Project 1

Coral Chen and Jessical Li

Apr 9, 2020

Abstract

In this project, our goal is to construct a multiple linear regression to explain the money theater earned by films with categorical and numerical explanatory variables, and to figure out possible reasons and contribution that each predictor could make. We start the model selection process by plotting EDA for each numerical explanatory variable against the response variabland ensure the most suitable transformation to use, and we delete insignificant levels for categorical variables. We then use the correlation matrix to discover correlation between explanatory variables and their interactions, and construct a rough model at this point. Next step we continue using the BSS to select seven models with different predictors in high R sequared range, and then use nested F-test to check which predictors to keep and named this model as the refined model. We checked the correlation and the p-value of the refined model to see whether any more predictors need to remove. Summing up, the final model is selected and to be used for interpretation. Our final model is $\widehat{log(Gross)} = 0.319colorColor + 0.258genresComedy - 0.276ratingPG - 13 - 0.744ratingR + 1.545Englishyes + 0.709USAyes - 0.0416titleYear - 0.611logCriticReviews + 1.376lognumVotedUsers - 0.000432sqrtCastFacebookLikes + 0.731lognumUserReviews + 0.427logBudget - 0.132lognumVotedUsers : lognumUserReviews + 0.150logNumCriticReviews : lognumUserReviews$ During our exploration, we found out that 12 variables and 2 interactions are important for predicting the money made in theater by a movie. We found that colored film, genres of comedy, English-language film, film made in the USA, number of voted user, number of user reviews and budget, are positively related and contributes to the gross box office of a movie, while the rating as PG-13, rating of R, title year, critic reviews, total cast facebook likes are inverse correlates with it. To make a movie with high box office, we would want it to be a PG-rated comedy in color made in the USA with language as English, and have as more number of voted user, the number of user reviews and budget as possible, and as small title year, critic reviews, and total cast Facebook likes as possible.

Data Cleaning

To prepare the dataset ready for further analysis, we decide to remove the missing data as the first step. We found 934 missing data points, and we removed them by using the code na.omit. Furthermore, we removed several unnecessary explanatory variables in order to ensure better accuracy in our EDA. We choose to remove director name[Column 2], actor2 name [7], actor1 name[11], movie title[12], actor3 name[15], plot keywords[17], and movie link[18].There are two kinds of variables we decided to remove. The first ones are names and the keywords because there are too many different kinds of them which do not help us to sort the data and form the model. The second kind we want to delete is the movie link because it does not relate to the response variable, and including it in the data can only complicate the analysis.

Exploratory Data Analysis (EDA)

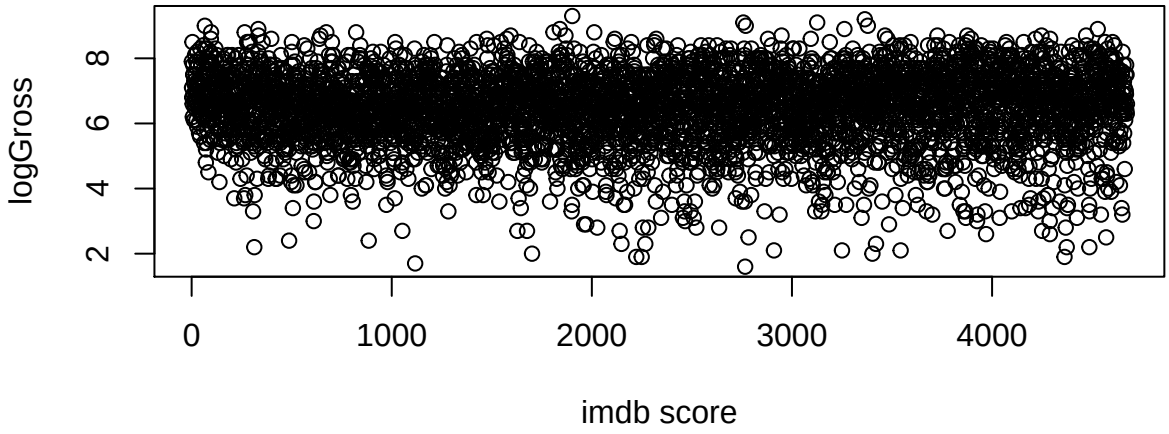
In the EDA process, we first log-transformed the response variable gross" because in convention it is normally log-transformed money related response variable, so as the variable “gross”, and it is conducive for more concise model construction.

