The Impact of Film's Properties on IMDB Ratings: A logistic Regression Approach

Yujie Tang, Jialu FU, Weiqing GUO, Bashiru Mukaila, Wanding Wang

2023-03-14

1 Introduction

The entertainment industry happens to be a competitive market and the film industry is an important aspect of it. The stakeholders in the film industry are interested in properties or features that make their films successful. One key aspect of a film's success is its rating on platforms such as IMDB, which can influence audience perception and drive revenue.

The aim of this project is to investigate the relationship between the properties of films and their IMDB ratings. Specifically, we want to use a logistic regression model to know which properties of a film influence whether a film is rated by IMDB as greater than 7 or not, using variables such as year of release, length, budget, number of votes, and genre.

Section 2 speaks to the exploratory data analysis of IMDP ratings and explores the relationship between the rating and the properties of the films. Section 3 contains the results from the logistic regression model. While Section 4 gives the final remark to the research.

2 Exploratory Data Analysis

The data contains 2847 observations and 7 variables. The variables year, length, budget, votes, and rating are numerical variables, and genre is a categorical variable. Table 1 shows the first 6 rows of the table. It is also worth of note that the length of some of the films is not recorded.

Table 1: Highlight of the IMDB data

film_id	year	length	budget	votes	genre	rating
5993	1943	65	15.5	42	Action	7.6
37190	1961	87	12.3	6	Drama	6.0
43646	1987	79	16.4	161	Action	7.5
28476	1976	NA	12.2	5	Documentary	8.0
23975	1982	88	12.5	97	Action	3.5
50170	1936	NA	7.0	146	Drama	4.4

From the summary statistics in table 2, we can see that the year of release of the films is between 1898 to 2005 and the year with the highest release of films rated is 2002. The minimum budget for film in the data set is 2.5 million while the maximum is 22.3 million. It is also clear that Action films is the highest genre in terms of frequency.

Table 2: Summary statistics of the numerical variables

year	length	budget	votes	rating
Min. :1898	Min. : 1.00	Min. : 2.50	Min. : 5	Min. :0.800
1st Qu.:1957	1st Qu.: 73.00	1st Qu.: 9.90	1st Qu.: 11	1st Qu.:3.700
Median :1982	Median : 90.00	Median :11.90	Median : 29	Median :4.600
Mean :1976	Mean : 82.22	Mean :11.85	Mean : 657	Mean :5.342
3rd Qu.:1997	3rd Qu.:101.00	3rd Qu.:13.70	3rd Qu.: 114	3rd Qu.:7.700
Max. :2005	Max. :480.00	Max. :22.30	Max. :149494	Max. :9.200
NA	NA's :131	NA	NA	NA

From the scatterplot in figure 1, we can see that there is a negative correlation between the length of the film and its rating. This implies that films with longer length tend to be rated low. There is a positive correlation between the budget of film production and its rating, however, the correlation is weak.

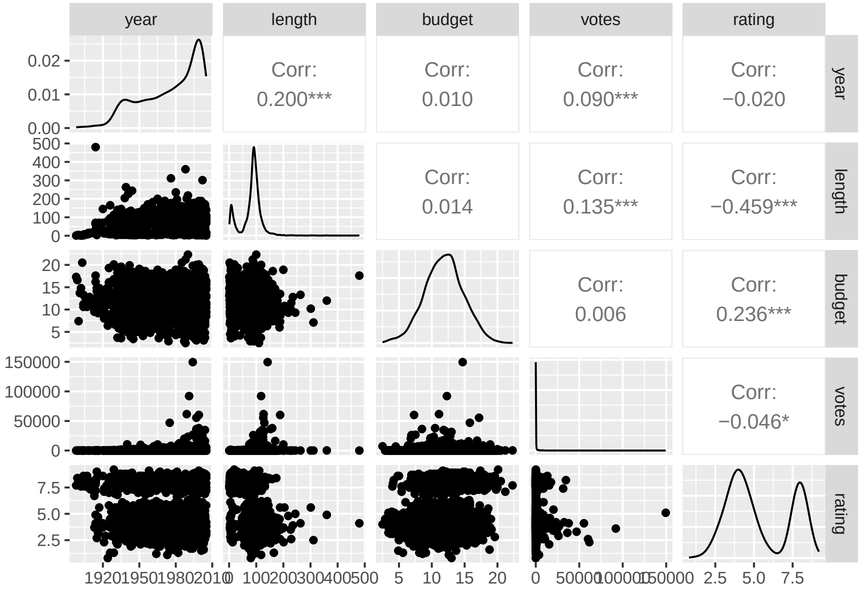

Figure 1: Scatterplot matrix of the numerical variables.

