Final Project

Christina Yu, Damian Kim

Introduction and data

With the rise in the age of information, the processing speeds of data are accelerated to impossible levels. One of the densest forms of media content out there is the transcription of words into pixels. A book that previously took 2 weeks to finish now fits in the span of 2 hours in the insanely digestible form of: movies. The movie industry is worth 95 billion dollars in the US alone, and indeed there is a great demand for quality movies. Though it is not easily apparent to tell what makes a movie great, there are some basic data that might help determine some of the factors of a great movie. We decided to try and tackle this heavy yet essential question. What factors might help make a movie successful? Throughout this project we can consider this question through the lens of a potential movie investor.

This dataset was scraped from IMDb (Internet Movie Database). There are 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following attributes: `budget`: the budget of a movie. Some movies don't have this, so it appears as 0. `genre`: main genre of the movie. `gross`: revenue of the movie. `rating`: rating of the movie (R, PG, etc.). `runtime`: duration of the movie. `score`: IMDb user rating. `votes`: number of user votes.

First we will explore missingness in our dataset (see appendix).

The dataset includes movies that haven't been published yet (which causes a missing gross value). In our project we are considering gross revenue of a movie as the defining indicator of success. Considering that, we removed all observations that haven't been published yet or are missing gross values, because we are interested only in movies with a quantitative measure of success.

In fact, we also decided to rid our dataset of missing values for the other predictor variables as well. About 28% of the observations were missing a value for budget, whereas the next most missing variable was gross, with 2%. From elementary missingness analysis there did not seem to be very strong, numerically significant relationships between the missingness of each variable with other variables of interest. We decided to do a complete case analysis.

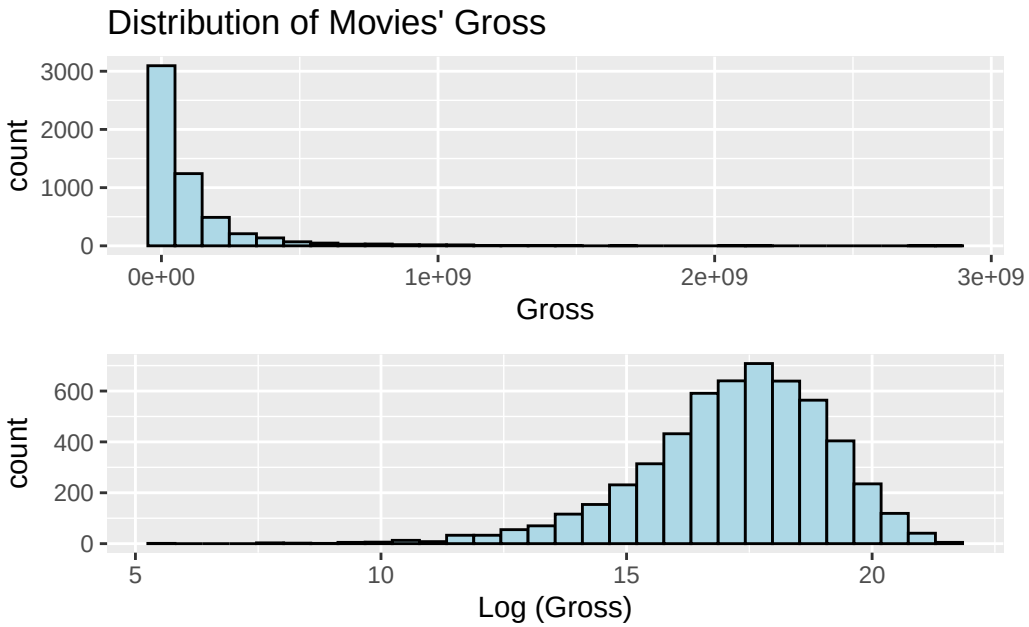
The Predictor Variables & The Response Variable

We will use the `budget`, `genre`, `rating`, `score`, `votes`, `runtime`, and `country` variables as predictors. Among them, `budget`, `score votes`, and `runtime` are numerical variables, while `genre`, `rating`, and `country` are categorical variables. As described earlier, the `gross` numerical variable is our response variable.

