IMDb Prediction Challenge

Presented to: Prof Juan Camilo Serpa

MGSC 401 - 001 - Group 8

Hector Acket (261080293)

Roy Cimber (261101047)

William Fecteau (261117998)

Sam Levkovsky (261179679)

Michael Sitcoff (261116272)

March 10, 2025

**Introduction**

In the rapidly evolving world of cinema, predicting how audiences will receive a film is

both an art and a science. While studios invest millions of dollars, and filmmakers invest significant time into production, marketing, and distribution of films, the ultimate sign of a movie’s success is often strongly related to its public reception post-release. In this project, we aim to develop a predictive statistical model that forecasts IMDb ratings for twelve upcoming blockbusters that will be released shortly. By utilizing a given dataset of approximately 2000 past movies, we will apply various statistical techniques to identify which are the key factors that influence a film’s rating, followed by a prediction of what those twelve film’s IMDb ratings will look like.

Predicting IMDb ratings is a challenging task due to the numerous variables that can impact an audience’s perception of the film. There are more traditional factors that could be at play, such as the film’s cast, director, and genre, as well as some less thought about variables, such as the film’s budget, release year, and even the number of unique faces featured in the theatrical release poster. Furthermore, one’s opinion of a film is purely subjective, and the criteria that makes a good film can vary from person to person, further adding to the complexity of this task. Our goal is to build a model, based on past results, that effectively balances both the traditional and less thought of predictors, while attempting to minimize common pitfalls such as collinearity and heteroskedasticity.

Throughout this project, we will apply various statistical foundations learned thus far in class while building our predictive model. While testing various combinations of predictors and assessing the trade-offs between having too many, or not enough, variables used in our model, we will obtain a final model that will allow us to predict the IMDb ratings of the upcoming twelve films, while minimizing the potential error in our model.

This project will also allow us to bridge the gap between the technical world, and non-technical world. We will gain first-hand experience in how using data and statistical models is important in nearly every industry, however, it is even more important to be able to understand and explain those statistical results in non-technical terms, so that the general public is able to make sense of the work done, and use it to drive successful results. In other words, this project serves as both a learning experience and a practical application of statistical modeling techniques, in a real-world context.

**Data Description**

What insights can nearly 2,000 films reveal about audience preferences? Our dataset provides insight into variables spanning production details, cast attributes, and content characteristics; all potential influences on how viewers ultimately rate a film. At the center of our analysis is the dependent variable, the IMDb score itself, a metric ranging from 1 to 10 that captures the collective judgment of audiences worldwide. Our initial correlation analysis highlighted several promising relationships between these scores and various film attributes, setting the stage for our deeper investigation. All variables were rigorously evaluated for statistical significance and predictive value, with non-contributory predictors excluded to minimize model complexity and avoid overfitting.

*Key Numerical Predictors and Their Relationships*

Among the numerical predictors, movie duration emerged as significant, showing a generally positive relationship with IMDb scores in our exploratory scatter plots. However, this relationship exhibited non-linearity in residual plots, suggesting the need for polynomial transformations. Similarly, movie budget demonstrated a notable positive correlation with IMDb scores, albeit with considerable variance. The distribution of movie budgets displayed significant right-skew, prompting us to apply a logarithmic transformation (log\_movie\_budget) to normalize its distribution and stabilize variance. While the number of news articles (nb\_news\_articles) initially showed a meaningful correlation with IMDb scores and was included in early models, our analysis revealed this variable would introduce bias when predicting scores for newly released films and was finally left discarded. New movies naturally have fewer news articles written about them compared to older films in our training dataset. Additionally, we suspected a form of survivor bias, where films achieving significant success tend to generate disproportionately high numbers of news articles post-release. These factors artificially deflated the predicted score of upcoming movies. For this reason, we excluded this variable from our final model. The number of faces (nb\_faces) detected in movie promotional materials represented another numerical predictor that required logarithmic transformation. Our exploration suggested potential nonlinear aspects to this relationship, leading us to examine polynomial forms of this variable.   
  