From figure 2, we can see that some genres tend to have higher ratings than others. For example, comedy tends to have higher ratings than romance films. While the majority of the short films are rated higher than 7.

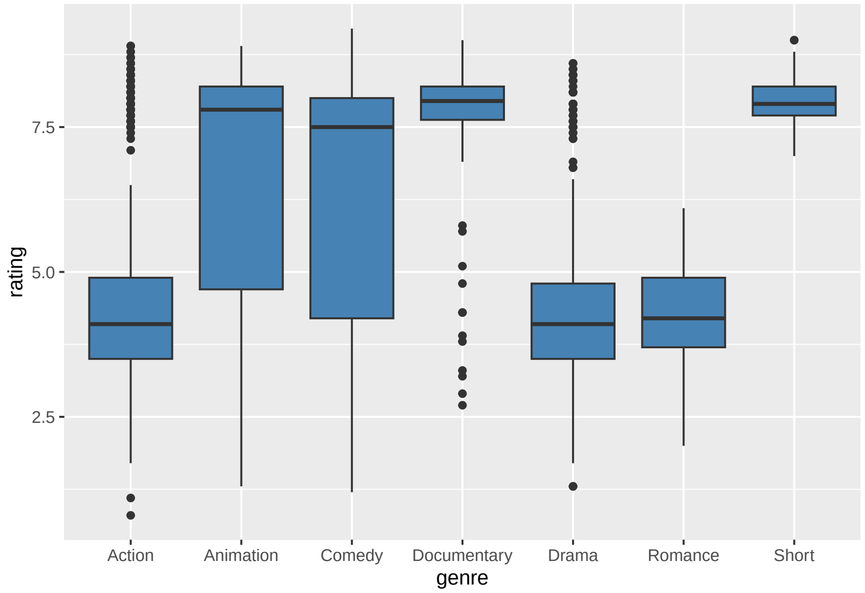

Figure 2: Boxplot of rating by genre.

From figure 3, the binary Boxplot of year by rating shows that the middle 50% of the ratings are between 1953 and 1999 for the ratings greater than 7 while it is between 1959 and 1996 for the ratings that are 7 and below. It is also evident that there is more variability in the length of ratings that are greater than 7 compared to the ratings that are 7 and below as depicted by the Binary Boxplot of length by rating. We can also see that the budget for ratings greater 7 seems to be higher that ratings of 7 and below.

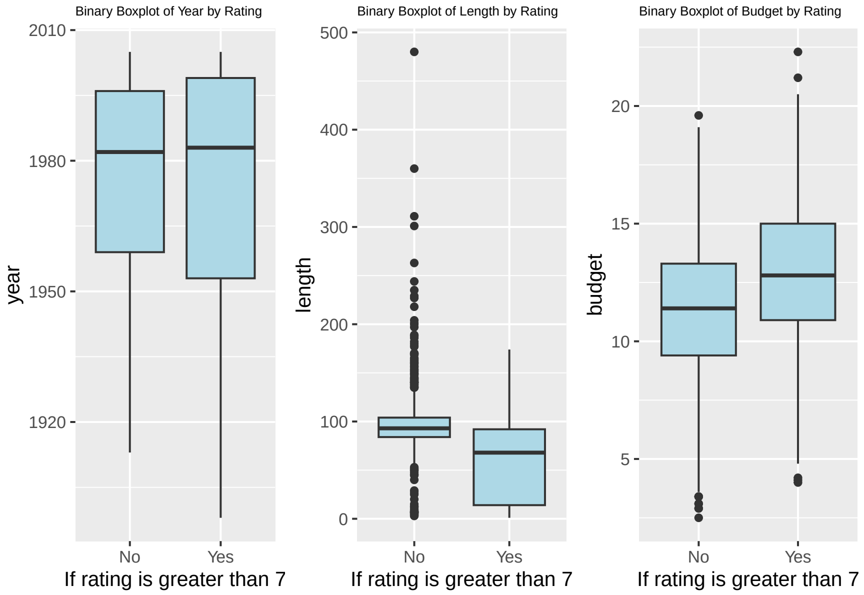

Figure 3: Binary Boxplot for years, Lenght and Budget.

