George Daher
Professor Zaino
STAT 216 – Applied Statistical Methods I
4/28/22

Examining a relationship IMDB Ratings and Movie Genre

With the consistent drip-feed of new blockbuster movie releases, whether it is the latest Disney movie based on existing intellectual property, a book, or a more independent project, movies are multi-million-dollar investments, and as a result, their success or failure can lead to profits or losses of the same scale.

I want to determine what factors may have a relationship with the success of movies. Specifically, I am considering the effect of the genre (or genres) of a movie on its success. This leads to the first two important questions that must be considered: how to determine the "success" of movies, and how to determine the genres of movies, as both can be disputed unless a standard metric is found.

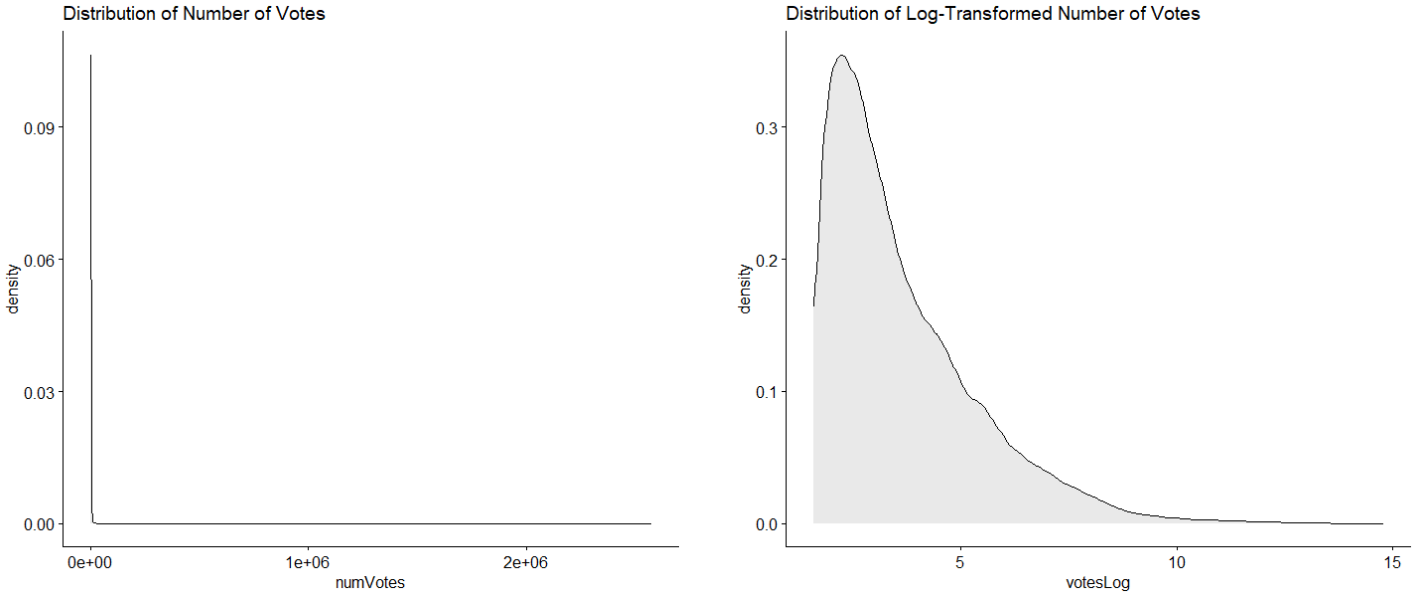
The question of determining the "success" of a movie can be considered in a couple of different ways. The first is financial success, but this can easily be confounded by the period during which and when a movie was released, as it would lead to the need to address inflation,

As a result, I chose to use the IMDB ratings to measure the success of movies for a few reasons. The first reason is due to the practicality of the data, the IMDB dataset on movie ratings was very comprehensive and easy to access. Additionally, the IMDB rating system is generally considered to be a credible metric for a movie's success. However, it is important to understand that it measures the success of movies in the eyes of critics and cannot be a general measure of success. This fact offers benefits and drawbacks, as one could easily argue that a movie's critical success is not an important component in a more general interpretation of success. But at the same time, the measure of critical success rather than more objective popularity means that the marketing of a movie would not have more of an influence on measured success rather than the content of the movie itself. It would be relatively simple to do another study in which one could analyze the relationship between IMDB scores and overall popularity or financial success, but this is not the objective of this study, which does not seek to determine the optimal metric for success for movies but rather seeks to determine if there is a significant relationship between the genre of a movie and its success.

