Analysis of IMDB data set

Group_06

1 Introduction

The study aims to investigate the relationship between various film attributes and IMDB ratings, drawing data from the IMDB film database allocated. The data set comprises of the factors such as film ID, release year, duration, budget, votes, genre, and IMDB rating. The research question focuses on examining the factors that impact IMDB ratings, particularly whether specific film properties contribute to ratings greater than seven. A Generalized Linear Model (GLM) analysis is conducted to derive the relationships between these properties and IMDB ratings.

2 Data Wrangling Methods

Before we begin the analysis of our data, let's transform the data using various tools. The process below describes the detailed data wrangling techniques that are used to get the desired data set. After having a glimpse of the data set, the 'genre' column is converted to a type factor type.

A check for missing values is conducted and it is found that 103 observations are missing from the column 'length'. Missing values are imputed with the median since median is a robust measure, less impacted by outliers as much as mean. The function *median( )* reveals the median to be 90 minutes. However, it is observed in Table 1 that the median lengths vary across the different genres. With this information, the missing lengths of films are replaced by median length of the respective genre.

Table 1: Median length by genre.

genre	median_length
Action	90.0
Animation	7.0
Comedy	91.0
Documentary	73.0
Drama	96.0
Romance	92.5
Short	13.5

As per the research question, a new column 'high_rating' containing binary variables corresponding to 'rating' values is created. This column takes a value of 1 for IMDB ratings greater than or equal to seven and 0 for IMDB ratings less than seven. Additionally, another categorical variable 'rate' conveying the same is also added.

3 Exploratory Data Analysis

3.1 View the data

The data set has 1937 rows and 9 columns, 7 of which are from the original data.

Let's have a look at the first five rows of the data frame.

Table 2: Glimpse of the first five rows in the IMDB data set.

film_id	year	length	budget	votes	genre	rating	high_rating	rate
31804	2002	18	9.6	15	Drama	8.0	1	Rating greater than 7
25453	2000	98	13.8	23	Action	3.3	0	Rating less than 7
5479	1989	81	11.5	57	Documentary	7.9	1	Rating greater than 7
44235	1995	100	7.5	32	Action	3.4	0	Rating less than 7
14580	2003	80	10.8	30	Action	2.6	0	Rating less than 7

The variables in Table 2 are defined as:

- **film.id** : The unique identifier for the film
- **year** : Year of release of the film in cinemas
- **length** : Duration (in minutes)
- **budget** : Budget for the films production (in $1000000s)
- **votes** : Number of positive votes received by viewers
- **genre** : Genre of the film
- **rating** :IMDB rating from 0 to 10
- **high_rating** : 1 for IMDB ratings greater than or equal to seven and 0 for ratings less than 7
- **rate** : 'Rating greater than 7' got high_rating = 1 and 'Rating less than 7' for high_rating = 0

3.2 Summary Statistics

Table 3: Summary statistics on the IMDB data by variables.

Variables	Mean	Median	Std. Dev	Min	Max	IQR	Sample Size
year	1,976.21	1,982.00	23.44	1,896.00	2,005.00	39.00	1,937.00
length	82.88	90.00	35.62	1.00	316.00	24.00	1,937.00
budget	12.03	12.00	2.92	3.20	21.20	3.90	1,937.00
votes	590.47	31.00	3,894.33	5.00	103,854.00	104.00	1,937.00
rating	5.29	4.60	2.05	0.60	9.30	4.00	1,937.00

The Table 3 shows that the summary for the columns year, length, budget, votes and rating.

- For the variable year, the years of the films ranges from 1896 to 2005.

- For the variable length, the films runs from 1 minute to 316 minutes.The median for length of films is 90 minutes.

- For variable budget, the budget of films is from 3.2 ($1000000s) to 21.2 ($1000000s).The median budget of a film is 12($1000000s).

- For variable votes, the votes of films ranges from 5 to 103,854 which suggests large variation. It can be observed the IQR is relatively large as well.

3.3 Correlation

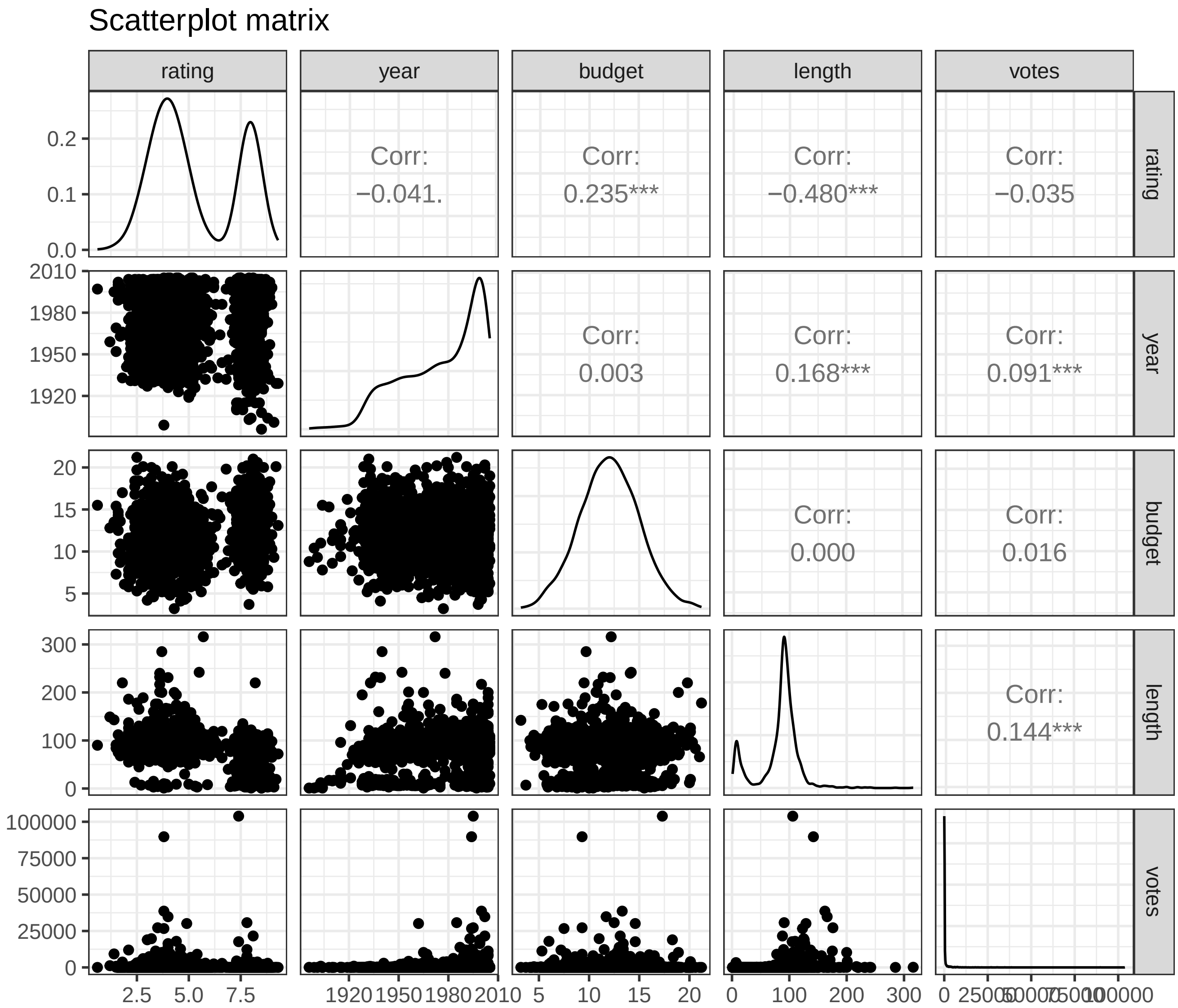

Figure 1: Scatterplot matrix between rating and explanatory variables.

The Figure 1 shows weak correlation between the variables. ‘length’ shows not so strong negative correlation with ‘rating’.

3.4 Visualization

3.4.1 Histograms

The Figure 2 shows that the data structures follow exponential distributions. The variable ‘votes’ displays skewness due to large difference in values of maximum and minimum values. To reduce this skewness and facilitate more robust analysis, a logarithmic transformation is used.