From figure 4, we can see that almost all short films are rated above 7 and there is no romance film that has a rating greater than 7. It is also very evident that comedy has the highest number of rating that is greater than 7.

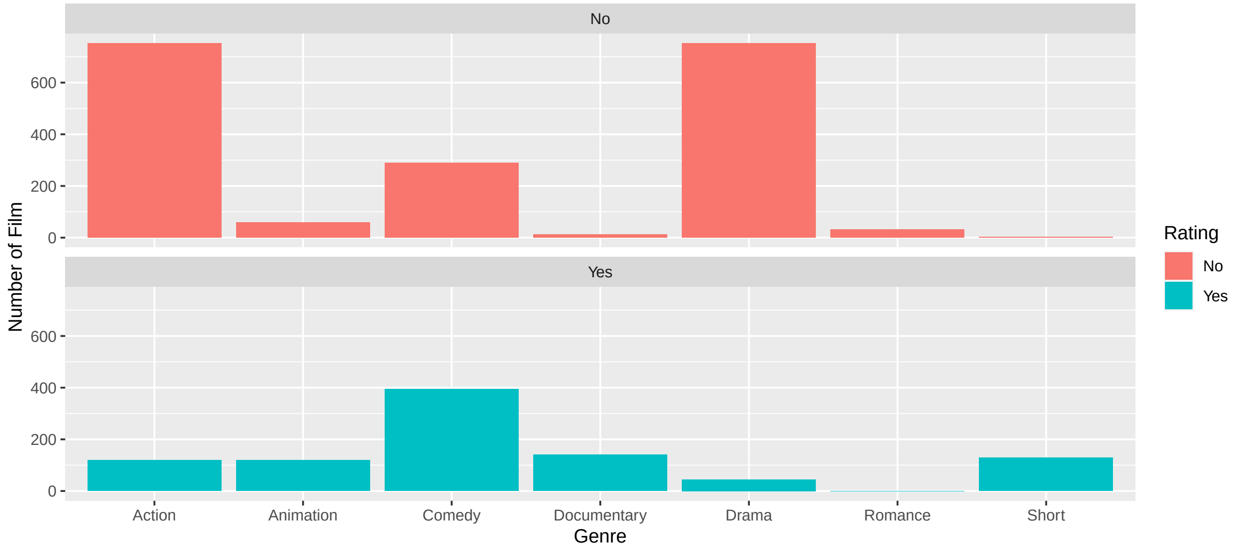

Figure 4: Genres of IMDB Rated Films and their rating status.

3 Formal Data Analysis

3.1 Logistic regression with one categorical explanatory variable

An IMDB rating of over 7 indicates an excellent and highly-regarded film. Consequently, investigating which properties influence movie ratings is an intriguing research question.

Consider determining whether the categorical variable *genre* has an effect on the movie rating above seven. Display the distribution by creating a barplot of *genre* and *Rating_above_7*:

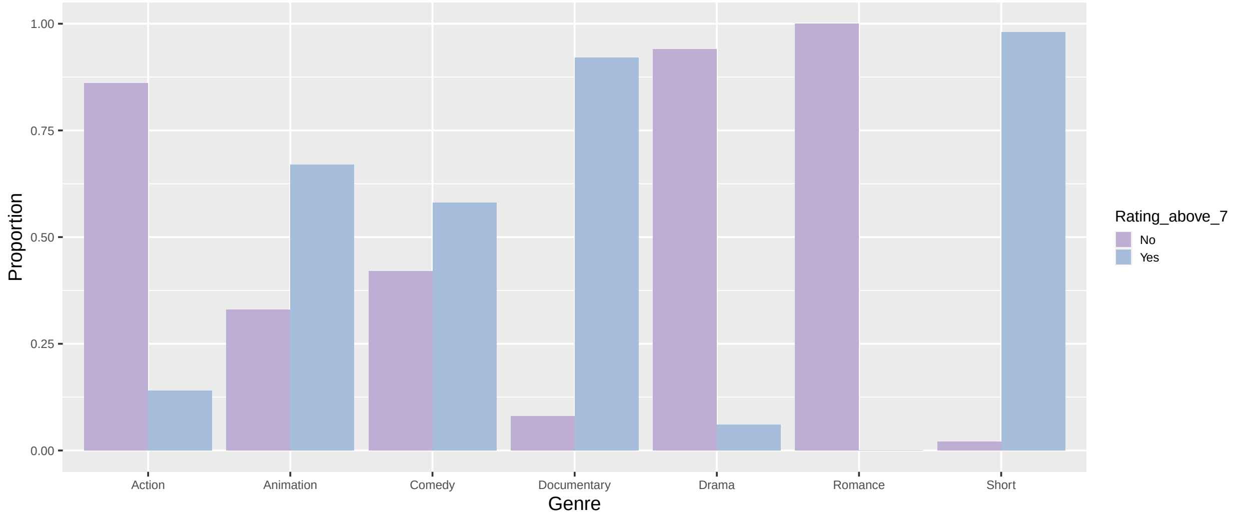

Figure 5: Proportions of Yes/No Ratings by Genre.

From Figure 5, we can observe that the proportion of low ratings for Action, Drama, and Romance genres is relatively high, exceeding 80%, while the proportion of high ratings for Documentary and Short genres is higher, exceeding 90%. The genre of Comedy has a relatively balanced proportion of high and low ratings.

3.1.1 Model specification and estimation

Therefore, we further investigate the impact of movie genres on movie ratings. Here $p_i = Prob(Rating\_above\_7 = Yes)$, with $genre_i$ being the type of the film for $i = 1, 2, ..., 2847$. The model we will consider is of the form:

$$g(p_i) = log(\frac{p_i}{1 - p_i}) = \alpha + \beta \cdot genre_i$$

and we fit it in R as follows:

```
##
## Call:
## glm(formula = Rating_above_7 ~ genre, family = binomial(link = "logit"),
##     data = imdb_group_08)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8921  -0.5417  -0.3409   0.4200   2.3976
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.84494    0.09865 -18.702  < 2e-16 ***
## genreAnimation     2.54653    0.18731  13.595  < 2e-16 ***
## genreComedy        2.15141    0.12537  17.160  < 2e-16 ***
## genreDocumentary   4.22875    0.30618  13.811  < 2e-16 ***
## genreDrama        -0.97113    0.18244  -5.323 1.02e-07 ***
## genreRomance     -13.72113  261.39713  -0.052    0.958
## genreShort         6.01161    0.71936   8.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3621.5  on 2846  degrees of freedom
## Residual deviance: 2309.4  on 2840  degrees of freedom
## AIC: 2323.4
##
## Number of Fisher Scoring iterations: 14
```

From the output, the estimated coefficients for *Action (Intercept)*, *Drama*, and *Romance* are negative, suggesting that movies of these three genres are more likely to receive ratings below 7, consistent with our previous observation. However, the coefficient for *Romance* is not significant, which may be due to a lack of observations with ratings above 7. All other estimated coefficients are significant at the 5% level of significance.

3.1.2 Model inference and interpretation

The baseline category for our binary response is *No*, and the baseline category for our explanatory variable is *Action*. Hence the estimates from the logistic regression model are on the log-odds scale and apply to movies with a rating above 7:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -1.84 + \beta \cdot \mathrm{II}_{genre}(\cdot)$$

...

where β is 2.55 for Animation, 2.15 for Comedy, 4.23 for Documentary, -0.97 for Drama, -13.72 for Romance, 6.01 for Short films

and $\mathrm{II}_{genre}(\cdot)$ is an indicator function, that takes a specific movie genre (excluding Action genre) as input and returns 1 if the input variable matches the genre, and 0 otherwise.