One of the main steps required in order to answer this question is to determine what type of statistical analyses and tests should be performed, but before even doing this, it is important to consider how the data should be examined and within what bounds. The first step of doing this is intuitive, and only movies and their ratings should be considered because IMDB also has comprehensive rating data on short films, tv shows, individual episodes of tv shows, and other types of digital media and are not within the scope of the study. Additionally, the dataset includes movies from 1915 to 2022 and a range of international blockbusters to extremely obscure titles. This could cause biases due to the limited number of ratings that could

theoretically skew the rating of the title. However, the dataset also includes data on the number of votes each movie had on its rating, which can allow us to eliminate movies with few responses that could be subject to non-response and voluntary response biases. Even if we decide to remove data points based on the number of rating votes they received due to the potential sources of bias it would eliminate, the threshold for this cut is not obvious. The extreme skewness can be seen in the following density plots.

Density Plots for the Number of Votes on Rating per Movie



Even after being log-transformed, the data remains very skewed (However, the non-transformed data is considered in the study, because the normality of predictor of variables is not a condition). but in setting a threshold, it is important to note that this is not an attempt to change the skew or normalize the data but rather the aforementioned biases due to non-response or limited voluntary response that exists for some of the obscure movies included in the dataset, some of which have received ratings fewer than ten times. As a result, setting a threshold should be considered with the idea of practicality rather than concerning statistical summaries. Appealing to practicality, I can set a threshold of 10,000 votes with the idea that any movie that received few than 10,000 IMDB ratings may be subject to non-response bias. This number could be chosen to be larger or smaller, depending on exactly the target of the experiment, but some level of thresholds makes it possible to avoid the probably larger sources of biases in the data.

The next step is to consider how genres are understood within the data. It can be noted again, that although the genre categories for movies are viewer submitted, this is another problem that should be addressed for the most part by setting a threshold for votes, as this assures more accurate genre labels as it will follow the level of agreement among those who submitted genre for any particular film. It is also important to note that the movies can be considered to have multiple genres (a fact that certainly makes anecdotal sense), and the number of genres for any

particular movie can vary from 1 to 3. This means that the data must be considered with 16 different indicator variables, that have values of either 1 or 0, depending on which of the 16 total genres being considered by the dataset individual movies fall into. On top of this, as one could expect, the number of occurrences of the different genre categories varies.

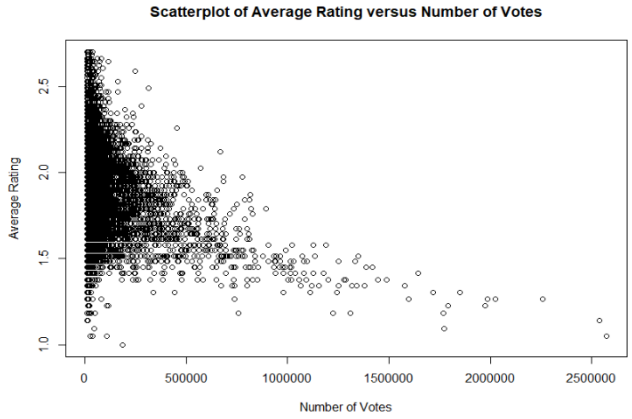
Most Frequent Genres among Movies in the Dataset