 The movie\_meter\_IMDBpro ranking also emerged as a significant predictor after logarithmic transformation, with lower values (indicating higher film prominence) correlating with higher IMDb scores. Examination of the correlation matrix among our quantitative variables revealed moderate correlations between several predictors.   
Specifically, movie budget showed correlations with both the number of news articles and the number of faces (coefficients of 0.49 and 0.37, respectively). This correlation structure indicated potential multicollinearity concerns requiring careful management in our model development.  
Linearity and the effects of logarithmic transformation on our predictors (Number of News Articles, Movie Budget, Number of Faces, Movie Duration) can be seen in **Figure 1 and Figure 2.**

*Categorical Variables and Content Factors*

Our dataset included binary indicators for various film genres, including action, adventure, sci-fi, thriller, musical, romance, western, sport, horror, drama, war, and animation. From our initial analysis, certain genres demonstrated stronger relationships with IMDb scores. Drama emerged as a particularly significant genre predictor, showing a positive association with ratings. Animation similarly displayed a notably positive relationship with scores. Conversely, the horror genre exhibited a negative correlation with IMDb scores, suggesting that horror films typically receive lower ratings on average. The romance and action genres also showed meaningful, though more moderate, associations with our dependent variable. The statistical significance of the different movie genres can be seen in **Figure 3**.

Our analysis suggested a complex relationship between release year and IMDb scores, with ratings varying across different time periods. This complexity led us to investigate polynomial transformations to capture non-linear temporal patterns in audience reception. Maturity ratings represented another classification factor that might influence audience expectations and subsequent ratings. Our analysis indicated that certain ratings, particularly NC-17, R, TV-14, and TV-G, showed statistically significant relationships with IMDb scores. This finding suggested that content restrictions and target audience segmentation play meaningful roles in how films are received and rated. Actor prominence, measured through IMDb Pro's star meter rankings (where lower values indicate higher prominence), offered another dimension for analysis. Our data included rankings for up to three leading actors per film. From these, we derived several potential predictors, including the lowest (best) star meter among a film's actors. This metric emerged as particularly informative, showing a significant relationship with IMDb scores. By evaluating initial and individual correlations, this allowed the narrowing down of all variables that were worth exploring. The table of individual correlations can be seen in **Figure 4.**

*Statistical Challenges and Data Transformations*

Many of our numerical predictors exhibited right-skewed distributions, which can complicate linear regression modeling by violating assumptions of linearity and homoskedasticity. We applied logarithmic transformations to several variables: movie\_budget, nb\_faces, duration, and movie\_meter\_IMDBpro (as seen in **Figure 2**). These transformations substantially improved the distributional characteristics, making them more suitable for regression analysis. The transformed variables generally showed more linear relationships with IMDb scores, though some still exhibited non-linear patterns warranting further exploration. Our diagnostic procedures identified several outlier observations (specifically observations 1806, 1581, 395, 191, 1123, and 1592) that exerted disproportionate influence on model parameters. After careful consideration, these observations were removed to improve model stability and reduce the impact of anomalous data points. This adjustment resulted in a refined dataset used for subsequent modeling. Analysis of residual patterns revealed the presence of heteroskedasticity in our models, particularly when using logarithmically transformed predictors. The non-constant variance test (ncvTest) confirmed this issue statistically. This finding informed our modeling approach, leading us to implement appropriate correction techniques in later stages. The presence of heteroskedasticity suggested that our model's prediction accuracy varies across different ranges of the predictors, a challenge requiring specific statistical adjustments. The finding and adjusting of the heteroskedasticity can be seen in **Figure 11 and Figure 12.**

Altogether, through our data exploration, we uncovered a complex landscape of variables influencing IMDb scores. Numerical predictors showed meaningful but often non-linear relationships with ratings, while genre classifications, temporal factors, maturity ratings, and actor prominence metrics added important qualitative dimensions. Through appropriate transformations like logarithmic transformation and statistical adjustments such as weight adjustments, we prepared these variables for incorporation into our predictive framework, addressing challenges of skewed distributions, outliers, and heteroskedasticity along the way.