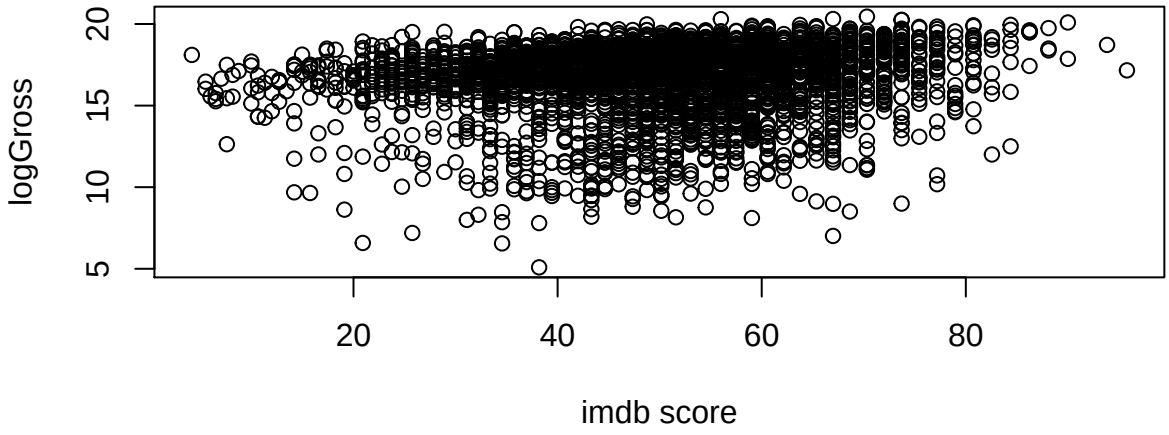
We made a correlation matrix to test the multicollinearity and found there are 8 really strong correlation, for which we use 0.6 as a multicollinearity cutoff: (1) lognum_voted_users & lognum_critic_for_reviews 0.77424736 (2) lognum_user_for_reviews & logNum_critic_for_reviews 0.77424736 (3) lognum_voted_users & lognum_user_for_reviews 0.862194677 (4) sqrtCast_total_facebook_likes & sqrtActor_3_facebook_likes 0.62709898 (5) sqrtActor_2_facebook_likes & sqrtActor_3_facebook_likes 0.68871420 (6) sqrtCast_total_facebook_likes & sqrtActor_1_facebook_likes 0.96307481 (7) sqrtActor_2_facebook_likes & sqrtActor_1_facebook_likes 0.58859505 (8) sqrtCast_total_facebook_likes & sqrtActor_2_facebook_likes 0.75574431. This means that we need to add the interactions between these pairs when constructing models.

For numerical variables, we plot each planatory variable against the response variable. We then transformed each plot, respectively, with square root transformation, log transformation, and second power polynomial

transformation on each explanatory variable, and compared them with the original plot by the calculated adjusted r square value for each. In this way, we choose the transformation, which has the highest adjusted r square value as the best fit model that is used for further analysis.

Take imdb score as an example, the original graph shown to be very unclear, and then we take on three types of the transformation to it. We found the adjusted R^2 for original, square root, log, and the second power polynomial transformation(p2) respectively are 0.00902, 0.00813, 0.00864, 0.00954, so we decide to go with the p2 transformation, which also helps better improved the linearlity.