1. Summary of the `gross` variable:

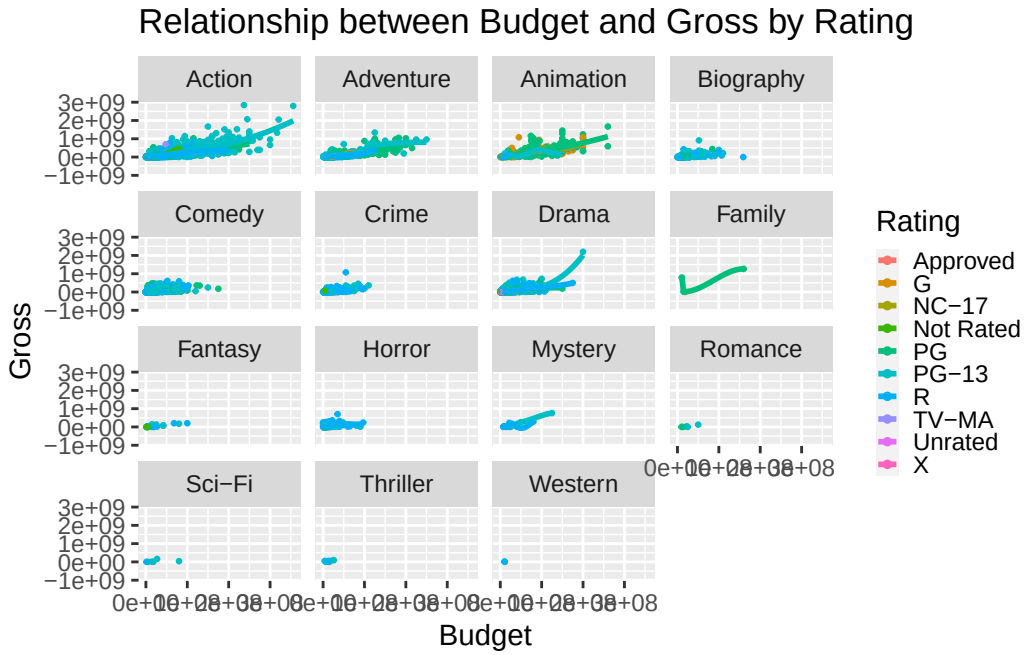
mean_gross	median_gross	sd_gross	min_gross	max_gross
103192280	36850101	187278279	309	2847246203

2. Log-transformation and Distribution of the `gross` variable:



Since the response variable is significantly right skewed, we apply a log-transformation to it and will use log(gross) as our new response variable in the future analysis. Now, our response variable is unimodal, following a roughly normal distribution, with a mean at 17.2102, and there exist some outliers on the left end.

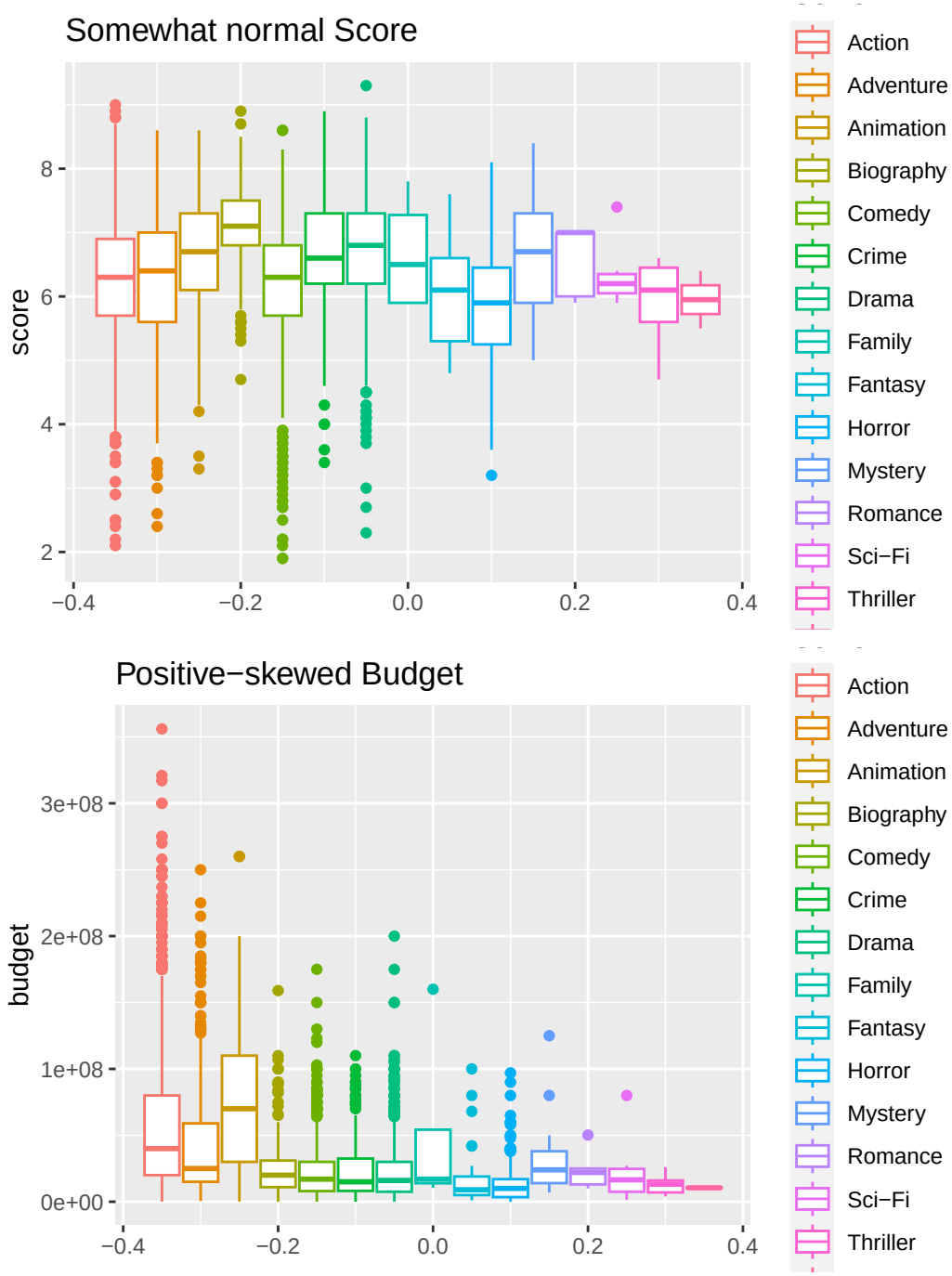
3. Relationship between Gross and Budget based on different Genres & Ratings:



We observe that the relationship between budget and gross is vastly different across genres: action, adventure, animation movies have a steep slope and generally high budget ranges, with outliers which have exceedingly high budget and relatively high gross values. Noticeably, there are also a lot more movies in these three categories, with the most in action. On the other hand, genres such as horror, mystery and romance have a visibly flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others, even when provided with additional funding. Thus, we're interested in further exploring the relationship between budget, genre, and success.

EDA: Visualizatioins and Summary Statistics

The dataset contains many interesting variables that we want to explore fully. We will first do some elementary exploratory data analysis on our datasets to show potential insights that we can explore further. Main effects are described here, additional relationships described in appendix.



As can be seen, the score variable is relatively unskewed, whereas the votes, budget, and runtime variables are heavily positively-skewed (refer to appendix).

We have some correlated variables between our predictor variables, but none are particularly strong (0.8+) so we don't have to remove any of these predictor variables on the basis of

correlation for our later models. Though there seem to be no particularly strong correlations between predictor variables, the strongest correlation is between votes and score. We may consider adding this into our model because it also fits with our domain knowledge that a movie that is rated by more people might generally have a higher score because, in our experience, people are more motivated to rate a movie they enjoy than one they dislike.

Methodology

As a potential movie investor, we're obviously curious about the prediction of movies' success, and the factors that lead to success. Therefore, we want to explore the following topics in detail in order for us to be better investors: 1) Prediction of movies' gross value 2) the factors that affect the success of the movie. Some of the methods by which we are going to tackle this is through linear regression, residual plots, linear mixed models, and repeated k-fold validation.

Linear Regression of all variables

Gross revenue is a continuous response variable, so we will first examine a linear regression model for our data. Since there are two many country variables and so does rating, we'll omit writing all the coefficient for all these variables in our liear regression result.

We've gained the equation of $log\_gross = 1.65 + 0.48 * log\_budget - 0.14 * score + 0.74 * votes + 0.62 * runtime - 0.86 * country\_Aruba - 0.69 * country\_Asutria.... - 1.31 * ratingG - 2.69 * ratingNC - 17... + 0.36 * genre\_Horror - 0.2 * genre\_Mystery.$

#Assumptions We may assume that each movie is independent of each other. We can see that the linear assumption is violated since the data is not symmetrically distributed observations around the horizontal axis (see appendix). It is visible that the constant variance and normality assumptions are not satisfied from our plots. Even after log transforming our gross response variable to fix the normality, the constant variance condition was still not met. The validity of our model results are thus going to be not perfectly reliable.

Linear Mixed Model & Random Effect

Because variables such as genre, rating, and country likely have some variance in their effect on the model across their categories, we decided to create random effects for these variables. Our fixed effects are ones which have constant effects on the model which doesn't change for different "categories", namely budget, score, votes, and runtime.

While adjusting for the random intercept based on country variable and holding all other variables constant, we noticed that the estimated variance is 0.1623 which means that all the country-specific intercepts are distributed around the model's overall intercept within

estimated variance of 0.1623. Given that our gross unit is log dollars, and our estimated variance is quite small, we may expect that the random effects of the country on our model is relatively similar. In terms of our interest in knowing whether a movie is successful or not, production of country may not be a big effect on the movie's success.

Additionally, we noticed that genre has an even smaller estimated variance (0.08254), and rating has a marginally higher variance. The variable effects of the different genres, countries, and ratings don't seem to be hugely related to a movie's success.

```
[1] 1.352356
```

```
[1] 1.357526
```

Our linear model has a smaller RMSE than our mixed effects model: $0.3523 < 0.3575$. The difference is very minor, however, we conclude that our linear model is a better fit for our data and has better performance.