**Model Selection**

Our objective in the model selection phase was to systematically identify the most predictive combination of variables, transformations, and modeling approaches to accurately predict IMDb ratings while minimizing the risks of overfitting. Our systematic approach to model building was fundamentally grounded in extensive cross-validation (K=20) driven grid searches. It was explicitly structured to explore and evaluate all possible combinations of predictor transformations to find the most predictive and cohesive model as a unified whole. These grid searches focused exclusively on the set of variables previously identified as highly significant through our exploratory analysis, often applying necessary transformations to address issues like collinearity, heteroskedasticity, and overfitting (as detailed in ‘*Statistical Challenges and Data Transformations’).*

We began this rigorous process by systematically exploring polynomial transformations, methodically testing combinations of polynomial degrees ranging from 1st to 5th degrees for each numeric predictor. Following the polynomial exploration, we conducted similar extensive grid searches focusing exclusively on spline-based models, assessing various spline configurations with 3 and 4 knots placed at quantiles. Lastly, we executed CV-driven grid searches for hybrid models combining polynomial and spline transformations, aiming explicitly to verify if a hybrid approach could further optimize predictive performance through complementary transformations. Our results revealed that the hybrid model did in fact marginally outperform the polynomial-only approach in terms of predictive accuracy. However, the incremental improvement gained was minimal (less than 0.005 R² increase) and was not worth the resulting loss of model interpretability and increased complexity. By prioritizing practicality and interpretability, we ultimately chose the simpler polynomial-only model approach for our final model.

Before finalizing our model, we generated initial forecasts for upcoming movies and observed unexpected inconsistencies (unusually low predicted ratings, 3~5). Upon investigation, we identified the “number of news articles” predictor as the source of systematic bias (as explained in *Key Numerical Predictors and Their Relationships)*. We therefore chose to exclude it, prioritizing generalizability to future films over an optimal historical fit. This deliberate adjustment reduced overfitting and enhanced the model’s robustness for predicting new releases. **This effect can be seen in Figure 7.** Our final model excludes “number of news articles” and incorporates polynomial transformations of movie budget (3rd degree), number of faces (2nd degree), duration (3rd degree), release year (4th degree), and IMDbPro popularity score (4th degree).   
It also includes categorical predictors: genre (action, romance, drama, animation), maturity rating (NC-17, R, TV-14, TV-G), and actor prominence (lowest star meter ranking).   
This model achieved an R-squared of 0.52 and a Mean Squared Error (MSE) of 0.624.

**Results**

After testing numerous different models with different combinations of predictors, we have arrived at a final model that included movie budget, number of faces in the promotional poster, the duration of the movie, the movie’s release year, the log-transformed meter score from IMDb pro, and finally, certain categorical variables such as the movie genre and it’s recommended age rating. More specifically for the movie’s genre, we wanted to note if it was an action, romance, drama, or animation movie, as we observed that historically, action and romance movies tend to have lower IMDb ratings, while drama and animation movies tend to have higher IMDb ratings.

After running our final model and analyzing the given results, we have established a set of predictions for each of the twelve upcoming blockbuster’s IMDb ratings. The full ranking can be seen in Table 1, with our highest rated movie being The Day the Earth Blew Up (6.30 rating), and the lowest rated movie being O’Dessa (4.72 rating). Our lowest rated movie, O’Dessa, and our second-highest rated movie, Novocaine, are separated by just four minutes in terms of duration at 106 and 110 minutes respectively. Similarly, our highest rated movie, The Day the Earth Blew Up, is just four minutes shorter than our third-lowest rated movie, Ash, at 91 and 95 minutes respectively. This indicates that although duration may be a powerful predictor, it still remains complex and definitely reliant on a combination of other predictors. However, certain predictors do clearly show a pattern, such as movie\_meter\_IMDbpro. The Day the Earth Blew Up and Novocaine received low scores in this predictor (which indicates a more popular movie), while O’Dessa and My Love Will Make You Disappear received relatively larger scores, which is consistent with our predictions. However, it is important to consider that there are almost always external factors influencing something so complex as a movie score. Snow White received the lowest movie\_meter\_IMDBpro rating, indicating that it should have a strong IMDb critic rating, however in reality, it was in the bottom half of our predictions.

One movie with a lower predicted score than expected, is the aforementioned Snow White. At first glance, Snow White is a very well known fairytale story, with the story being passed down between two centuries of generations. Therefore, at first glance, one might think that a modern day remake of the classic story would score well with critics. However after running our model, that was not the case as Snow White has a projected IMDb rating of 5.09, which ranks eighth among the twelve blockbusters. Snow White’s lower than expected score can be due to to the fact that it does not fit into the drama or animation categories, and there are also factors that we were unable to consider in our model, such as the notion that remakes and/or reboots tend to have lower critical reception, as they are typically not as well received as the original story[1]. Therefore, Snow White illustrates quite well the limitations of purely data-driven predictions, as certain societo-industrial trends and audience biases, such as nostalgia for the original production or skepticism towards remakes, are difficult to quantify but still play a key role in a film’s overall reception.

