s work on evaluating Chinese audiences’ attitudes towards Korean popular culture flows [Yecies et al., 2016] uses movie reviews to learn about the preferences of Chinese audiences on Korean films. This study uses text analysis and only aim to learn about preferences related to Korean movies.

A more comprehensive analysis using ratings data is necessary to fully understand an audience group. This paper takes a step further to compare Chinese and American audiences by using data from two movie markets.

1.2.2 Prediction v.s. interpretation

Most studies using ratings data focus on predictions or improved recommendation systems using machine learning approaches such as neural networks and random forest [Hsu et al., 2014, Cui, 2018]. Although these methods can give accurate predictions of responses such as box offices, their results are difficult to interpret, due to the complexity of features generated in the learning process. To better understand how each feature impacts audiences’ ratings of movies, regression analysis using econometric methods is necessary. This paper will use such methods to enable better interpretations of the results.

1.2.3 Larger Data Sets

Many studies under inspection use data sets with less than 100 films
[Yecies et al., 2016, Han and Ji-Fan, 2017, Yang and Yecies, 2016]. Some use survey data [Yang and Yecies, 2016] that is too idiosyncratic to some specific groups of audiences (e.g. students at a certain college). For the purpose of this paper, large-enough Douban and IMDb raw data sets are necessary. Douban has anti-scraping configurations that block IP addresses to prevent continuous scraping, which adds to the difficulty in mass-collecting data from the website. IMDb provides a subset of its data to the public, but the selection criterion is unknown, and many features are missing. This research scraped data of 10,000 most popular films on Douban and 1,000 most popular films on IMDb, as well as a combined common data set of 773 films. These data sets have a tolerable number of missing values, and will be large enough to conduct the regression analysis as well as prediction analysis.
2. Data

2.1 Overview of the Data

All data used in this study were obtained from Douban and IMDb websites directly, from February to April 2019, using python (for Douban) and a scraping software called ScrapeStorm (for IMDb). To make sure that the films to be selected in the data sets have been rated by enough audience (so that the ratings are significant), the study only includes films that are "popular enough" in terms of the number of raters (i.e. "the number of votes").

Thus, a Douban data set with the 8200 most popular films (db_raw) and an IMDb data set with the 1000 most popular films (imdb_raw) are obtained. After data cleansing and filtering, a list of 773 films are found in both data sets. They form a combined full data set, which is called commondb3. This data set contains features of the movies from Douban and IMDb, such as "title_imdb", "rating_imdb", "rating_db", "year", "us_box_office", "title_douban", and "rating_douban". The study only includes films that have been rated by at least 1000 raters, so that the ratings are significant.
"language", "num_votes_imdb", "length", "rating_db", etc., as well as some dummy variables on genres and countries.

The main data sets used to run regressions are `commondb3` and `commondb4`. The `commondb4` data set combines some of the dummy variables in `commondb3` by categories. Models based on these two data sets might have different performances and generate different useful results for analysis.

Refer to Section 2.2.1 for a full description of data.

### 2.1.1 Data Collection Methods

Many previous studies use relatively small data sets of films that are manually and arbitrarily selected. This can be problematic due to high variance of the observations and a lack of possibility to generalize the results. Specifically, with smaller data sets, the number of variables that researchers can include in the models are limited, which might lead to omitted variable bias. In general, at least 10 observations for each predictor is preferred. Therefore, this study makes sure that each data set suffices this criteria.

### 2.1.2 Limitations to the Data and the Solutions

**Separating TV series and films data within Douban Data**

The web-scraped Douban data contain both films and TV series, while IMDb data contain only films. However, since only the films that appear on both lists will be used, the observations in the data sets only contain films. Therefore,
this will not cause problems.

Dealing with the time bias on ratings due to different establishment dates of Douban and IMDb

One potential problem with Douban data is that the website is relatively new (established in 2005). The ratings given to movies that are produced in earlier years might suffer from a "time bias": since these movies are rated a long time after they were produced, the preference of audiences might have changed. For example, audiences might tend to like more recent films due to the advanced technologies in special effects. This bias should be less significant in IMDb data, since the website was established in 1990, which is 15 years earlier than Douban. By adding "year of release" as a feature in regressions, this bias is taken into account.

Isolating U.S. Raters from international ones on IMDb

One potential problem with acquiring data from IMDb is that the audience population is international, although mostly American. Therefore, it is necessary to carefully isolate U.S. raters from the rest during the data collection process. There is one data set provided by IMDb; however, the data set is not on U.S. level. Therefore, the study uses an original data set created through scraping, isolating U.S. audiences data from the global data. The IMDb data set used in this study not only includes the number of U.S. raters and the corresponding ratings of each film, but also the median and average scores that the U.S. raters give. This isolation will help limit the attention to comparing Chinese and American audience groups much better.
Using A Balanced Panel instead of an Unbalanced Panel

Another potential problem for the data collection method is that the raw data is an unbalanced panel: there are 10,000 films from IMDb, and only 1,000 from Douban. This is due to the difficulty of bypassing the anti-scraping mechanism of Douban website. In order to make scraping easier, the study uses Douban’s list of the most rated films, so that the IP address will not be blocked repeatedly due to too many searches.

Since there are enough observations to use, and it makes sense to only focus on the most popular films, which will likely be rated by a lot of people in both countries, the study combines raw data sets into a balanced panel by only including the films that appear in both Douban and IMDb data sets. Using this method would include 773 (out of 1,000) films, which is sufficient for the analysis. The following section will further discuss whether this selection of data will introduce a bias.

2.2 Descriptive Statistics of the Data Sets

2.2.1 A Description of the Data Sets

Below are the detailed descriptions of the data sets:

**commondb3:**

**Abstract:** Ratings and influencing factors for 773 films on IMDb and Douban

**Source:** Douban and IMDb websites
Date of Generation: Apr 2019

Number of Variables: 56

Number of Observations: 1546

Description: a pre-processed data set of 773 films that are common to the Douban and the IMDb raw data sets. There are 56 variables in total. It also includes an indicator variable db boo, which is equal to 0 if the rating is from IMDb, and equal to 1 if the rating is from Douban.
**commondb4:**

**Abstract:** Ratings and influencing factors for 773 films on IMDb and Douban, modifying some of the dummy variables in **commondb3**

**Source:** Douban and IMDb websites

**Date of Generation:** Apr 2019

**Number of Variables:** 49

**Number of Observations:** 1546

**Description:** a further-processed data set of 773 films in **commondb3**. It re-groups the country and language dummy variables to reduce dimensions of the regression problems. There are 49 variables after the consolidation. However, reducing the dimension is not guaranteed to produce better interpretation insights or prediction results. So we will run regressions on both **commondb3** and **commondb4**, compare the results, and then determine which model to choose.

### 2.2.2 Advantages of Using the Subset of Raw Data

Since the raw data contains 9 times more IMDb films than Douban films, considering the fact that the data is sufficient, this study takes a subset of the raw data by only including the films that show up in both websites.

This treatment has its advantages: by using this common subset of films, the less rated films in IMDb raw data will be eliminated (since we shrink the IMDb data set from 10,000 to 1000 observations), resulting in more significant
IMDb ratings that are comparable to those for Douban. At the same time, the common data set only includes the films that are well-known to both Chinese and American audiences, reducing some bias caused by unfamiliarity of the films. This will limit the attention of this study to other deeper reasons that cause the difference of perceptions and preferences of the two audience groups.

In order to make sure that the subset (i.e. the balanced panel data, which includes 773 films that show up in both IMDb and Douban raw data) well represents the original raw data population of our samples, below is a comparison between the films that are included in the subset of data \textit{(commondb3, commondb4)} with the most rated 773 films in IMDb and Douban raw data sets, respectively.

First the raw Douban and IMDb data sets are sorted by the number of votes for each film, and select the 773 most rated films to match the number of films in the balanced data sets. Then the frequency of the number of votes are plotted and used to compare the population of films that are most rated on both websites with that of the most rated films on either website (See Figure 2.1). These steps are repeated on the ratings in the different populations (See Figure 2.2).

From the two figures, the IMDb raw population is quite similar to the population of common films, in terms of both ratings and numbers of raters. However, the Douban raw films are different from the common population: first, the average number of raters for films on Douban reduces by half from 265267 to 136752 after subsetting; second, the mean of Douban ratings increases from 7.91 to 8.293 after subsetting. It seems that taking the common subset of films eliminates some of the popular films on Douban with relatively lower ratings. This is acceptable as explained before, and it makes sense as
some of the popular Douban films (e.g. localized and culturally-rooted films) are expected to be less known to U.S. audiences, but less so vice versa.

Therefore, the subset of the raw data is acceptable to use as an proxy of the raw data sample from Douban and IMDb. The next chapter will introduce the methods and regression models used in this study.
Figure 2.2: The ratings for each film
3. Methods

This chapter will describe the models used in the study, both the "naive" models that are ran at the first stages of the analysis, and the "mature" models that generate useful and reliable results. Results of the less appropriate models will not be reported.

As described in the previous chapter (and in the appendix), common\texttt{db3} and common\texttt{db4} are similar data sets, with different categorizations of countries and languages, resulting in different numbers of dummy variables. The common\texttt{db4} data set groups some of the dummies according to their similarities (e.g. grouping Asian countries into one single dummy) and has less variables than common\texttt{db3}. Regressions are ran on both data sets, and the results are compared for insights from different perspectives.
3.1 Hypotheses

hypothesis I: In both data sets, the variables explain the ratings well, and the number of votes should be significant in predicting the film ratings.

In general, it is expected that the variables would explain ratings fairly well. Specifically, the number of votes for the film should be significant in predicting the popularity and the quality of the film, which will positively influence the ratings. It is worth pointing out that Douban and IMDb were established in different years, and the variable "year" should be significant. It is necessary to include this variable to adjust for the time bias, although it can be difficult to use the coefficient of this variable to interpret the true impact of the year of release on a film’s ratings.

hypothesis II: Models on the two data sets will not produce contrasting results, although the performances of the models might be different.

The two data sets (commondb3, commondb4) are quite similar. The only difference are in the dummy variables. Therefore, it is expected to get similar results from the two data sets.
hypothesis III: Budget will positively impact both Douban and IMDb ratings.

Since films with lower budgets might not be able to employ expensive actors or use extensive special effects, it is natural to believe that films with higher budgets will have higher qualities and will be liked more by audiences in both China and the U.S. The models with interaction terms will also show whether such an impact, if exists, is different across different audience groups.

hypothesis IV: Length of the films will not impact the ratings.

Although length of the films may not impact the ratings in a systematical way, it is likely that Chinese and American audiences have different preferences over the length of films. For example, Chinese audiences might prefer shorter films comparing to their American counterparts. The models with interaction terms will tell whether this is the case.

hypothesis V: Douban ratings are higher than IMDb ratings in general.

If this hypothesis is true, then when one wants to know about whether a film is liked by Chinese and American audiences, he/she need to keep in mind that a lower Douban score does not show that Chinese audiences like this film less than their American counterparts. The arguments are similar if the result is the other way around.
hypothesis VI: Chinese audiences and American audiences prefer different genres, languages and production countries of foreign films.

Different audience groups might have different favourite film genres. For example, Chinese audiences may like animation films better, while American audiences might like action and adventure films better.

hypothesis VII: For foreign films, Chinese audiences and American audiences prefer different languages/production countries.

Cultural similarities and differences, familiarity with the language of subtitles and speeches, as well as other factors, may lead to different preferences of movie genres and production countries. For example, Chinese audiences might prefer the films produced by Asian countries, while American audiences might prefer those produced in Europe and U.K.

hypothesis VIII: Audiences in both countries prefer their domestic productions.

[Feng, 2018, Han and Ji-Fan, 2017].

It makes sense to assume that the films that audiences feel most comfortable with are their domestic productions. Therefore, it is expected that Chinese audiences give systematically higher ratings to Chinese films, and American audiences give higher ratings to American films.
hypothesis IX: Audiences in both countries are more likely to give extreme low ratings to foreign films.

Since audiences might be less familiar with the cultural concepts or the language of foreign films comparing to domestic ones, they might be more likely to "hate" these films and give extreme low ratings. This paper will run separate regressions on Douban and IMDb data sets to test whether this is true for either country.

The following section will exhibit the models used to test these hypotheses. Most models are ran on both commondb3 and commondb4, and the appropriate models are selected after obtaining their significance level using F-tests.

3.2 Models

3.2.1 "Naive" Multiple Linear Regressions (MLR)

Dummy variables are created for genres, countries, and languages in commondb3 and commondb4. Now the fastest and easiest way to use these data is to run a multiple linear regression on all possible variables, including the dummies. However, this method is too "naive". It carries a significant multicollinearity issue, which will be addressed and fixed in later models.

Multiple linear regression models are of the form:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \]

where \( y \) is the dependent variable, \( x_i \)'s are the independent variables (or
features), \( \beta \)'s are the coefficients of the variables, and \( p \) are the number of features (including a bias term). Since both data sets contain more than 30 variables, this section will not write out the models in equation form. Instead, the variables that are included in the model will be listed for clarity. Models in their full forms are included in the appendix.

"Naive" Full model on commondb3

The model on **commondb3** regresses the variable "rating" on the following 36 variables: `num_votes_imdb_us`, `num_votes_db`, `year`, `budget`, `length`, `db_boo`, genre dummies (action, adventure, animation, biography, comedy, crime, drama, horror, mystery), country dummies (china, france, germany, india, italy, japan, russia, southkorea, spain, uk, usa), and language dummies (chinese, english, french, french, german, hindi, italian, japanese, korean, spanish).

Specifically, `db_boo` is a dummy variable that indicates whether the observation is for Douban rating or IMDb rating. If `db_boo` is set to 1, then the "rating" variable represents the Douban rating for the film; if `db_boo` is set to 0, then the "rating" variable represents the IMDb rating for the film. This design includes Douban and IMDb ratings in a single regression, and compares their systematic difference. Other dummy variables in the model are designed in similar ways.

"Naive" Full model on commondb4

The model on **commondb4** regresses the variable "rating" on the following 28 variables: `num_votes_imdb_us`, `num_votes_db`, `year`, `budget`, `length`,
db_boo, genre dummies (action, adventure, animation, biography, comedy, crime, drama, horror, mystery), country dummies (china, usa, india, uk, europe, asia, other-america), and language dummies (english, indian, chinese, european, asian, american).

Note that the two models are quite similar. The only difference is that commondb4 reduces the number of dummy variables by grouping similar ones together.

3.2.2 Adding Two-Way Interaction Terms

In order to better understand how the same set of variables have different impact on Douban and IMDb ratings, it is necessary to add to the models the interaction terms between db_boo and other variables. In other words, it is necessary to include the multiples of db_boo with other variables in the regressions.

In R, interaction variables can be easily added to the MLR models on commondb3 and commondb4 described in 3.2.1. The interaction terms are named in the form of "db_boo:xxx". For example, "db_boo:length" is the multiple of db_boo and length. It shows the additional length coefficient when db_boo is equal to 1.

Although the models with interaction terms are powerful, they may still suffer from multicollinearity. If so, the estimates of coefficients will be less precise, and the models will be less stable, than under the case without multicollinearity. Next section will illustrate how to test for multicollinearity and fix this issue using feature selection.
3.2.3 Detecting and Fixing the Multicollinearity Issue

As explained before, the data set might suffer from multicollinearity because the country-dummies and the language dummies would be highly correlated. For example, a film produced in Italy will likely to have Italian as its primary speech language.

Therefore, it may be necessary to select only a subset of the variables, in order to get rid of the highly correlated variables. One possible way is to only use the country dummies (and exclude the language dummies) in the regressions. Another way is to only use the language dummies in the regressions. Although these two methods seem similar, they may actually produce different results.