The point estimate and the corresponding 95% confidence interval for the log-odds of each movie genre can be obtained, as shown in the Figure 6:

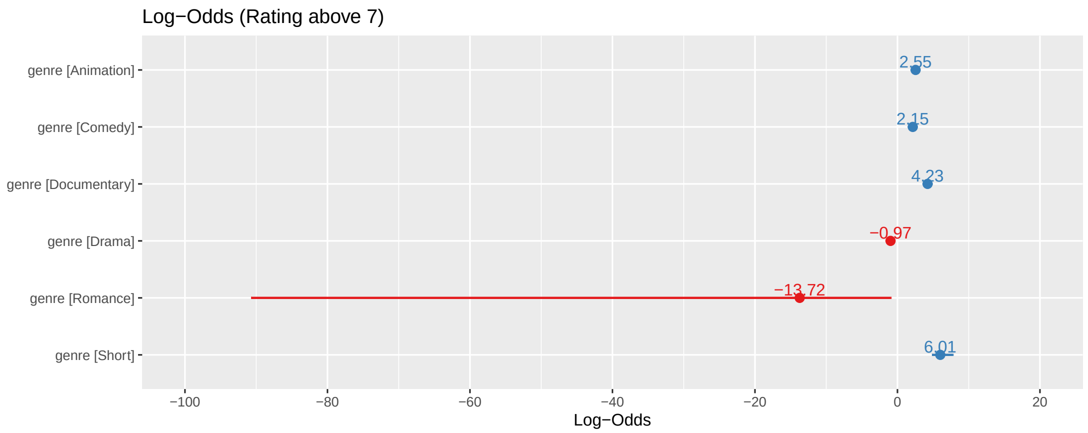

Figure 6: Log-Odds (Rating above 7).

Consider using the estimated coefficients to quantify the effect of genre, the regression coefficients on the odds scale are given by:

```
##      (Intercept)   genreAnimation      genreComedy genreDocumentary
##     1.580345e-01     1.276271e+01     8.596986e+00     6.863154e+01
##       genreDrama     genreRomance       genreShort
##     3.786541e-01     1.098982e-06     4.081387e+02
```

Similarly, the point estimate and the corresponding 95% confidence interval for the odds scale of each movie genre can be obtained, as shown in the Figure 7:

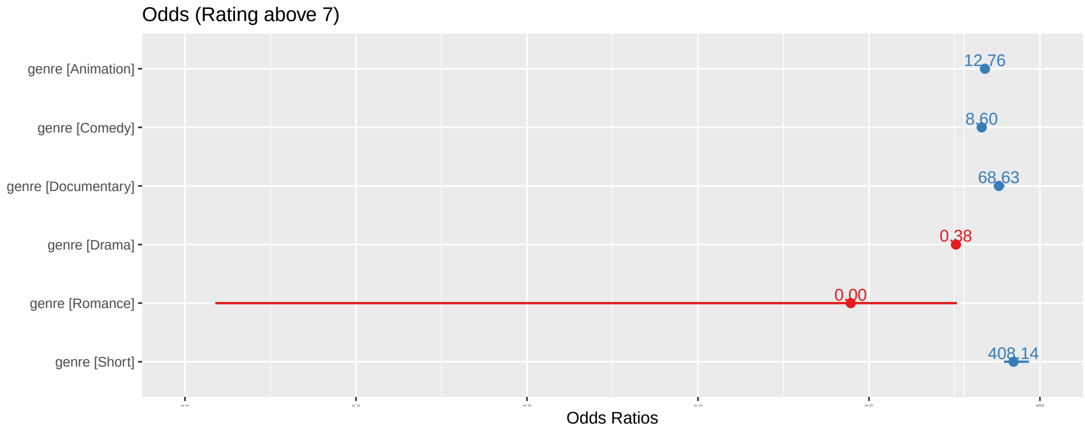

Figure 7: Odds (Rating above 7).

The *Action (Intercept)* gives us in the Action movie genre, the odd of a high rating is approximately 0.16 times that of a low rating. And the odds of the rating being above 7 for *Animation* are 12.76 times greater than the odds if the movies are *Action*. For the movie genres *Documentary* and *Short*, the odds of having a high rating is 68.63 and 408.14 times higher than that of *Action*, respectively.