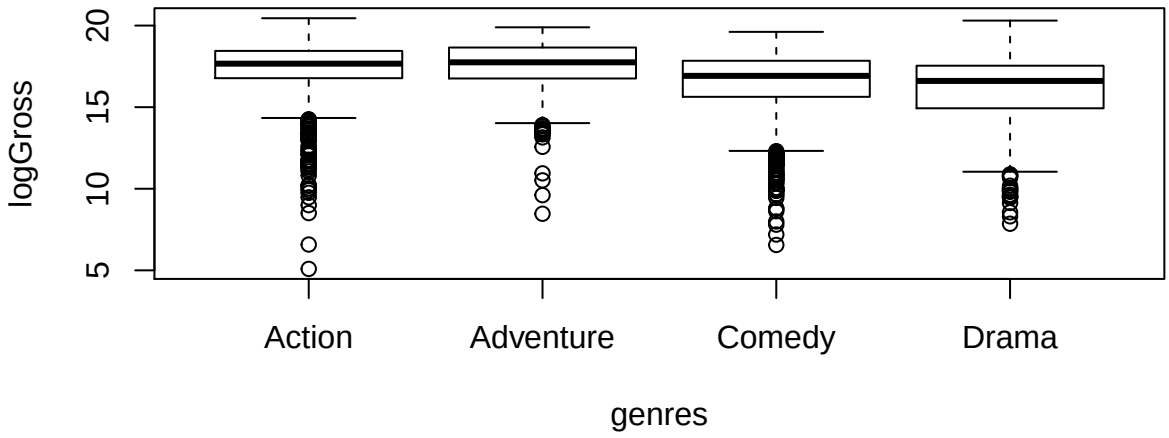


After comparing between three kind of transformation, we decided that number_of_critic_for_review, num_voted_users, num_user_for_reviews, and budget need log transformation; actor_3_facebook_likes, actor_1_facebook_likes, actor_2_facebook_likes, cast_facebook_likes,movie_facebook_likes, and facenumber_in_poster need squared root transformation; duration, director_facebook_likes, imdb_score, and aspect_ratio need polynomial transformation.

In this dataset, there are quite a few outliers. For numerical variables, all the Facebook likes (actors, cast) have some outliers at the far right of the graph, and it makes sense because some actors can be much more famous than others and some people do not use Facebook at all. This is a problem.

After comparing with the correlation matrix, we realized that there are strong correlations between actor1,2,3 and cast facebook likes, besides, in common sense, we would think the cast total Facebook likes covers each actor's Facebook likes. To solve the outliers problem and to simplify the further modeling as the five interactions between is too complicated to compute, we decided to use Cast_total_facebook_likes to represent actors facebook like and delete the explanatory variable as actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes.

For categorical variables, we at first did not change anything and run the rough model, and we figured that there are too many insignificant levels in some of the categorical variables (language and country). Therefore, we decided to use droplevels function to delete those levels which have less than 5 observations in country, language, content_rating, and genres. This change did help our adjusted R squared to be higher.



At this point, we built a model that includes all of the modified categorical variables(dropped levels) and the transformed numerical variables to have a rough overview on the accuracy of the modifications and the transformations and the necessity for each predictors. We named this model as "RoughModel". The regression line for RoughModel is:

$\widehat{log(Gross)} = color + genres + rating + language + country + titleYear + logNumCriticReviews + p2duration + p2directorFacebook + lognumVotedUsers + sqrtCastFacebookLikes + sqrtFaceInPoster + lognumUserReviews + logBudget + p2imdbScore + p2aspectRatio + sqrtMovieFacebook + logNum_criticfor_reviews : lognum_votedusers + lognum_votedusers : lognum_userfor_reviews + logNum_criticfor_reviews : lognum_userfor_reviews$