One reason for this is the difference between country and language categories. For example, U.K., U.S. and Australia all speak English. However, U.K. and U.S. produce significant amount of films, so they are listed as different country dummy variables. Since there are only a couple of Australia films that are popular enough to be included in the data sets, Australia as a country is not listed in the country variables.

Another possible reason is that languages have influence on audiences’ perceptions and preferences of films, which is independent of the cultural influence that can be explained by production countries. Due to this potential influence, if the multicollinearity issue is not severe, one may still want to look at the "naive" full models that includes both country and language dummies, in order to see if such an influence exist.
Detecting Multicollinearity

Variance Inflation Factor (VIF) is used detect multicollinearity. VIF measures how much the variance of a coefficient is inflated due to multicollinearity in a model. In general, variables with high VIF should be excluded from the model.

By calculating VIF using the "naive" full models on commondb3 and commondb4, all of the genre, country, and language dummy variables are shown to be highly correlated to other variables in the models. Therefore, it makes sense to exclude all the country dummy variables (or exclude all the language dummy variables).

Note that for models with interaction terms, it is normal to have high multicollinearity, since the interaction terms are generated from existing variables.

Fixing Multicollinearity

In general, exclude all language dummies (or all country dummies) from the full model on commondb3 (or commondb4), and then further exclude the variables (often some genre dummies) with high VIF scores. Even after excluding all the country dummy variables from commondb3, the genre dummy variables still have very high VIF scores (see Figure 3.1). Remove the genre dummies with VIF>10 (genre_comedy, genre_crime, genre_drama), and check for multicollinearity again: now most of the terms have VIF scores less than 2 (see Figure 3.2). Thus the multicollinearity issue is fixed for the model on commondb3.
Then take a similar approach to exclude all the language dummies from \texttt{commondb3}. This time the variables with the highest VIF scores are \texttt{genre_action} and \texttt{genre_drama}. We remove these two variables from the regression.

Repeat similar steps on \texttt{commondb4}, and obtain two models: the first model excludes the high-VIF terms including all language dummies, \texttt{genre_drama}, and \texttt{genre_action}. The second model excludes the high-VIF terms including all country dummies, \texttt{language_english}, and \texttt{genre_action}.

Four models are derived using this method: a model that primarily excludes the language dummies from the full model on \texttt{commondb3}, a model that primarily excludes the country dummies from the full model on
\texttt{commondb3}, a model that primarily excludes the language dummies from the full model on \texttt{commondb4}, and a model that primarily excludes the country dummies from the full model on \texttt{commondb4}.

### 3.2.4 Models for Extreme Ratings Analysis

To study extreme rating behaviors (e.g. 1 out of 5 stars for Douban, and 1-2 out of 10 for IMDb ratings) and its influencing factors, data on the proportion of extreme low votes (instead of the average ratings) are used as dependent variables in the regressions.

Since models on \texttt{commondb4} usually perform better, three-way interaction models are only ran on \texttt{commondb4}. Different definitions for "extreme low ratings" are used: 1-2 out of 5 stars for Douban and 1-4 out of 10 for IMDb ratings, 1-3 out of 5 stars for Douban and 1-6 out of 10 for IMDb ratings, etc. Half of the models use country dummies, and the others use language dummies.

### 3.2.5 Three-Way Interaction Models

To further study the interactions between production countries (or, almost equivalently, languages) and genres on Douban and IMDb ratings, it may be helpful to run three-way interaction models. Specifically, the interaction terms of country/language dummies, genre dummies, and db boo are added to the above models. Note that lower-level terms must be included in the models as well.
As interaction terms are added to the models, many variables become multicollinear (and even perfectly collinear). Therefore, it is necessary to remove the variables with high VIF scores (a threshold of VIF=10 is chosen) from the models. Finally, two models are generated: one that uses country dummies, and the other uses language dummies.

The next part will exhibit the results of the regression models, analyze the potential reasons, and discuss the insights drawn from the results.
4. Results

4.1 A Note on the Models

The previous chapter derived many models from the two data sets commondb3 and commondb4. The preliminary models are the "naive" full models which just include all possible variables in the data sets. These models not only suffer from multicollinearity, but also lack explanatory power when it comes to how different audience groups would be influenced differently by individual factors, such as length of the film, and the language of the film, etc. The second type of models add interaction terms to the "naive" models, which has the ideal explanatory power. However, these models still suffer from multicollinearity. In fact, there is even more multicollinearity in such models since each interaction term is generated from the original data sets.

The first set of models used for analysis are the full models with two-way interaction terms. These models are used to test most hypotheses and provide an understanding of the different impacts of the factors across
Chinese and American audiences. The second set of models are the ones using the proportions of extreme low ratings as dependent variables. This set of models mainly tests the last hypothesis. The third set of models add three-way interaction terms to the first set, in order to further understand the complex interactions between genres and production countries/languages across Douban and IMDb.

The following sections will discuss the findings of testing the hypotheses of this study, as well as other interesting regression results.

4.2 Revisiting the Hypotheses

In Section 3.1, 9 hypotheses were made about the data sets:

**hypothesis I:** In both data sets, the variables explain the ratings well, and the number of votes should be significant in predicting the film ratings.

This hypothesis is validated. In two-way interaction models, R-squared is around 0.50. Given the random nature of the quality of films and their ratings, explaining about 50 percent of the total variation in ratings is satisfactory.

The number of votes is indeed significant in predicting the film ratings: `num_votes_imdb_us` is the number of raters for a film, which indicates the "popularity" of the film in terms of how many people have viewed and rated it. In all of the 4 models, this variable is statistically significant at a 0.001 significance level. Its coefficient is always greater than 0, which means
that the more "popular" a film is, the more likely it has a higher rating.

**hypothesis II:** Models on the two data sets will not produce contrasting results, although the performances of the models might be different.

This hypothesis is also validated. Although different models have slightly different R-squares, F-statistics and significance levels for variables when used on the two data sets, they imply similar conclusions in general.

**hypothesis III:** Budget will positively impact both Douban and IMDb ratings.

This hypothesis is shown to be wrong! Surprisingly, the regression results indicate that budget has no significant influence on either Douban or IMDb ratings. This suggests that for the relatively high quality films (this is the nature of the population of the data sets), higher budget films are no longer guaranteed to be preferred by audiences than lower budget ones. In fact, many lower budget films achieve high ratings on both data sets.

**hypothesis IV:** Length of the films will not impact the ratings.

This hypothesis is validated. In all models with two-way interaction terms, length is not shown to be a significant variable.
hypothesis V: Douban ratings are higher than IMDb ratings in general.

This hypothesis is validated. In all the models used in this study, including those without interaction terms (see appendix), the coefficient of the dummy that indicates Douban ratings, db_boo, is always significantly positive with a 0.001 significance level. Therefore, for an average film in the data sets, its Douban rating is expected to be higher than its IMDb rating. This is just a numerical difference, and does not indicate the different level of preferences by American and Chinese audiences.

hypothesis VI: Chinese audiences and American audiences prefer different genres of films.

This hypothesis is validated. For a certain genre, if the interaction term of its genre dummy with db_boo is significant and has a positive coefficient, then such a genre is considered to be liked more by Chinese (than by American) audience. If it the coefficient is significant and negative, then such a genre is considered to be liked less by Chinese (than by American) audience. In particular, Chinese audiences like biography, animation, comedy films better, comparing to their American counterparts. These three film genres have significantly positive coefficients in multiple models. On the other hand, American audiences like action, horror films better than Chinese audiences.

Chinese audiences might also like adventure films and crime films better than American audiences. However, we are not very confident in the signs of these two variables, as the coefficients are only significant in one of the models.
hypothesis VII: For foreign films, Chinese audiences and American audiences prefer different languages/production countries.

There is not enough evidence to determine whether this hypothesis is correct. Overall, the two-way interaction models do not show strong evidence that the influence of country of production (or somewhat equivalently, the language of the films) is significant different across Chinese and American audiences. One possible reason is that Chinese and American audiences’ preferences are impacted by country of production/languages in the same way; other reasons include regression design, data set coverage, etc.

hypothesis VIII: Audiences in both countries prefer their domestic productions.

There is not enough evidence to determine whether an average audience would prefer domestic productions, for similar reasons as above. It is possible that Chinese and American audiences do not have strong preferences for their domestic productions. However, a difference in preferences is seen on extreme low ratings given by American audiences, as discussed below:

hypothesis IX: Audiences in both countries are more likely to give extreme low ratings to foreign films.

This hypothesis is partially correct. Specifically, American audiences are more likely to give extreme low ratings to foreign films (those produced outside of U.S.) than to domestic films. This phenomenon is not seen on Chinese audiences to films produced outside of China.
Regressions are ran separately using proportions of Douban and IMDb extreme low ratings as dependent variables. Audiences are considered to be more likely to give extreme low ratings to a certain country if this country dummy is significant and has a positive coefficient. Similar for language dummies. Results show that in the Douban regressions, country/language dummies are mostly not significant; in the IMDb regressions, most country/language dummies are significant, and the domestic dummy is often lower in value comparing to other dummies. This means that if a film is produced in U.S. instead of other countries, it would likely have a lower proportion of extreme low ratings.

This is an interesting result. One possible explanation is that Chinese audiences are more open to different cultures and therefore their films, while there is a part of American audiences who absolutely hate exotic films. Another explanation, as some previous study [Feng, 2018, Han and Ji-Fan, 2017] argues, is that Chinese audiences are more moderate and reserved in attitudes, so they would hesitate to give ratings that are more extreme.

Section 4.3.3 will expand on the two explanations and use further regressions results to discuss the extreme rating behaviors of Chinese and American audiences:
4.3 Other Findings

4.3.1 Audiences Hate Early Films

As expected, Douban and IMDb ratings are influenced to different extents by the year of release of a film. In general, earlier films are rated lower by both websites. Specifically, Douban ratings penalize earlier films with even lower ratings, with a significance level of 0.001. One possible reason is that Chinese audiences hate earlier films more than their American counterparts do.

Another reason may be that the different establishment dates of Douban and IMDb result in different penalties of earlier films. For example, assume that modern audiences would give a film lower ratings comparing to the audiences of the time. Then the ratings for earlier films on Douban will easily be even lower than those on IMDb, because Douban was established later than IMDb, and Douban ratings are on average given by more modern audiences.

4.3.2 Takeaways From the Two-way and Three-way Interaction Models

The two-way models show valuable results for the genre dummies, but not for the country/language dummies. This might due to some genre-country/language interactions that the two-way models fail to capture. Although some of these interactions are captured in the extreme rating models, this is one of the motivations for the three-way models.
However, the three-way models are not satisfying. After creating the interaction terms, it is necessary to remove some three-degree (as well as some lower-level) terms to take care of the multicollinearity issue. Due to the lacking terms, and given that only a few of the variables are significant, the models are difficult to interpret.

For example, in one of the three-way models, the interaction term of db_boo, genre_crime, and language_european is significant at 0.01 and has a negative coefficient. Since the interaction coefficient should not be analyzed by itself, it is necessary to also look at the lower-level interactions of db_boo and language_european, as well as of genre_crime and language_european. However, these two interaction terms are not significant in the regression. No confident conclusion can be drawn here.

4.3.3 Takeaways From the Extreme Rating Models

Regressions are also ran using the proportions of extreme high Douban/IMDb ratings as dependent variables. Douban and IMDb audiences exhibit different extreme rating behaviors:

There is not enough evidence to show that Douban audiences give extreme low ratings (1 or 2 stars) to films under the influence of the production countries. However, they tend to give extreme high ratings (5 stars) to domestic films. This might be a counterexample for the argument that Chinese audiences are more reserved to show extreme views.

The situation is different for IMDb audiences: they are more likely to give extreme low ratings (1-4) to foreign films, as well as more likely to give extreme high ratings (9-10) to domestic films.
4.4 Conclusion

This study shows that there is indeed a Country-of-Origin (COO) effect on the ratings of films, in a sense that audiences originated from different countries would be impacted differently by a fixed set of factors when they rate a film. Factors that influence film ratings, especially genres, behave differently for different audience groups. Factors such as production countries show more differences in the extreme rating models. The two-way interaction models and the extreme rating models are useful in this study; however, the three-way interaction models are difficult to interpret due to the multicollinearity and the lack of significant variables.

Chinese audiences like biography, animation, comedy films better than American audiences, while American audiences like action, horror films better than Chinese audiences. There is no strong evidence for country-of-production influencing average film ratings in different ways across audience groups; however, such evidence exist for extreme rating behaviors. American audiences are more likely to give extreme low ratings to foreign films, and more likely to give extreme high ratings to domestic films. Chinese audiences also give domestic films more extreme high ratings, but they are not so sensitive to the country of production when they give extreme low ratings. Surprisingly, there is no evidence that budget of a film influences the film’s rating. No evidence shows that length is a significant influencing factor, either.
5. Discussion

This part will discuss the contributions and limitations of this study, and suggest future steps to take in order to deeper understand this issue.

5.1 Contributions

While previous research mostly focus on predicting box offices or use machine learning models to design recommendation systems, this study contributes to the current literature by studying the dynamics of film ratings and its influencers across audience groups of different countries.

This study proposes three sets of models to analyze film ratings: two-way interaction models, three-way interaction models, and extreme rating models. Many models generate valuable insights (such as the different preferences of genres across Chinese and American audiences, and the lack of significance of a film’s budget in predicting its ratings).
This paper proposes a feasible framework for film analysis using big data, comparing film ratings across audience groups. Specifically, it shows one possible way to combine data obtained from different websites. It also shows how to create dummy variables to study categorical variables such as genres and production countries, as well as how to detect and fix multicollinearity issues.

5.2 Limitations and Next Steps

As discussed in Section 2.1.2, there are some limitations to the data sets used in this study. However, most are not severe and can be fixed with careful treatments and interpretations of data.

A potential way to generate better results is to include more data and more variables in the regressions. Including more characteristics of films as factors would generate more insights and possibly more significant variables, and more data is needed if the number of variables is to be increased.

Another potential next step to take is to include more film rating websites (e.g. Rotten Tomatoes) in the analysis. It would also be interesting to do similar analysis on three or more audience groups (e.g. Chinese, Japanese, Korean).
Bibliography


A. Appendix
Data Analysis Using R

Xiaotong Chen

05/06/2019

For the sake of this report, most R codes for pre-processing data sets are hidden. Only the most important models are included. The first section reports the two-way interaction models on commondb3 and commondb4. Four models are reported: two using the language dummies, and two using the country dummies. The second section reports the extreme rating models. Multiple models are reported, using different dependent variables consisting of proportions of extreme (low/high) ratings. The third section reports the three-way interaction models. Two models are reported: one using the language dummies, and one using the country dummies.

