Project 1

Frank Liu & David Qu

Apr 10, 2020

Abstract

The goal of this project is to analyze and build a model for the prediction of the box office of certain movies by using several predictors. We remove the predictors that cannot be used and that are insignificant, clean not available data, transform some of the variables, check for multicolliearity, adjust levels of several categorical data, and change several numeric variables to categorical variables to get a rough model. We use both Best Subset Selection and Nested-F test to finalize our model. And our model is shown as following:

$$log\hat{G}ross = -1.76934 + 1.00372 \times logVoted + 0.21554 \times logUserReviews + 0.36638 \times logBudget + 0.74854 \times imdb_score - 0.07079 \times imdb_score^2 - 0.20929 \times AspectRatio2.35above + 0.41048 \times ratingPG - 0.49100 \times ratingR + 1.11785 \times GenresDocumentary + 0.76935 \times GenresHorror + 0.27583 \times GenresOther - 1.08641 \times LanguageNonEnglish + 0.62023 \times Countryyes - 0.03972 \times logCriticReviews : logVoted + 0.04874 \times logCriticReviews$$

Conclusion: Although this model has a high R^2_{adj} of 0.6365, it fails to achieve the assumption of constant variance and normality of residuals. We should be careful to use this model to predict the box office of movies, but it provides us certain insights about movie-related datasets.

Data Cleaning

Firstly, log transformation is applied to our response variable, gross. This is because the scale of the data is so large, and it can only take positive values. Typically we will log the response variables if it is money.

We then removed 7 predictors, including director_name, actor_1_name, actor_2_name, actor_3_name, plot_keywords, movie_imdb_link, movie_title. We removed movie title and imdb link because each film has unique title and link, so it could not be a explanatory variable. We also removed the names of directors and actors. Although the famous directors or actors may effect the box office, we have too many different names in this variable. We removed plot keywords with a similar reason. Most films have different keywords. If we do not remove them, we will be having thousands levels of categorical variables, which will make our analysis super complicated.

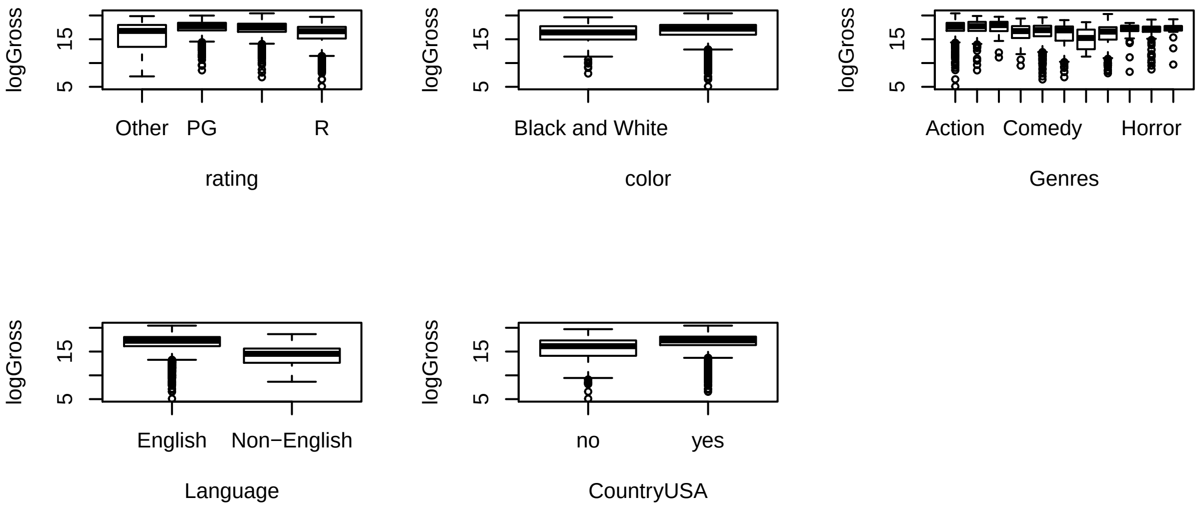
We removed all the individuals with missing data. And we also removed a single data point which also contains missing information. There are only 3753 objects in our new dataset. Comparing to 4674 objects in original dataset, there are 921 data points with missing information (N/A).