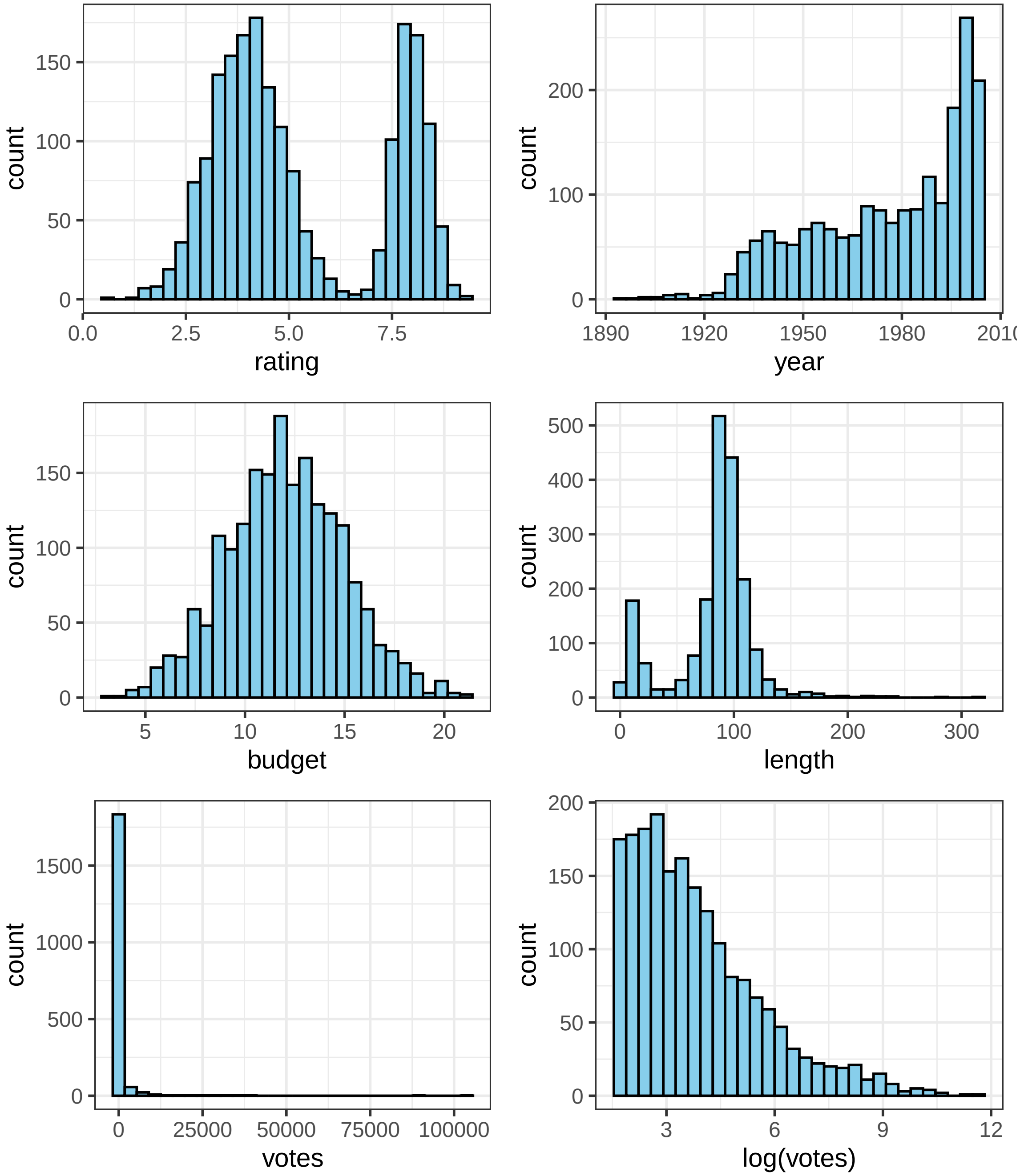

Figure 2: Histograms of statistical distribution for varibles

3.4.2 Scatter plot for rating vs explanatory variables

The Figure 3 suggests that there seems to be no linear relationship between the response variable and the explanatory variables which justifies the weak correlation observed earlier.

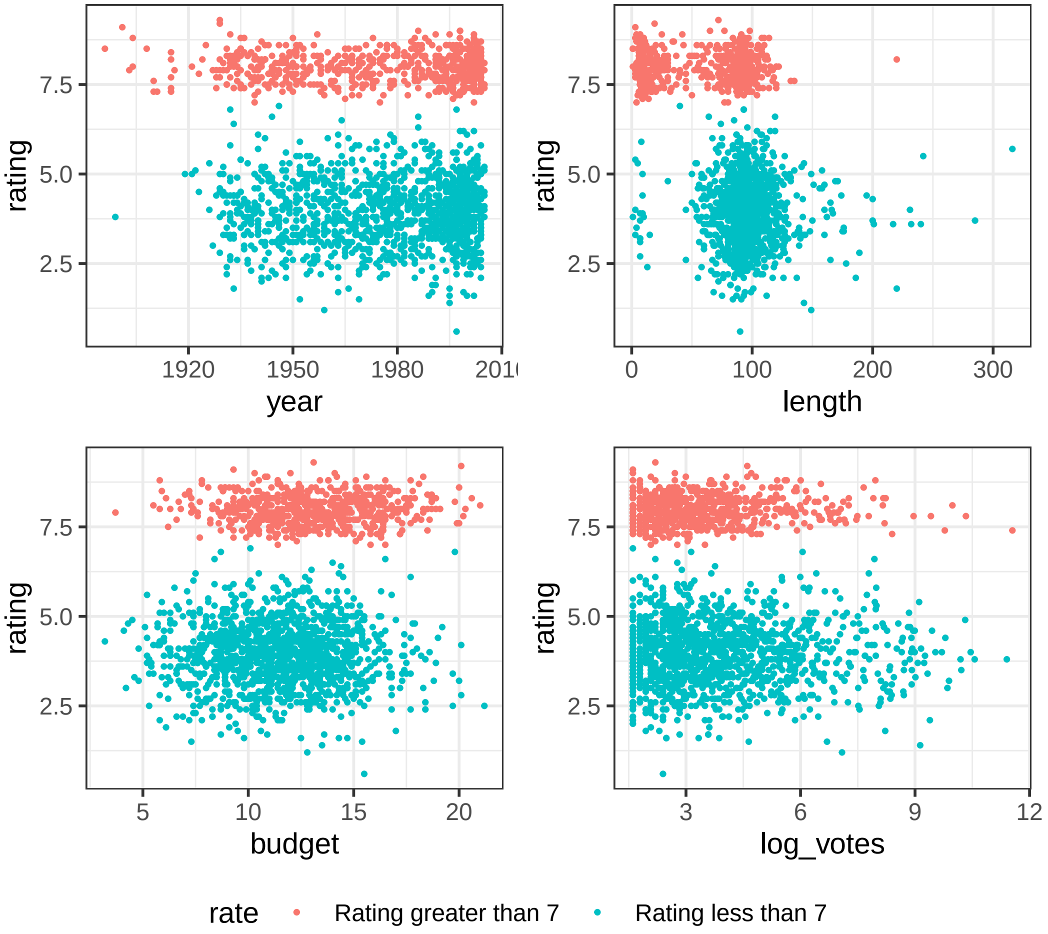

Figure 3: Scatterplots between rating and four explanatory variables.

3.4.3 Boxplot for genre

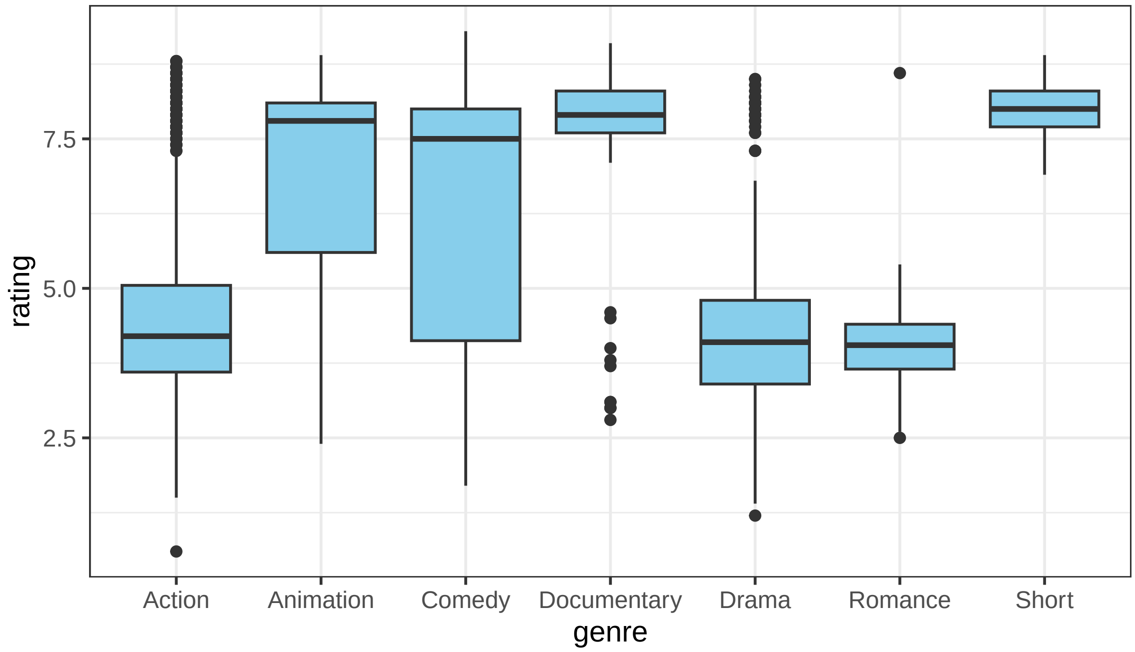

Figure 4: Boxplot of ratings by genre.

The Figure 4 distribution of rating genre wise. Outliers for ratings can be seen for the genre's Action, Documentary, Drama and Romance.

3.5 The relationship response and explanatory variable

3.5.1 Variable 1: Length

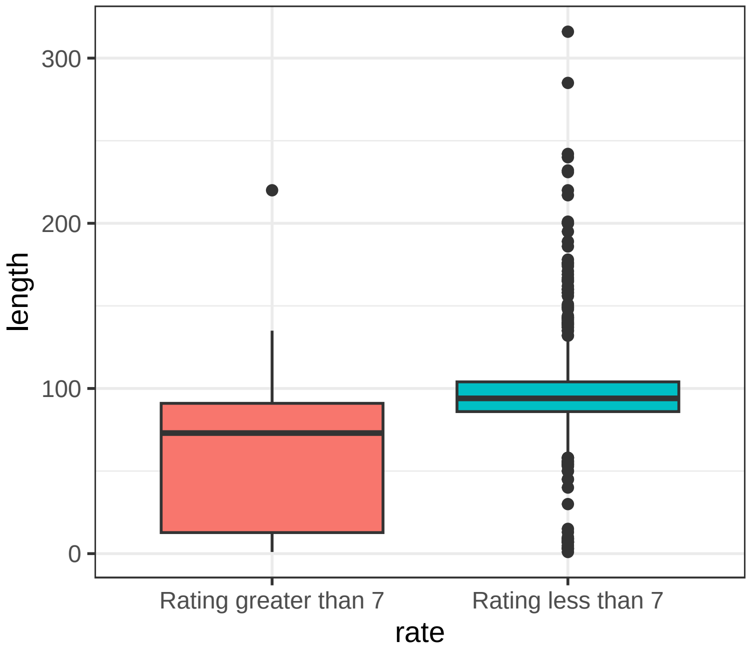

Figure 5: Boxplot of length by rating.