Next, we calculate the estimated probability of a rating above 7 for each movie genre using the following formula:

$$\hat{p} = \frac{\exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}{1 + \exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}$$

The estimated probabilities of each type of movie reaching a score of 7 or higher were obtained, and the probabilities were compared with the frequencies of the data to produce the following graphs.

```
## $genre
```

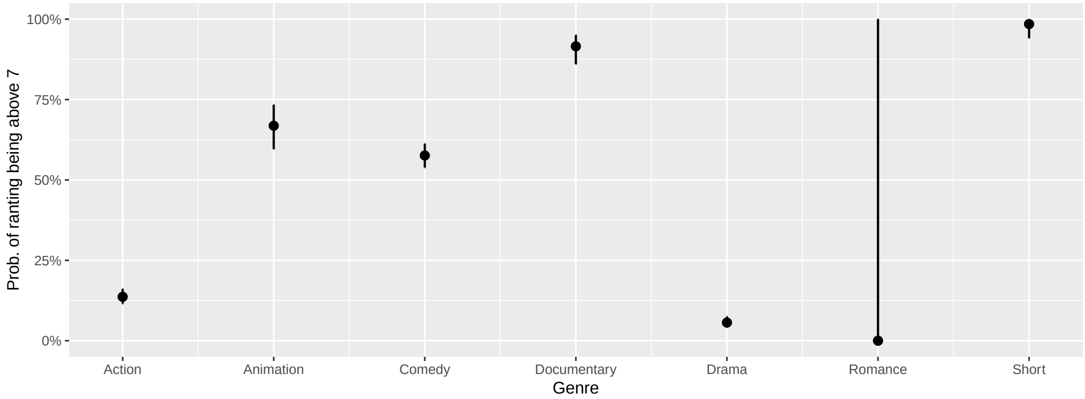

Figure 8: Estimated probability of ranting being above 7 by genre.

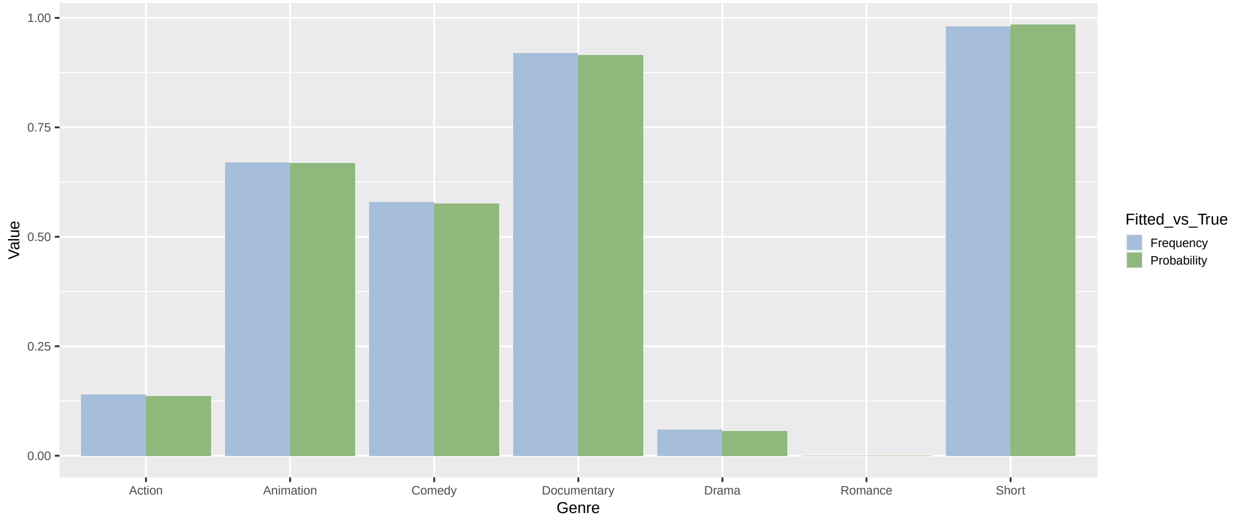

Figure 9: Frequencies and estimated probabilities by Genre.

The Figure 9 shows that the estimated probabilities are consistent with the frequencies, indicating that the model has a good fit to the data.