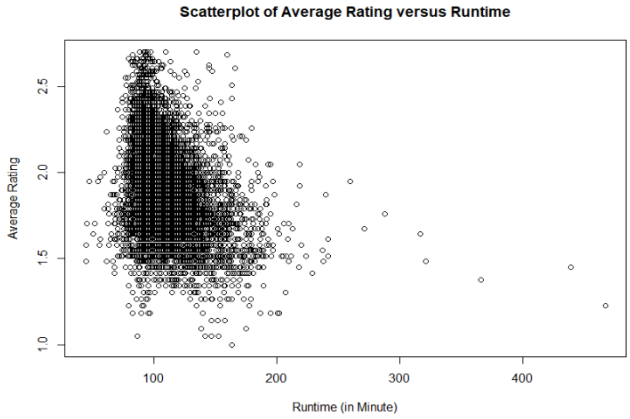
Genre	Number of Occurrences
Drama	5485
Comedy	3592
Action	2388
Crime	2022
Romance	1668
Thriller	1649
Horror	1193
Mystery	1002

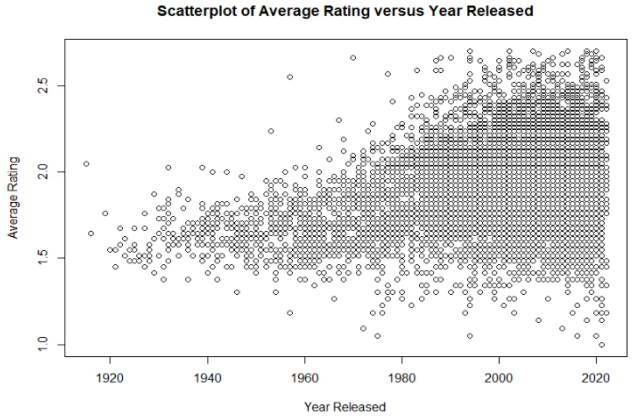
Genre	Number of Occurrences
Sci-Fi	703
Biography	694
Animation	457
Music	416
History	363
War	224
Documentary	209
Western	112

The next step in order to determine whether a regression model is appropriate is to determine the linearity of each quantitative variable in relation to the response variable. The three quantitative variables included in the data are not of direct interest to the study, but their existence should be considered as covariates and as a result, their linearity as part of a regression model must be checked.

Scatterplot of Quantitative Predictors vs. Average







These scatter plots show a weak linear relationship between each of the variables and the response, and this would not be sufficient to justify a linear regression using them as variables, however, they are not of direct interest and should still be considered as covariates.

Before any of the specific statistical analysis, the dataset ends up having 16 binary indicator variables for genre, the number of votes on rating a title received, its runtime in minutes, the release year, and of course the proposed response variable of IMDB rating. Due to the number of variables, the most appropriate procedure for statistical analysis is multiple regression. But the specific multiple regression model that should be used is unclear, as a completely additive model for the data cannot be assumed and there may be an interaction effect between variables. This becomes especially clear when one reconsiders the overlap between genres that occurs, as certain specific genres may work poorly or well together and have a multiplicative effect.

Most Frequent Combination of Genres among Movies in the Dataset

Genre Intersection	Number of Occurrences
Drama and Comedy	1420

Genre Intersection (CONT.)	Number of Occurrences
Drama and Biography	601

Drama and Romance	1188
Drama and Crime	1169
Comedy and Romance	893
Drama and Action	802
Crime and Action	760
Drama and Thriller	664

Drama (only)	541
Comedy (only)	529
Drama and Mystery	529
Comedy and Crime	524
Comedy and Action	515
Thriller and Crime	500

Note: There are other intersections between genres among movies in the dataset, but these are the most frequent.

This graph shows the distribution in the overlap between different genres, the more common genres must have their interaction effects considered and checked for statistical significance. The possibility of a multiplicative interaction effect becomes even clearer when one examines an initial multiple regression model that does not account for any interaction effect between variables.

There are a few ways to begin, the first of which is a comparison of models that only additively factors genre with no interaction effect between genres. This model was labeled Model A. However, the dataset also includes potential covariates that re the year the movie was released, its runtime, and the number of total votes that it received for its rating. Considering these covariates is important as they may influence the model that should be accounted for. Because of this, Model A can be considered unnecessary as part of the statistical analysis, but it helps to show the effect the covariates have on the ability of regression to model to predict error within the data.