Exploratory Data Analysis

Categorical Variables:

We have five categorical variables in our data, but some of the categorical variables have levels with only few data points. We consider this has the potential of over-fitting (tuning the model to suit a few data points). Therefore, we consider to ignore or combine some particular levels.

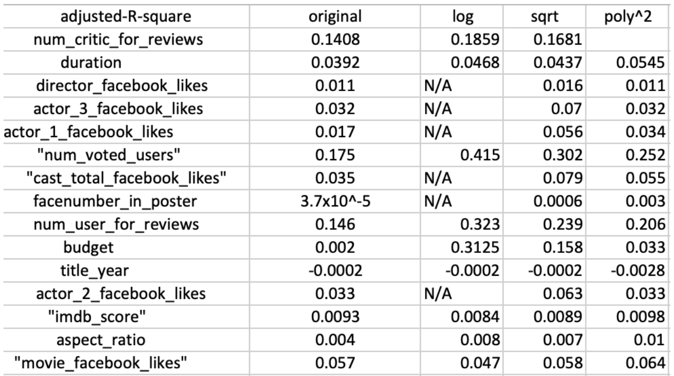
Content Rating: Most of the data points are PG, PG-13, R. They are also three standard levels of MPAA rating system. Therefore, we combine G and the rest of levels to “Others”. Color: Although the majority of data points are “color”, the “black and white” still has more than 100 data points. Therefore, we are trying to save this variable. Genres: There are multiple levels for this variable. However, the data points are distributed in all levels evenly. Therefore, we do not change the level for this categorical variable. Language: There are many languages in this data set, but roughly above 90% of them are “English”. Instead of using each language as a level of predictor variable, we want an indicator variable for English. We create a binary variable that takes on the value “English” for English movies and “Non-English” for other languages. Country: This variable is similar to language. Most of the films are made in USA. Therefore, we create a binary variable that takes on the value “USA” for American movies and “Foreign” for movies made in other countries.



After plotting the categorical variables against the response variable, response variable values seem to be considerably different for different levels of each predictor, this is a good indication that this categorical variables may be an important predictor to consider. We will not delete any categorical variable. Also, note that all USA films speak English, so Language and Country might be strongly correlated. We decide to add an interaction between these two variables.

Numerical Variables:

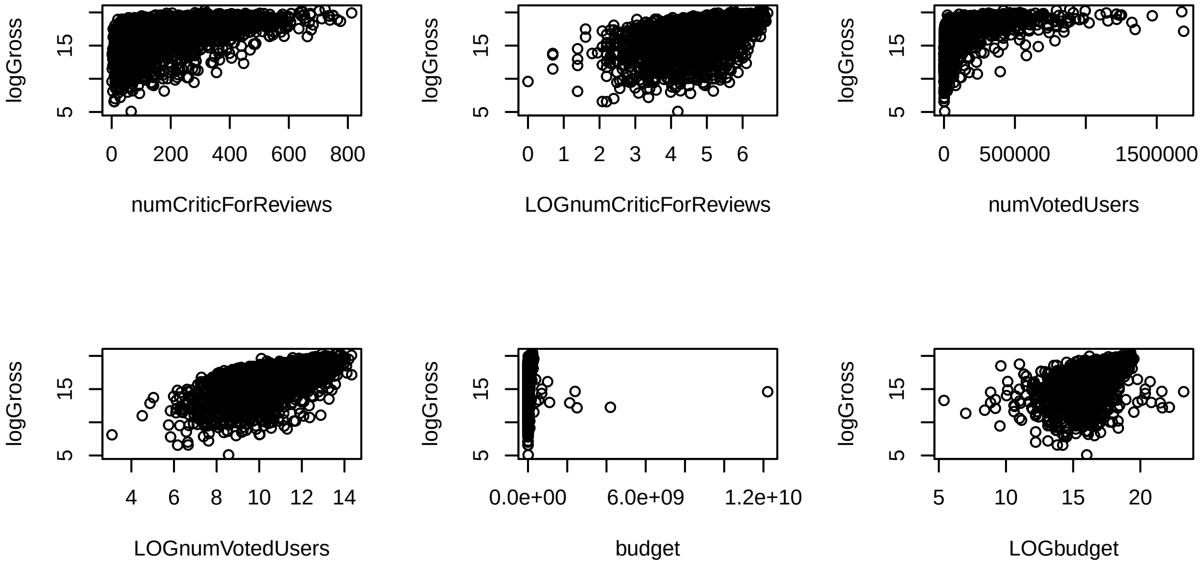
Visualization and Transformation



adjusted-R-square	original	log	sqrt	poly^2
num_critic_for_reviews	0.1408	0.1859	0.1681	
duration	0.0392	0.0468	0.0437	0.0545
director_facebook_likes	0.011	N/A	0.016	0.011
actor_3_facebook_likes	0.032	N/A	0.07	0.032
actor_1_facebook_likes	0.017	N/A	0.056	0.034
"num_voted_users"	0.175	0.415	0.302	0.252
"cast_total_facebook_likes"	0.035	N/A	0.079	0.055
facenumber_in_poster	3.7x10^-5	N/A	0.0006	0.003
num_user_for_reviews	0.146	0.323	0.239	0.206
budget	0.002	0.3125	0.158	0.033
title_year	-0.0002	-0.0002	-0.0002	-0.0028
actor_2_facebook_likes	0.033	N/A	0.063	0.033
"imdb_score"	0.0093	0.0084	0.0089	0.0098
aspect_ratio	0.004	0.008	0.007	0.01
"movie_facebook_likes"	0.057	0.047	0.058	0.064

After plotting very explanatory variable with different transformation, we decide to delete the year variable, because it has a negative adjusted R^2 for all different kinds of transformation we try. It means it makes the model even worse if we add this variable.

Also, we decide to delete all the variables related to facebook likes except the cast total facebook likes, director_facebook_likes, and movie facebook likes, and there are several reasons. First, these variables are likely to be correlated with each other. It is intuitive that the cast total face book likes is the sum of all the facebook likes of the actors. Second, we observed a great amount data points with missing information, which means, the number of their facebook likes is 0. Although the cast total facebook likes only has one data point is 0, other variables have hundreds or thousands of 0. This greatly effects our regression and it will cost us too much to remove all these data points. So we either remove it or convert it to a categorical variable.

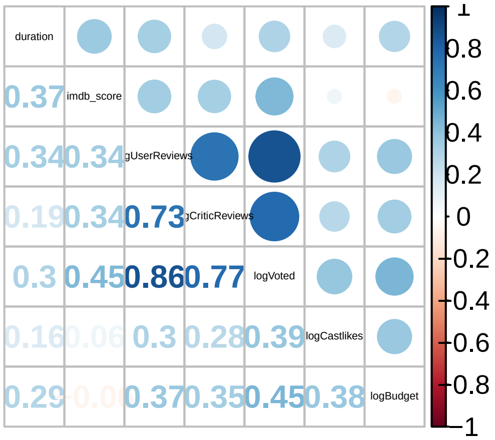


As we visualized it in our graph above, we decide to apply log transformation to number_critic_for_reviews, num_voted_users, cast_total_facebook_likes, num_user_for_reviews, and budget, because they all have the highest R^2_{adj} for log transformation.

We decide to apply polynomial transformation to duration and imdb_score, because they have the highest R^2_{adj} for polynomial transformation.

For aspect_ratio, facenumber_in_poster, director_facebook_likes and movie_facebook_likes, we decide to convert these variables to categorical. We convert director_facebook_likes and movie_facebook_likes because it has lots of missing data for old movies. Therefore, we will divide it into three levels, which are none, low, and high. We convert the other two because most films have the same numerical value.