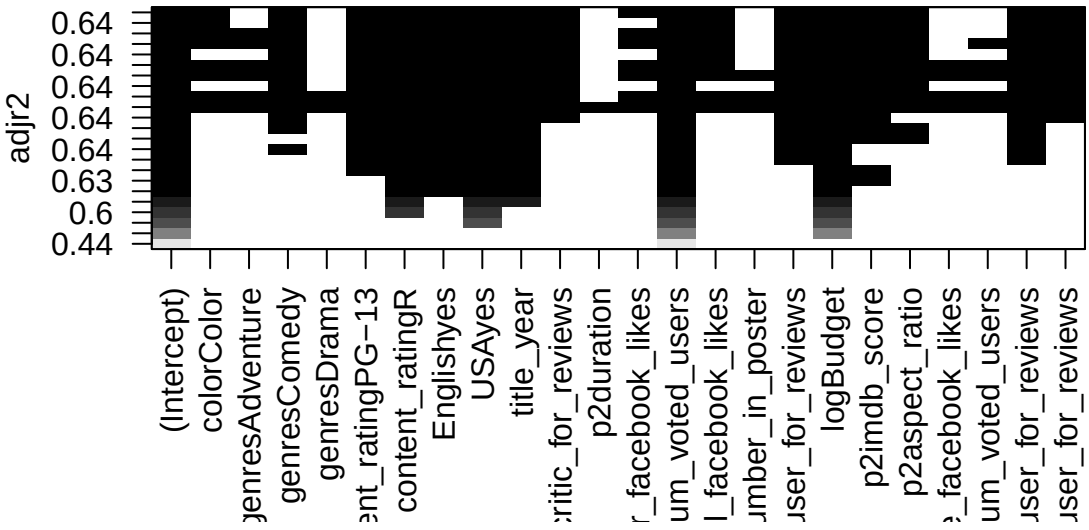
After summarizes the rough model and computes for relevant indicators,we found 20 predictors have convincing evidence too be kept in the model if using a=0.05 as the p value cutoff. However, as this result cut off too many predictors and this is only the beginning stage of our model selection, so we decide to keep unconvincing predictors by t-test first and then cross-compare with the result coming out of the BSS selection to decide whether we should keep or delete a certain variable.

Model Selection

In this section, We choose to do both nested F-tests and subset selection because after we make the subset selection, there are many model with the similar adjusted R^2 value, so we decided to use nested F-tests and compare the most suited ones.

To better perform a BSS, we respectively change the variable "language" and "country" into two binary terms as "English or non-English" and "USA and non-USA." We did this out of two reasons: the first one is that we found BSS couldn't run with language and country as it is too complicated to run with so much data points, so we need some simplifications; the second one is that there are many levels within these two categorical variables which have very little data points, and the level "English" and "USA" have significantly more data points compared to the rest of the levels, so we decided to make them a predictor rather than language and country.

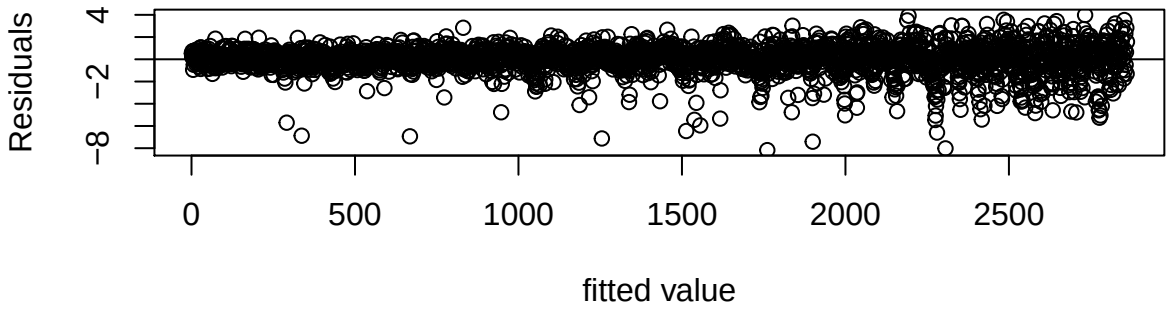
With this modification, we perform the BSS techniques and get a group of 35 subsets with different variables, and there are 22 models of them that have the adjusted r square rounded up to be 0.63. We selected the 13th, 17th, 20th, 22nd, 26th, 30th, 35th model and named them as Model 1 - Model 7 to perform the nested F-tests to see which predictors are rather significant. By comparing the p-value for each model, we decided to use model 6 as its p-value of 0.001793 shows convincing evidence to have these more predictors, and the p-value of model 7 0.893181 shows and compelling evidence to add any more.Below are the variables in our refined model: colorColor, genresAdventure, genresComedy,genresDrama,content_ratingPG-13, content_ratingR, Englishyes, USAyes,title_year, logNum_critic_for_reviews, lognum_voted_users, sqrtCast_total_facebook_likes, lognum_user_for_reviews, logBudget, lognum_voted_users:lognum_user_for_reviews, logNum_critic_for_reviews:lognum_