When including the covariates as well as all the variables for the genre, there are many different options for multiple regression models that include a varying combination of these predictors, so we must determine the best of these models using a stepwise or best subset multiple regression. Performing best subset regression yields that the best model in terms of adjusted r-squared and Mallow's Cp criterion is one with 15 predictors. This model is shown below, labeled Model B.

However, this model does not account for any interaction effects between the variables as it is purely a linear regression model and the existence of interaction between variables would be accounted for by some multiplicative effect between certain predictors.

There are a few different ways to account for interaction effects, but the due to overlapping nature of the categorical genre data, analyzing different models that analyze all the interactions is not only impractical as it follows exponential complexity but unnecessary. Instead, it is important to consider interaction effects that would seem probable according to more general knowledge about the data. While someone with more knowledge may choose to examine different interaction effects, I chose to refer to the most common overlap between genres, more specifically the top 10 interactions between genres, as I presumed that the common intersection of the genres could possibly be related to a change in success in such movies. Using another stepwise regression, I determined the best multiple regression model that combined each of the predictors linearly but also considered the multiplicative effects between the most frequent intersection in genre that I labeled below as Model C.

While there are many more models one could consider involving more interaction effects or eliminating certain variables due to inherent knowledge held about the domain, these three models are what I determined the best. In comparing the various multiple regression models, it is important to seek to maximize adjusted r-squared and minimize Mallow's cp criterion, but also follow the Principle of Parsimony, and favor simple models over complicated ones.

Details on Regression Models A,B and C

Note: All of these models were selected using stepwise regression with considerations of all the variables indicated.

Model A: Genre and No Interaction Effects Considered

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.55226    0.03385 193.541  < 2e-16 ***
Documentary  0.85962    0.06971  12.332  < 2e-16 ***
Drama        0.47816    0.02421  19.754  < 2e-16 ***
Horror      -0.54565    0.03343 -16.320  < 2e-16 ***
Comedy      -0.23810    0.02558  -9.310  < 2e-16 ***
Western      0.43563    0.08960   4.862 1.18e-06 ***
Thriller    -0.09709    0.02839  -3.420 0.000628 ***
Crime        0.05411    0.02510   2.156 0.031092 *
Action      -0.22086    0.02607  -8.472  < 2e-16 ***
Romance     -0.11023    0.02775  -3.972 7.18e-05 ***
Mystery     -0.07406    0.03372  -2.196 0.028089 *
SciFi       -0.06901    0.03886  -1.776 0.075768 .
Biography    0.17200    0.03962   4.341 1.43e-05 ***
Animation    0.42449    0.04753   8.931  < 2e-16 ***
History      0.07617    0.05225   1.458 0.144942
War          0.34381    0.06402   5.370 8.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9281 on 9847 degrees of freedom
Multiple R-squared:  0.1767,    Adjusted R-squared:  0.1754
F-statistic: 140.8 on 15 and 9847 DF,  p-value: < 2.2e-16
```

Model B: All Predictors and No Interaction Effects Considered

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.939e+01  9.196e-01  42.830  < 2e-16 ***
runtime      1.026e-02  4.084e-04  25.127  < 2e-16 ***
yearPast    -1.714e-02  4.604e-04 -37.227  < 2e-16 ***
numVotes     1.835e-06  5.357e-08  34.243  < 2e-16 ***
Drama        5.195e-01  2.035e-02  25.533  < 2e-16 ***
Documentary  1.294e+00  5.939e-02  21.783  < 2e-16 ***
Animation    6.561e-01  4.080e-02  16.080  < 2e-16 ***
Horror      -2.863e-01  2.899e-02  -9.877  < 2e-16 ***
Action      -2.001e-01  2.207e-02  -9.065  < 2e-16 ***
Biography    2.184e-01  3.329e-02   6.560 5.63e-11 ***
SciFi       -1.736e-01  3.336e-02  -5.205 1.98e-07 ***
Crime        8.593e-02  2.149e-02   3.999 6.40e-05 ***
Romance     -7.704e-02  2.350e-02  -3.279  0.00105 **
War          1.518e-01  5.515e-02   2.752  0.00594 **
Comedy      -3.313e-02  2.002e-02  -1.655  0.09804 .
Western      1.247e-01  7.729e-02   1.613  0.10678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7991 on 9847 degrees of freedom
Multiple R-squared:  0.3897,    Adjusted R-squared:  0.3887
F-statistic: 419.1 on 15 and 9847 DF,  p-value: < 2.2e-16
```