Predicting moves' gross

In order to make our findings applicable for future use, we'd like to make a prediction on movies' gross based on our current variables. However, we also want to test to see if scaling our predictor variables happens to create a more predictive model than without scaling. We thought that even though a linear model with log transformed predictors (for the ones that are skewed) would fit our data better, it might not necessarily be a better predictive model.

Results

After testing a linear model with and without log-transformed predictor variables, with 10-fold validation with five repeats, the average RMSE with log transformed predictor variables was 1.0831, whereas the average RMSE without transformed predictor variables was 1.3736. The transformed predictor variables model has better predictive power, and thus we determine that our final, best model is our linear model with log-transformed predictor variables.

```
[1] 20.45502
```

We decided to try and predict the results of Avatar 2, a very popular movie. Our model predicts Avatar: Way of Water to have a gross of 764700000. This is quite off from the real gross of 2300000000. However, Avatar is an atypical movie because it holds the 1st and 3rd highest gross of all movies ever.

Interested Hypothesis Test 1:

Question 1: Is there evidence to suggest that log(Budget) has an effect on the success of the movies? Null hypothesis: $p_1 = 0$ There isn't sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: $p_1 \neq 0$ There is sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is 30.26 and the p-value is $<$ 2e-16 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to log(budget) does have an effect on the success of the movies (log(gross)).

Interested Hypothesis Test 2:

Aim 2: Is there evidence to suggest that log(Score) has an effect on the success of the movies? Null hypothesis: $p_2 = 0$ There isn't sufficient evidence to suggest that Score is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: $p_2 \neq 0$ There is sufficient evidence to suggest that Score is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is -4.537 and the p-value is 5.84e-06 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to score does have an effect on the success of the movies.

Summary based on Hypothesis:

We choose to use linear regression model since the RMSE result shows that it performs better than the other model. In terms of quantitatively addressing this issue: in our model, we expect to see that every 1 million increase in our budget, the gross value is expected to increase by $e^{0.5} = 1.648$ million dollars; every 1 point increase in our score, the gross value is expected to increase by $e^{-0.525} = 0.59$ million dollars. Through our hypothesis tests, we also concluded that both budget and score has an effect on the success of the movies.

Sensitivity Matrix

We're curious about the sensitivity, specificity, positive and negative predicted values. In order to make log_gross variable binary, we convert it into "Successful Movies" (gross >= mean_gross) and "Not Successful Movies" (gross < mean_gross)

```
                   0    1
Successful      2430 2974
Successful NOT    19    0
```

The sensitivity is 2974/(2974+ 0)= 1; the specificity is 19/(2430+19) = 0.007; positive predicted value is 2974/(2430+2974) = 0.55; and the negative predicted value is 19/(19+0) = 1. The prevalence is 0.548, which means that 54.8% of the movies are successful, and the classifier correctly identify 100% of the successful observations.

Discussion

The broad objective of this project is to try to discern what factors might be correlated with success in the film industry. We learned that many numeric data surrounding movies, such as gross, votes, budget, and runtime were heavily right-skewed. Along that line, after transforming gross (our measure of success) in order to follow normality assumptions, we found that the log of budget indeed has a directly related, significant effect on the success of a movie. We also found that score also had a significant effect on the success of a movie.

Determining the factors that correlate to a successful movie is purely an observational study, and not an experiment. Thus, even though the log of budget and score are correlated with successful movies, they can't be isolated and said to causally determine the success of a movie. We attempted to analyze the predictive power of our model, and used this along with root mean squared error to determine the best model for our data. This happened to be the linear model with log transformed predictor variables and gross. We thus learned that for the context of our data, having random effects for the variables country, genre, and rating did not lead to better fit. This is quite interesting, especially for country and genre of movie, since there were not too many observations per level of the variable and we assumed that different levels would have different relationships with the success of a model. We thought a mixed model would be a better fit for our data.

Finally, we predicted the results for a recent, hugely popular movie known as Avatar: The Way of Water. Avatar and Avatar 2 take the 1st and 3rd slots for highest grossing films of all time. Our model predicted that the gross would be around 800 million, as opposed to the actual gross of 2.3 billion. Though this prediction is not very accurate, we found this result to be okay because Avatar 2 had the extreme bias and imminent popularity that came with being the sequel to the highest grossing film of all time, a variable that was unaccounted for in our data. This goes to show that even if a model is made based off thousands of data entries, linear regression models predict the average effect of predictor variables, and can't really be generalized to an individual datapoint with great predictive accuracy.

Limitations

Our dataset was created from a compilation of scraped data from IMDb, a database for films. The data arbitrarily had 220 movies per year, it is difficult to examine if there was random selection for these 220 movies or if there was some selection bias in the scraping process. For variables with missingness such as budget (with 28%) it is somewhat unlikely that the data was MCAR (perhaps there was survivorship bias, where movies with higher budget had a higher chance of having published budgets). A better way to deal with missing variables was the process of imputation, which should be explored further.