Multicolliearity



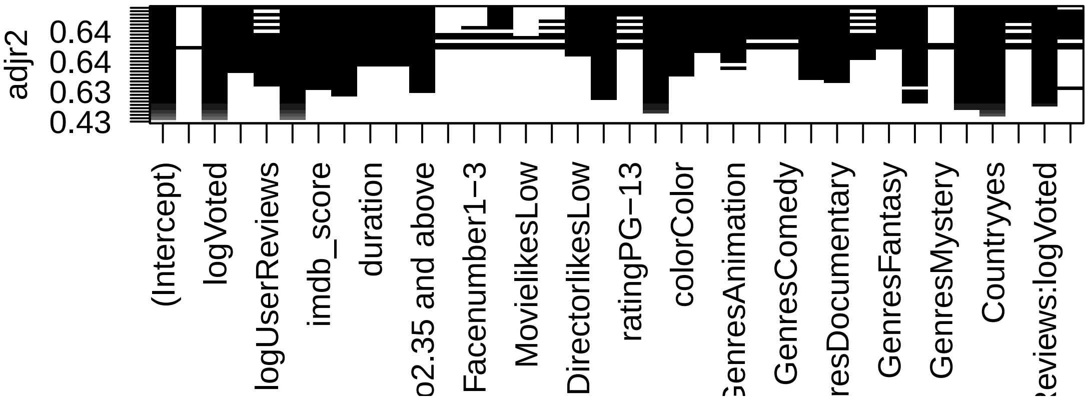
We set our multicollinearity cutoff to 0.6. We found a large correlation between logVoted, logCriticReviews, and logUserReviews. Therefore, we assume that there exists multicollinearity, and we add interaction terms between these variables. Therefore, our final rough model contains 16 variables: imdb_score, duration, AspectRatio, Facenumber, Movielikes, Directorlikes, rating, color, genres, Language, Country, logCriticReviews, logVoted, logCastlikes, logUserReviews, Budget (35 predictors including levels and interaction).

Outliers

We perform the simple linear model for each of the transformed numerical variable. We found that there are 29 data points are extreme outliers. After looking at these data points, we decide to remove those data points. Most of these data points have a extremely large negative residual, which means that we over-estimate their box office. The main reason is that these films may not be made for commercial purposes, or lots of them are not even available in the cinema. For example, there are some award-wining films made by fairly famous directors with a fair amount of reviews, but the film was not liked by the majority. Also, some films are just made for online purposes. Most of these non-commercial films are inappropriate for our sample. Therefore, it is reasonable to ignore these outliers.

Model Selection

In this process, we refine our model and decide final variables to put in our model. We apply the technique of best subset selection (BSS). We fit a separate least squares regression for every subset of predictors. We then look at all of the models, and use some metric to choose among them. We have 35 predictors in total, we build every possible model with these predictors.