Model C: All Predictors and Some Interaction Effects Considered

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.949e+01  9.163e-01  43.093  < 2e-16 ***
yearPast       -1.724e-02  4.583e-04 -37.619  < 2e-16 ***
runtime         1.025e-02  4.042e-04  25.351  < 2e-16 ***
Documentary     1.303e+00  6.149e-02  21.189  < 2e-16 ***
Drama           6.797e-01  2.740e-02  24.807  < 2e-16 ***
Horror         -2.716e-01  2.794e-02  -9.722  < 2e-16 ***
Western         1.413e-01  7.674e-02   1.841 0.065619 .
Thriller        1.394e-01  3.038e-02   4.590 4.49e-06 ***
Crime           1.442e-01  3.975e-02   3.627 0.000288 ***
Action         -1.582e-01  2.980e-02  -5.311 1.12e-07 ***
Romance        -1.819e-02  4.203e-02  -0.433 0.665149
SciFi          -1.532e-01  3.304e-02  -4.638 3.57e-06 ***
Biography       5.211e-01  8.903e-02   5.852 5.00e-09 ***
Animation       7.050e-01  4.141e-02  17.027  < 2e-16 ***
War             1.368e-01  5.483e-02   2.494 0.012647 *
numVotes        1.856e-06  5.344e-08  34.729  < 2e-16 ***
Drama:Romance  -1.123e-01  4.963e-02  -2.262 0.023692 *
Drama:Crime    -1.499e-01  4.398e-02  -3.408 0.000658 ***
Drama:Action   -1.505e-01  4.163e-02  -3.614 0.000303 ***
Crime:Action    7.469e-02  4.555e-02   1.640 0.101073
Drama:Thriller -2.573e-01  4.552e-02  -5.653 1.62e-08 ***
Drama:Biography -3.713e-01  9.561e-02  -3.884 0.000104 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7959 on 9841 degrees of freedom
Multiple R-squared:  0.3948,    Adjusted R-squared:  0.3935
F-statistic: 305.7 on 21 and 9841 DF,  p-value: < 2.2e-16
```

For comparing different models adjusted r-squared and Mallow's criterion are sufficient, making it clear in this case that Model A is weaker due to a larger residual standard error and noticeably smaller adjusted r-squared. By this same logic, and the principle of parsimony, the added interaction effects in Model C as compared to Model B do not account for a significant amount of the error in the model, as the residual standard error of 0.7959 in Model C is close and slightly smaller than that of 0.7991 in Model B. This holds for the adjusted r-squared as well, for Model B it is 0.3887 which is only slightly smaller than Model C which has an adjusted r-squared of 0.3935 despite the interaction effects included. However, this is not enough to consider Model B an acceptable regression model for the average rating of the movies, even though this shows that comparatively, Model B is better than the others. It is also important to consider the assumptions necessary for the model to be generally considered statistically sound.