The Figure 5 shows that the median film length of films with 'Rating greater than 7' is less than that of 'Rating less than 7' films. It can be observed IQR of 'Rating less than 7' is smaller but has many outliers.

Table 4: Summary statistics on length by rating.

rate	Mean	Median	St.Dev	Min	Max	IQR	Sample Size
Rating greater than 7	57.05	73.00	39.49	1.00	220.00	78.25	644.00
Rating less than 7	95.74	94.00	25.03	1.00	316.00	18.00	1, 293.00

The Table 4 The median length film with 'Rating greater than 7' is (73 minutes) lower than that with 'Rating less than 7' (95.74 minutes).

3.5.2 Variable 2 : budget

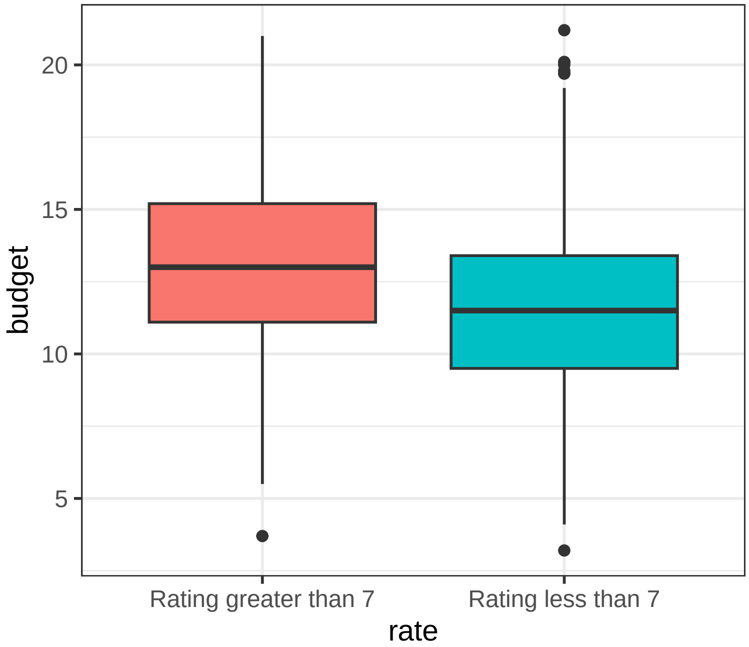

Figure 6: Boxplot of budget by rating.

The Figure 6 shows that the median budget film of ‘Rating greater than 7’ is slightly higher than that of ‘Rating less than 7’ films. There are 9 outliers.

Table 5: Summary statistics on budget by rating.

rate	Mean	Median	Std. Dev	Min	Max	IQR	Sample Size
Rating greater than 7	13.09	13.00	2.84	3.70	21.00	4.10	644.00
Rating less than 7	11.51	11.50	2.82	3.20	21.20	3.90	1, 293.00

The Table 5 shows that the mean and median for ‘Rating greater than 7’ is almost equal. Similarly, it can be observed for and ‘Rating less than 7’ as well. This suggests a normal distribution. The variability is also equivalent for the 2 categories.

3.5.3 Variable 3 : log_votes

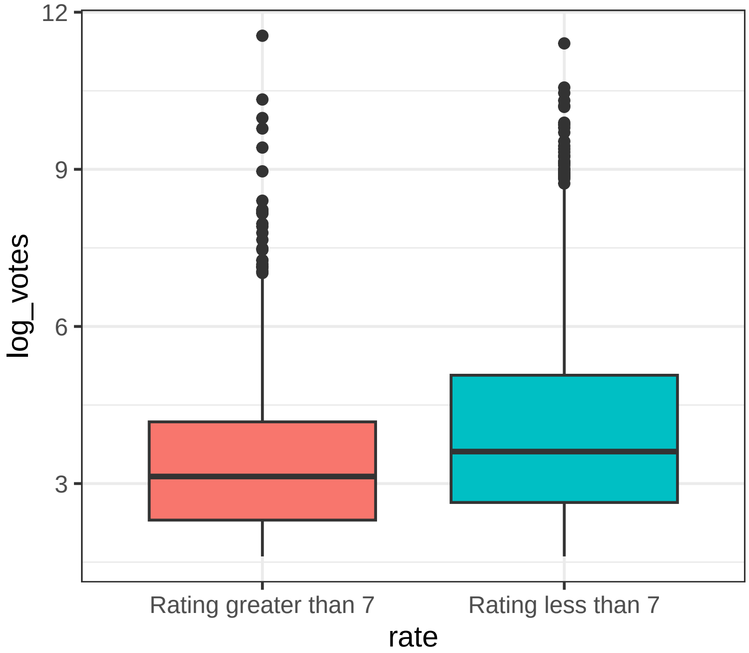

Figure 7: Boxplot of log_votes by rating.

The Figure 7 shows that the median log_votes film of 'Rating greater than 7' films is lower than that of 'Rating less than 7' films.

Table 6: Summary statistics of votes(log) by rating.

rate	Mean	Median	Std. Dev	Min	Max	IQR	Sample Size
Rating greater than 7	3.47	3.14	1.58	1.61	11.55	1.88	644.00
Rating less than 7	4.03	3.61	1.86	1.61	11.40	2.43	1, 293.00

The Table 6 shows that the mean and median for 'Rating greater than 7' is almost equal.

3.5.4 Variable 4 : genre

The ratio of ratings above 7 to ratings below 7 and sample sizes for each type

Table 7: Summary statistics of genre

genre	Rating greater than 7	Rating less than 7	genre_sum_count
Action	15.9% (92)	84.1% (487)	579
Animation	73.5% (86)	26.5% (31)	117
Comedy	59.1% (266)	40.9% (184)	450
Documentary	89.9% (89)	10.1% (10)	99
Drama	5.7% (34)	94.3% (561)	595
Romance	5.0% (1)	95.0% (19)	20
Short	98.7% (76)	1.3% (1)	77

It can be seen the size for the genre 'Romance' is only 20, which is comparatively small. We observe the following: - Animation, Documentary, Short - have more films with 'Rating greater than 7' - Action, Drama, Romance - have more films with 'Rating less than 7' - Comedy - films with 'Rating greater than 7' are moderately higher than 'Rating less than 7'

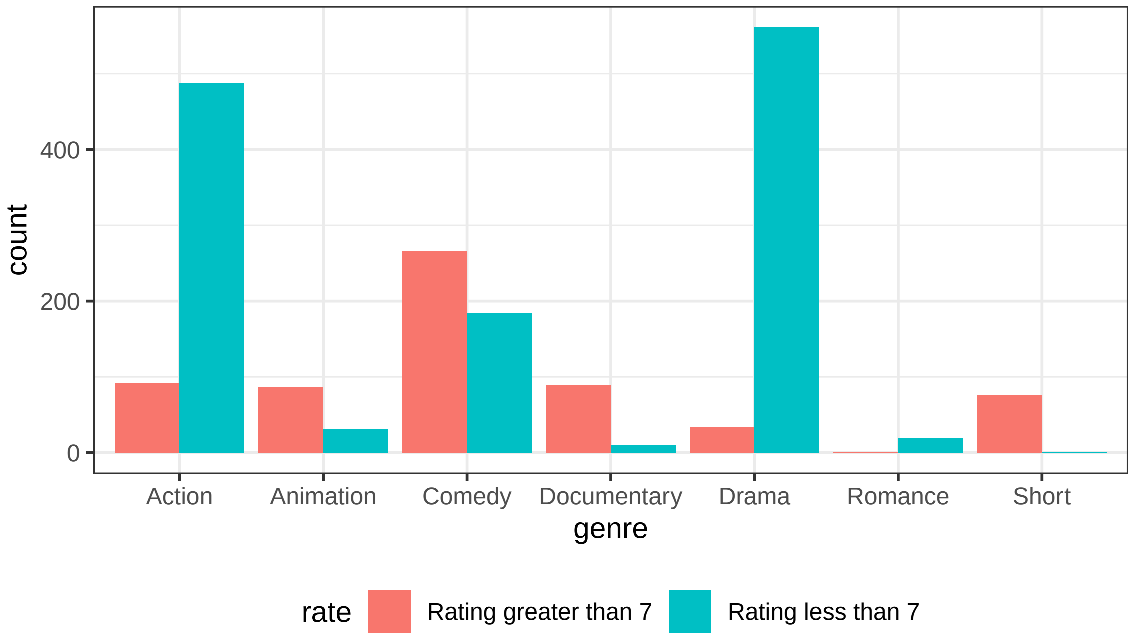

Figure 8: Dodged barplot of genre by rating.

The Figure 8 displays the information in the table Table 7

3.5.5 Outliers

It can be observed from the scatter plot that there are outliers present especially for length and votes. Then extreme outlier values are replaced by threshold values corresponding to specific percentiles -

- length - 10th and 90th percentiles
- budget- 5th percentile and 95th percentile
- log_votes- Since logarithmic transformation had already removed a considerable amount of outlier, the threshold was set to 10th and 90th percentiles

This replacement strategy aims to mitigate the impact of outliers on the analysis while retaining the overall distribution of the data. The approach ensures that extreme values are transformed to less extreme values, thereby improving the robustness of subsequent statistical analyses.