1. Data Description

1.1 From commondb3 to commondb4

Below are the list of variables included in commondb3 and commondb4:

### Variables in commondb3

```r
colnames(commondb3)
```

```r
## [1] "id_imdb"            "rating"             "num_votes_imdb_us"
## [2] "num_votes_db"       "year"               "budget"
## [3] "us_box_office"      "length"             "genre_action"
## [4] "genre_adventure"    "genre_animation"    "genre_biography"
## [5] "genre_comedy"       "genre_crime"        "genre_drama"
## [6] "genre_horror"       "genre_mystery"      "country_china"
## [7] "country_france"     "country_germany"    "country_india"
## [8] "country_italy"      "country_japan"      "country_russia"
## [9] "country_southkorea" "country_spain"       "country_uk"
## [10] "country_usa"        "language_chinese"  "language_english"
## [11] "language_french"    "language_german"    "language_hindi"
## [12] "language_italian"   "language_japanese"   "language_korean"
## [13] "language_spanish"   "genre_imdb"         "production.country"
## [14] "language"           "db_boo"             "rating_imdb_10"
## [15] "rating_imdb_9"      "rating_imdb_8"      "rating_imdb_7"
## [16] "rating_imdb_6"      "rating_imdb_5"      "rating_imdb_4"
## [17] "rating_imdb_3"      "rating_imdb_2"      "rating_imdb_1"
## [18] "star_db_5"          "star_db_4"          "star_db_3"
## [19] "star_db_2"          "star_db_1"
```

### Variables in commondb4

```r
colnames(commondb4)
```

```r
## [1] "id_imdb"            "rating"             "num_votes_imdb_us"
## [2] "num_votes_db"       "year"               "budget"
## [3] "us_box_office"      "length"             "genre_action"
## [4] "genre_adventure"    "genre_animation"    "genre_biography"
## [5] "genre_comedy"       "genre_crime"        "genre_drama"
## [6] "genre_horror"       "genre_mystery"      "country_china"
## [7] "country_france"     "country_germany"    "country_india"
## [8] "country_italy"      "country_japan"      "country_russia"
## [9] "country_southkorea" "country_spain"       "country_uk"
## [10] "country_usa"        "language_chinese"  "language_english"
## [11] "language_french"    "language_german"    "language_hindi"
## [12] "language_italian"   "language_japanese"   "language_korean"
## [13] "language_spanish"   "genre_imdb"         "production.country"
## [14] "language"           "db_boo"             "rating_imdb_10"
## [15] "rating_imdb_9"      "rating_imdb_8"      "rating_imdb_7"
## [16] "rating_imdb_6"      "rating_imdb_5"      "rating_imdb_4"
## [17] "rating_imdb_3"      "rating_imdb_2"      "rating_imdb_1"
## [18] "star_db_5"          "star_db_4"          "star_db_3"
## [19] "star_db_2"          "star_db_1"
```
Note that commondb4 is created from commondb3 by reducing the number of country dummies down to 7 dummies: China USA India UK Europe: Russia, France, Italy, Germany, Ireland, Denmark, Sweden, Poland, Belgium, Austria, Netherlands, Serbian Asia: Japan, South Korea, Indonesia, Dari, Romany America: Mexico, Argentina, Brazil, Canada, Spain Others: South Africa, Iran, Turkey, Israel, Australia, New Zealand, Maya

Also reduce the number of language dummies down to 6 dummies: English Indian: Hindi, Telugu Chinese: Cantonese, Mandarin European: French, Italian, German, Portuguese, Danish, Swedish, Dutch, Serbian, Romanian, Russian, Polish Asian: Korean, Japanese, Indonesian American: Spanish, Persian Others: Turkish, Bosnian, Hebrew, Mayan, no dialogue

1.2 Genre-level and country-level Visualizations of commondb4

Below are some visualizations to help understand the genre dummies and production country dummies in commondb4 data set. Note that IMDb data are in red, and Douban data are in green color.

```r
# commondb4 data structure
## [1] "id_imdb" "rating" "num_votes_imdb_us"
## [4] "num_votes_db" "year" "budget"
## [7] "us_box_office" "length" "genre_action"
## [10] "genre_adventure" "genre_animation" "genre_biography"
## [13] "genre_comedy" "genre_crime" "genre_drama"
## [16] "genre_horror" "genre_mystery" "genre_imdb"
## [19] "production.country" "language" "db_boo"
## [22] "rating_imdb_10" "rating_imdb_9" "rating_imdb_8"
## [25] "rating_imdb_7" "rating_imdb_6" "rating_imdb_5"
## [28] "rating_imdb_4" "rating_imdb_3" "rating_imdb_2"
## [31] "rating_imdb_1" "star_db_5" "star_db_4"
## [34] "star_db_3" "star_db_2" "star_db_1"
## [37] "country_china" "country_usa" "country_india"
## [40] "country_uk" "country_europe" "country_india"
## [43] "country_america" "language_english" "language_indian"
## [46] "language_chinese" "language_european" "language_asian"
## [49] "language_american"
```
# red: imdb; green: db
# genre-level visualization

ggplot(commondb4, aes(x = genre_imdb, y = rating, fill = as.factor(db_boo))) + geom_boxplot() + xlab("Genre") + ylab("Film Rating (0-10)") + scale_fill_discrete(name = "Website", labels = c("IMDb", "Douban"))+ theme(axis.text.x = element_text(angle = 90, hjust = 1))
2. Two-way interaction models

The R code below detects multicollinearity from the “naive” full models with all dummy variables, and justifies that some variables should be removed from the regressions:

```r
# for commondb3
vif(naive_lm_3)
```
## num_votes_imdb_us        num_votes_db              year
##           1.841813           1.534672           1.363443
##          budget             length             db_boo
##           1.046306           1.332667           1.000000
##    genre_action        genre_adventure        genre_animation
##          12.861402         7.085063            9.491202
##    genre_biography        genre_comedy        genre_crime
##          8.499156         10.188447           10.119858
##    genre_drama        genre_horror        genre_mystery
##          18.823040          2.069650           2.104310
## country_china        country_france        country_germany
##          14.461384          3.003724           2.789494
## country_india        country_italy        country_japan
##           5.626317          4.564188          22.215657
## country_russia        country_southkorea        country_spain
##           1.261457         14.364213           2.187062
## country_usa        country_india        country_uk
##          2.469499         16.966499          11.852442
## language_english        language_french        language_german
##          8.344399          3.638279           3.117709
## language_hindi        language_italian        language_japanese
##          5.469898          4.745220          22.998047
## language_korean        language_spanish
##          14.641341           2.546291

# for commondb4
vif(naive_lm_4)

## num_votes_imdb_us        num_votes_db              year            budget
##           1.833407           1.494026          1.317101          1.042740
##            length            db_boo        genre_action        genre_adventure
##           1.311237           1.000000          12.128662          6.647351
## genre_animation        genre_biography        genre_comedy        genre_crime
##          8.856367         7.972678           9.678115          9.535700
## genre_drama        genre_horror        genre_mystery        country_china
##          17.685918          1.960160           2.035460          14.896317
## country_usa        country_india        country_uk        country_europe
##          11.852442         16.966499          4.054063          8.220541
## country_asia        country_america        language_english        language_indian
##          11.706212          3.018512          21.946091          4.157388
## language_chinese        language_european        language_asian        language_american
##          15.822744         14.003176          15.686322          4.157388

In the below four models, different choices of variables are excluded to fix the multicollinearity issue:

### 2.1 Exclude language dummies to fix multicollinearity

On commondb3:
## Exclude language dummies, and genre_comedy, genre_crime, genre_drama:

```r
nolanguage_interaction_lm_3 <- lm(rating ~ num_votes_imdb_us + num_votes_db + db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_action + db_boo*genre_adventure + db_boo*genre_animation + db_boo*genre_biography + db_boo*genre_horror + db_boo*genre_mystery + db_boo*country_china + db_boo*country_france + db_boo*country_germany + db_boo*country_india + db_boo*country_italy + db_boo*country_japan + db_boo*country_russia + db_boo*country_southkorea + db_boo*country_spain + db_boo*country_uk + db_boo*country_usa, 
                                    data=commondb3)
```

```r
summary(nolanguage_interaction_lm_3)
```
## Call:
```r
lm(formula = rating ~ num_votes_imdb_us + num_votes_db + db_boo * year + db_boo * length + db_boo * budget + db_boo * genre_action + db_boo * genre_adventure + db_boo * genre_animation + db_boo * genre_biography + db_boo * genre_horror + db_boo * genre_mystery + db_boo * country_china + db_boo * country_france + db_boo * country_germany + db_boo * country_india + db_boo * country_italy + db_boo * country_japan + db_boo * country_russia + db_boo * country_southkorea + db_boo * country_spain + db_boo * country_uk + db_boo * country_usa, data = commondb3)
```
## Residuals:
<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.70497</td>
<td>-0.19375</td>
<td>0.01722</td>
<td>0.21338</td>
<td>1.23047</td>
<td></td>
</tr>
</tbody>
</table>
## Coefficients:
|                                | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------------|----------|------------|---------|---------|
| (Intercept)                    | 2.328e+01 | 1.294e+00  | 17.988  | < 2e-16 *** |
| num_votes_imdb_us              | 2.273e-06 | 2.351e-07  | 9.672   | < 2e-16 *** |
| num_votes_db                   | 7.954e-07 | 5.819e-08  | 13.669  | < 2e-16 *** |
| db_boo                         | 4.516e-01 | 1.794e+00  | 0.252   | 0.801281  |
| year                           | -7.771e-03 | 6.493e-04  | -11.968 | < 2e-16 *** |
| length                         | 6.478e-04 | 5.313e-04  | 1.219   | 0.222901  |
| budget                         | -1.335e-04 | 1.640e-04  | -0.814  | 0.415786  |
| genre_action                   | 2.739e-02 | 3.677e-02  | 0.745   | 0.456460  |
| genre_adventure                | 3.480e-02 | 4.876e-02  | 0.714   | 0.475467  |
| genre_animation                | 2.487e-02 | 4.967e-02  | 0.501   | 0.616670  |
| genre_biography                | 5.453e-02 | 4.408e-02  | 1.237   | 0.216234  |
| genre_horror                   | 1.261e-01 | 1.148e-01  | 1.099   | 0.272148  |
| genre_mystery                  | -6.529e-02 | 1.086e-01  | -0.601  | 0.547834  |
| country_china                  | -2.042e-01 | 1.019e-01  | -2.004  | 0.045267  * |
| country_france                 | -1.219e-01 | 6.899e-02  | -1.767  | 0.077387 . |
| country_germany                | -6.114e-02 | 8.489e-02  | -0.720  | 0.471535  |
| country_india                  | -1.720e-01 | 9.530e-02  | -1.805  | 0.071298 . |
| country_italy                  | -8.978e-02 | 8.966e-02  | -1.001  | 0.316844  |
| country_japan                  | -1.657e-02 | 7.416e-02  | -0.223  | 0.823197  |
| country_russia                 | -2.754e-01 | 1.155e-01  | -2.386  | 0.017173 * |
| country_southkorea             | 8.219e-02  | 1.006e-01  | 0.817   | 0.414166  |
| country_spain                  | -9.068e-02 | 9.709e-02  | -0.934  | 0.350466  |
| country_uk                     | -1.250e-01 | 6.279e-02  | -1.992  | 0.046592  * |
| country_usa                    | -9.966e-02 | 4.600e-02  | -2.167  | 0.030428  * |
| db_boo:year                    | -4.114e-05 | 9.003e-04  | -0.406  | 0.693562  |
| db_boo:length                  | 9.368e-04  | 7.421e-04  | 1.262   | 0.207048  |
| db_boo:budget                  | 1.001e-04  | 2.317e-04  | 0.432   | 0.665797  |
| db_boo:genre_action            | -3.142e-01 | 5.135e-02  | -6.118  | 1.21e-09 *** |
| db_boo:genre_adventure         | -5.492e-02 | 6.879e-02  | -0.798  | 0.424820  |
| db_boo:genre_animation         | 1.318e-01  | 7.007e-02  | 1.881   | 0.060187  |
| db_boo:genre_biography         | 1.093e-01  | 6.244e-02  | 1.756   | 0.079302  |
| db_boo:genre_horror            | -7.752e-01 | 1.620e-01  | -4.784  | 1.89e-06 *** |
| db_boo:genre_mystery           | -1.622e-01 | 1.535e-01  | -1.057  | 0.290898  |
| db_boo:country_china           | -1.061e-01 | 1.432e-01  | -0.741  | 0.458673  |
| db_boo:country_france          | -3.374e-03 | 9.733e-02  | -0.035  | 0.972354  |
## db_boo:country_germany    -1.088e-01  1.200e-01  -0.906 0.364865
## db_boo:country_india       1.523e-01  1.343e-01   1.135 0.256762
## db_boo:country_italy       2.293e-01  1.265e-01   1.813 0.070100 .
## db_boo:country_japan       1.013e-01  1.044e-01   0.970 0.332287
## db_boo:country_russia      2.922e-01  1.632e-01   1.790 0.073650 .
## db_boo:country_southkorea  -2.188e-01  1.420e-01  -1.541 0.123545
## db_boo:country_spain      -7.638e-02  1.373e-01  -0.556 0.578074
## db_boo:country_uk         -1.361e-01  8.830e-02  -1.541 0.123509
## db_boo:country_usa        -2.392e-01  6.410e-02  -3.732 0.000197 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3367 on 1502 degrees of freedom
## Multiple R-squared:  0.4685, Adjusted R-squared:  0.4532
## F-statistic: 30.79 on 43 and 1502 DF,  p-value: < 2.2e-16

On commondb4:

```r
## Exclude language dummies, and genre_drama, genre_action
nolanguage_interaction_lm_4 <- lm(rating~db_boo*num_votes_imdb_us + db_boo*num_votes_db + db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure + db_boo*genre_animation + db_boo*genre_biography + db_boo*genre_comedy + db_boo*genre_crime + db_boo*genre_horror + db_boo*genre_mystery + db_boo*country_china + db_boo*country_usa + db_boo*country_india + db_boo*country_spain + db_boo*country_europe + db_boo*country_asia + db_boo*country_america, data=commondb4)
summary(nolanguage_interaction_lm_4)
```

48
## Call:
library(tidyverse)
## lm(formula = rating ~ db_boo * num_votes_imdb_us + db_boo * num_votes_db +
##     db_boo * year + db_boo * length + db_boo * budget + db_boo *
##     genre_adventure + db_boo * genre_animation + db_boo * genre_biography +
##     db_boo * genre_comedy + db_boo * genre_crime + db_boo * genre_horror +
##     db_boo * genre_mystery + db_boo * country_china + db_boo *
##     country_usa + db_boo * country_india + db_boo * country_uk +
##     db_boo * country_europe + db_boo * country_asia + db_boo *
##     country_america, data = commondb4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.64744 -0.18690  0.01135  0.19258  1.06980
##
## Coefficients:
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)              2.049e+01  1.235e+00  16.598  < 2e-16 ***
## db_boo                   7.148e+00  1.746e+00   4.094 4.47e-05 ***
## num_votes_imdb_us        4.411e-06  3.134e-07  14.073  < 2e-16 ***
## num_votes_db             7.143e-08  7.848e-08   0.910 0.362892
## year                     -6.378e-03  6.169e-04  -10.340  < 2e-16 ***
## length                   3.841e-04  5.204e-04   0.738 0.460639
## budget                   -1.764e-04  1.583e-04  -1.114 0.265533
## genre_adventure          7.179e-03  4.701e-02   0.153 0.878632
## genre_animation          5.457e-02  4.580e-02   1.191 0.233645
## genre_biography          3.747e-02  4.267e-02   0.878 0.379940
## genre_comedy             7.179e-03  4.701e-02   0.153 0.878632
## genre_crime              -3.037e-02  3.888e-02  -0.781 0.434830
## genre_horror             4.143e-03  1.103e-01   0.376 0.707297
## genre_mystery            -8.100e-02  1.047e-01  -0.773 0.439358
## country_china            -1.369e-02  1.127e-01  -0.121 0.903322
## country_usa              -7.338e-02  7.182e-02  -1.022 0.307098
## country_india            -3.340e-02  1.096e-01  -0.305 0.760700
## country_uk               -7.101e-02  8.250e-02  -0.861 0.389523
## country_europe           -2.950e-03  7.682e-02  -0.038 0.969371
## country_asia             -6.746e-02  1.071e-01  -0.630 0.529447
## country_america          -4.722e-06  4.433e-07   1.063 0.286589
## db_boo:num_votes_imdb_us -4.722e-06  4.433e-07  -10.653  < 2e-16 ***
## db_boo:num_votes_db       1.482e-06  1.110e-07  13.356  < 2e-16 ***
## db_boo:year              -3.479e-03  8.724e-04  -3.988 6.98e-05 ***
## db_boo:length             1.919e-03  7.360e-04   2.607 0.009219 **
## db_boo:budget             2.508e-04  2.239e-04   1.120 0.262865
## db_boo:genre_adventure    4.688e-02  6.648e-02   0.705 0.480789
## db_boo:genre_animation    1.685e-01  6.477e-02   2.602 0.009361 **
## db_boo:genre_biography    1.856e-01  6.034e-02   3.076 0.002133 **
## db_boo:genre_comedy       1.099e-01  5.499e-02   1.998 0.045900 *
## db_boo:genre_crime        -4.431e-02  5.369e-02  -0.825 0.409395
## db_boo:genre_horror      -5.552e-01  1.560e-01  -3.559 0.000384 ***
## db_boo:genre_mystery     -9.316e-02  1.481e-01  -0.629 0.529447
## db_boo:country_china     -5.355e-01  1.593e-01  -3.361 0.000796 ***
## db_boo:country_usa       -2.150e-01  1.016e-01  -2.117 0.034446 *
## db_boo:country_india     -7.154e-02  1.551e-01  -0.461 0.644620
## db_boo:country_europe    -1.021e-02  1.021e-01  -0.099 0.921741
## db_boo:country_asia      -1.351e-01  1.270e-01  -1.069 0.285057
## db_boo:country_america   -4.688e-02  8.250e-02  -0.571 0.568473
## db_boo:country_america   -1.021e-02  1.021e-01  -0.099 0.921741
## db_boo:country_europe    -1.351e-01  1.270e-01  -1.069 0.285057
## db_boo:country_asia      -1.636e-01  1.523e-01  -1.073 0.283088
## db_boo:country_america   -2.221e-02  9.098e-02  -0.244 0.805192