After testing numerous different models based on different combinations of variables, our final model has an R-squared value of 0.52, indicating that approximately 52% of the variability in IMDb ratings is explained by our selected predictors. While 0.52 may be a relatively moderate level of predictive power, it is important to consider that giving a rating to a movie is very subjective and one’s criteria for what qualities make up a good movie vary from person to person. Therefore, it would be nearly impossible to obtain a high R-squared value in this situation. Moreover, an R-squared value of 0.52 was settled on because it is generally considered a more than acceptable standard in research and since further improvements were beginning to become difficult to achieve and marginal in nature.

Additionally, our final model had an MSE of 0.62, meaning that on average, the squared deviation of the predicted ratings from the true ratings is 0.62. For interpretability, by rooting this number, we obtain a mean error value of 0.784, meaning that on average, our predicted ratings deviate from the true result by 0.784 which is excellent for this case since it the score is on a scale of 1 to 10 – our error is therefore narrow. Since MSE measures the average squared error in our model, the relatively low MSE that we obtained indicates strong predictive accuracy. However, the effectiveness of our MSE result should also be considered with other factors, such as the complexity of the model and the potential for overfitting. It is important to ensure that our model is not too complex, as using too many predictors can drown out the true patterns present in the sample data.  
 One important consideration: given the nature of the dataset, our model was precisely optimized to predict ratings based on comprehensive historical data, and it excels at this task. However, forecasting ratings for upcoming movies presents a subtly different challenge. As previously discussed, certain variables behave differently before and after a film's release, inherently altering the dynamics on which our model was trained. Consequently, although our model performs robustly on historical data, we do acknowledge that its predictions for future releases may consistently underestimate actual ratings due to these subtle differences.

Finally, while our model can provide valuable insights, predicting audience and critic reception remains challenging due to the subjective nature of movie preferences and surrounding buzz. What one person likes may vary greatly from another, and cultural trends can shift how films are received over time and across regions. Additionally, external factors such as marketing strategies and star power of actors and/or directors can significantly influence a movie’s perception before it is even released. These variables can make it difficult to perfectly predict how a movie will be received, despite the insights provided in our model. However, in terms of overall predictive power and relative error, our model demonstrates strong performance.

**Appendix**

Table 1: Predicted IMDb Ratings

| **Movie Title** | **Predicted IMDb Rating (rounded to two decimals)** | **Rank** |
| --- | --- | --- |
| The Day the Earth Blew Up | 6.30 | 1 |
| Novocaine | 6.21 | 2 |
| A Working Man | 5.97 | 3 |
| The Alto Knights | 5.84 | 4 |
| Black Bag | 5.60 | 5 |
| The Woman in the Yard | 5.59 | 6 |
| Locked | 5.54 | 7 |
| Snow White | 5.09 | 8 |
| My Love Will Make You Disappear | 5.08 | 9 |
| Ash | 5.05 | 10 |
| High Rollers | 4.73 | 11 |
| O’Dessa | 4.72 | 12 |

Figure 1: Issues with linearity with the following predictors: Number of News Articles, Movie Budget, Number of Faces, Movie Duration.