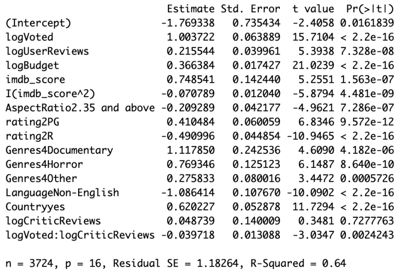


We look at the models with highest R^2_{adj}. However, the graph tells us that there are tons of models with similar R^2_{adj}. Since a smaller model is easier to interpret, we will look at those high R^2_{adj} models with less predictors. We will compare 4 smallest models with a R^2_{adj} higher than 0.63.

```
## Analysis of Variance Table
##
## Model 1: logGross ~ logVoted + logUserReviews + logBudget + imdb_score +
##     I(imdb_score^2) + AspectRatio + rating2 + Genres2 + Language +
##     Country + logCriticReviews:logVoted + logCriticReviews
## Model 2: logGross ~ logVoted + logUserReviews + logBudget + imdb_score +
##     I(imdb_score^2) + AspectRatio + rating2 + Genres3 + Language +
##     Country + logCriticReviews:logVoted + logCriticReviews
## Model 3: logGross ~ logVoted + logUserReviews + logBudget + imdb_score +
##     I(imdb_score^2) + AspectRatio + rating2 + Genres4 + Language +
##     Country + logCriticReviews + logCriticReviews:logVoted
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3710 5222.3
## 2   3709 5202.8  1    19.444 13.902 0.0001955 ***
## 3   3708 5186.2  1    16.621 11.883 0.0005726 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The smallest model has 11 predictors. However, since is has an interaction term, we have to add the original variable back. Therefore, we have 13 variables in total. The best model with 12 variables also has an interaction term, so it has 13 variables. By looking at the p-values of their predictors, we found that many predictors for the first model lacks evidence to be significant. Therefore, we decide to choose the second model. By applying anova analysis, we found that the model with 15 predictors is better, because the additional predictors add a non-zero R^2_{adj} to the model. Although the increase in R^2_{adj} is not high, we still want to add these predictors because they are the levels for genres. We think if we make a level with a small amount of data points to be indicator could result in over-fitting. However, it is possible they have some strong attributes that makes them very different. We should either treat them as outliers or put other levels in our model. Therefore, we will choose the largest model, which has 15 predictors. Hence, our final model is

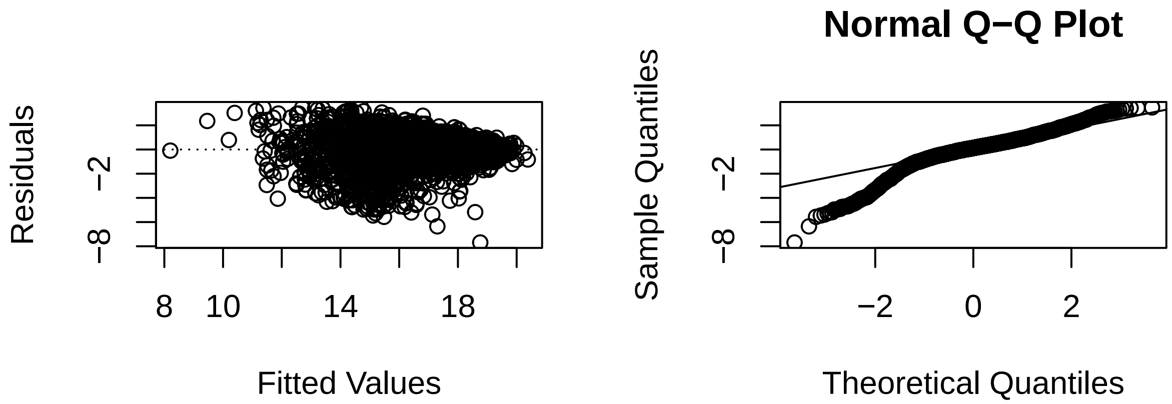
$$\hat{logGross} = -1.76934 + 1.00372 \times logVoted + 0.21554 \times logUserReviews + 0.36638 \times logBudget + 0.74854 \times imdb_score - 0.07079 \times imdb_score^2 - 0.20929 \times AspectRatio2.35above + 0.41048 \times ratingPG - 0.49100 \times ratingR + 1.11785 \times GenresDocumentary + 0.76935 \times GenresHorror + 0.27583 \times GenresOther - 1.08641 \times LanguageNonEnglish + 0.62023 \times Countryyes - 0.03972 \times logCriticReviews : logVoted + 0.04874 \times logCriticReviews$$