There are a few conditions and assumptions for the regression model to be considered appropriate and statistically sound. All of the models follow the assumption of independence (because I did not collect the data, this is presumed). On top of this, three conditions must be met: minimal collinearity between the predictor's variables; homoscedasticity, meaning that there is no pattern among the residuals compared to the predicted values; and multivariate normality, meaning that the residuals of the regression model are normally distributed. The first condition of minimal collinearity can be checked by calculating variance inflation factors, a factor that measures the correlation between different predictors.

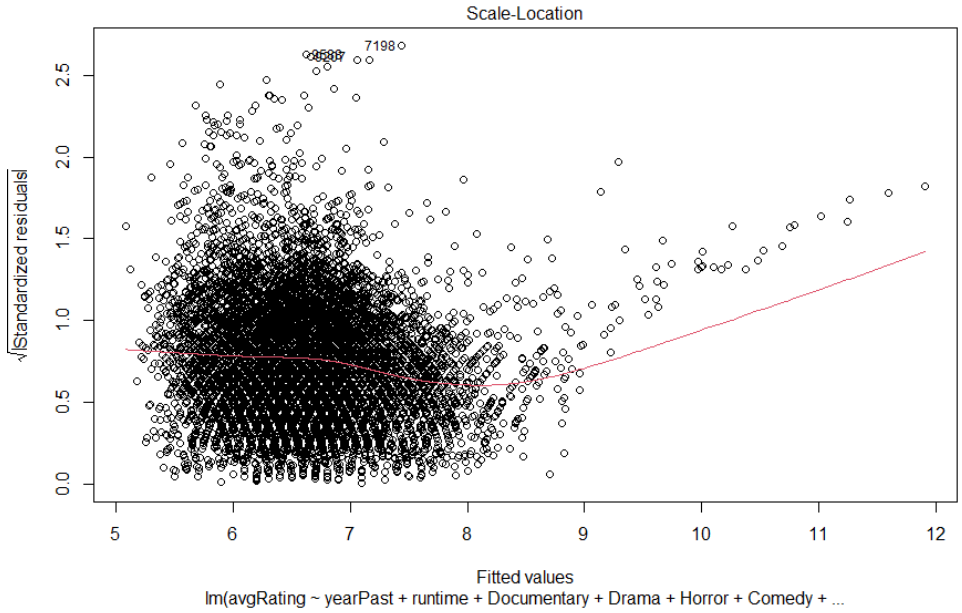
VIF Values for Each of the Predictors

Predictor	VIF value
Years Passed	1.058
Runtime	1.26
Number of Votes	1.10
Documentary	1.13
Drama	1.58
Horror	1.38
Comedy	1.43
Western	1.03

Predictor	VIF value
Crime	1.16
Action	1.38
Romance	1.20
SciFi	1.14
Biography	1.12
Animation	1.14
War	1.04

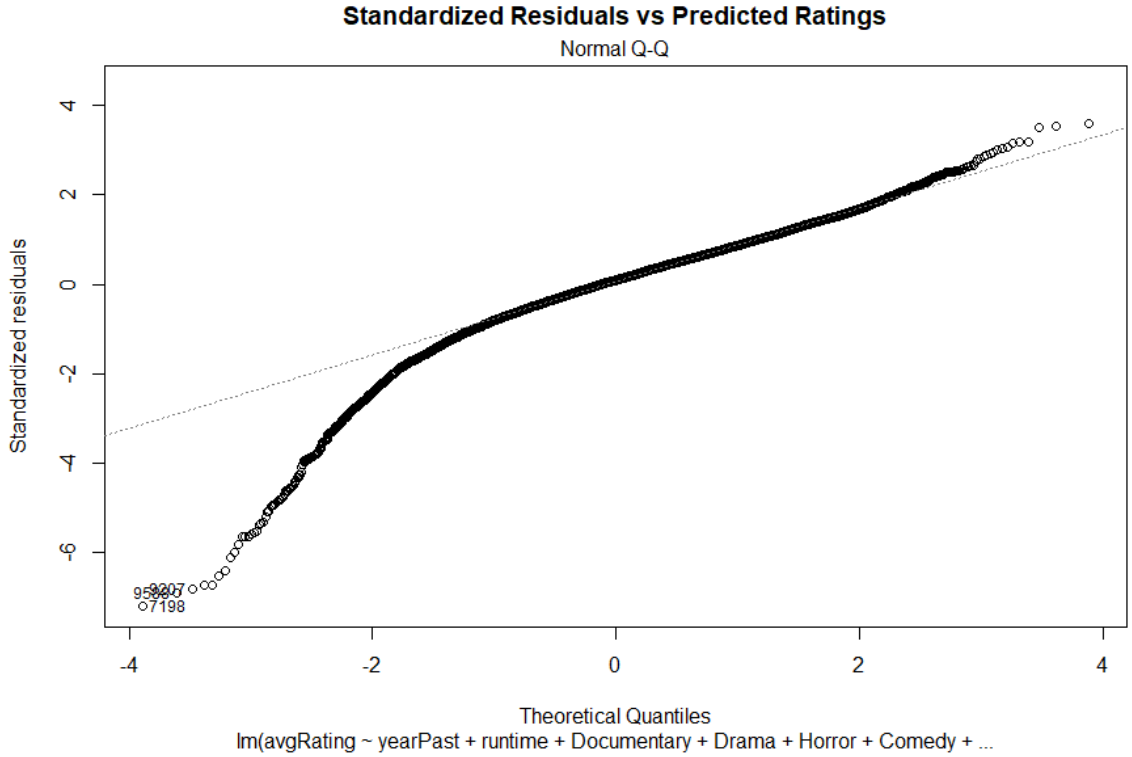
These results indicate that while there is some level of correlation to be considered among the predictor variables due to the variance inflation factors greater than 1, it is not enough for the predictors to be considered collinear, so this condition is met by the model. The condition of homoscedasticity can be checked by spread location plot.

Spread-Location Plot for Model B



The spread location plot shows that the assumption for homoscedasticity does not hold due to the fact the spread of residuals changes as the fitted values changes, meaning the data is heteroscedastic rather than homoscedastic However, the fact the failure of this condition can be understood since the ratings are limited on a scale from 1 to 10, even though a regression model would not include this threshold, and would round any values above 10 back down to 10 as the upper limit of the rating scale. If the data points to the right of the fitted value of 10 are understood in this way and this rule was applied to the model, then the data could potentially satisfy this condition. However it still fails the assumption of homoscedasticity, so the model cannot be considered reliable, we can still consider the final assumption that must be satisfied: multivariate normality, which can be checked using a normal probability plot of the regression.

Normal Probability Plot for Model B



This shows that while generally, the data follows the line, meaning the residuals are approximately normally distributed, there is a clear departure from the line. This failure can possibly be explained in the same prior failed assumption can: by the fact that the scale is restricted from 1 to 10 any fitted values to the left could be rounded up to improve the model, but this still does not satisfy the assumption.