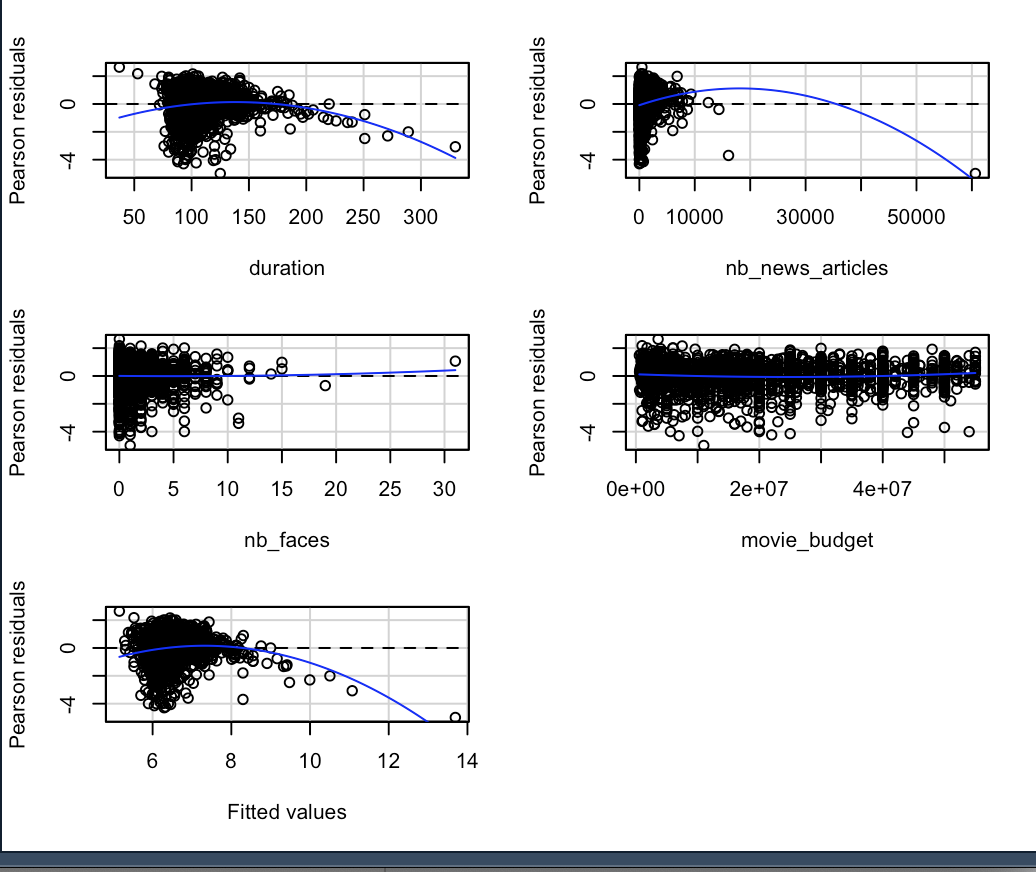


Figure 2: The effects of logarithmic transformation on Number of News Articles, Movie Budget, Number of Faces, Movie Duration.