4 Formal Data Analysis

The response variable 'high_rating' is the rating for 1937 films taking the values 1 for 'Rating greater than 7' and 0 for 'Rating less than 7'. Predictors include the properties of films 'year', 'length', 'budget', 'log_votes' and 'genre'. It is assumed that $high\_rating_i \sim \text{Bin}(1, p_i)$ where p_i is the probability of film with 'Rating greater than 7' for the ith film. A logistic regression model is fitted.

4.1 Fitting the Model

Baseline category for our binary response high_rating is 0 i.e 'Rating less than 7'

4.1.1 Saturated Model

A full model with all the continuous independent variables 'year', 'length', 'budget', 'log_votes' and categorical explanatory variable 'genre' is explored:

$$\begin{aligned}
\ln\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta_1 \cdot \text{year}_i + \beta_2 \cdot \text{length}_i + \beta_3 \cdot \text{budget}_i + \beta_4 \cdot \text{log\_votes}_i \\
&\quad + \beta_{\text{Animation}} \cdot \mathbb{1}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{1}_{\text{Comedy}}(i) \\
&\quad + \beta_{\text{Document}} \cdot \mathbb{1}_{\text{Document}}(i) + \beta_{\text{Drama}} \cdot \mathbb{1}_{\text{Drama}}(i) \\
&\quad + \beta_{\text{Romance}} \cdot \mathbb{1}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{1}_{\text{Short}}(i)
\end{aligned}$$

$$\mathbb{1}_{\text{genre}}(i) = \begin{cases} 1 & \text{if genre of } i\text{th observation is in genre,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{genre} = \{\text{Animation, Comedy, Documentary, Drama, Romance, Short}\}$$

Table 8: Summary for Saturated Model

term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-21.841	7.289	-2.996	0.003	-36.239	-7.641
year	0.009	0.004	2.548	0.011	0.002	0.017
length	-0.070	0.005	-13.995	0.000	-0.080	-0.061
budget	0.560	0.039	14.357	0.000	0.485	0.638
log_votes	0.019	0.063	0.306	0.760	-0.104	0.142
genreAnimation	-0.425	0.413	-1.028	0.304	-1.241	0.382
genreComedy	3.079	0.217	14.170	0.000	2.663	3.515
genreDocumentary	5.064	0.460	11.012	0.000	4.204	6.015
genreDrama	-1.631	0.279	-5.844	0.000	-2.198	-1.101
genreRomance	-2.433	1.854	-1.312	0.189	-6.270	0.323
genreShort	3.226	1.072	3.010	0.003	1.521	6.160

Baseline category for explanatory variable ‘genre’ is “Action”

From Table 8 the following can be observed:

- **Variable Selection** log_votes has a p-value of 0.760. This suggests that there this parameter should not be included in the model.
- **Hypothesis Testing** Since log-votes has p-value > 0.05, it is not statistically significant and does not contribute in explaining the variation in the response variable.
- **95% Confidence Interval** The approximate 95% confidence interval of log_votes contains zero, it can be concluded log_votes is not statistically significant.

Analysis of Deviance Table

Table 9: Anova Table

Df	Deviance	Resid. Df	Resid. Dev
NA	NA	1,936.00	2,463.54
1.00	3.01	1,935.00	2,460.53
1.00	588.93	1,934.00	1,871.60
1.00	189.57	1,933.00	1,682.03
1.00	3.20	1,932.00	1,678.83
6.00	660.86	1,926.00	1,017.97

In Table 9 each row represents a term (predictor variable) added to the model. It can be observed that largest reduction in residual deviance comes when adding ‘genre’ and the smallest when adding ‘year’ and ‘log_votes’. A model without ‘year’ and ‘log_votes’ could be tried.

Goodness-of-fit

1. Deviance : for a GLM model that fits the data well the approximate deviance D is $\chi^2(m-p)$ where m is the number of parameters in the saturated model (full model) and p is the number of parameters in the model of interest. In the above model, 2463.5-1018.0 is larger than 95th percentile of the $\chi^2(1936-1926)$. There is no evidence of lack of fit.
2. Hosmer-Lemeshow goodness of fit test: For a model with binary responses,

   H_0 = the model fits the data well,

   H_1 = the model does not fit the data well

```
#Deviance#
m0_q1 <- qchisq(df=10, p=0.95) # 18.30704
# Hosmer-Lemeshow goodness of fit test
m0_hl <- HLTest(m0_model, g = 6)
```

A large p-value indicates no lack of fit. From the above output there is no evidence of lack of fit.

Assumptions

1. The dependent variable is binary
2. Independence of observations
3. The independent variable do not correlate too strongly with each other
4. Linearity of continuous explanatory variables and the log-odds outcome
5. No outliers

The first assumption is fulfilled as the the ’rating; has been converted to a binary variable in accordance. For asuumption 2, since the observation belong to independent fils, it is satisfied. For assumption 3, Figure 1 justifies that there are no strong correlations between the independent variables.

Assumption 4: Check linearity of continuous variables against log odds of the dependent variable

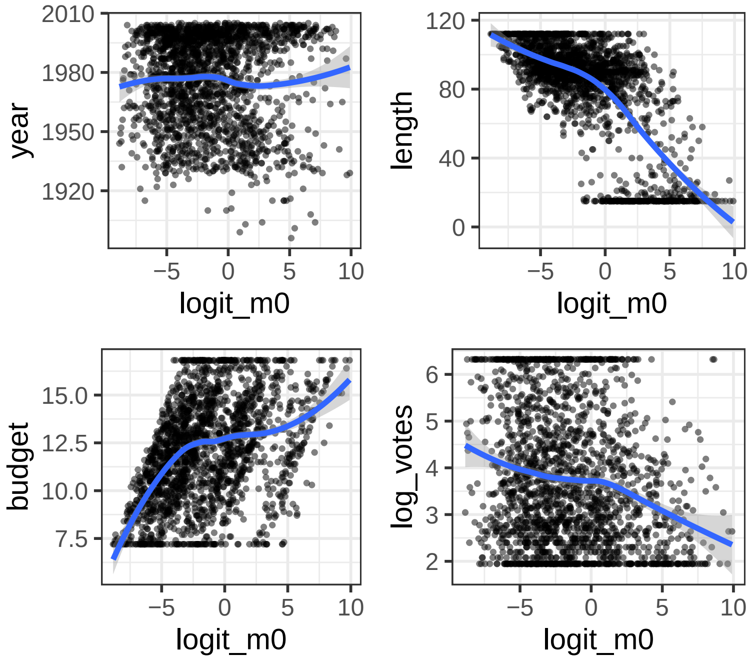

Figure 9: Checking Linearity for m0_model

The relationship between continuous variables against log-odds seems to be fairly linear.

Assumption 5: Checking for outliers

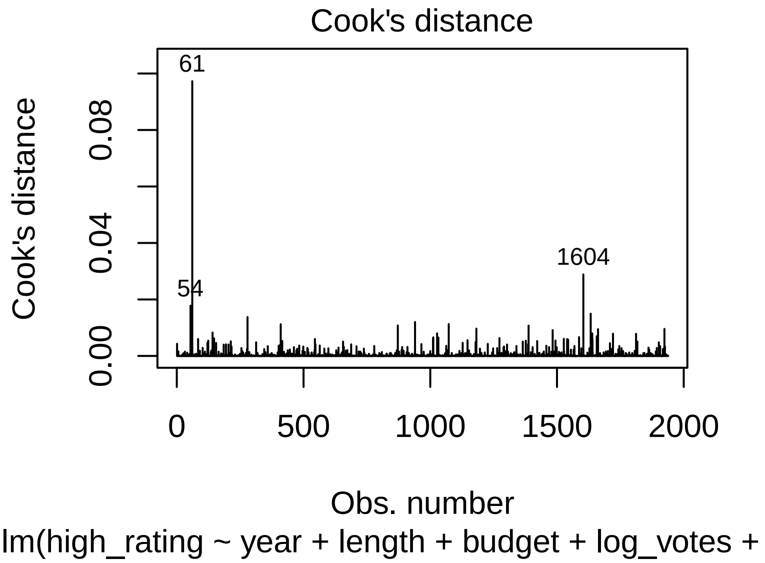

Figure 10: Checking Outliers for m0_model

The model has outliers which can be observed in Figure 10 .This is due to the outliers present in the data set.

4.1.2 Model 1

A model with continuous independent variables with 'year', 'length', 'budget' and categorical explanatory variable is explored:

$$\begin{aligned}\ln\left(\frac{p_i}{1-p_i}\right) = \alpha &+ \beta_1 \cdot \text{year}_i + \beta_2 \cdot \text{length}_i + \beta_3 \cdot \text{budget}_i \\ &+ \beta_{\text{Animation}} \cdot \mathbb{1}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{1}_{\text{Comedy}}(i) \\ &+ \beta_{\text{Document}} \cdot \mathbb{1}_{\text{Document}}(i) + \beta_{\text{Drama}} \cdot \mathbb{1}_{\text{Drama}}(i) \\ &+ \beta_{\text{Romance}} \cdot \mathbb{1}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{1}_{\text{Short}}(i)\end{aligned}$$

$$\mathbb{1}_{\text{genre}}(i) = \begin{cases} 1 & \text{if genre of } i\text{th observation is in genre,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{genre} = \{\text{Animation, Comedy, Documentary, Drama, Romance, Short}\}$$

Table 10: Summary for m1_model

term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-22.196	7.194	-3.085	0.002	-36.415	-8.190
year	0.010	0.004	2.645	0.008	0.003	0.017
length	-0.070	0.005	-14.259	0.000	-0.080	-0.061
budget	0.559	0.039	14.355	0.000	0.485	0.638
genreAnimation	-0.407	0.409	-0.995	0.320	-1.215	0.391
genreComedy	3.090	0.214	14.413	0.000	2.679	3.521
genreDocumentary	5.054	0.459	11.019	0.000	4.197	6.003
genreDrama	-1.629	0.279	-5.842	0.000	-2.196	-1.100
genreRomance	-2.414	1.846	-1.308	0.191	-6.240	0.334
genreShort	3.226	1.072	3.010	0.003	1.521	6.159

From Table 10 the following can be observed:

- **Variable Selection** All the estimates are statistically significant with a p-value < 0.05.
- **Hypothesis Testing** Since p-values < 0.05 for all the parameters, the predictors are statistically significant and contributes to explaining the variation in the response variable.
- **95% Confidence Interval** The approximate 95% confidence interval for all the parameters do not contain 0, it can be concluded they are statistically significant.