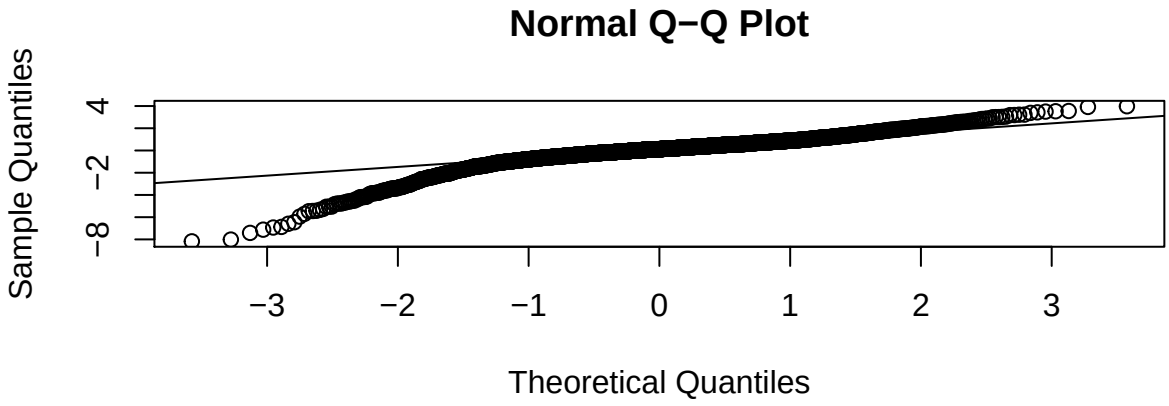


Conditions for Inference

Linearity: This condition is satisfied because We made sure the linearity for each numerical variable in the EDA section by plotting and transforming. 2.Zero means: This condition is satisfied as we calculated the mean value to be 1.848957e-17, which is very close to being 0. 3. Constant variability: This condition is violated because we find from graph that each data points doesn't have a constant distance to the zero lines.4.Normality: This

condition is violated because the QQ plot indicates that the right part deviates from the baseline, and the data falls below to a certain extent, which means that the model is left-skewed and not follow a normal distribution.