Even after determining what could be comparatively determined to be the best multiple regression model in terms of the predictor variables available while still adhering to the principle of parsimony, the model is not reliable because it does not meet the necessary assumption of homoscedasticity and multivariate normality. This does not mean that the data cannot be modeled or predicted with the predictor variables provided in the dataset, but it seems that an additive effect linear regression is not appropriate for this data, even if interaction effects are considered.

Even if we consider that the model is reliable, it still does not explain a majority of the error found within the data. This is where the practical application of such a model is important to consider again. Is a relationship that explains 35% of the randomness enough to be considered important in this specific instance? While it is possible that the answer to this question could be yes, the fact that this was an observational study rather than an experiment means that the weak relationship cannot be considered causal, meaning the relationship has even less practical meaning. Additionally, by looking at the coefficients of the variables, while many of the coefficients are considered statistically significant, meaning that there is statistical evidence that

they differ from zero, they are still very small, and a predicted rating change of a tenth of the rating point may not be considered important.

In conclusion, because it fails to satisfy the assumptions for a reliable multiple regression model, there is not sufficient evidence to reject the null hypothesis that there is a relationship between genre and the IMDB rating of a movie. Even if the assumptions for the model were met, then the strength of the relationship would not be immediately certain or necessarily practically important. It is possible that other types of regression would show a relationship, or to only consider one main genre for each movie, as the overlapping categories made regression for the dataset particularly difficult. A preferrable course of action however, would be to examine other datasets, whether using existing datasets of box office data or other movie ratings metrics, or possibly even collecting new data. However, from this data alone, there is not sufficient evidence to show a relationship between IMDB rating, which we have accepted as a metric for movie success, and the genres of a movie.