Goodness-of-fit

1. Deviance : In the above model, 2463.5-1018.1 = 1445.4 is larger than 95th percentile of the $\chi^2(1936-1927) = 16.92$. There is no evidence of lack of fit.
2. Hosmer-Lemeshow goodness of fit test: From the output, there is no evidence of lack of fit.

```
#Deviance#
m1_q <- qchisq(df=9, p=0.95) # 16.91898
#Hosmer-Lemeshow goodness of fit test#
m1_hl <- HLTest(m1_model, g = 6) #p-value = 0.1181 >0.05
```

Assumptions

1. The dependent variable is binary
2. Independence of observations
3. The independent variable do not correlate too strongly with each other
4. Linearity of continuous explanatory variables and the log-odds outcome
5. No outliers

The first assumption is fulfilled as the the ’rating; has been converted to a binary variable in accordance. For assumption 2, since the observation belong to independent films, it is satisfied. For assumption 3, Figure 1 justifies that there are no strong correlations between the independent variables.

Assumption 4: Check linearity of continuous variables against log odds of the dependent variable

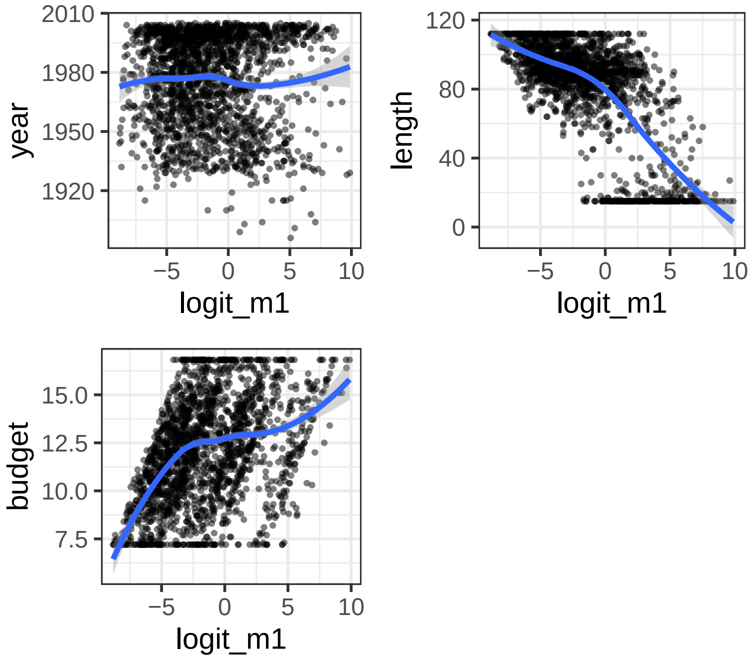

Figure 11: Checking Linearity for m1_model

The relationship between continuous variables against log-odds seems to be fairly linear for all the variables in Figure 11

Assumption 5:Checking for outliers

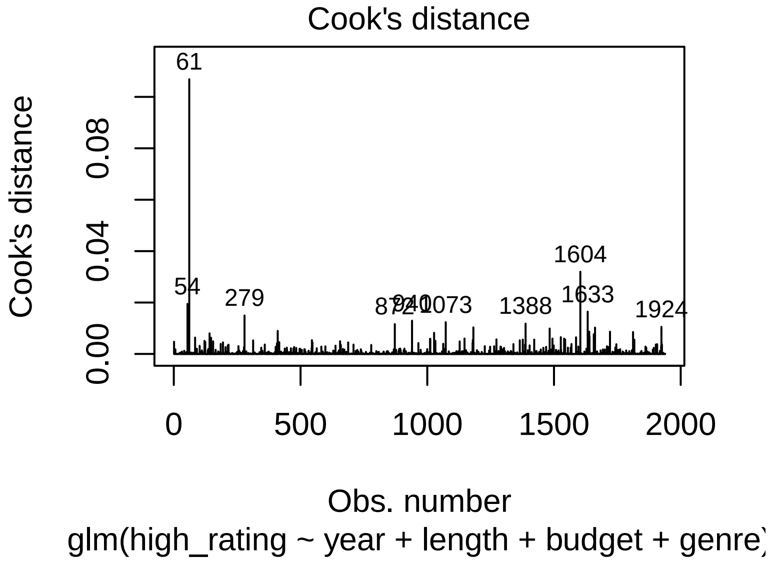

Figure 12: Checking Outliers for m1_model

The model has outliers which can be observed in Figure 12. This is due to the outliers present in the data set.

4.1.3 Model 2

$$\begin{aligned}\ln\left(\frac{p_i}{1-p_i}\right) = \alpha &+ \beta_1 \cdot \text{length}_i + \beta_2 \cdot \text{budget}_i \\ &+ \beta_{\text{Animation}} \cdot \mathbb{1}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{1}_{\text{Comedy}}(i) \\ &+ \beta_{\text{Document}} \cdot \mathbb{1}_{\text{Document}}(i) + \beta_{\text{Drama}} \cdot \mathbb{1}_{\text{Drama}}(i) \\ &+ \beta_{\text{Romance}} \cdot \mathbb{1}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{1}_{\text{Short}}(i)\end{aligned}$$

$$\mathbb{1}_{\text{genre}}(i) = \begin{cases} 1 & \text{if genre of } i\text{th observation is in genre,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{genre} = \{\text{Animation, Comedy, Documentary, Drama, Romance, Short}\}$$

Table 11: Summary for m2_model

term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-3.244	0.551	-5.888	0.000	-4.337	-2.175
length	-0.067	0.005	-14.305	0.000	-0.077	-0.058
budget	0.554	0.039	14.363	0.000	0.480	0.632
genreAnimation	-0.409	0.408	-1.003	0.316	-1.214	0.387
genreComedy	3.056	0.213	14.346	0.000	2.648	3.484
genreDocumentary	5.154	0.456	11.300	0.000	4.302	6.099
genreDrama	-1.584	0.274	-5.790	0.000	-2.139	-1.064
genreRomance	-2.362	1.705	-1.385	0.166	-6.043	0.237
genreShort	3.428	1.069	3.206	0.001	1.730	6.359

From the Table 11 following can be observed:

- **Variable Selection** All the estimates are statistically significant with a p-value < 0.05. However, it Table 9 computed that the smallest reduction in residual deviance comes from 'year'.
- **Hypothesis Testing** Since p-values < 0.05 for all the parameters, the predictors are statistically significant and contributes to explaining the variation in the response variable.
- **95% Confidence Interval** The approximate 95% confidence interval for all the parameters do not contain 0, it can be concluded they are statistically significant.

Goodness-of-fit

1. Deviance : In the above model, 2463.5-1025.2 = 1438.3 is larger than 95th percentile of the $\chi^2(1936-1928) = 16.92$. There is no evidence of lack of fit.
2. Hosmer-Lemeshow goodness of fit test: For a model with binary responses,

   H_0 = the model fits the data well,, H_1 = the model does not fit the data well

```r
#Deviance#
m2_q <- qchisq(df=9, p=0.95) # 16.91898
# Hosmer-Lemeshow goodness of fit test
m2_hl <- HLTest(m2_model, g = 6) # 0.3674
```

A large p-value indicates no lack of fit. From the above output there is no evidence of lack of fit.

Assumptions

1. The dependent variable is binary
2. Independence of observations
3. The independent variable do not correlate too strongly with each other
4. Linearity of continuous explanatory variables and the log-odds outcome
5. No outliers

The first assumption is fulfilled as the the 'rating; has been converted to a binary variable in accordance. For assumption 2, since the observation belong to independent films, it is satisfied. For assumption 3, Figure 1 justifies that there are no strong correlations between the independent variables.