3.2 Logistic regression with several explanatory variables

3.2.1 Model selection and estimation

We now consider a generalized linear model that takes into account multiple variables. To identify the factors that influence movie ratings, we will use a stepwise approach and select the best model based on AIC as the selection criterion.

```
##
## Call:
## glm(formula = Rating_above_7 ~ year + length + budget + genre,
##     family = binomial(link = "logit"), data = imdb_group_08)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7225  -0.3744  -0.1209   0.1962   3.4412
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -23.864236   5.767450  -4.138 3.51e-05 ***
## year                 0.010238   0.002940   3.483 0.000497 ***
## length              -0.056869   0.003537 -16.077  < 2e-16 ***
## budget               0.509979   0.030117  16.933  < 2e-16 ***
## genreAnimation      -0.078708   0.320139  -0.246 0.805793
## genreComedy          3.110781   0.179174  17.362  < 2e-16 ***
## genreDocumentary     5.604906   0.442282  12.673  < 2e-16 ***
## genreDrama          -1.556649   0.239136  -6.509 7.54e-11 ***
## genreRomance       -14.607631 391.828859  -0.037 0.970261
## genreShort           4.005562   0.796669   5.028 4.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3470.8  on 2715  degrees of freedom
## Residual deviance: 1456.6  on 2706  degrees of freedom
##   (131 observations deleted due to missingness)
## AIC: 1476.6
##
## Number of Fisher Scoring iterations: 15
```

The best model removed the *votes* variable from the full model, indicating that its impact on the rating was not significant when considering other variables, which resulted in a relatively poor model fit. Additionally, due to 131 missing values in *length*, it was removed from the analysis. The model we obtained is of the form:

$$g(p_i) = log(\frac{p_i}{1 - p_i}) = \alpha + \beta_1 \cdot year_i + \beta_2 \cdot length_i + \beta_3 \cdot budget_i + \beta_4 \cdot genre_i$$

From the output, the estimated coefficients for *length* is negative, suggesting that movies of longer duration are more likely to receive ratings below 7. The estimated coefficients for *year* and *budget* are positive, indicating that as the year of the movie's release becomes more recent and the budget increases, the movie's rating also tends to be higher. All three numerical variables were significant at the 5% level of significance, while the categorical variable *genre* showed that the 'Animation' and 'Romance' genres were not significant.

3.2.2 Model inference and interpretation

The estimates from the logistic regression model are on the log-odds scale and apply to movies with a rating above 7:

$$g(\hat{p_i}) = log(\frac{\hat{p_i}}{1 - \hat{p_i}}) = -23.86 + 0.01 \cdot year_i - 0.06 \cdot length_i + 0.51 \cdot budget_i + \hat{\beta_4} \cdot \mathrm{II}_{genre}(\cdot)$$

The point estimate and the corresponding 95% confidence interval for the log-odds can be obtained similarly, as shown in the Figure 10:

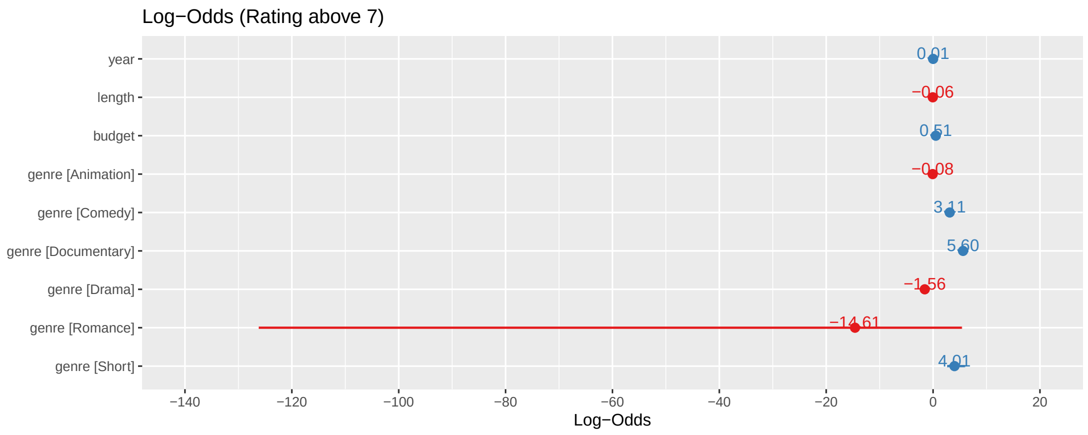

Figure 10: Log-Odds (Rating above 7) for the step model.