In our predictor variables there are likely to be correlated variables that weren't fully considered (particularly any interaction effect such as two-way, or even four-way, like perhaps between budget, votes, rating, and score). Including such interaction effects would've led to increased model complexity and less easily interpretable results, however this should be better explored in the future.

We didn't include the variables such as director, actors, and company in our variables of interest. If we had included these, we would have random effects for correlated variables like director, actors, or company. For these variables however it would have created too many different effects (there were thousands of directors and actors, many different companies too). Because the director/actors are not very generalizable this would have led to misleading predictions as well. Additionally, for our numeric variables such as budget and rating, it would have been a good idea to scale our predictor variables so our model's coefficients could have been standardized. Future analysis could consider standardizing the numeric variables by subtracting means and dividing by standard deviations to create more readily comparable coefficients.

We were planning on using LASSO selection to further identify variables of interest, however, in the context of our data considering we had categorical variables with 2+ categories (for example, country had way more than 2 categories), instead of going with common literature practices such as group lasso with dummy coding, contrast coding, or Helmert coding, we decided to treat each categorical variable as a all-or-nothing feature (where if even one level of a feature was shrunk to zero, that the entire categorical variable should be removed), which unsurprisingly suggested in the removal of all the categorical variables from our model. For future analysis, a variable selection with group Lasso would seem appropriate to better create a meaningful model. We disregarded the results of our LASSO approach of variable selection however because again, we had such a low ratio of predictor variables to observations that we didn't need to worry much about overfitting.

Our models failed the constant variance assumption of linearity as well as normality, so our residuals suffered from heteroscedasticity. In essence, our linear model coefficients, p values, and confidence intervals may not have been accurate for our explanatory variables. We attempted to adjust the normality by log scaling our response variable, gross, which resulted in better normality, however even after trying to scale by various means, constant variance was

still somewhat violated. A future model should aim to find a transformation that results in constant variance for the model.

Sources

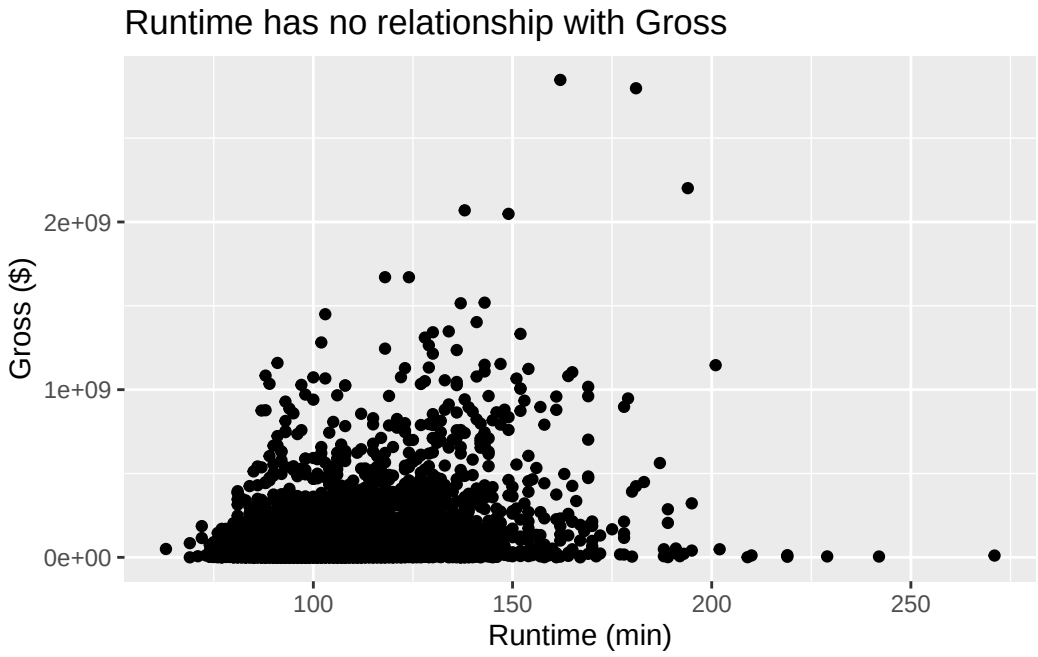
https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a