## db_boo:country_uk        -1.683e-01  1.167e-01  -1.442 0.149429
## db_boo:country_europe    -3.128e-02  1.086e-01  -0.288 0.773421
## db_boo:country_asia      -1.976e-01  1.178e-01  -1.677 0.093796 .
## db_boo:country_america   -9.066e-02  1.336e-01  -0.678 0.497644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.3233 on 1506 degrees of freedom
## Multiple R-squared:  0.5085, Adjusted R-squared:  0.4958
## F-statistic: 39.95 on 39 and 1506 DF,  p-value: < 2.2e-16

2.2 Exclude country dummies to fix multicollinearity

On commondb3:

```r
## The full model without countries, excluding genre_comedy, genre_crime, genre_drama:
nocountry_interaction_lm_3 <- lm(rating~db_boo*num_votes_imdb_us + db_boo*num_votes_db +
db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure + db_boo*genre_animation +
db_boo*genre_biography + db_boo*genre_comedy + db_boo*genre_crime + db_boo*genre_horror +
db_boo*genre_mystery + db_boo*country_china + db_boo*country_france + db_boo*country_germany +
db_boo*language_chinese + db_boo*language_english + db_boo*language_french + db_boo*language_german +
db_boo*language_hindi + db_boo*language_italian + db_boo*language_japanese + db_boo*language_korean +
db_boo*language_spanish,
data=commondb3)

summary(nocountry_interaction_lm_3)
```
## Call:
```r
lm(formula = rating ~ db_boo * num_votes_imdb_us + db_boo * num_votes_db +
    db_boo * year + db_boo * length + db_boo * budget + db_boo *
    genre_adventure + db_boo * genre_animation + db_boo * genre_biography +
    db_boo * genre_comedy + db_boo * genre_crime + db_boo * genre_horror +
    db_boo * genre_mystery + db_boo * country_china + db_boo *
    country_france + db_boo * country_germany + db_boo * language_chinese +
    db_boo * language_english + db_boo * language_french + db_boo *
    language_german + db_boo * language_hindi + db_boo * language_italian +
    db_boo * language_japanese + db_boo * language_korean + db_boo *
    language_spanish, data = commondb3)
```
## Residuals:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-1.69404</td>
<td>-0.17963</td>
<td>0.01214</td>
<td>0.19454</td>
</tr>
<tr>
<td>1Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td></td>
<td></td>
<td>1.07917</td>
</tr>
</tbody>
</table>

## Coefficients:

|                                      | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------------------|----------|------------|---------|---------|
| (Intercept)                          | 2.000e+01| 1.214e+00  | 16.476  | < 2e-16 *** |
| db_boo                               | 5.892e+00| 1.716e+00  | 3.433   | 0.000613 *** |
| num_votes_imdb_us                    | 4.393e-06| 3.100e-07  | 14.171  | < 2e-16 *** |
| num_votes_db                         | 7.957e-08| 7.864e-08  | 1.012   | 0.311760 |
| year                                 | -6.126e-03| 6.107e-04  | -10.030 | < 2e-16 *** |
| length                               | 4.454e-04| 5.225e-04  | 0.852   | 0.394190 |
| budget                               | -1.578e-03| 1.574e-04  | -1.002  | 0.316290 |
| genre_adventure                      | 1.324e-02| 4.680e-02  | 0.283   | 0.777250 |
| genre_animation                      | 5.324e-02| 4.800e-02  | 1.109   | 0.267528 |
| genre_biography                      | 4.362e-02| 4.300e-02  | 1.014   | 0.310581 |
| genre_comedy                         | -2.393e-02| 3.928e-02  | -0.609  | 0.542443 |
| genre_crime                          | -1.289e-02| 3.805e-02  | -0.339  | 0.734927 |
| genre_horror                         | 8.217e-02| 1.100e-01  | 0.747   | 0.455067 |
| genre_mystery                        | -6.937e-02| 1.045e-01  | -0.664  | 0.507055 |
| country_china                        | -1.323e-01| 3.263e-01  | -0.405  | 0.685259 |
| country_france                       | -6.119e-02| 8.461e-02  | -0.723  | 0.469645 |
| country_germany                      | 3.064e-02| 1.120e-01  | 0.274   | 0.784490 |
| language_chinese                     | 1.158e-01| 3.421e-01  | 0.339   | 0.734955 |
| language_english                     | -8.680e-02| 5.358e-02  | -1.620  | 0.105438 |
| language_french                      | 6.698e-02| 1.005e-01  | 0.666   | 0.505398 |
| language_german                      | -6.090e-02| 1.303e-01  | -0.467  | 0.640382 |
| language_hindi                       | -9.934e-02| 1.063e-01  | -0.935  | 0.350189 |
| language_italian                     | 2.558e-02| 9.615e-02  | 0.266   | 0.790271 |
| language_japanese                    | 1.089e-01| 7.675e-02  | 1.419   | 0.156055 |
| language_korean                      | 1.713e-01| 1.026e-01  | 1.670   | 0.095166 |
| language_spanish                     | 3.583e-02| 9.537e-02  | 0.376   | 0.707226 |
| db_boo:num_votes_imdb_us             | -4.719e-06| 4.384e-07  | -10.765 | < 2e-16 *** |
| db_boo:num_votes_db                  | 1.482e-06| 1.112e-07  | 13.323  | < 2e-16 *** |
| db_boo:year                          | -2.807e-03| 8.637e-04  | -3.250  | 0.001180 ** |
| db_boo:length                        | 1.474e-03| 7.399e-04  | 1.994   | 0.046290 * |
| db_boo:budget                        | 3.131e-04| 2.226e-04  | 1.407   | 0.159746 |
| db_boo:genre_adventure               | 6.918e-02| 6.619e-02  | 1.045   | 0.296103 |
| db_boo:genre_animation               | 1.337e-01| 6.789e-02  | 1.969   | 0.049157 * |
| db_boo:genre_biography               | 1.890e-01| 6.082e-02  | 3.108   | 0.001919 ** |
## Data Analysis Using R

On commondb4:

```R
# Exclude country dummies, language_english, and genre_action
nocountry_interaction_lm_4 <- lm(rating ~ db_boo*num_votes_imdb_us + db_boo*num_votes_db +
                                  db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure + db_boo*genre_animation + db_boo*genre_biography + db_boo*genre_comedy + db_boo*genre_crime + db_boo*genre_drama + db_boo*genre_horror + db_boo*genre_mystery + db_boo*language_indian + db_boo*language_chinese + db_boo*language_european + db_boo*language_asian + db_boo*language_american,
                                  data=commondb4)

summary(nocountry_interaction_lm_4)
```

---

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.3226 on 1496 degrees of freedom
## Multiple R-squared:  0.5139, Adjusted R-squared:  0.498
## F-statistic: 32.27 on 49 and 1496 DF,  p-value: < 2.2e-16
```

```R
# Exclude country dummies, language_english, and genre_action
nocountry_interaction_lm_4 <- lm(rating ~ db_boo*num_votes_imdb_us + db_boo*num_votes_db +
                                  db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure + db_boo*genre_animation + db_boo*genre_biography + db_boo*genre_comedy + db_boo*genre_crime + db_boo*genre_drama + db_boo*genre_horror + db_boo*genre_mystery + db_boo*language_indian + db_boo*language_chinese + db_boo*language_european + db_boo*language_asian + db_boo*language_american,
                                  data=commondb4)

summary(nocountry_interaction_lm_4)
```
## Call:
```
lm(formula = rating ~ db_boo * num_votes_imdb_us + db_boo * num_votes_db +
    db_boo * year + db_boo * length + db_boo * budget + db_boo *
    genre_adventure + db_boo * genre_animation + db_boo * genre_biography +
    db_boo * genre_comedy + db_boo * genre_crime + db_boo * genre_drama +
    db_boo * genre_horror + db_boo * genre_mystery + db_boo *
    language_indian + db_boo * language_chinese + db_boo * language_european +
    db_boo * language_asian + db_boo * language_american, data = commondb4)
```
## Residuals:
```
       Min        1Q   Median        3Q       Max
-1.80430 -0.17941  0.01213  0.19688  1.11840
```
## Coefficients:
```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.993e+01  1.182e+00   16.857  < 2e-16 ***
  db_boo                     5.367e+00  1.672e+00    3.210  0.00136 **
  num_votes_imdb_us          4.463e-06  3.118e-07   14.315  < 2e-16 ***
  num_votes_db               6.284e-08  7.720e-08    0.814  0.41575
  year                      -6.141e-03  5.943e-04   -10.334  < 2e-16 ***
  length                     3.471e-04  5.136e-04    0.676  0.49922
  budget                     -1.533e-04  1.549e-04   -0.989  0.32264
  genre_adventure            3.222e-02  5.134e-02    0.628  0.53035
  genre_animation            6.408e-02  5.024e-02    1.275  0.20236
  genre_biography            6.104e-02  4.865e-02    1.255  0.20982
  genre_comedy               3.222e-02  5.134e-02    0.628  0.53035
  genre_crime                6.720e-03  4.400e-02    0.153  0.87863
  genre_drama                3.143e-02  3.691e-02    0.851  0.39463
  genre_horror               9.671e-02  1.106e-01    0.875  0.38192
  genre_mystery              -5.516e-02  8.836e-02   -0.522  0.60251
  language_indian            4.603e-02  8.836e-02    0.521  0.60251
  language_chinese           8.351e-02  9.309e-02    0.897  0.39683
  language_european          7.155e-02  3.769e-02    1.898  0.05783 .
  language_asian             2.203e-01  1.250e-01    1.756  0.07940 .
  db_boo:year                6.616e-03  4.409e-07   14.315  < 2e-16 ***
  db_boo:length              -4.386e-06  4.409e-07   -0.989  0.32264
  db_boo:budget              -2.760e-03  8.405e-04   -3.284  0.00105 **
  db_boo:genre_adventure     2.256e-01  7.261e-02    3.106  0.00193 **
  db_boo:genre_animation     3.309e-01  7.105e-02    4.657  3.49e-06 ***
  db_boo:genre_biography     3.570e-01  6.881e-02    5.188  2.41e-07 ***
  db_boo:genre_comedy        2.725e-01  6.422e-02    4.243  2.34e-05 ***
  db_boo:genre_crime         1.158e-01  5.219e-02    2.228  0.02689 .
  db_boo:genre_drama         2.665e-01  5.219e-02    5.105  3.7e-07 ***
  db_boo:genre_mystery       -3.284e-04  1.564e-01   -2.116  0.03493 *
  db_boo:language_indian     9.022e-02  1.493e-01    6.064  0.54586
  db_boo:language_chinese    -2.871e-01  1.316e-01   -2.181  0.02933 *
  db_boo:language_european   1.616e-01  5.330e-02    3.031  0.00248 **
  db_boo:language_asian      7.143e-03  7.103e-02    0.101  0.91991
```
3. Extreme Rating Analysis

Below are the models ran using the proportions of extreme votes as dependent variables. Each regression makes use either Douban or IMDb observations. From Section 2 commondb4 seems to perform better than commondb3. So this section only use commondb4.

Also, this section does not include the interaction terms, since the interaction terms are about the boolean that indicates Douban population. Note that db boo needs to be removed from the regressions.

```r
# define db4, imdb4 (db/imdb datasets that are subsets of commondb4)
db4 <- commondb4[commondb4$db_boo==1,]
imdb4 <- commondb4[commondb4$db_boo==0,]
```

3.1 Douban Extreme Rating Analysis

3.1.1 1-star votes as dependent variable

Excluding language dummies and certain genre dummies:

```r
## remove genre_drama, genre_action
db4_1star_nolanguage_lm <- lm(star_db_1~num_votes_imdb_us + num_votes_db + year+budget + length + genre_adventure + genre_animation + genre_biography + genre_comedy + genre_cinema + genre_horror + genre_mystery + country_china + country_usa + country_india + country_uk + country_europe + country_asia + country_america, 
                                data=db4)

summary(db4_1star_nolanguage_lm)
```
## Call:
\[
\text{lm(formula = star\_db\_1 ~ num\_votes\_imdb\_us + num\_votes\_db + year + budget + length + genre\_adventure + genre\_animation + genre\_biography + genre\_comedy + genre\_crime + genre\_horror + genre\_mystery + country\_china + country\_usa + country\_india + country\_uk + country\_europe + country\_asia + country\_america, data = db4)}
\]

## Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>-0.005418</td>
<td>-0.001945</td>
<td>-0.000866</td>
<td>0.000730</td>
<td>0.034431</td>
</tr>
</tbody>
</table>

## Coefficients:

|                          | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------|----------|------------|---------|---------|
| (Intercept)              | -6.833e-02 | 1.427e-02 | -4.790  | 2.01e-06 *** |
| num\_votes\_imdb\_us    | 9.910e-10 | 3.621e-09 | 0.274   | 0.78444 |
| num\_votes\_db          | -5.259e-09 | 9.067e-10 | -5.800  | 9.74e-09 *** |
| year                    | 3.688e-05 | 7.127e-06 | 5.174   | 2.94e-07 *** |
| budget                  | -1.861e-07 | 1.829e-06 | -0.102  | 0.91901 |
| length                  | -5.732e-07 | 6.013e-06 | -0.095  | 0.92408 |
| genre\_adventure        | 5.903e-05 | 5.431e-04 | 0.109   | 0.91348 |
| genre\_animation        | -1.723e-03 | 5.292e-04 | -3.255  | 0.00118 ** |
| genre\_biography        | -2.157e-05 | 4.930e-04 | -4.376  | 0.00105 ** |
| genre\_comedy           | -1.314e-05 | 4.493e-04 | -0.029  | 0.97667 |
| genre\_crime            | 9.869e-05 | 4.387e-04 | 0.225   | 0.82206 |
| genre\_horror           | 4.192e-03 | 1.275e-03 | 3.288   | 0.00105 ** |
| genre\_mystery          | -1.375e-04 | 1.210e-03 | -0.114  | 0.90953 |
| country\_china          | 8.839e-04 | 1.302e-03 | 0.679   | 0.49735 |
| country\_usa            | -1.597e-04 | 8.298e-04 | -0.192  | 0.84745 |
| country\_india          | 1.444e-03 | 1.267e-03 | 1.140   | 0.25477 |
| country\_uk             | -6.287e-04 | 9.533e-04 | -0.660  | 0.50977 |
| country\_europe         | 2.550e-04 | 8.876e-04 | 0.287   | 0.77398 |
| country\_asia           | 1.819e-04 | 9.626e-04 | 0.189   | 0.85019 |
| country\_america        | -1.563e-03 | 1.092e-03 | -1.432  | 0.15268 |

### Signif. codes:  
0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.003736 on 753 degrees of freedom  
Multiple R-squared:  0.1223, Adjusted R-squared:  0.1001  
F-statistic: 5.52 on 19 and 753 DF, p-value: 7.214e-13

Excluding country dummies and certain genre dummies:

```r
# remove language_english and genre_action
db4_1star_nocountry_lm <- lm(star\_db\_1\-num\_votes\_imdb\_us + num\_votes\_db + year+budget + length + genre\_adventure + genre\_animation + genre\_biography + genre\_comedy + genre\_crime + genre\_drama + genre\_horror + genre\_mystery + language\_indian + language\_chinese + language\_european + language\_asian + language\_american, data=db4)
```

summary(db4_1star_nocountry_lm)
## Call:
```
lm(formula = star_db_1 ~ num_votes_imdb_us + num_votes_db + year +
    budget + length + genre_adventure + genre_animation + genre_biography +
    genre_comedy + genre_crime + genre_drama + genre_horror +
    genre_mystery + language_indian + language_chinese + language_european +
    language_asian + language_american, data = db4)
```
## Residuals:
```
       Min        1Q    Median        3Q       Max
-0.006495 -0.001873 -0.000841  0.000673  0.034207
```
## Coefficients:
```
                        Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)         -6.513e-02  1.383e-02   -4.709 2.97e-06 ***
num_votes_imdb_us    1.941e-10  3.647e-09    0.053 0.957574
num_votes_db        -5.103e-09  9.031e-10   -5.651 2.26e-08 ***
year                3.541e-05  6.952e-06    5.093 4.46e-07 ***
budget              -1.380e-07  1.812e-06   -0.076 0.939332
length              -1.782e-07  6.008e-06   -0.030 0.976344
genre_adventure     -5.061e-04  6.006e-04   -0.843 0.399646
genre_animation     -2.210e-03  5.877e-04   -3.760 0.000183 ***
genre_biography     -2.666e-03  5.691e-04   -4.684 3.33e-06 ***
genre_comedy        -5.896e-04  5.312e-04   -1.110 0.267388
genre_crime         -4.273e-04  5.147e-04   -0.830 0.406646
genre_drama         -8.462e-04  4.317e-04   -1.960 0.050359 .
genre_horror        3.570e-03  1.293e-03    2.760 0.005921 **
genre_mystery       -5.902e-04  1.235e-03   -0.478 0.632932
language_indian     1.923e-03  1.034e-03    1.860 0.063262 .
language_chinese    8.522e-04  1.089e-03    0.783 0.434087
language_european   7.028e-04  4.409e-04    1.594 0.111338
language_asian     4.543e-04  5.875e-04    0.773 0.439578
language_american  -1.420e-03  8.893e-04   -1.597 0.110725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
## Residual standard error: 0.00372 on 754 degrees of freedom
## Multiple R-squared:  0.1286, Adjusted R-squared:  0.1078
## F-statistic: 6.181 on 18 and 754 DF, p-value: 2.821e-14

### 3.1.2 1-star + 2-star votes as dependent variable

Excluding language dummies and certain genre dummies:
## add one more column to db4
star_db_12 <- db4$star_db_1 + db4$star_db_2
db4 <- cbind(db4, star_db_12)

## remove genre_drama, genre_action
db4_12star_nolanguage_lm <- lm(star_db_12~num_votes_imdb_us + num_votes_db + year+budget + length + genre_adventure + genre_animation + genre_biography + genre_comedy + genre_crime + genre_horror + genre_mystery + country_china + country_usa + country_india + country_uk + country_europe + country_asia + country_america,
   data=db4)

summary(db4_12star_nolanguage_lm)
## Call:
```
lm(formula = star_db_12 ~ num_votes_imdb_us + num_votes_db +
    year + budget + length + genre_adventure + genre_animation +
    genre_biography + genre_comedy + genre_crime + genre_horror +
    genre_mystery + country_china + country_usa + country_india +
    country_uk + country_europe + country_asia + country_america,
    data = db4)
```

## Residuals:
```
       Min        1Q    Median        3Q       Max
-0.035007 -0.009252 -0.003371  0.005108  0.163798
```

## Coefficients:
```
                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)       -4.935e-01  6.382e-02  -7.733  3.38e-14 ***
num_votes_imdb_us  1.075e-08  1.620e-08   0.664   0.507032
num_votes_db      -3.335e-08  4.056e-09  -8.222  8.74e-16 ***
year               2.615e-04  3.188e-05   8.200  1.03e-15 ***
budget            -2.533e-06  8.184e-06  -0.310   0.757022
length            -3.318e-05  2.690e-05  -1.233   0.217797
genre_adventure   -1.002e-03  2.430e-03  -0.413   0.680022
genre_animation   -8.919e-03  2.367e-03  -3.768  0.000178 ***
genre_biography   -1.068e-02  2.205e-03  -4.845  1.54e-06 ***
genre_comedy      -8.072e-04  2.010e-03  -0.402   0.688061
genre_crime       -1.879e-04  1.962e-03  -0.096   0.923746
genre_horror      2.576e-02  5.702e-03   4.517  7.28e-06 ***
genre_mystery      7.526e-03  5.823e-03   2.051   0.040612 *
genre_drama       -4.611e-03  4.885e-03  -0.944   0.345449
```

---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.01671 on 753 degrees of freedom

## Multiple R-squared:  0.1958, Adjusted R-squared:  0.1755

## F-statistic: 9.648 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:

```
# remove language_english and genre_action
db4_12star_nocountry_lm <- lm(star_db_12~num_votes_imdb_us + num_votes_db + year + budget + length + genre_adventure + genre_animation + genre_biography + genre_comedy + genre_crime + genre_drama + genre_horror + genre_mystery + language_indian + language_chinese + language_european + language_asian + language_american,
                           data=db4)

summary(db4_12star_nocountry_lm)
```
3.1.3 1-star + 2-star + 3-star votes as dependent variable

Excluding language and certain genre dummies:
## add one more column to db4
star_db_123 <- db4$star_db_1 + db4$star_db_2 + db4$star_db_3
db4 <- cbind(db4, star_db_123)

## remove genre_drama, genre_action
db4_123star_nolanguage_lm <- lm(star_db_123~num_votes_imdb_us+num_votes_db+year+budget+length
                                   +genre_adventure+genre_animation+genre_biography+genre_comedy
                                   +genre_crime+genre_horror+genre_mystery
                                   +country_china+country_america+country_india+country_uk+country_europe
                                   +country_asia+country_america,
                                   data=db4)
summary(db4_123star_nolanguage_lm)
## Call:
```
lm(formula = star_db_123 ~ num_votes_imdb_us + num_votes_db +
    year + budget + length + genre_adventure + genre_animation +
    genre_bio + genre_comedy + genre_crime + genre_horror +
    genre_mystery + country_china + country_usa + country_india +
    country_uk + country_europe + country_asia + country_america,
    data = db4)
```
## Residuals:
```
   Min     1Q Median     3Q    Max
-0.2051 -0.0589 -0.0112  0.0423  0.4282
```
## Coefficients:
```
                           Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)               -3.313e+00  3.211e-01  -10.317  < 2e-16 ***
num_votes_imdb_us        8.800e-08  8.152e-08   1.079  0.28075
num_votes_db             2.634e-07  2.041e-08  -12.903  < 2e-16 ***
year                     1.786e-03  1.604e-04   11.135  < 2e-16 ***
length                   -3.458e-05  4.118e-05  -0.840  0.40135
genre_adventure         -1.060e-02  1.223e-02  -0.867  0.38636
genre_animation          5.783e-02  1.183e-02   3.551  0.00035 **
genre_bio                3.377e-03  2.743e-03  -0.123  0.90213
genre_comedy            -1.178e-02  1.011e-02  -1.165  0.24429
genre_crime             -5.454e-03  9.875e-03  -0.552  0.58087
genre_horror            -1.203e-01  2.869e-02  -4.191  3.09e-05 ***
genre_mystery           -1.628e-01  2.899e-02  -5.570  1.47e-08 ***
```
## Residual standard error: 0.0841 on 753 degrees of freedom
## Multiple R-squared:  0.3268, Adjusted R-squared:  0.3098
## F-statistic: 19.24 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:
# remove language_english and genre_action

db4_123star_nocountry_lm <- lm(star_db_123~num_votes_imdb_us+num_votes_db+year+budget+length
+genre_adventure+genre_animation+genre_biography+genre_comedy
+genre_crime+genre_drama+genre_horror+genre_mystery
+language_indian+language_chinese+language_european+language_asian
+language_american,
data=db4)

summary(db4_123star_nocountry_lm)

##
## Call:
## lm(formula = star_db_123 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_drama +
##     genre_horror + genre_mystery + language_indian + language_chinese +
##     language_european + language_asian + language_american, data = db4)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -0.19397 -0.05631 -0.01292  0.04213  0.45871
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   -2.882e+00  3.067e-01  -9.397  < 2e-16 ***
## num_votes_imdb_us  6.591e-09  8.087e-08   0.082 0.935063
## num_votes_db     -2.531e-07  2.002e-08 -12.640  < 2e-16 ***
## year             1.610e-03  1.541e-04  10.443  < 2e-16 ***
## budget           -4.016e-05  4.018e-05  -1.000 0.317827
## length           -3.349e-04  1.332e-04  -2.514 0.012150 *
## genre_adventure  -5.092e-02  1.332e-02  -3.823 0.000142 ***
## genre_animation  -7.236e-02  1.303e-02  -5.552 3.91e-08 ***
## genre_biography  -8.952e-02  1.262e-02  -7.094 3.00e-12 ***
## genre_comedy     -4.883e-02  1.178e-02  -4.145 3.78e-05 ***
## genre_crime      -3.022e-02  1.141e-02  -2.648 0.008265 **
## genre_drama      -5.939e-02  9.573e-03  -6.204 9.06e-10 ***
## genre_horror     -7.619e-02  2.868e-02  -2.657 0.008057 **
## genre_mystery   -1.767e-02  2.739e-02  -0.645 0.519007
## language_indian  -3.096e-02  2.292e-02  -1.351 0.177181
## language_chinese 5.780e-02  2.414e-02  2.394 0.016905 *
## language_european -4.070e-02  9.775e-03  -4.163 3.50e-05 ***
## language_asian   -3.975e-02  1.303e-02  -3.051 0.002360 **
## language_american -3.697e-02  1.972e-02  -1.875 0.061195 .
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08248 on 754 degrees of freedom
## Multiple R-squared:  0.3516, Adjusted R-squared:  0.3361
## F-statistic: 22.71 on 18 and 754 DF,  p-value: < 2.2e-16

3.1.4 5-star as dependent variable
Excluding language dummies and certain genre dummies:

```r
## remove genre_drama, genre_action

db4_5star_nolanguage_lm <- lm(star_db_5~num_votes_imdb_us+num_votes_db+year+budget+length
 +genre_adventure+genre_animation+genre_biography+genre_comedy
 +genre_crime+genre_horror+genre_mystery
 +country_china+country_usa+country_india+country_uk+country_europe
 +country_asia+country_america,
 data=db4)

summary(db4_5star_nolanguage_lm)
```
## Call:
```r
lm(formula = star_db_5 ~ num_votes_imdb_us + num_votes_db + year +
    budget + length + genre_adventure + genre_animation + genre_biography +
    genre_comedy + genre_crime + genre_horror + genre_mystery +
    country_china + country_usa + country_india + country_uk +
    country_europe + country_asia + country_america, data = db4)
```
## Residuals:
```
  Min       1Q   Median       3Q      Max
-0.26813 -0.07642 -0.00494  0.07161  0.38360
```
## Coefficients:
```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.974e+00  4.068e-01  14.684  < 2e-16 ***
num_votes_imdb_us  -2.882e-08  1.033e-07  -0.279 0.780260
num_votes_db       4.806e-07  2.586e-08  18.585  < 2e-16 ***
year              -2.861e-03  2.033e-04 -14.076  < 2e-16 ***
budget             9.539e-07  5.217e-05   0.018 0.985416
length             7.539e-04  1.715e-04   4.397 1.26e-05 ***
genre_adventure    1.398e-02  1.549e-02   0.902 0.367087
genre_animation    6.171e-02  1.509e-02   4.089 4.79e-05 ***
genre_biography    4.828e-02  1.406e-02   3.435 0.000626 ***
genre_comedy       2.800e-02  1.281e-02   2.185 0.029166 *
genre_crime        -2.208e-02  1.251e-02  -1.765 0.077916 .
genre_horror       -1.119e-01  3.635e-02  -0.307 0.762879
genre_mystery     -6.671e-02  3.451e-02  -1.933 0.053583 .
country_china     -1.425e-01  3.712e-02  -0.383 0.701357
country_usa       -9.022e-02  2.366e-02  -3.813 0.000149 ***
country_india     -2.036e-02  3.613e-02  -0.564 0.573246
country_uk        -7.921e-02  2.718e-02  -2.914 0.003675 **
country_europe    -7.915e-03  2.531e-02  -0.313 0.754593
country_asia      -1.937e-02  2.745e-02  -0.706 0.480577
country_america   -2.086e-02  3.114e-02  -0.670 0.503083
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.1065 on 753 degrees of freedom
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4619
## F-statistic: 35.87 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:
```
# remove language_english and genre_action
db4_5star_nocountry_lm <- lm(star_db_5~num_votes_imdb_us+num_votes_db+year+budget+length+
    +genre_adventure+genre_animation+genre_biography+genre_comedy+
    +genre_crime+genre_drama+genre_horror+genre_mystery+
    +language_indian+language_chinese+language_european+language_asian
summary(db4_5star_nocountry_lm)
```