Consider using the estimated coefficients to quantify the effect of these four variables, the regression coefficients on the odds scale are given by:

```
##      (Intercept)            year          length          budget
##     4.324085e-11    1.010290e+00    9.447176e-01    1.665257e+00
##   genreAnimation     genreComedy genreDocumentary     genreDrama
##     9.243095e-01    2.243857e+01    2.717564e+02    2.108414e-01
##     genreRomance      genreShort
##     4.528834e-07    5.490269e+01
```

The point estimate and the corresponding 95% confidence interval for the odds scale can be obtained, as shown in the Figure 11:

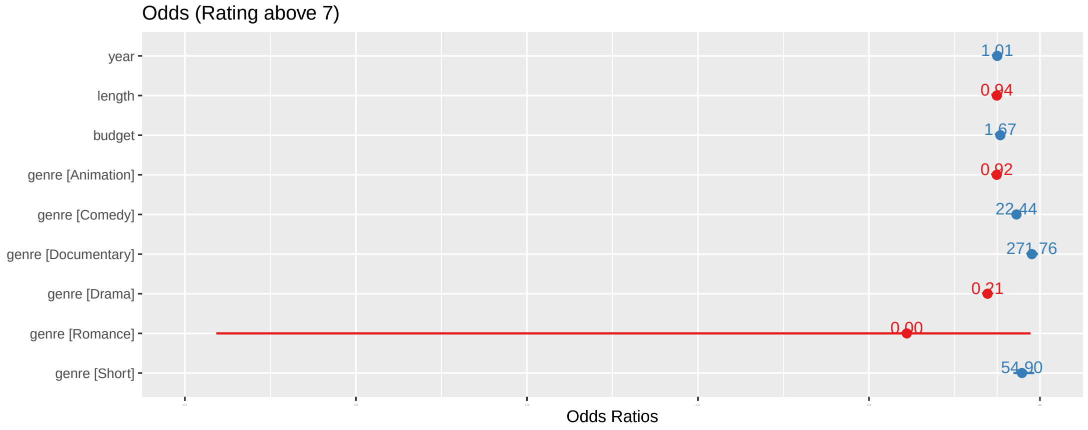

Figure 11: Odds (Rating above 7) for the step model.

This suggests that for two movies released one *year* apart, the odds of a new movie having a rating above 7 is 1.01 times that of an old movie. Likewise, for every additional minute (*length*) in the duration of a movie, the odds of the movie having a rating above 7 decreases by a factor of 0.94, indicating a slight decrease. As for the *budget*, for every additional one million US dollars invested in the production of a movie, the odds of the movie having a rating above 7 increase by a factor of 1.67. The influence of the genre is similar to what was discussed earlier.

4 Conclusion

After a movie is released, it is important to pay attention to the audience's word of mouth, which is reflected in the ratings given by the general public on IMDB. We consider movies with ratings of 7 or higher as high-rated movies and explore which movie attributes affect the rating results.

The following conclusion was drawn by modelling binary logistic regression with one categorical variable as an explanatory variable and by mixed modeling with multiple numerical variables as explanatory variables:

- The *genres* of Action, Drama, and Romance have fewer high-rated movies, while Documentary and Short genres have more high-rated movies. This indicates that audiences prefer movies with logic and depth, rather than formulaic commercial films.
- The *genres* of Animation and Comedy receive mixed ratings, but overall, there are more high-rated movies, which is in line with our impression of popular trends among the general audience.

- In terms of numerical factors, the *year*, *length*, and *budget* are taken into account. However, the *votes*, representing the number of positive votes received by viewers, does not have a significant impact on the final rating of the movie, which may be due to its subjectivity.
- The release year and production budget of a movie have a positive impact on its rating, indicating that people prefer new movies that match their current tastes. Moreover, the budget invested in a movie can greatly influence its production quality, providing people with new and exciting experiences and thus gaining more popularity. However, the length of a movie has a negative impact on its rating, indicating that people no longer enjoy the long narrative of old movies, which is related to today's fast-paced culture.

Future work could include:

- In addition to the movie attributes mentioned in the study, other movie attributes such as box office, investment in art and music, etc., may also have an impact on movie ratings. Not only the movie's inherent attributes, but other cultural factors such as language and actors may also cause certain differences. If more variables can be included, it will be more effective in improving the accuracy of the classifier's functionality.
- In this study, the response variable was set to two categories, greater than 7 points or lower. If the response variable can be divided into more categories, such as high-rated movies, average-rated movies, and low-rated movies, it will be more discriminative.
- The distribution of genres in the sample used in the study is uneven, as there were no *Romance* films with ratings above 7, which has led to errors in the model. Obtaining a more comprehensive sample would result in better model performance.