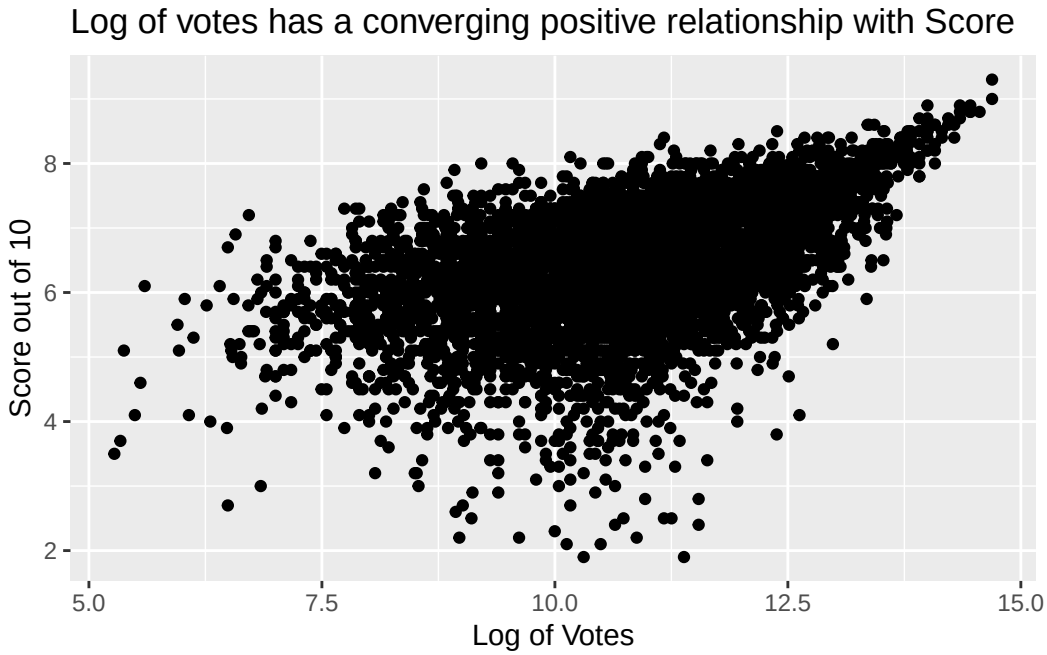
https://stackoverflow.com/questions/13646654/root-mean-square-error-in-r-mixed-effect-model