Assumption 4: Check linearity of continuous variables against log odds of the dependent variable

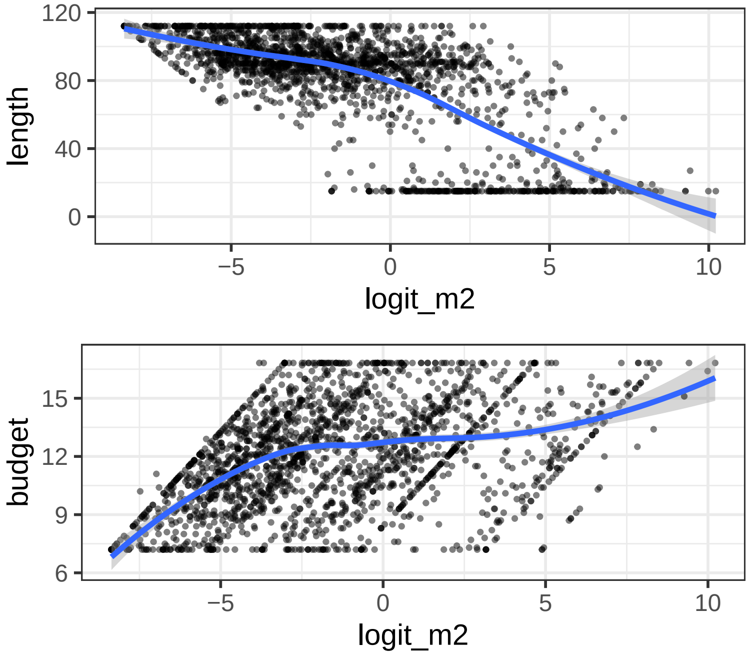

Figure 13: Checking Linearity for m2_model

The relationship between continuous variables against log-odds seems to be fairly linear for all the variables in Figure 13

Assumption 5: Checking for outliers

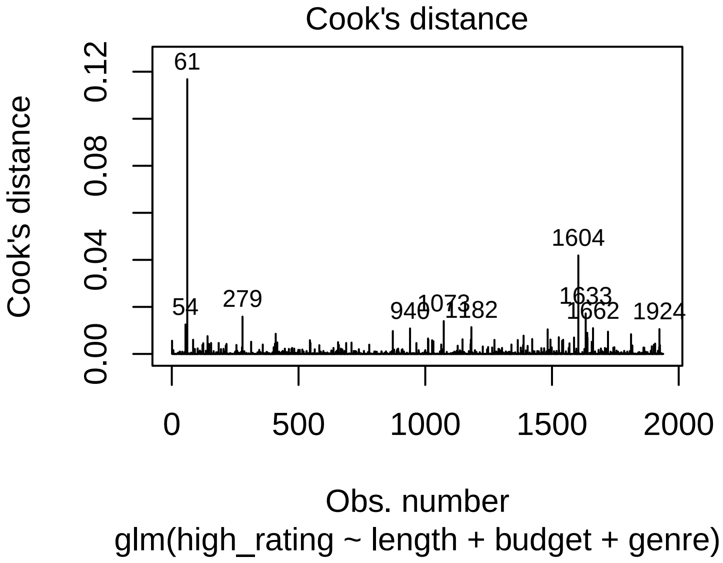

Figure 14: Checking Outliers for m2_model

The model has outliers which can be observed in Figure 14 . This is due to the outliers present in the data set.

4.1.4 Log-Odds, Odds and Probabilities

Log-Odds

The baseline category for out binary response variables is 'Rating less than 7'. This implies from the logistic regression model are on the log-odds scale for 'Rating greater than 7' in comparison to the baseline'. 'Rating less than 7'.

The coefficients are extracted from the Table 11 and are found to be:

- **Intercept (-3.244):** This represents the log-odds of the outcome when all predictor variables are zero.
- **Length (-0.067):** For every one-unit increase in the "length" variable, holding all other variables constant, the log-odds of the outcome decrease by 0.067.
- **Budget (0.554):** For every one-unit increase in the "budget" variable, holding all other variables constant, the log-odds of the outcome increase by 0.554 .
- **Genre Animation (-0.409):** Observations belonging to the "Animation" genre have log-odds 0.409 lower than observations belonging to the "Action" genre, holding all other variables constant.
- **Genre Comedy (3.056):** Observations belonging to the "Comedy" genre have log-odds 3.056 higher than observations belonging to the "Action" genre, holding all other variables constant.

- **Genre Documentary (5.154):** Observations belonging to the "Documentary" genre have log-odds 5.154 higher than observations belonging to the "Action" genre, holding all other variables constant.
- **Genre Drama (-1.584):** Observations belonging to the "Drama" genre have log-odds 1.584 lower than observations belonging to the "Action" genre, holding all other variables constant.
- **Genre Romance (-2.362):** Observations belonging to the "Romance" genre have log-odds 2.362 lower than observations belonging to the "Action" genre, holding all other variables constant.
- **Genre Short (3.428):** Observations belonging to the "Short" genre have log-odds 3.428 higher than observations belonging to the "Action" genre, holding all other variables constant.

The equation are as follows:

For the "Action" genre (reference category):

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i$$

For the "Animation" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i - 0.409$$

For the "Comedy" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i + 3.056$$

For the "Documentary" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i + 5.154$$

For the "Drama" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i - 1.584$$

For the "Romance" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i - 2.362$$

For the "Short" genre:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -3.244 - 0.067 \times \text{length}_i + 0.554 \times \text{budget}_i + 3.428$$

where p = Prob(Rating Greater than 7)

and *1-p* = Prob(Rating less than 7)

95% confidence interval for these log-odds can be found in Table 11

Hence the point estimate for the log-odds can be displayed graphically in Figure 15 with there corresponding 95% confidence interval.

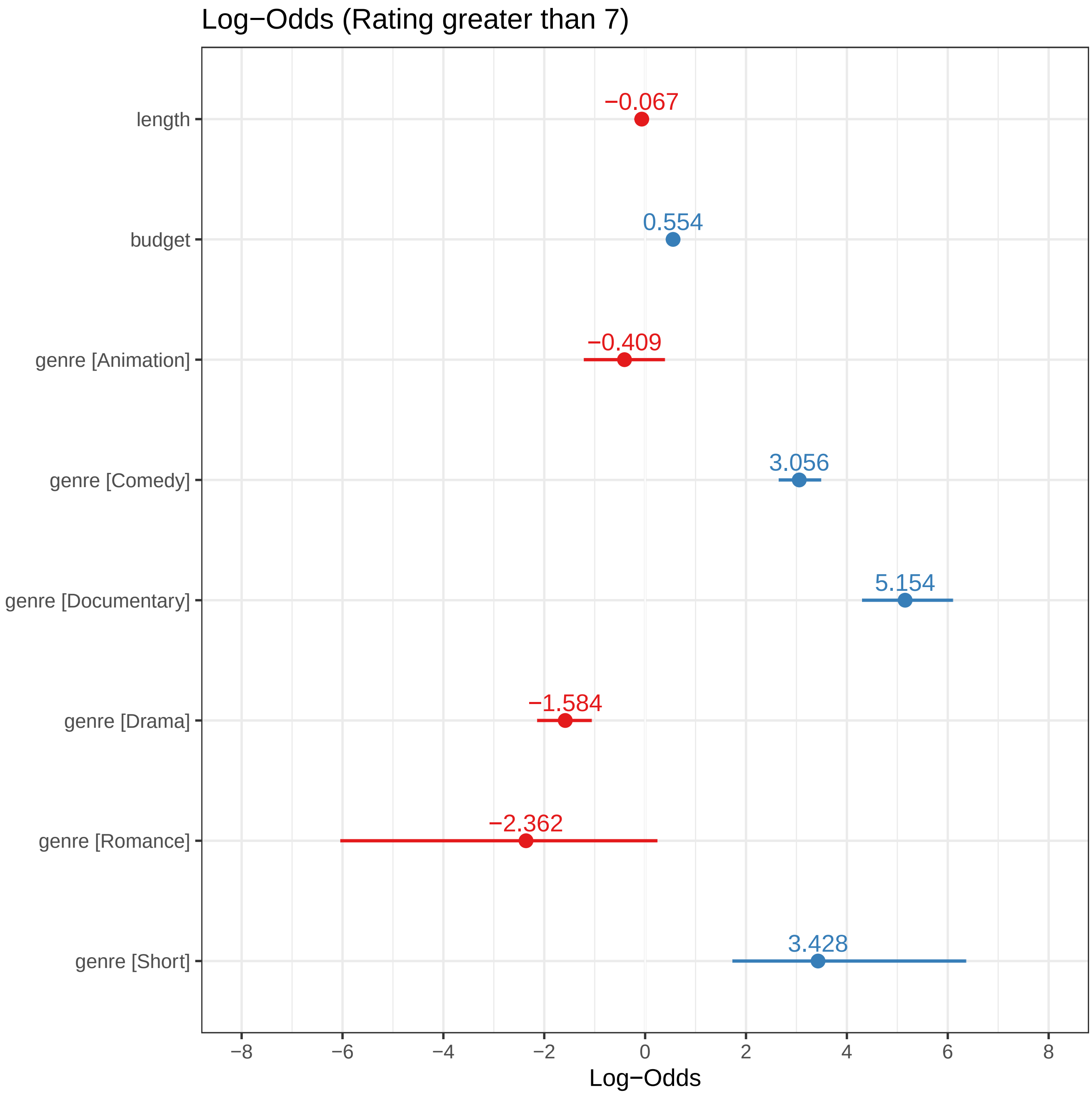

Figure 15: Plot for Log-Odds

The estimates of the log-odds were added to thr data set.

Odds

$$
\begin{aligned}
\text{Odds}(p_i) &= \frac{p_i}{1 - p_i} \\
&= \exp\big(\alpha + \beta_1 \cdot \text{length}_i + \beta_2 \cdot \text{budget}_i \\
&\quad + \beta_{\text{Animation}} \cdot \mathbb{1}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{1}_{\text{Comedy}}(i) \\
&\quad + \beta_{\text{Document}} \cdot \mathbb{1}_{\text{Document}}(i) + \beta_{\text{Drama}} \cdot \mathbb{1}_{\text{Drama}}(i) \\
&\quad + \beta_{\text{Romance}} \cdot \mathbb{1}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{1}_{\text{Short}}(i)\big)
\end{aligned}
$$

On the **odds** scale the regression coefficients are given by:

Table 12: Odds Ratios for m2_model

Variable	Odds Ratio
(Intercept)	0.039
length	0.935
budget	1.740
genreAnimation	0.664
genreComedy	21.234
genreDocumentary	173.181
genreDrama	0.205
genreRomance	0.094
genreShort	30.816

On the odds scale,

- The intercept coefficient (0.039) represents the odds of the film being in the "high_rating" category when all predictor variables are zero. Given this is not viable, the intercept value is very close to zero.
- For each one unit increase in "length", the odds of the film being in the "high_rating" category decrease by a factor of approximately 0.935.
- For each one unit increase in "budget", the odds of the film being in the "high_rating" category increase by a factor of approximately 1.740.
- For each film belonging to "Animation" genre, the odds of the film being in the "high_rating" category approximately decrease by the factor 0.664 times the odds of the reference category ("Action").
- For each film belonging to "Comedy" genre, theodds of the film being in the "high_rating" category approximately increase by the factor of 21.234 the odds of the reference category.
- For each film belonging to "Documentary" genre, the odds of the film being in the "high_rating" category approximately increase by the factor of 173.181 the odds of the reference category.
- For each film belonging to "Drama" genre odds, the odds of the film being in the "high_rating" category approximately decrease by the factor of 0.205 the odds of the reference category.
- For each film belonging to "Romance" genre, the odds of the film being in the "high_rating" category approximately decrease by the factor of 0.094 the odds of the reference category.