```r
64
```
## Call:
\[
\text{lm(formula = star\_db\_5} \sim \text{num\_votes\_imdb\_us} + \text{num\_votes\_db} + \text{year} + \\
\text{budget} + \text{length} + \text{genre\_adventure} + \text{genre\_animation} + \text{genre\_biography} + \\
\text{genre\_comedy} + \text{genre\_crime} + \text{genre\_drama} + \text{genre\_horror} + \\
\text{genre\_mystery} + \text{language\_indian} + \text{language\_chinese} + \text{language\_european} + \\
\text{language\_asian} + \text{language\_american}, \text{data = db4})
\]

## Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>-0.27746</td>
<td>-0.07382</td>
<td>-0.00718</td>
<td>0.06771</td>
<td>0.39608</td>
</tr>
</tbody>
</table>

## Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 5.290e+00 | 3.888e-01  | 13.605  | < 2e-16 *** |
| num\_votes\_imdb\_us | 6.945e-08 | 1.025e-07  | 0.677   | 0.4983   |
| num\_votes\_db      | 4.689e-07 | 2.539e-08  | 18.472  | < 2e-16 *** |
| year                | -2.583e-03 | 1.954e-04  | -13.217 | < 2e-16 *** |
| budget              | 1.199e-05  | 5.094e-05  | 0.235   | 0.8139   |
| length              | 6.875e-04  | 1.689e-04  | 4.071   | 5.18e-05 *** |
| genre\_adventure    | 6.832e-02  | 1.688e-02  | 4.047   | 5.73e-05 *** |
| genre\_animation    | 1.073e-01  | 1.652e-02  | 6.495   | 1.50e-10 *** |
| genre\_biography    | 1.004e-01  | 1.600e-02  | 6.275   | 5.90e-10 *** |
| genre\_comedy       | 7.667e-02  | 1.493e-02  | 5.134   | 3.62e-07 *** |
| genre\_crime        | 2.478e-02  | 1.447e-02  | 1.712   | 0.0872   |
| genre\_drama        | 7.899e-02  | 1.214e-02  | 6.509   | 1.38e-10 *** |
| genre\_horror       | -5.156e-02 | 3.636e-02  | -1.418  | 0.1566   |
| genre\_mystery      | -9.762e-03 | 3.472e-02  | -0.281  | 0.7787   |
| language\_indian    | 7.609e-02  | 2.905e-02  | 2.619   | 0.0090 ** |
| language\_chinese   | -3.968e-02 | 3.061e-02  | -1.296  | 0.1952   |
| language\_european  | 7.867e-02  | 1.239e-02  | 6.348   | 3.77e-10 *** |
| language\_asian     | 7.650e-02  | 1.652e-02  | 4.632   | 4.26e-06 *** |
| language\_american  | 3.920e-02  | 2.500e-02  | 1.568   | 0.1172   |

---

## Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.1046 on 754 degrees of freedom

## Multiple R-squared: 0.4936, Adjusted R-squared: 0.4816

## F-statistic: 40.84 on 18 and 754 DF, p-value: < 2.2e-16

### 3.2 IMDb Extreme Rating Analysis

#### 3.2.1 12-rating votes as dependent variable

Here 12-rating refers to films rated either 1 or 2. Similar naming practice for the sections below.

Excluding language dummies and certain genre dummies:
## add one more column to imdb4
rating_imdb_12 <- imdb4$rating_imdb_1 + imdb4$rating_imdb_2
imdb4 <- cbind(imdb4, rating_imdb_12)

## remove genre_drama, genre_action
imdb4_12rating_nolanguage_lm <- lm(rating_imdb_12~num_votes_imdb_us+num_votes_db+year+budget+length +genre_adventure+genre_animation+genre_biography+genre_comedy +genre_crime+genre_horror+genre_mystery +country_china+country_usa+country_india+country_uk+country_europe +country_asia+country_america, data=imdb4)
summary(imdb4_12rating_nolanguage_lm)
## Call:
```r
lm(formula = rating_imdb_12 ~ num_votes_imdb_us + num_votes_db +
    year + budget + length + genre_adventure + genre_animation +
    genre_biography + genre_comedy + genre_crime + genre_horror +
    genre_mystery + country_china + country_usa + country_india +
    country_uk + country_europe + country_asia + country_america,
    data = imdb4)
```

## Residuals:
```r
Min 1Q Median 3Q Max
-0.022531 -0.004896 -0.001530 0.003583 0.036034
```

## Coefficients:
```r
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.194e-01 2.937e-02 10.874 < 2e-16 ***
num_votes_imdb_us -8.980e-09 7.456e-09 -1.205 0.228763
num_votes_db -1.619e-10 1.867e-09 -0.087 0.930902
year -1.504e-04 1.467e-05 -10.253 < 2e-16 ***
length 5.095e-05 1.238e-05 4.116 4.29e-05 ***
genre_adventure -1.638e-03 1.118e-03 -1.465 0.143285
genre_animation -2.832e-03 1.089e-03 -2.600 0.009504 **
genre_biography -2.833e-03 1.015e-03 -2.792 0.005377 **
genre_comedy -1.019e-03 9.249e-04 -1.102 0.270766
genre_crime -1.131e-03 9.031e-04 -1.252 0.210961
genre_horror -3.402e-03 2.624e-03 -1.297 0.195156
genre_mystery -4.882e-03 2.491e-03 -1.960 0.050390 .
country_china -5.192e-03 2.680e-03 -1.937 0.053082 .
country_usa -8.785e-03 1.708e-03 -5.143 3.46e-07 ***
country_india 1.375e-02 2.608e-03 5.271 1.77e-07 ***
country_uk -7.459e-03 1.962e-03 -3.801 0.000156 ***
country_europe -4.292e-03 1.827e-03 -2.349 0.019094 *
country_asia -4.425e-03 1.982e-03 -2.233 0.025840 *
country_america -5.617e-03 2.248e-03 -2.499 0.012676 *
```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.007691 on 753 degrees of freedom
## Multiple R-squared:  0.3086, Adjusted R-squared:  0.2911
## F-statistic: 17.69 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:
# remove language_english and genre_action
imdb4_12rating_nocountry_lm <- lm(rating_imdb_12~num_votes_imdb_us+num_votes_db+year+budget+length
+genre_adventure+genre_animation+genre_biography+genre_comedy
+genre_crime+genre_drama+genre_horror+genre_mystery
+language_indian+language_chinese+language_european+language_asian
+language_american,
data=imdb4)
summary(imdb4_12rating_nocountry_lm)

##
## Call:
## lm(formula = rating_imdb_12 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_drama +
##     genre_horror + genre_mystery + language_indian + language_chinese +
##     language_european + language_asian + language_american, data = imdb4)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -0.023145 -0.005338 -0.001488  0.003409  0.034655
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.703e-01  2.902e-02   9.312  < 2e-16 ***
## num_votes_imdb_us      -7.733e-09  7.652e-09  -1.011  0.31254
## num_votes_db           -4.049e-10  1.895e-09  -0.214  0.83085
## year                   -1.307e-04  1.459e-05  -8.960  < 2e-16 ***
## budget                 -4.232e-06  3.802e-06  -1.113  0.26603
## length                 5.073e-05  1.261e-05   4.024 6.30e-05 ***
## genre_adventure       -1.473e-03  1.233e-03  -1.194  0.23269
## genre_animation      -1.321e-03  1.194e-03  -1.106  0.26913
## genre_biography       -1.321e-03  1.194e-03  -1.106  0.26913
## genre_comedy          -1.473e-03  1.233e-03  -1.194  0.23269
## genre_crime           -1.321e-03  1.194e-03  -1.106  0.26913
## genre_drama           -1.473e-03  1.233e-03  -1.194  0.23269
## genre_horror          -1.473e-03  1.233e-03  -1.194  0.23269
## genre_mystery         -1.473e-03  1.233e-03  -1.194  0.23269
## language_indian       2.160e-02  2.169e-03   9.959  < 2e-16 ***
## language_chinese      4.621e-03  2.285e-03   2.023  0.04348 *
## language_european     4.911e-03  9.250e-04   5.309 1.45e-07 ***
## language_asian       3.841e-03  1.233e-03   3.115  0.00191 **
## language_american    2.669e-03  1.866e-03   1.430  0.15308
##
## ---
## Signif. codes:  < 0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 1
##
## Residual standard error: 0.007805 on 754 degrees of freedom
## Multiple R-squared:  0.2869, Adjusted R-squared:  0.2699
## F-statistic: 16.85 on 18 and 754 DF,  p-value: < 2.2e-16

3.2.2 1234-rating votes as dependent variable

Excluding language dummies and certain genre dummies:

```r
## add one more column to imdb4
rating_imdb_1234 <- imdb4$rating_imdb_1 + imdb4$rating_imdb_2 +
                    imdb4$rating_imdb_3 + imdb4$rating_imdb_4
imdb4 <- cbind(imdb4, rating_imdb_1234)

## remove genre_drama, genre_action
imdb4_1234rating_nolanguage_lm <- lm(rating_imdb_1234 ~ num_votes_imdb_us + num_votes_db +
                                       year + budget + length +
                                       genre_adventure + genre_animation +
                                       genre_biography + genre_comedy +
                                       genre_crime + genre_horror +
                                       genre_mystery +
                                       country_china + country_usa + country_india +
                                       country_uk + country_europe +
                                       country_asia + country_america,
                                       data = imdb4)
summary(imdb4_1234rating_nolanguage_lm)
```
## Call:
## lm(formula = rating_imdb_1234 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_horror +
##     genre_mystery + country_china + country_usa + country_india +
##     country_uk + country_europe + country_asia + country_america,
##     data = imdb4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028278 -0.008389 -0.002198  0.006815  0.046532
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.787e-01  4.721e-02   8.022 3.98e-15 ***
## num_votes_imdb_us -4.001e-08  1.198e-08  -3.339 0.000882 ***
## num_votes_db      -4.842e-09  3.000e-09  -1.614 0.106991
## year              -1.700e-04  2.358e-05  -7.207 1.40e-12 ***
## budget            -6.000e-06  6.054e-06  -0.991 0.321951
## length             7.034e-05  1.990e-05   3.535 0.000432 ***
## genre_adventure   -2.772e-03  1.797e-03  -1.543 0.123345
## genre_animation   -5.277e-03  1.751e-03  -3.014 0.002669 **
## genre_biography   -7.568e-03  1.631e-03  -4.639 4.12e-06 ***
## genre_comedy      -5.535e-04  1.487e-03  -0.372 0.709758
## genre_crime       -3.174e-03  1.452e-03  -2.187 0.029067 *
## genre_horror      -2.993e-03  4.218e-03  -0.710 0.478147
## genre_mystery     -6.147e-03  4.004e-03  -1.535 0.125152
## country_china     -6.726e-03  4.307e-03  -1.561 0.118844
## country_usa      -1.041e-02  2.746e-03  -3.793 0.000161 ***
## country_india     7.536e-03  4.192e-03   1.798 0.072629 .
## country_uk       -9.775e-03  3.154e-03  -3.099 0.002015 **
## country_europe   -5.790e-03  2.937e-03  -1.971 0.049061 *
## country_asia     -7.270e-03  3.185e-03  -2.282 0.022741 *
## country_america  -9.992e-03  3.613e-03  -2.765 0.005823 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01236 on 753 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2061
## F-statistic: 11.55 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:
# remove language_english and genre_action
imdb4_1234rating_nocountry_lm <- lm(rating_imdb_1234~num_votes_imdb_us+num_votes_db+year+
budget+length+
  +genre_adventure+genre_animation+genre_biography+genre_comedy+
  +genre_crime+genre_drama+genre_horror+genre_mystery+
  +language_indian+language_chinese+language_european+language_asian+
  +language_american,
data=imdb4)
summary(imdb4_1234rating_nocountry_lm)

## Call:
## lm(formula = rating_imdb_1234 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_drama +
##     genre_horror + genre_mystery + language_indian + language_chinese +
##     language_european + language_asian + language_american, data = imdb4)
##
## Residuals:
##    Min     1Q    Median     3Q    Max
##-0.026724 -0.008450 -0.002208  0.006861  0.046804
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.277e-01  4.617e-02 7.097 2.94e-12 ***
## num_votes_imdb_us  3.609e-08  1.217e-08 -2.965  0.003127 **
## num_votes_db  -5.367e-09  3.014e-09 -1.781   0.075393 .
## year  -1.505e-04  2.320e-05 -6.488 1.58e-10 ***
## budget    5.241e-06  6.048e-06  0.867   0.386401
## length    6.980e-05  2.005e-05  3.481  0.000529 ***
## genre_adventure   8.280e-05  2.005e-03   4.1    0.967063
## genre_animation -2.956e-03  1.962e-03 -1.507  0.132287
## genre_biography -5.193e-03  1.900e-03 -2.734  0.006410 **
## genre_comedy   1.497e-03  1.773e-03  0.844  0.398844
## genre_crime    1.080e-03  1.718e-03 -0.629  0.529723
## genre_drama   3.482e-03  1.441e-03  2.416  0.015910 *
## genre_horror  6.122e-04  4.317e-03 -0.142  0.887270
## genre_mystery  3.051e-03  4.123e-03 -0.740  0.459561
## language_indian   1.739e-02  3.450e-03  5.042  5.78e-07 ***
## language_chinese  5.259e-03  3.635e-03  1.447  0.148347
## language_european  4.991e-03  1.471e-03  3.392  0.000730 ***
## language_asian   2.636e-03  1.961e-03  1.344  0.179223
## language_american -5.164e-04  2.968e-03 -0.174  0.861914
## ---
## Signif. codes:  < 0.001 ***  0.001 **  0.05 *  0.1  ' '  1
##
## Residual standard error: 0.01242 on 754 degrees of freedom
## Multiple R-squared:  0.2179, Adjusted R-squared:  0.1992
## F-statistic: 11.67 on 18 and 754 DF,  p-value: < 2.2e-16

3.2.3 123456-rating votes as dependent variable
Excluding language dummies and certain genre dummies:

```r
## add one more column to imdb4
rating_imdb_123456 <- imdb4$rating_imdb_1 + imdb4$rating_imdb_2+
imdb4$rating_imdb_3 + imdb4$rating_imdb_4+
imdb4$rating_imdb_5 + imdb4$rating_imdb_6
imdb4 <- cbind(imdb4, rating_imdb_123456)

## remove genre_drama, genre_action
imdb4_123456rating_nolanguage_lm <- lm(rating_imdb_123456~num_votes_imdb_us+num_votes_db+
                                      +year+budget+length
                                      +genre_adventure+genre_animation+genre_biography+genre_comedy+
                                      +genre_crime+genre_horror+genre_mystery
                                      +country_china+country_usa+country_india+country_uk+country_europe
                                      +country_asia+country_america,
                                      data=imdb4)
summary(imdb4_123456rating_nolanguage_lm)
```
```
## Call:
## lm(formula = rating_imdb_123456 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_horror +
##     genre_mystery + country_china + country_usa + country_india +
##     country_uk + country_europe + country_asia + country_america,
##     data = imdb4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.076132 -0.019152  0.000222  0.018761  0.107713
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.152e-01  1.075e-01  -1.072 0.284024
## num_votes_imdb_us -2.858e-07  2.729e-08 -10.473  < 2e-16 ***
## num_votes_db      -3.180e-08  6.833e-09  -4.654 3.84e-06 ***
## year               1.326e-04  5.371e-05   2.468 0.013801 *
## budget            -3.580e-07  1.379e-05  -0.26   0.797291
## length            -1.146e-05  4.531e-05  -0.25   0.800338
## genre_adventure   -9.201e-04  4.093e-03  -0.225 0.822174
## genre_animation   -9.378e-03  3.987e-03  -2.352 0.018930 *
## genre_biography   -2.222e-02  3.715e-03  -5.983 3.39e-09 ***
## genre_comedy      -1.625e-03  3.466e-03  -0.468 0.640447
## genre_crime       -9.619e-03  3.305e-03  -2.910 0.003721 **
## genre_horror      -1.774e-03  7.183e-03  -0.246 0.806880
## genre_mystery     -5.811e-03  9.809e-03  -0.590 0.556670
## country_china     6.458e-03  9.809e-03   0.66   0.509679
## country_india    -3.205e-02  9.546e-03  -3.35   0.000823 ***
## country_usa       7.328e-03  6.253e-03   1.17   0.241581
## country_europe    -2.857e-03  6.688e-03  -0.427 0.670171
## country_asia      -9.374e-03  7.253e-03  -1.29   0.196606
## country_america   -1.797e-02  8.228e-03  -2.18   0.029237 *
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02815 on 753 degrees of freedom
## Multiple R-squared:  0.2764, Adjusted R-squared:  0.2581
## F-statistic: 15.14 on 19 and 753 DF,  p-value: < 2.2e-16
```
# remove language_english and genre_action
imdb4_123456rating_nocountry_lm <- lm(rating_imdb_123456 ~ num_votes_imdb_us + num_votes_db +
  year + budget + length +
  genre_adventure + genre_animation + genre_biography + genre_comedy +
  genre_crime + genre_drama +
  language_indian + language_chinese + language_european + language_asian +
  language_american,
  data = imdb4)
summary(imdb4_123456rating_nocountry_lm)

## Call:
## lm(formula = rating_imdb_123456 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_drama +
##     genre_horror + genre_mystery + language_indian + language_chinese +
##     language_european + language_asian + language_american, data = imdb4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##-0.07463 -0.01915  0.00038  0.01858  0.10875
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02943e-02 1.050e-01 -0.280 0.779407
## num_votes_imdb_us -2.807e-07 2.769e-08 -10.136  < 2e-16 ***
## num_votes_db -3.189e-08 6.857e-09 -4.651 3.90e-06 ***
## year 9.299e-05 5.279e-05  1.761 0.078562 .
## budget -1.474e-06 1.376e-05 -0.107 0.914692
## length -1.055e-05 4.562e-05 -0.231 0.817195
## genre_adventure -2.752e-03 4.560e-03 -0.603 0.546365
## genre_animation -9.429e-03 4.463e-03 -2.113 0.034944 *
## genre_biography -2.320e-02 4.322e-03 -5.368 1.06e-07 ***
## genre_comedy -6.189e-04 4.034e-03 -0.153 0.878106
## genre_crime -1.054e-02 3.908e-03 -2.697 0.007143 **
## genre_drama -1.282e-03 3.278e-03 -0.391 0.695891
## genre_horror -3.039e-03 9.821e-03 -0.309 0.757111
## genre_mystery -7.030e-03 9.380e-03 -0.749 0.453823
## language_indian -3.653e-02 7.848e-03 -4.654 3.84e-06 ***
## language_chinese -1.387e-03 8.269e-03 -0.168 0.866817
## language_european -1.049e-02 3.348e-03 -3.135 0.001785 **
## language_asian -1.680e-02 4.461e-03 -3.765 0.000179 ***
## language_american -2.445e-02 6.752e-03 -3.621 0.000313 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02825 on 754 degrees of freedom
## Multiple R-squared:  0.2705, Adjusted R-squared:  0.2531
## F-statistic: 15.53 on 18 and 754 DF,  p-value: < 2.2e-16

3.2.4 10-rating votes as dependent variable
Excluding language dummies and certain genre dummies:

```r
## remove genre_drama, genre_action
imdb4_10rating_nolanguage_lm <- lm(rating_imdb_10 ~ num_votes_imdb_us + num_votes_db + year + budget + length +
                                     genre_adventure + genre_animation + genre_biography + genre_comedy +
                                     genre_crime + genre_horror + genre_mystery +
                                     country_china + country_usa + country_india + country_uk + country_europe +
                                     country_asia + country_america,
                                     data = imdb4)
summary(imdb4_10rating_nolanguage_lm)
```
## Call:
\[
\text{lm(formula = rating\_imdb\_10 ~ num\_votes\_imdb\_us + num\_votes\_db + }
\text{ year + budget + length + genre\_adventure + genre\_animation + }
\text{ genre\_biography + genre\_comedy + genre\_crime + genre\_horror + }
\text{ genre\_mystery + country\_china + country\_usa + country\_india + }
\text{ country\_uk + country\_europe + country\_asia + country\_america, }
\text{ data = imdb4})
\]

## Residuals:

\begin{verbatim}
 Min  1Q Median  3Q Max
-0.182430 -0.032873 -0.005144  0.026694  0.230363
\end{verbatim}

## Coefficients:

\begin{verbatim}
 (Intercept) 2.912e+00  1.868e-01  15.586  < 2e-16 ***
 num\_votes\_imdb\_us 6.056e-07  4.743e-08  12.768  < 2e-16 ***
 num\_votes\_db  7.139e-08  1.188e-08   6.011 2.87e-09 ***
 year        -1.390e-03  9.335e-05 -14.888  < 2e-16 ***
 budget      -2.361e-05  2.396e-05  -0.985 0.324827
 length       5.247e-04  7.875e-05   6.663 5.18e-11 ***
 genre\_adventure -6.511e-03  7.114e-03  -0.915 0.360324
 genre\_animation  2.353e-02  6.931e-03   3.395 0.000723 ***
 genre\_biography  2.353e-02  6.931e-03   3.395 0.000723 ***
 genre\_comedy    1.207e-03  5.884e-03  2.050 0.041993
 genre\_crime    -1.449e-02  5.745e-03  -2.522 0.011870 *
 genre\_horror    3.288e-03  1.669e-02   0.019 0.843938
 genre\_mystery   -2.739e-02  1.585e-02  -1.728 0.084331 .
 country\_china  -4.159e-02  1.705e-02  -2.439 0.014938 *
 country\_usa    -7.673e-02  1.087e-02  -7.060 3.80e-12 ***
 country\_india   1.449e-01  1.659e-02   8.731  < 2e-16 ***
 country\_uk     -7.295e-02  1.249e-02  -5.843 7.62e-09 ***
 country\_europe -4.000e-02  1.163e-02  -3.441 0.000612 ***
 country\_asia   -3.471e-02  1.261e-02  -2.753 0.0056041 **
 country\_america -3.126e-02  1.430e-02  -2.186 0.029112 *
\end{verbatim}

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.04893 on 753 degrees of freedom
## Multiple R-squared:  0.5258, Adjusted R-squared:  0.5138
## F-statistic: 43.94 on 19 and 753 DF,  p-value: < 2.2e-16

Excluding country dummies and certain genre dummies:
# remove language_english and genre_action
imdb4_10rating_nocountry_lm <- lm(rating_imdb_10~num_votes_imdb_us+num_votes_db+year+budget+length
+genre_adventure+genre_animation+genre_biography+genre_comedy
+genre_crime+genre_drama+genre_horror+genre_mystery
+language_indian+language_chinese+language_european+language_asian
+language_american,
data=imdb4)

summary(imdb4_10rating_nocountry_lm)

## Call:
## lm(formula = rating_imdb_10 ~ num_votes_imdb_us + num_votes_db +
##     year + budget + length + genre_adventure + genre_animation +
##     genre_biography + genre_comedy + genre_crime + genre_drama +
##     genre_horror + genre_mystery + language_indian + language_chinese +
##     language_european + language_asian + language_american, data = imdb4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.18186 -0.03364 -0.00642  0.02536  0.29674
##
## Coefficients:
##                           Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                2.506e+00  1.898e-01  13.204  < 2e-16 ***
## num_votes_imdb_us           5.945e-07  5.005e-08  11.877  < 2e-16 ***
## num_votes_db                7.007e-08  1.239e-08   5.653 2.23e-08 ***
## year                        -1.227e-03  9.541e-05  -12.861  < 2e-16 ***
## budget                      -1.608e-05  2.487e-05  -0.647       0.517990
## length                      5.369e-04  8.245e-05   6.512 1.36e-10 ***
## genre_adventure             4.748e-03  8.242e-03   0.576       0.564731
## genre_animation             3.149e-02  8.066e-03   3.904 0.000103 ***
## genre_biography             7.569e-04  7.811e-03   0.097       0.922825
## genre_comedy                7.522e-03  5.925e-03   1.294       0.200338
## genre_crime                 1.139e-02  1.775e-02   0.647       0.517990
## genre_horror                4.748e-03  8.242e-03   0.576       0.564731
## genre_mystery               3.149e-02  8.066e-03   3.904 0.000103 ***
## language_indian             7.569e-04  7.811e-03   0.097       0.922825
## language_chinese            7.522e-03  5.925e-03   1.294       0.200338
## language_european           1.139e-02  1.775e-02   0.647       0.517990
## language_asian              3.149e-02  8.066e-03   3.904 0.000103 ***
## language_american           7.569e-04  7.811e-03   0.097       0.922825
## ---
## Signif. codes:  < 0.001 ***  0.001 **  0.01 *  0.05 .  0.1   1
##
## Residual standard error: 0.05105 on 754 degrees of freedom
## Multiple R-squared:   0.4831, Adjusted R-squared:  0.4707
## F-statistic: 39.15 on 18 and 754 DF,  p-value: < 2.2e-16
4. Three-way interaction models

In order to further understand the interactions between genre and production country dummies, add three-term interaction terms and the related lower level terms into the models.

After the `lm` objects are created, use VIF function to check for multicollinearity, and remove the high-VIF terms to fix the multicollinearity problem.

4.1 Model using language dummies (instead of country dummies)

High-VIF terms are already removed in the model below:
three_term_nocountry_lm <- lm(rating~db_boo*num_votes_imdb_us + db_boo*num_votes_db + db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure*language_indian + db_boo*genre_animation*language_indian + db_boo*genre_biography*language_indian + db_boo*genre_comedy*language_indian + db_boo*genre_crime*language_indian + db_boo*genre_drama*language_indian + db_boo*genre_horror*language_indian + db_boo*genre_mystery*language_indian + db_boo*genre_comedy*language_chinese + db_boo*genre_mystery*language_chinese + db_boo*genre_animation*language_chinese + db_boo*genre_biography*language_chinese + db_boo*genre_horror*language_chinese + db_boo*genre_comedy*language_chinese + db_boo*genreCrime*language_chinese + db_boo*genre_animation*language_chinese + db_boo*genre_biography*language_chinese + db_boo*genre_horror*language_chinese + db_boo*genre_comedy*language_chinese + db_boo*genre_adventure*language_chinese + db_boo*genre_animation*language_chinese + db_boo*genre_biography*language_chinese + db_boo*genre_horror*language_chinese + db_boo*genre_comedy*language_chinese + db_boo*genre_adventure*language_chinese + db_boo*genre_animation*language_chinese + db_boo*genre_biography*language_chinese + db_boo*genre_horror*language_chinese + db_boo*genre_comedy*language_chinese + db_boo*genre_adventure*language_chinese + db_boo*genre_animation*language_chinese + db_boo*gene...
## Call:
```
# lm(formula = rating ~ db boo * num votes imdb us + db boo * num votes db +
#     db boo * year + db boo * length + db boo * budget + db boo *
#     genre_adventure * language indian + db boo * genre animation *
#     language indian + db boo * genre biography * language indian +
#     db boo * genre comedy * language indian + db boo * genre crime *
#     language indian + db boo * genre drama * language indian +
#     db boo * genre horror * language indian + db boo * genre mystery *
#     language indian + db boo * genre adventure * language indian +
#     db boo * genre animation * language chinese + db boo * genre biography *
#     language chinese + db boo * genre comedy * language chinese +
#     db boo * genre crime * language chinese + db boo * genre drama *
#     language chinese + db boo * genre horror * language chinese +
#     db boo * genre mystery * language chinese + db boo * genre adventure *
#     language european + db boo * genre animation * language european +
#     db boo * genre biography * language european + db boo * genre comedy *
#     language european + db boo * genre crime * language european +
#     db boo * genre drama * language european + db boo * genre horror *
#     language european + db boo * genre mystery * language european +
#     db boo * genre adventure * language asian + db boo * genre animation *
#     language asian + db boo * genre biography * language asian +
#     db boo * genre comedy * language asian + db boo * genre crime *
#     language asian + db boo * genre drama * language asian +
#     db boo * genre horror * language asian + db boo * genre mystery *
#     language asian + db boo * genre adventure * language american +
#     db boo * genre animation * language american + db boo * genre biography *
#     language american + db boo * genre comedy * language american +
#     db boo * genre crime * language american + db boo * genre drama *
#     language american + db boo * genre horror * language american +
#     db boo * genre mystery * language american - genre adventure:language indian -
#     language indian:genre animation - language indian:genre biography -
#     language indian:genre horror - language indian:genre mystery -
#     genre adventure:language chinese - genre animation:language chinese -
#     genre biography:language chinese - genre horror:language chinese -
#     genre mystery:language european - genre horror:language european -
#     genre mystery:language chinese - genre horror:language european -
#     genre mystery:language european - genre biography:language american -
#     genre comedy:language asian - genre horror:language asian -
#     genre mystery:language asian - genre animation:language american -
#     genre horror:language american - db boo:genre adventure:language indian -
#     db boo:language indian:genre animation - db boo:language indian:genre biography -
#     db boo:language indian:genre horror - db boo:language indian:genre mystery -
```

## Residuals:
##    Min     1Q  Median     3Q    Max
## -1.8438 -0.1782   0.0100  0.1926  1.1773
##
## Coefficients:
##                               Estimate   Std. Error   t value
## (Intercept)                  2.256e+01   8.442e-01   26.716
## num_votes_imdb_us            4.159e-06   2.935e-07   14.170
## num_votes_db                  1.063e-07   7.763e-08   1.370
## year                         -7.487e-03   4.245e-04  -17.640
## length                       1.193e-03   3.672e-04    3.249
## budget                       -1.967e-04   1.510e-04  -0.130
## genre_adventure              -1.059e-02   5.150e-02  -0.206
## genre_animation              7.586e-02   5.458e-02    1.390
## genre_biography              3.875e-02   4.789e-02    0.809
## genre_comedy                 -6.601e-04   4.687e-02  -0.014
## genre_crime                  -3.844e-02   4.630e-02  -0.830
## genre_drama                  1.216e-02   3.588e-02    0.339
## genre_horror                 6.994e-02   1.096e-01    0.638
## language_chinese             4.267e-02   1.099e-01    0.388
## language_asian               2.497e-01   8.972e-02    2.783
## db_boo:num_votes_imdb_us      1.354e-06   3.917e-07   34.677
## db_boo:num_votes_db           3.360e-04   1.078e-07   12.565
## db_boo:budget                 2.786e-01   6.851e-02    4.067
## db_boo:genre_adventure        2.965e-01   7.151e-02    4.146
## db_boo:genre_animation        3.953e-01   6.176e-02    6.401
## db_boo:genre_biography        2.671e-01   6.037e-02    4.424
## db_boo:genre_comedy           2.704e-01   1.519e-01   -2.112
## db_boo:genre_drama            2.103e-01   6.079e-02    3.460
## language_indian:genre_comedy  3.237e-01   2.302e-01    1.406
## language_indian:genre_drama   2.885e-01   4.130e-02    6.985
## db_boo:language_chinese       2.371e-01   3.407e-01    0.696
## genre_adventure:language_chinese 1.389e-02   2.525e-01    0.055
## genre_drama:language_chinese  2.704e-01   6.531e-02    4.140
## language_indian:language_drama 3.363e-02   9.043e-02    0.367
## genre_drama:language_asian    2.631e-02   1.229e-01    2.122
## db_boo:language_asian         3.886e-02   2.468e-01    0.015
## genre_animation:language_asian 7.644e-02   1.179e-01   -0.659
## genre_crime:language_asian   -1.518e-02   1.872e-01   -0.081
## genre_drama:language_asian               -1.338e-01  1.413e-01  -0.946
## genre_adventure:language_american         5.504e-02  3.228e-01   0.171
## genre_biography:language_american         1.316e-01  3.223e-01   0.408
## genre_comedy:language_american            2.431e-02  1.890e-01   0.129
## genre_crime:language_american             4.138e-01  3.249e-01   1.274
## genre_drama:language_american             1.723e-01  7.094e-02   2.428
## genre_mystery:language_american          -2.531e-02  3.374e-01  -0.075
## db_boo:language_indian:genre_comedy       3.845e-01  1.807e-01   2.128
## db_boo:language_indian:genre_crime        3.425e-01  3.250e-01   1.054
## db_boo:language_indian:genre_drama        5.961e-01  2.213e-01   2.721
## db_boo:genre_comedy:language_chinese      9.982e-02  4.841e-01   2.070
## db_boo:genre_crime:language_chinese       6.043e-01  4.824e-01   1.253
## db_boo:genre_drama:language_chinese       5.704e-02  3.556e-01   0.143
## db_boo:genre_adventure:language_european  1.838e-01  2.765e-01   0.665
## db_boo:genre_animation:language_european  -7.403e-02  2.766e-01  -0.268
## num_votes_imdb_us                         < 2e-16 *** 
## num_votes_db                               < 2e-16 *** 
## year                                      < 2e-16 *** 
## length                                    0.001186 **
## budget                                    0.193219
## genre_adventure                           0.837068
## genre_animation                            0.164780
## genre_biography                            0.418472
## genre_comedy                               0.988765
## genre_crime                                0.406551
## genre_drama                                0.734662
## genre_horror                               0.523605
## genre_mystery                              0.498358
## language_chinese                           0.697785
## language_asian                             0.065448 **
## db_boo:num_votes_imdb_us                   < 2e-16 *** 
## db_boo:num_votes_db                        < 2e-16 *** 
## db_boo:budget                              0.098634 .
## db_boo:genre_adventure                     5.02e-05 ***
## db_boo:genre_animation                     3.85e-05 ***
## db_boo:genre_crime                         82.07e-10 ***
## db_boo:genre_comedy                        1.04e-05 ***
## language_indian:genre_comedy               0.268911
## db_boo:genre_crime                         0.000556 ***
## language_indian:genre_crime                0.160011
## db_boo:genre_drama                       4.29e-12 ***  
## language_indian:genre_drama              0.254027  
## db_boo:genre_horror                      0.024070 *   
## db_boo:genre_mystery                     0.238715  
## genre_comedy:language_chinese            0.486535  
## genre_crime:language_chinese             0.964141  
## genre_drama:language_chinese             0.956133  
## db_boo:language_european                 3.66e-05 *** 
## genre_adventure:language_european         0.528953  
## genre_animation:language_european         0.541380  
## genre_biography:language_european         0.916463  
## genre_comedy:language_european            0.693436  
## genre_crime:language_european             0.216242  
## genre_drama:language_european             0.956133  
## db_boo:language_asian                    0.830594  
## genre_adventure:language_asian            0.874880  
## genre_animation:language_asian            0.510207  
## genre_crime:language_asian                0.935411  
## genre_drama:language_asian                0.344134  
## db_boo:language_indian:genre_comedy       0.033492 *   
## db_boo:language_indian:genre_crime        0.292179  
## db_boo:language_indian:genre_drama        0.027071 *   
## db_boo:genre_comedy:language_chinese      0.835736  
## db_boo:genre_crime:language_chinese       0.210512  
## db_boo:genre_drama:language_chinese       0.886536  
## db_boo:genre_adventure:language_european   0.506281  
## db_boo:genre_animation:language_european   0.789004  
## db_boo:genre_biography:language_european  0.874318  
## db_boo:genre_comedy:language_european     0.845171  
## db_boo:genre_crime:language_european      0.007838 ** 
## db_boo:genre_adventure:language_asian     0.730570  
## db_boo:genre_animation:language_asian     0.700452  
## db_boo:genre_crime:language_asian         0.404358  
## db_boo:genre_drama:language_asian         0.417771  
## db_boo:genre_adventure:language_american   0.495382  
## db_boo:genre_biography:language_american  0.987912  
## db_boo:genre_comedy:language_american      0.395793  
## db_boo:genre_crime:language_american      0.561469  
## db_boo:genre_mystery:language_american    0.598785  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## Residual standard error: 0.3193 on 1474 degrees of freedom 
## Multiple R-squared:  0.5309, Adjusted R-squared:  0.5083 
## F-statistic: 23.49 on 71 and 1474 DF, p-value: < 2.2e-16
4.2 Model using country dummies (instead of language dummies)

High-VIF terms are already removed in the model below:

```R
three_term_nolanguage_lm <- lm(rating~db_boo*num_votes_imdb_us + db_boo*num_votes_db + db_boo*year + db_boo*length + db_boo*budget + db_boo*genre_adventure*country_china + db_boo*genre_animation*country_china + db_boo*genre_biography*country_china + db_boo*genre_comedy*country_china + db_boo*genre_crime*country_china + db_boo*genre_horror*country_china + db_boo*genre_mystery*country_china + db_boo*genre_adventure*country_usa + db_boo*genre_animation*country_usa + db_boo*genre_biography*country_usa + db_boo*genre_comedy*country_usa + db_boo*genre_crime*country_usa + db_boo*genre_horror*country_usa + db_boo*genre_mystery*country_usa + db_boo*genre_adventure*country_india + db_boo*genre_animation*country_india + db_boo*genre_biography*country_india + db_boo*genre_comedy*country_india + db_boo*genre_crime*country_india + db_boo*genre_horror*country_india + db_boo*genre_mystery*country_india + db_boo*genre_adventure*country_asia + db_boo*genre_animation*country_asia + db_boo*genre_comedy*country_asia + db_boo*genre_horror*country_asia + db_boo*genre_mystery*country_asia + db_boo*genre_adventure*country_europe + db_boo*genre_animation*country_europe + db_boo*genre_comedy*country_europe + db_boo*genre_crime*country_europe + db_boo*genre_mystery*country_europe + db_boo*genre_adventure*country_america + db_boo*genre_animation*country_america + db_boo*genre_comedy*country_america + db_boo*genre_horror*country_america + db_boo*genre_mystery*country_america + country_usa - db_boo:year - db_boo:length - db_boo:genre_adventure - db_boo:genre_comedy - db_boo:genre_crime - db_boo:genre_horror - country_usa, data=commondb4)
```

summary(three_term_nolanguage_lm)

84
## Call:
```r
lm(formula = rating ~ db_boo * num_votes_imdb_us + db_boo * num_votes_db +
    db_boo * year + db_boo * length + db_boo * budget + db_boo *
    genre_adventure * country_china + db_boo * genre_animation *
    country_china + db_boo * genre_biography * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_horror * country_china + db_boo *
    genre_mystery * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_horror * country_china + db_boo *
    genre_mystery * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
    country_china + db_boo * genre_animation * country_china +
    db_boo * genre_comedy * country_china + db_boo * genre_adventure *
```

# Residuals:
##    Min   1Q Median   3Q     Max
## -1.66494 -0.19113  0.00377  0.20021  1.15893

# Coefficients:
##                Estimate Std. Error t value
## (Intercept)    2.320e+01  8.892e-01  26.095
## num_votes_imdb_us  3.713e-06  2.971e-07  12.500
## num_votes_db      9.023e-08  7.982e-08   1.130
## year             -7.765e-03  4.480e-04 -17.333
## length           1.365e-03  4.808e-04   3.581
## budget           -4.863e-04  1.488e-04  -3.268
## country_china    -2.583e-02  1.016e-01  -0.254
## country_india    1.351e-02  1.371e-01   0.099
## country_america  2.064e-02  8.477e-02   0.244
## db Boo:num_votes_imdb_us -3.548e-06  4.065e-07  -8.729
## db Boo:num_votes_db    1.417e-06  1.112e-07  12.738
## db Boo:budget       8.935e-04  1.875e-04   4.765
## db Boo:country_china -3.325e-01  1.428e-01  -2.328
## country_china:genre_biography -2.123e-01  3.461e-01  -0.614
## country_china:genre_comedy   2.149e-01  3.462e-01   0.621
## country_china:genre_crime    1.167e-02  3.473e-01   0.034
## genre_adventure:country_usa -7.756e-02  5.779e-02  -1.342
## genre_animation:country_usa  7.648e-03  5.567e-02   0.137
## genre_biography:country_usa -3.282e-02  5.228e-02  -0.621
## genre_comedy:country_usa    -7.787e-02  4.715e-02  -1.652
## genre_crime:country_usa     -8.503e-02  4.612e-02  -1.844
## genre_horror:country_usa    6.843e-02  1.366e-01   0.501
## genre_mystery:country_usa   -1.265e-01  1.187e-01  -1.065
## db Boo:country_india      5.399e-02  1.922e-01   0.281
## genre_comedy:country_india  1.000e-01  1.844e-01   0.057
## genre_crime:country_india  6.534e-02  2.342e-01   0.279
## db Boo:country_uk         3.452e-02  7.190e-02   0.480
## genre_adventure:country_uk 5.327e-02  1.064e-01   1.950
## genre_animation:country_uk 7.483e-03  1.061e-01   0.016
## genre_comedy:country_uk   4.835e-03  1.266e-01   0.038
## genre_crime:country_uk    5.785e-02  1.666e-01   0.347
## genre_horror:country_uk   3.172e-02  3.324e-01   0.955
## db Boo:country_europe    3.742e-01  4.709e-02   7.947
## genre_adventure:country_europe 2.304e-03  1.064e-01   0.014
## genre_animation:country_europe 1.435e-01  1.280e-01   1.121
## genre_biography:country_europe 5.198e-02  9.348e-02   0.198
## genre_comedy:country_europe 8.848e-02  9.079e-02  -0.975
## genre_crime:country_europe 8.494e-02  8.457e-02  -1.004
## db Boo:country_asia     2.797e-01  6.936e-02   4.033
## genre_adventure:country_asia 2.260e-01  2.349e-01   0.962
## genre_animation:country_asia            1.855e-01  6.789e-02   2.732
## genre_comedy:country_asia              -1.407e-01  2.352e-01  -0.598
## genre_crime:country_asia                1.214e-01  1.665e-01   0.729
## genre_horror:country_asia               8.997e-02  3.316e-01   0.271
## db_boo:country_america                  2.414e-01  1.186e-01   2.036
## genre_adventure:country_america        -2.037e-02  2.482e-01  -0.082
## genre_biography:country_america         4.251e-02  3.412e-01   0.125
## genre_comedy:country_america            1.165e-01  1.850e-01   0.630
## genre_crime:country_america             4.431e-01  2.489e-01   1.780
## genre_mystery:country_america           -1.834e-01  3.410e-01  -0.538
## db_boo:country_china:genre_biography    1.207e+00  4.886e-01   2.471
## db_boo:country_china:genre_comedy       3.250e-01  4.894e-01   0.664
## db_boo:country_china:genre_crime        6.684e-01  4.909e-01   1.362
## db_boo:genre_adventure:country_usa      1.864e-01  8.107e-02   2.299
## db_boo:genre_animation:country_usa      1.604e-01  7.679e-02   2.088
## db_boo:genre_biography:country_usa      3.101e-01  7.291e-02   4.253
## db_boo:genre_comedy:country_usa         1.764e-01  6.521e-02   2.706
## db_boo:genre_crime:country_usa          1.155e-01  6.498e-02   1.778
## db_boo:genre_horror:country_usa        -5.921e-01  1.922e-01  -3.081
## db_boo:genre_adventure:country_uk       1.389e-01  1.642e-01   0.846
## db_boo:genre_biography:country_uk       2.435e-01  1.641e-01   1.484
## db_boo:genre_comedy:country_uk          2.392e-01  1.910e-01   1.252
## db_boo:genre_crime:country_uk           -2.841e-02  2.443e-01  -0.116
## db_boo:genre_adventure:country_asia    -2.091e-01  3.381e-01  -0.618
## db_boo:genre_animation:country_asia     1.186e-02  1.148e-01   0.103
## db_boo:genre_comedy:country_asia       -7.333e-03  3.385e-01  -0.022
## db_boo:genre_crime:country_asia        -3.361e-01  2.437e-01  -1.379
## db_boo:genre_horror:country_asia       -8.152e-01  4.730e-01  -1.723
## db_boo:genre_adventure:country_america  1.891e-01  3.508e-01   0.539
## db_boo:genre_biography:country_america  1.500e-01  4.823e-01   0.311
## db_boo:genre_comedy:country_america    -9.497e-01  2.616e-01  -3.563
## db_boo:genre_crime:country_america     -2.784e-01  3.518e-01  -0.791
## db_boo:genre_mystery:country_america   -4.023e-01  4.821e-01  -0.834
## (Intercept)                             < 2e-16 ***
## num_votes_imdb_us                        < 2e-16 ***
## num_votes_db                             0.258493
## year                                    < 2e-16 ***
## length                                  0.000354 ***
## budget                                  0.001109 **
## country_china                           0.0799377
## country_india                           0.921536
## country_america                         0.807644
## db_boo:num_votes_imdb_us                < 2e-16 ***
## db_boo:num_votes_db                      < 2e-16 ***
## db_boo:budget 2.07e-06 ***
## db_boo:country_china 0.020034 *
## country_china:genre_biography 0.539635
## country_china:genre_comedy 0.534886
## country_china:genre_crime 0.973201
## genre_adventure:country_usa 0.179784
## genre_animation:country_usa 0.890761
## genre_biography:country_usa 0.530282
## genre_comedy:country_usa 0.098847 .
## genre_crime:country_usa 0.065409 .
## genre_horror:country_usa 0.616577
## genre_mystery:country_usa 0.286832
## db_boo:country_india 0.778782
## genre_comedy:country_india 0.233673
## genre_crime:country_india 0.780268
## db_boo:country_usa 0.631209
## genre_adventure:country_usa 0.232460
## genre_biography:country_usa 0.986865
## genre_comedy:country_usa 0.969533
## genre_crime:country_usa 0.728416
## genre_horror:country_usa 0.923975
## db_boo:country_europe 3.78e-15 ***
## genre_adventure:country_europe 0.988961
## genre_animation:country_europe 0.262618
## genre_biography:country_europe 0.843340
## genre_comedy:country_europe 0.329961
## genre_crime:country_europe 0.315370
## db_boo:country_asia 5.79e-05 ***
## genre_adventure:country_asia 0.336112
## genre_animation:country_asia 0.006374 **
## genre_comedy:country_asia 0.549899
## genre_crime:country_asia 0.465857
## genre_horror:country_asia 0.786154
## db_boo:country_america 0.041956 *
## genre_adventure:country_america 0.934595
## genre_biography:country_america 0.900857
## genre_comedy:country_america 0.528769
## genre_crime:country_america 0.075264 .
## genre_mystery:country_america 0.590793
## db_boo:country_china:genre_biography 0.013578 *
## db_boo:country_china:genre_comedy 0.506810
## db_boo:country_china:genre_crime 0.173560
## db_boo:genre_adventure:country_usa 0.021629 *
## db_boo:genre_animation:country_usa 0.036943 *
## db_boo:genre_biography:country_usa 2.24e-05 ***
## db_boo:genre_comedy:country_usa 0.006891 **
## db_boo:genre_crime:country_usa 0.075602 .
## db_boo:genre_horror:country_usa 0.002099 **
## db_boo:genre_mystery:country_usa 0.575105
## db_boo:genre_comedy:country_india 0.859836 .
## db_boo:genre_crime:country_india 0.465003
## db_boo:genre_adventure:country_uk 0.397681
## db_boo:genre_biography:country_uk 0.138032
## db_boo:genre_comedy:country_uk 0.210740
## db_boo:genre_crime:country_uk          0.907437
## db_boo:genre_horror:country_uk         0.714433
## db_boo:genre_adventure:country_europe  0.416279
## db_boo:genre_animation:country_europe  0.855266
## db_boo:genre_comedy:country_europe     0.545170
## db_boo:genre_crime:country_europe      0.015037 *
## db_boo:genre_adventure:country_asia    0.536421
## db_boo:genre_animation:country_asia    0.917718
## db_boo:genre_comedy:country_asia       0.982721
## db_boo:genre_crime:country_asia        0.168156
## db_boo:genre_horror:country_asia       0.085028 .
## db_boo:genre_adventure:country_america 0.589987
## db_boo:genre_biography:country_america 0.755930
## db_boo:genre_comedy:country_america    0.716606
## db_boo:genre_crime:country_america     0.428966
## db_boo:genre_mystery:country_america   0.404187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.3307 on 1464 degrees of freedom
## Multiple R-squared:  0.5003, Adjusted R-squared:  0.4727
## F-statistic:  18.1 on 81 and 1464 DF,  p-value: < 2.2e-16