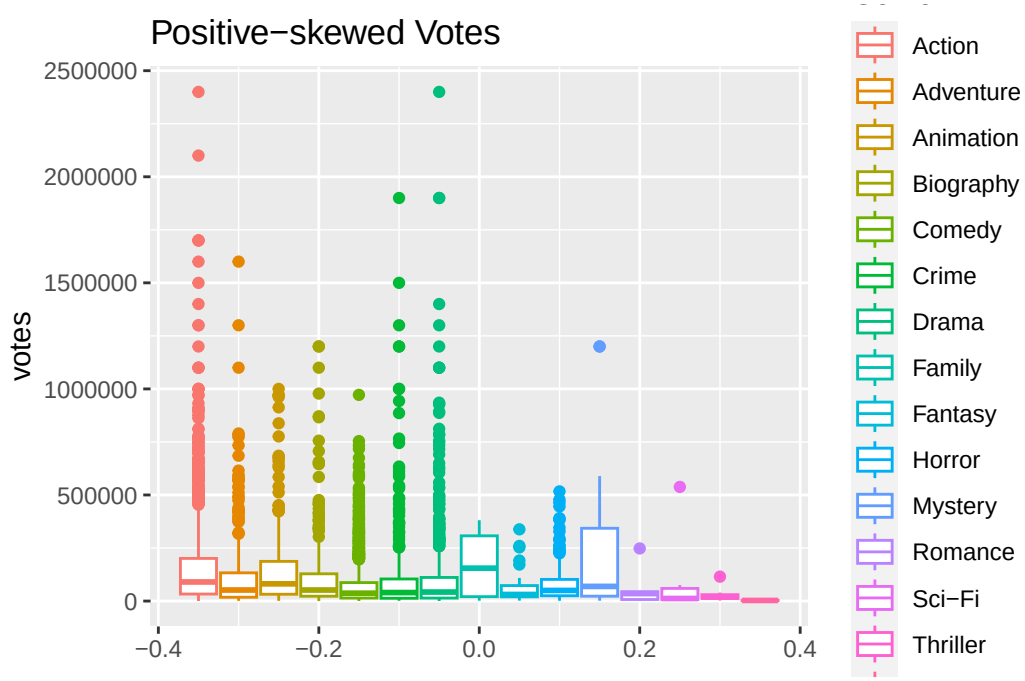
https://psyarxiv.com/wc45u/

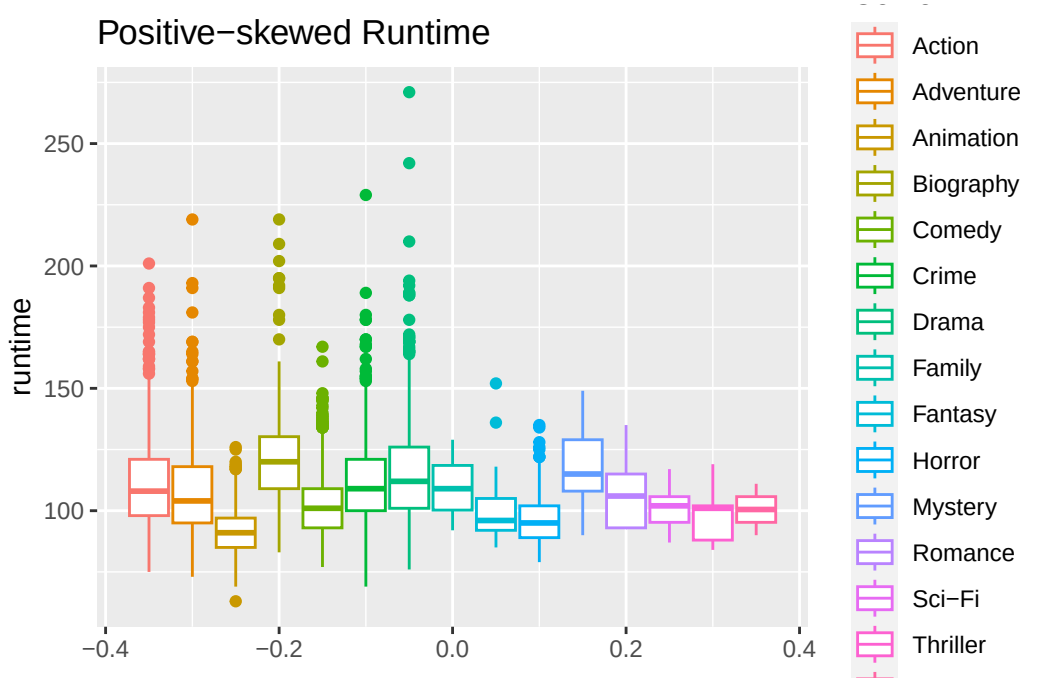
Appendix

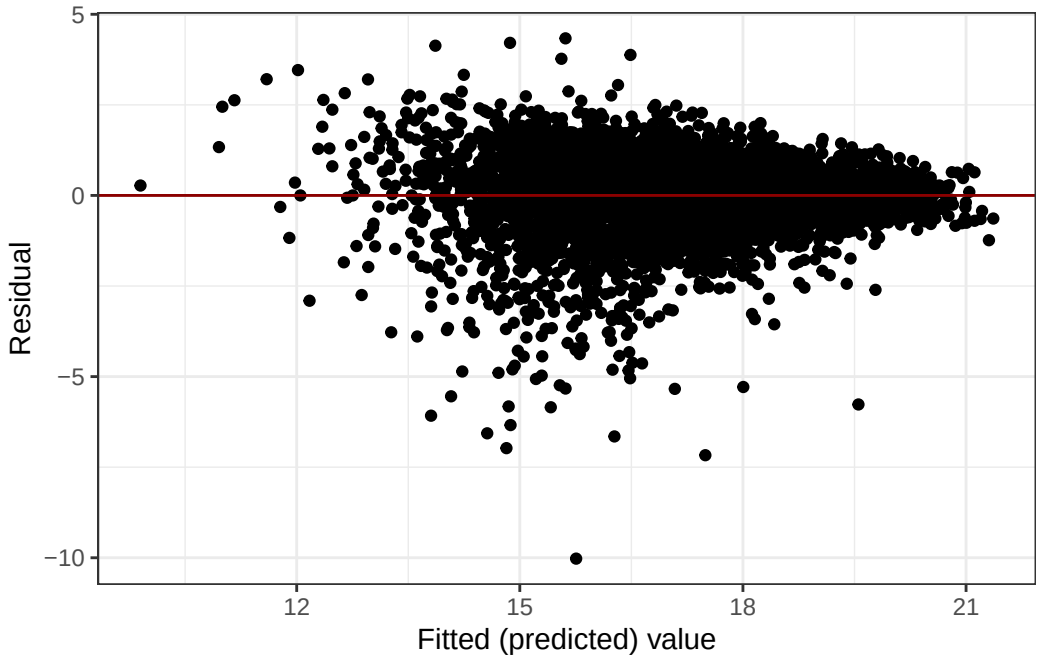
This dataset was scraped from IMDb (Internet Movie Database). There are 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following attributes: `budget`: the budget of a movie. Some movies don't have this, so it appears as 0. `company`: the production company. `director`: the director. `genre`: main genre of the movie. `gross`: revenue of the movie. `name`: name of the movie. `rating`: rating of the movie (R, PG, etc.). `released`: release date (YYYY-MM-DD). `runtime`: duration of the movie. `score`: IMDb user rating. `votes`: number of user votes. `star`: main actor/actress. `writer`: writer of the movie. `year`: year of release.