- For each film belonging to “Short” genre, the odds of the film being in the “high_rating” category approximately increase by the factor of 30.816 the odds of the reference category.

The 95% confidence interval for the odds can be obtained by applying exponential to the log odds interval:

Table 13: Odds Ratios CI for m2_model

Odds Ratio Confidence Intervals

2.5 %	97.5 %
0.013	0.114
0.926	0.943
1.617	1.881
0.297	1.472
14.123	32.574
73.854	445.210
0.118	0.345
0.002	1.267
5.643	577.912

Hence the point estimate for the odds can be displayed graphically in Figure 16 with there corresponding 95% confidence interval.

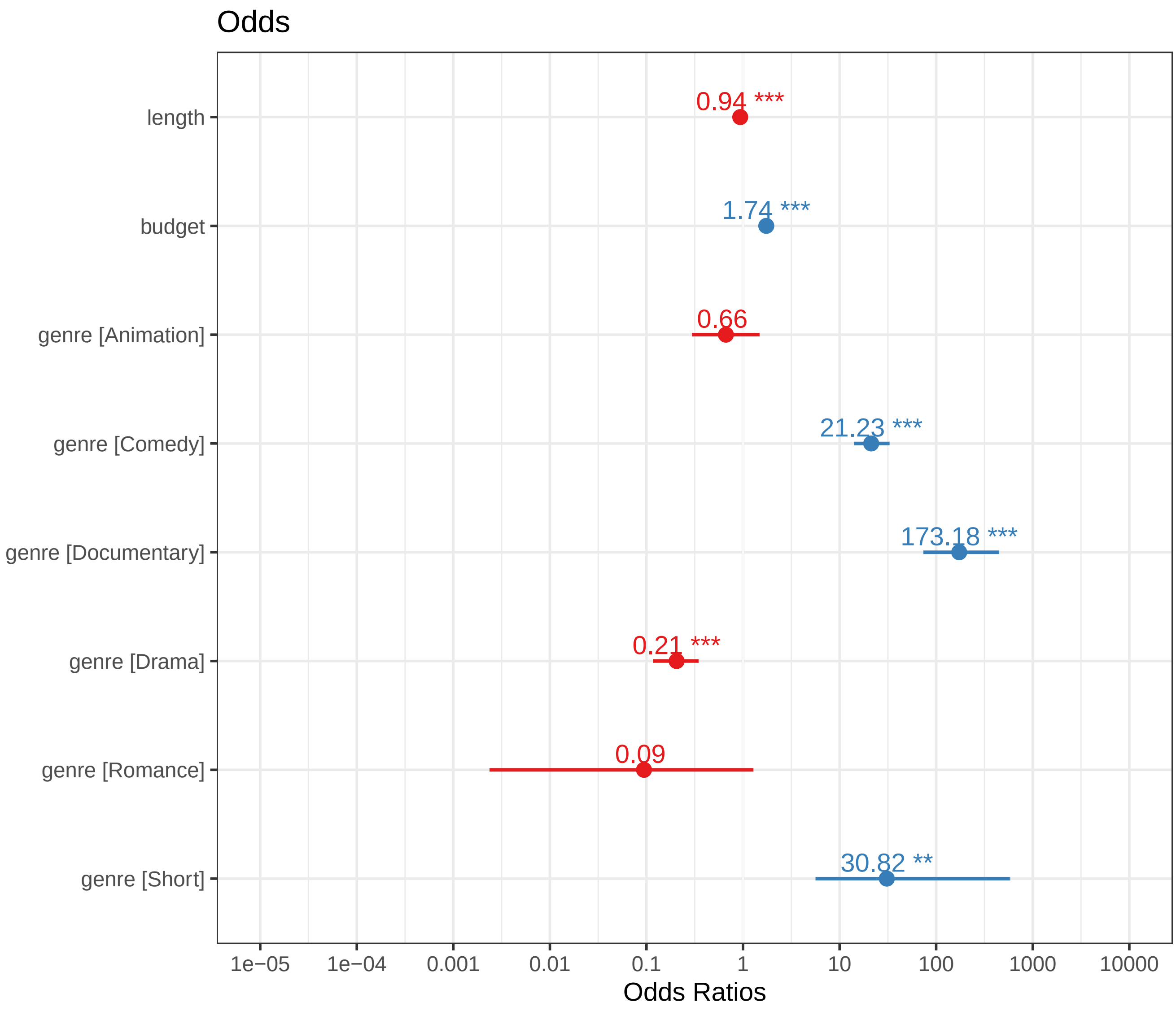

Figure 16: Plot for Odds

The odds estimates were added to the data set.

Probabilities

Probabilities added to the data set which have been formulated by using the fitted() function

$$p = \frac{\text{odds}}{\text{odds} + 1}$$

The predicted probabilities of high rating against the films ‘length’ and ‘budget’ by the films ‘genre’ is shown in Figure 17 .

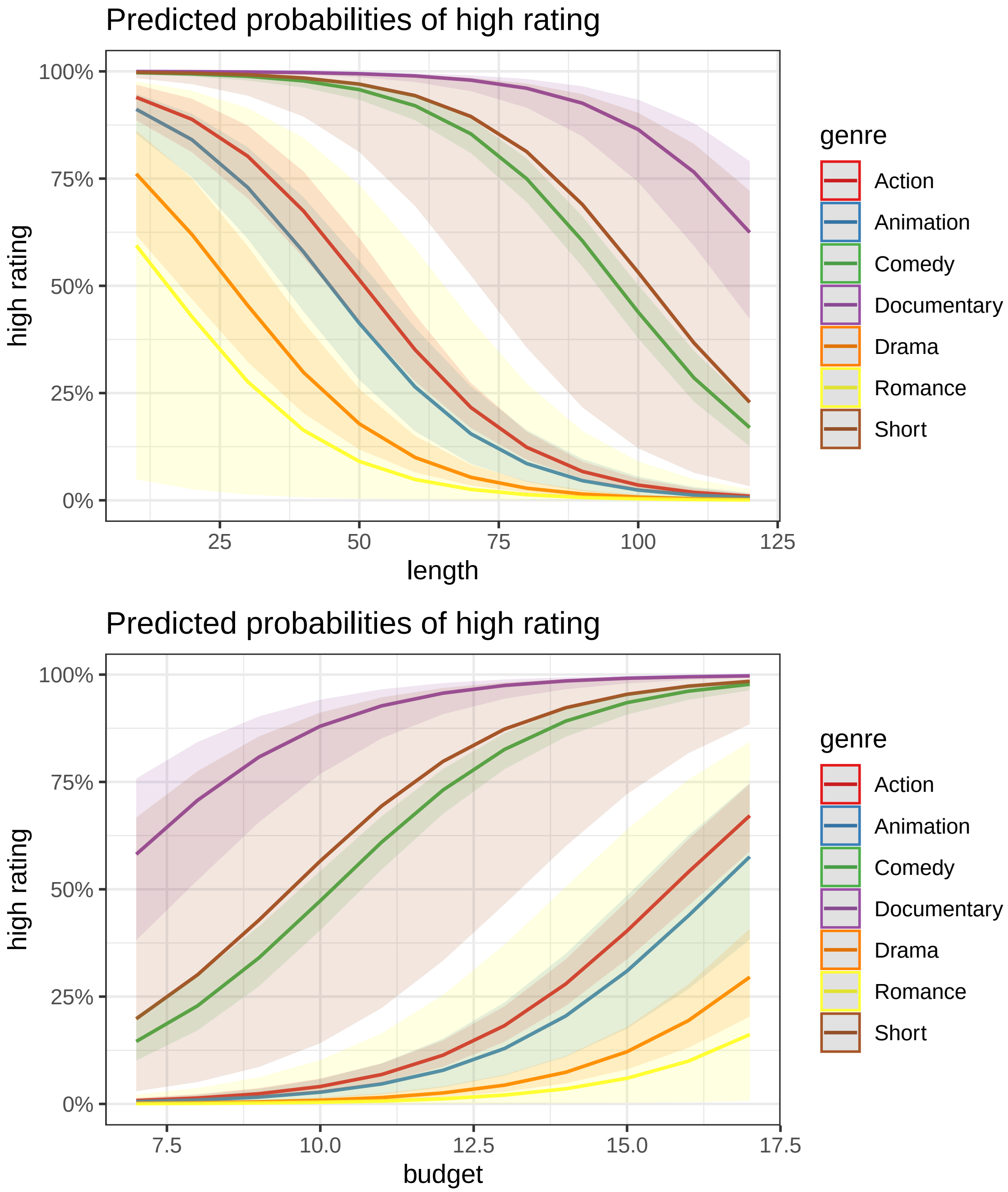

Figure 17: Plot for Predicted Probabilties

The Figure 17 shows that the probability of a film to have 'Rating greater than 7' increases with the decrease in length of the film and increases with the increase in budget of the film for all the genres.