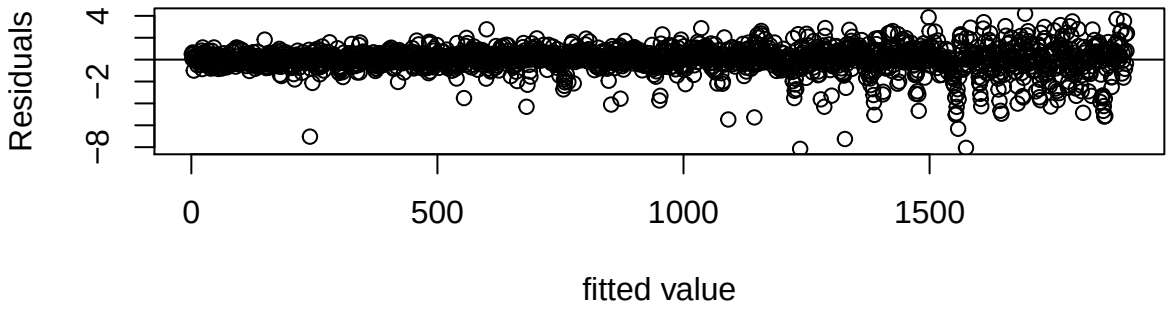


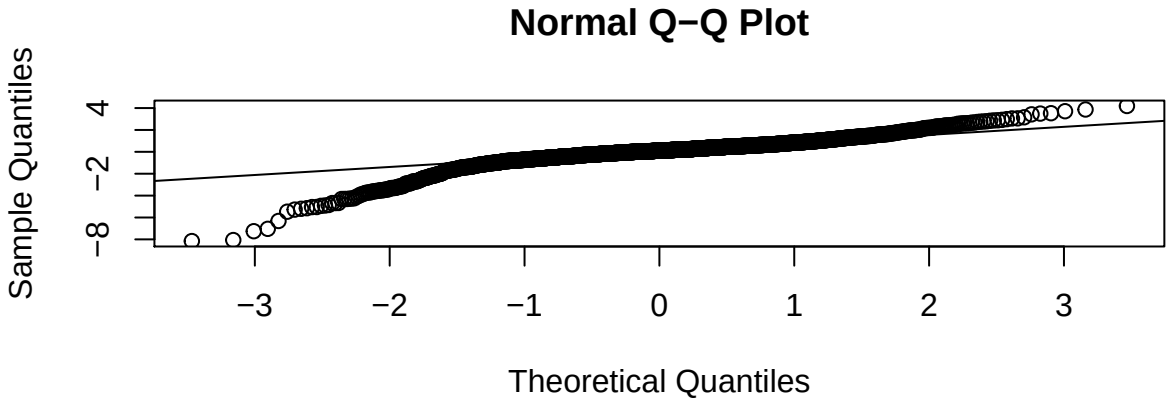


We made a few changes to improve the model. First, we deleted more levels in genres like "Adventure" and "Drama," and only kept action and Comedy in order to let the data be more consistent and thereby adjust the left-skewed issue for normality. Second, we removed the variable of aspect ratio, since we used the p-value of 0.05 as a cutoff and figured it for aspect ratio is 0.134033, which is greater than 0.05 and shows that this predictor doesn't have convincing evidence to be an influential predictor. In addition, we compared two models, one with IMDB score and another one without. The adjusted R squared difference is very small, which indicates the IMDB score is not an important predictor, so we removed it too.

For the final model, the outliers in the low right corner in the residual plot decreased significantly. Therefore, zero mean and constant variance do improve. However, this improvement is not a major one and violations still exist.

Below are our regression line: $\widehat{log(Gross)} = 0.319colorColor + 0.258genresComedy - 0.276ratingPG - 13 - 0.744ratingR + 1.545EnglishyES + 0.709USAyes - 0.0416titleYear - 0.611logCriticReviews + 1.376lognumVotedUsers - 0.000432sqrtCastFacebookLikes + 0.731lognumUserReviews + 0.427logBudget - 0.132lognumVotedUsers : lognumUserReviews + 0.150logNumCriticReviews : lognumUserReviews$





Analysis

This model we fitted includes predictors we think are important to estimate the money a movie made. By fitting the model, we can tell the level of importance that each predictor contributes to the total money made. There are some predictors associated with higher scores. For example, a movie made in English is a significant factor to consider because, because we would expect films in English gross to increase by multiplier of 4.688. Furthemore, our model also shows movies made in the USA will expect to increase gross by multiplier of 2.032. Higher gross means we can make more money in theater and help us determine what kind of movie would be favored by the audience. For the predictors that can lower gross like rating R that we would expect for each movie made in R, we would expect gross to decrease by a multiplier of 2.104, which makes sense because the age restriction does limit some people to see it.

We are comfortable with the final model as we selected it through layers of process, and the adjusted R square value of 0.6343 indicates a high level of preciseness as around 63% of the variability could be explained by this model. I think our final model we predict well and tell filmmakers what kind of movie is more popular with the audience and what the general public want to see in a movie. Using our model can efficiently eliminate some money losses.

As there are so many aspects that could potentially affect the movie, we would like to have more variables about the condition of a certain film. For example, the number of theaters screening the movie. Besies, the days of screening could also be an important predictor, as there would be more tickets sold if there are more days to screen the movie. These are the predictors we would like to have.