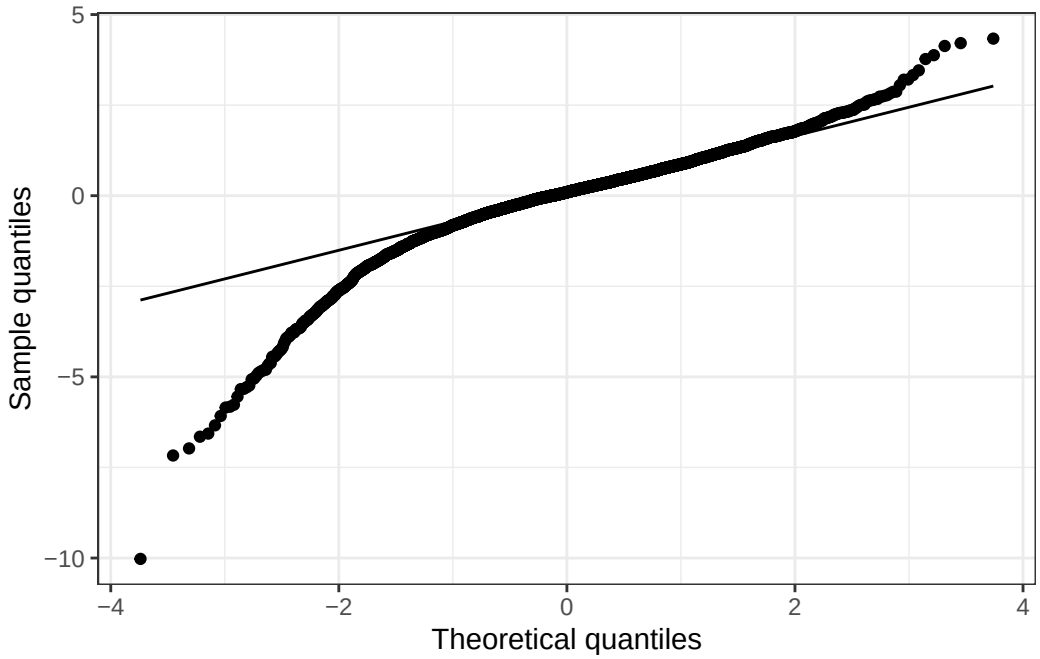


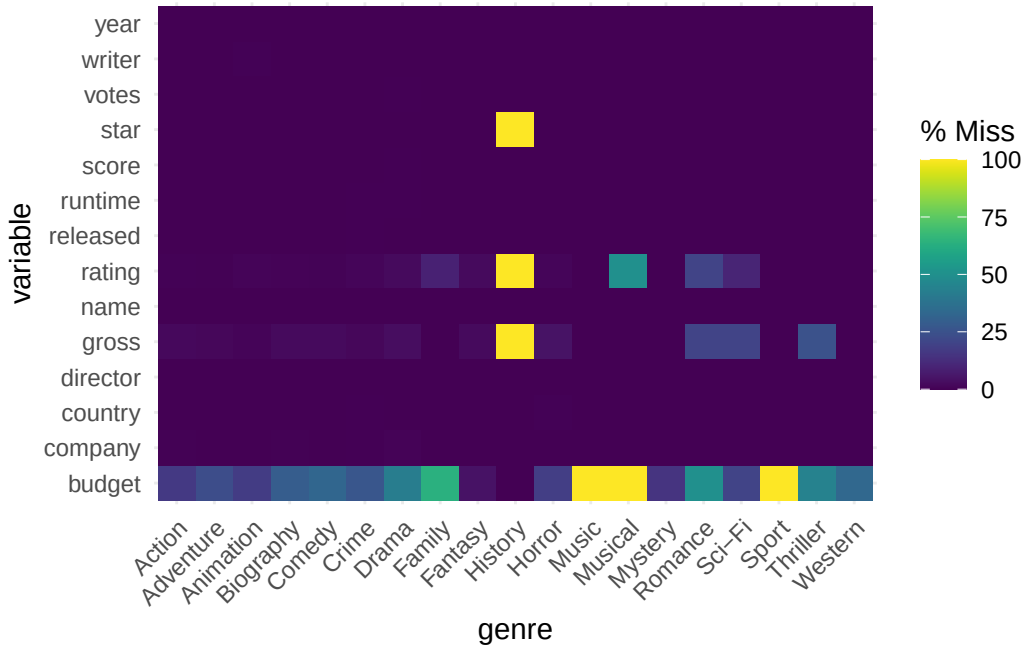

Log of votes has a converging positive relationship with Score
Score out of 10
8
6
4
2
5.0
7.5
10.0
12.5
15.0
Log of Votes


Positive−skewed Votes
votes
2500000
2000000
1500000
1000000
500000
0
−0.4
−0.2
0.0
0.2
0.4
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
Sci−Fi
Thriller


Positive−skewed Runtime
runtime
250
200
150
100
−0.4
−0.2
0.0
0.2
0.4
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
Sci−Fi
Thriller


Residual
5
0
−5
−10
12
15
18
21
Fitted (predicted) value


Sample quantiles
Theoretical quantiles
5
0
−5
−10
−4
−2
0
2
4


variable
year
writer
votes
star
score
runtime
released
rating
name
gross
director
country
company
budget
genre
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci−Fi
Sport
Thriller
Western
% Miss
100
75
50
25
0