4.2 Model Checking and Diagnostics (#Sec-mcd)

4.2.1 Model Selection

1. Likelihood Ratio Chi-Squared Statistic Test

Table 14: Likelihood Ratio Chi-Squared Test

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1926	1017.973	NA	NA	NA
1927	1018.067	-1	-0.09352632	0.759741145
1928	1025.170	-1	-7.10301538	0.007695438

This test suggests that m1_model could the best choice based on Likelihood ration test. However, m0_model has insignificant term 'log_votes' and smallest reduction in residual deviance when adding 'year' and 'log_votes'. A model without 'year' and 'log_votes' would be more suitable.

2. Residuals

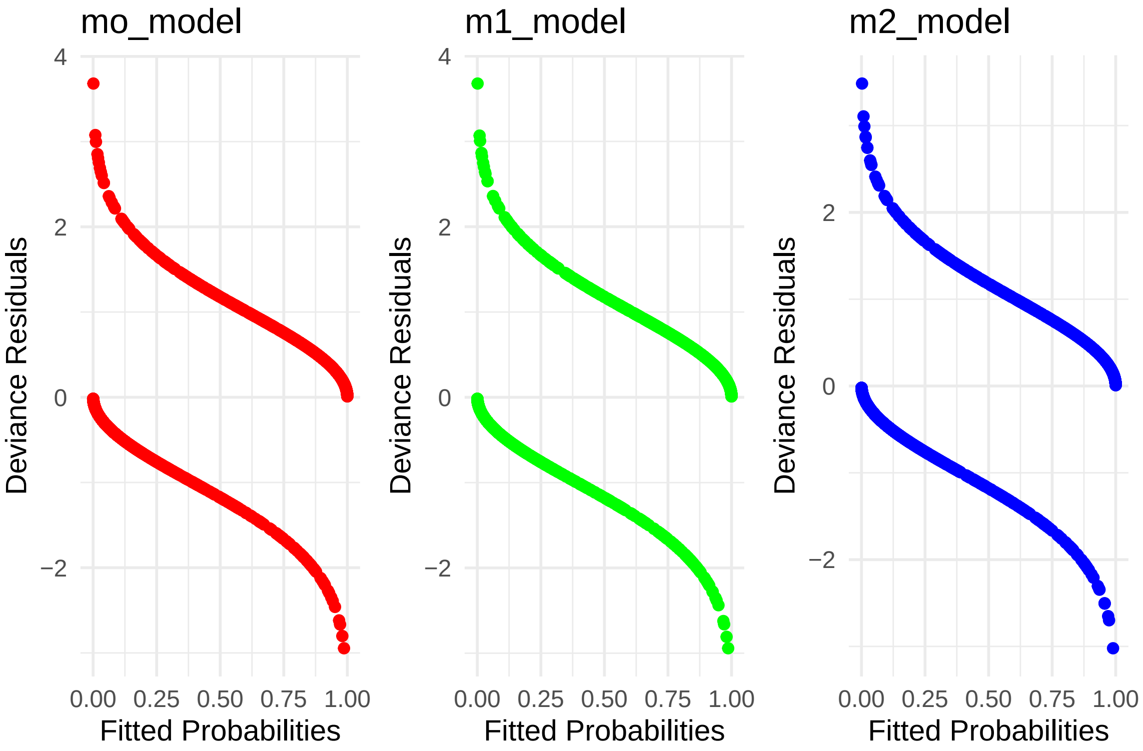

Figure 18: Plot for Residuals

Table 15: Comparison of Models

Model Comparison Table

Model	Mean_Deviance_Residual	Mean_Pearson_Residual	Mean_Fitted_Probabilities
m0_model	-0.04961650	-0.006920083	0.3324729
m1_model	-0.04945432	-0.006295522	0.3324729
m2_model	-0.05060707	-0.009906284	0.3324729

A comparison of the Mean Deviance Residual (MDR) all of the models fit the data well, with small difference suggesting **m2_model** is a better fit.

Mean Pearson Residual (MPR) varies slightly among the three models, with m0_model and m1_model having closer mean values and m2_model having the lowest mean value (-0.0099), suggesting that the **m2_model** is a better fit than the the other two models.

The mean value of Mean Fitted Probabilities (0.332) is same across all models, suggesting that the average prediction probability of a film receiving ‘Rating greater than 7’ is consistent across models.

3. ROC curve and AUC

Comparison of ROC curves, it can be observed from the figure that the ROC curves of the three models are very close to each other, almost overlapping, and all three curves are very tightly fitted to the upper left corner, indicating that all three models have good predictive ability This suggests that in terms of the balance between the sensitivity (true rate) and the specificity (false positive rate), these models have similar classification ability. This is also confirmed by the AUC values of the models, with m0_model having the highest AUC (9,951897), but the differences with m1_model (0.9518471) and m2_model (0.9512016) are very slight.

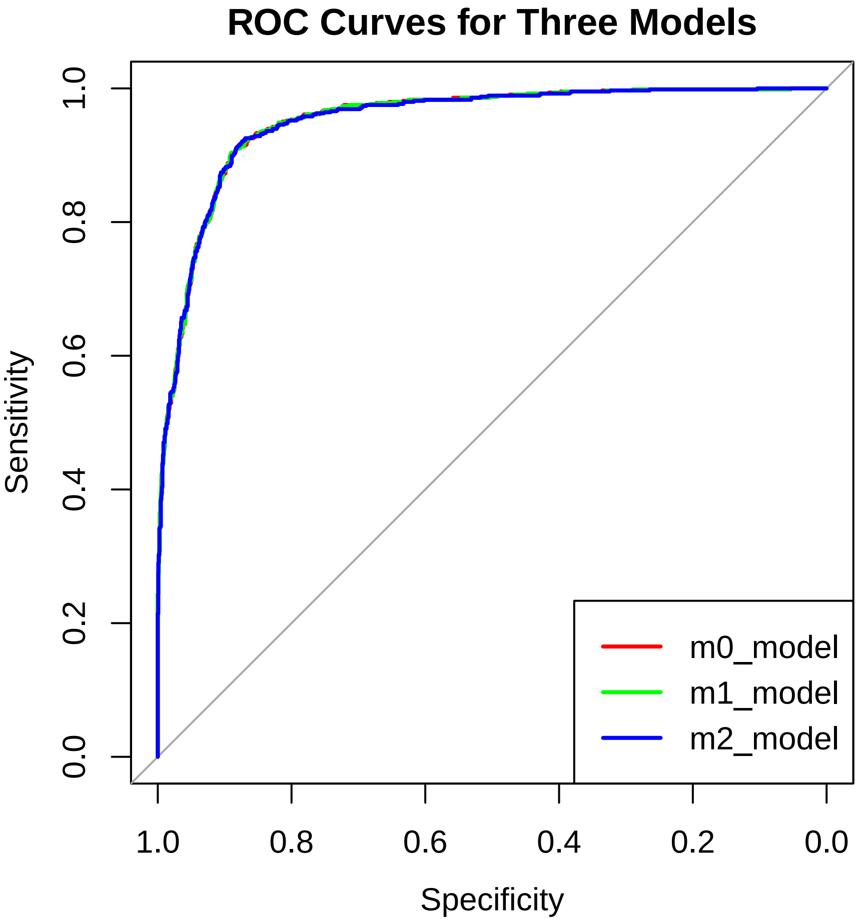

Figure 19: Plot for Predicted Probabilties

Table 16: AUC values

Model	AUC
m0_model	0.9519078
m1_model	0.9518664
m2_model	0.9512197

Comparison of AUC values shows that all models have very high predictive performance, with AUC values most 0.95, implying that the models work well in distinguishing between 'Rating greater than 7' and 'Rating less than 7' films. m0_model shows the best performance, but the difference with m1_model and m2_model is negligible, suggesting that there is little or no loss in prediction.

4. **AIC & BIC**

Table 17: AIC of Models

Model	df	AIC
m0_model	11	1039.973
m1_model	10	1038.067
m2_model	9	1043.170

Based on the AIC criterion for model selection, **m1_model** (with log_votes removed) provided the best fit to the data (AIC = 1038.203) compared to m0_model with all variables included (AIC = 1040.110). Therefore, m1_model is the preferred model.

Table 18: BIC of Models

Model	df	BIC
m0_model	11	1101.231
m1_model	10	1093.756
m2_model	9	1093.290

According to the BIC criterion, the m0_model shows the highest BIC value (1101.368), suggesting that this model may not be the most preferred choice. **m2_model** has a lowest BIC (BIC = 1093.426), suggesting that it is a better choice in a statistical perspective.

5 Conclusions

In summary, after comparing the GLM regression results and residual plots, we find that the GLM model **high_rating ~ length + budget + genre** is the most appropriate. This model effectively captures the relationship between the dependent and independent variables . The saturated model exhibit notable flaws, such as p-values above the 5% confidence level and confidence intervals containing zero.

In comparison to the other models, this model has the minimum BIC. Since BIC places a stronger penalty on model complexity than AIC for smaller data sets, simpler model is chosen. This model also performed slightly better in residual tests. All the parameters of the model are also statistically significant.It is also noted that even though m1_model performed better in likelihood ration test and AIC, it is not preferred since it has the variable ‘year’ which has the smallest reduction in residual deviance when adding to the model.

In conclusion, it can be observed that ‘length’, ‘budget’ and ‘genre’ are the properties of film that influence the IMDB ratings to be greater than 7 on not.

5.0.1 Limitations

There are few limitations to the analysis:

- Sensitivity to outliers
- Risk of overfitting the data
- Residuals may not be informative if the response is binary and if n_k is small for most covariate patterns
- The power of the Hosmer-Lemeshow test can be too small to detect lack of fit.