```
                             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                 -1.769338   0.735434  -2.4058 0.0161839
logVoted                     1.003722   0.063889  15.7104 < 2.2e-16
logUserReviews               0.215544   0.039961   5.3938 7.328e-08
logBudget                    0.366384   0.017427  21.0239 < 2.2e-16
imdb_score                   0.748541   0.142440   5.2551 1.563e-07
I(imdb_score^2)             -0.070789   0.012040  -5.8794 4.481e-09
AspectRatio2.35 and above   -0.209289   0.042177  -4.9621 7.286e-07
rating2PG                    0.410484   0.060059   6.8346 9.572e-12
rating2R                    -0.490996   0.044854 -10.9465 < 2.2e-16
Genres4Documentary           1.117850   0.242536   4.6090 4.182e-06
Genres4Horror                0.769346   0.125123   6.1487 8.640e-10
Genres4Other                 0.275833   0.080016   3.4472 0.0005726
LanguageNon-English         -1.086414   0.107670 -10.0902 < 2.2e-16
Countryyes                   0.620227   0.052878  11.7294 < 2.2e-16
logCriticReviews             0.048739   0.140009   0.3481 0.7277763
logVoted:logCriticReviews   -0.039718   0.013088  -3.0347 0.0024243

n = 3724, p = 16, Residual SE = 1.18264, R-Squared = 0.64
```

The baseline for rating is other; the baseline for Genres is Crime; the baseline for Language is English; the baseline for Country is no (Non-USA). The R^2_{adj} is 0.6365.



Conditions for Inference

Now we have selected our final model, which is the largest model that have 15 predictors. We are now going to check the conditions of inference to see if additional transformation was needed.

Linearity: We have made scatter plots for response variables and each numeric variables that are in our model. Each explanatory variables are linearly related to the response variable.

Zero Mean: The mean of residuals is almost zero, which means we meet the condition of zero mean.

Constant Variance: By seeing from the graph, we can see that the data is more spread out in the beginning and start to crowd together as fitted value increases. We don't meet the condition for this inference, and probably this is due to the diverge range of data point when the log gross is ow.

Normality: By seeing from our qqplot, we can conclude that the assumption of normality is not achieved.The residuals follow a bell-shape distribution, but is heavily skewed to the left.

Independence and Randomness: This dataset is movies in the US released from 1916 to 2016, it is safe to assume this is a random sample and the residuals are independent from one another.

Analysis

After deciding our final model, we find that our response variable, logGross, is related to log number voted, log user reviews, log budget, imdb score, the square of imdb score, aspect ratio, movie rating, genres, country, language, and log critic reviews. In particular, an aspect ratio of 2.35 and above, a rating of R and other, the square of imdb score, the language of non-English has a negative correlation with the total box office. The rest of variables relate to a higher box office. It is interesting that a high imdb score doesn't actually guarantee the box office, because a great amount of good movies are not made for commercial purposes. It is counter-intuitive that the aspect ratio of 2.35 and above has a negative correlation, because people should prefer a wider screen. I think this categorical variable is likely to be correlated with some numerical variable. Likewise, documentary has a large positive correlation, but we can see from previous plot that these films actually have a lower box office than average. These categorical variables probably contain multicollinearity with other numerical variables. Some other predictors match with our intuition. For example, an R-rated movie often has a lower box office; the popularity of a film (number voted and user reviews) is positively correlated with the box office.

Although our model has a high R^2_{adj} of 0.6365, I still think it is not a perfect model and we have to be careful to make any conclusion, because we fail to match all the conditions of linear regression. We've observed that the variance of residuals is higher for films with lower box office. I think this divergence of low box office films could be explained by intuition. There are lots of non-commercial films in this dataset, which has a huge influence on our final model. These films probably are not available in cinema, but they could be famous award-winning films in festivals like Cannes or Venice. This means we need more predictors. For example, the information about the film's presence in a particular film festival could be helpful. There could also be films with extremely low budget and popularity yet achieving a high box office. For example, some independence films like *The Purge* could be favored by the public. I think a variable about the release date will be helpful because lots of high box office films are released during summer. The cinema schedule and the film's marketing campaign could be influential as well. The premiere of a film is very important to their first-week box office. Therefore, there are many potentially useful variables.