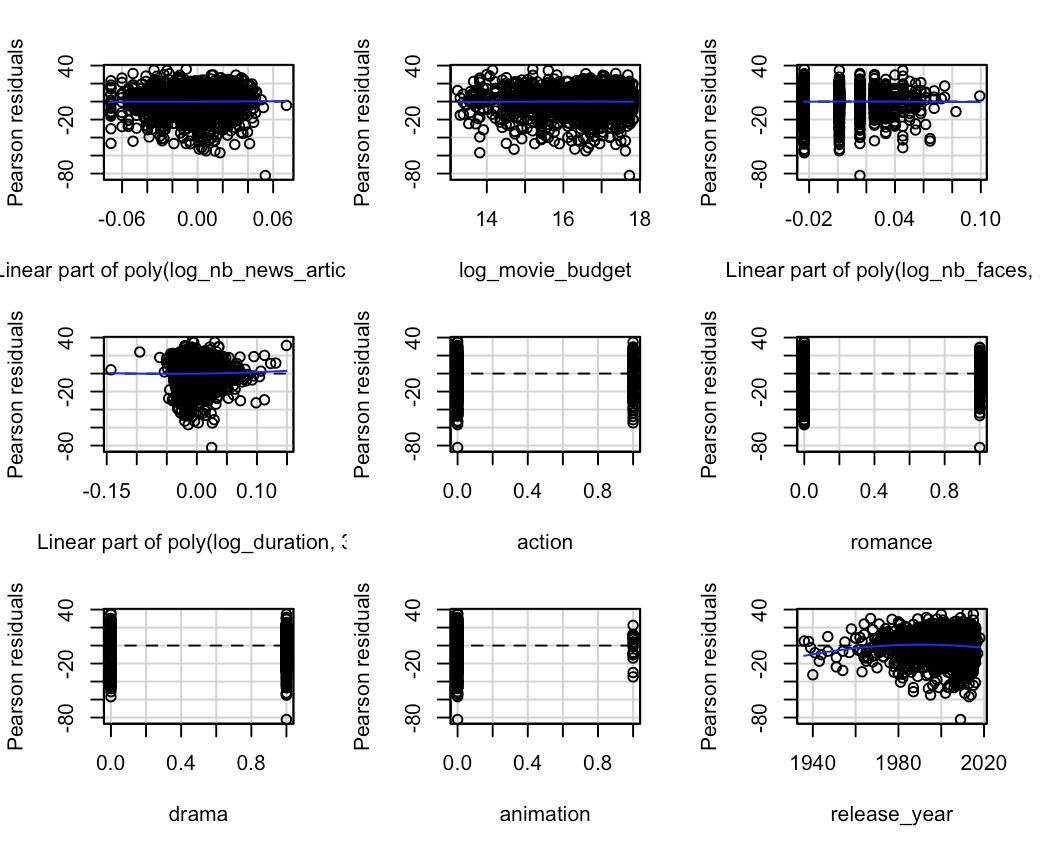


Figure 3: The statistical significance of the Genres

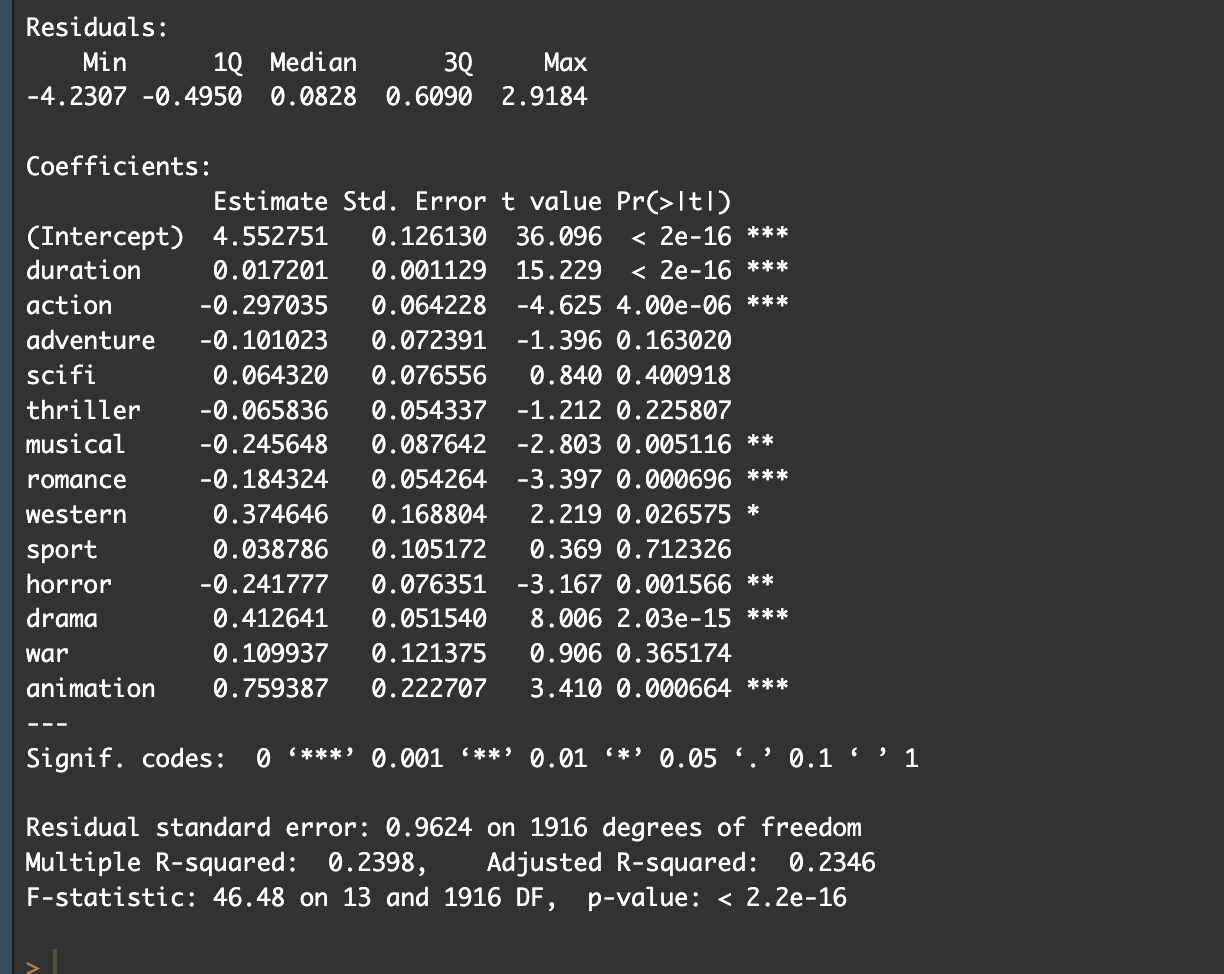


Figure 4: Correlations and significance for individual variables predicting IMDb score

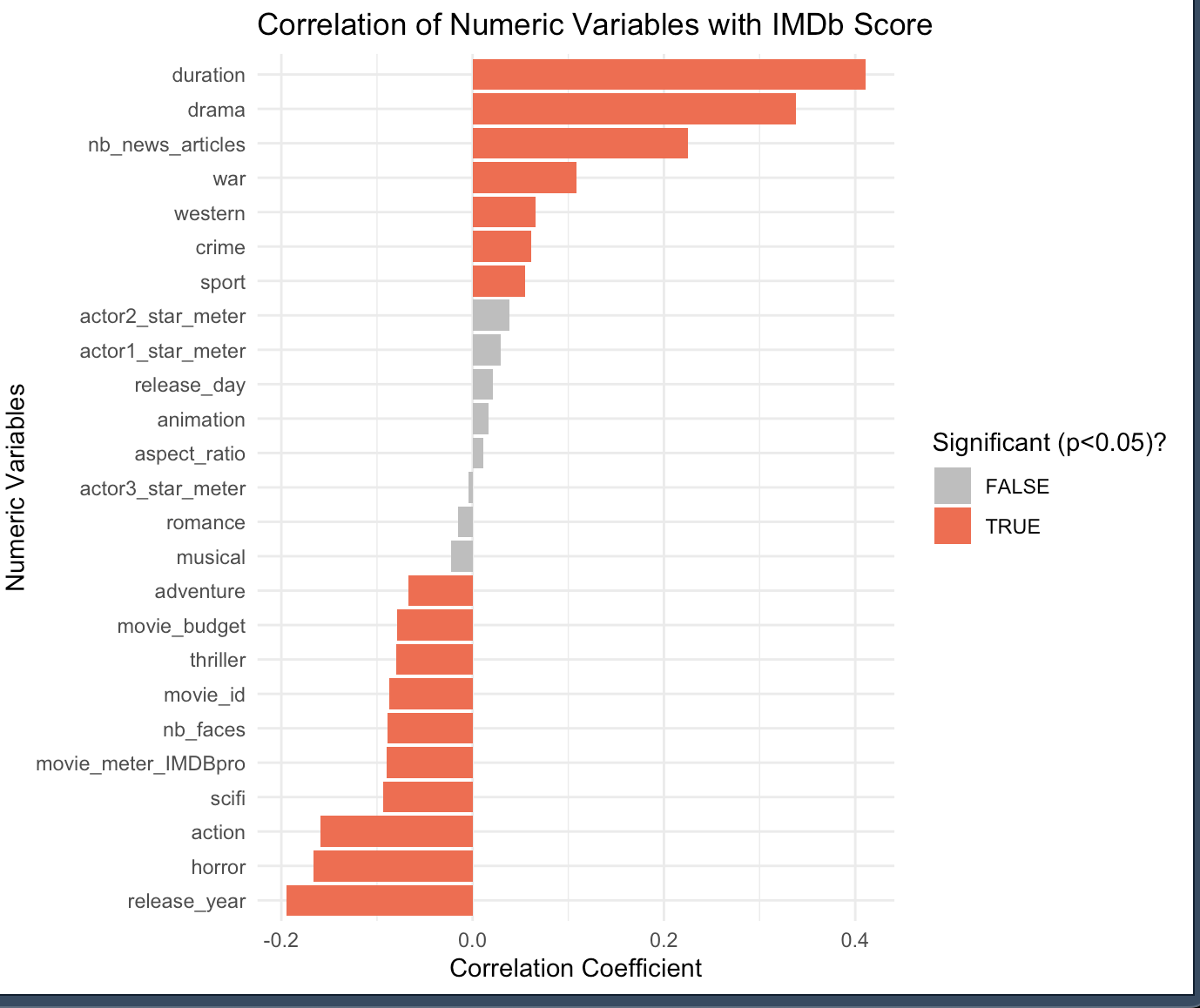


Figure 5: Right-Skewed Distribution of movie\_meter\_IMDBpro (Pre-Log Transformation)

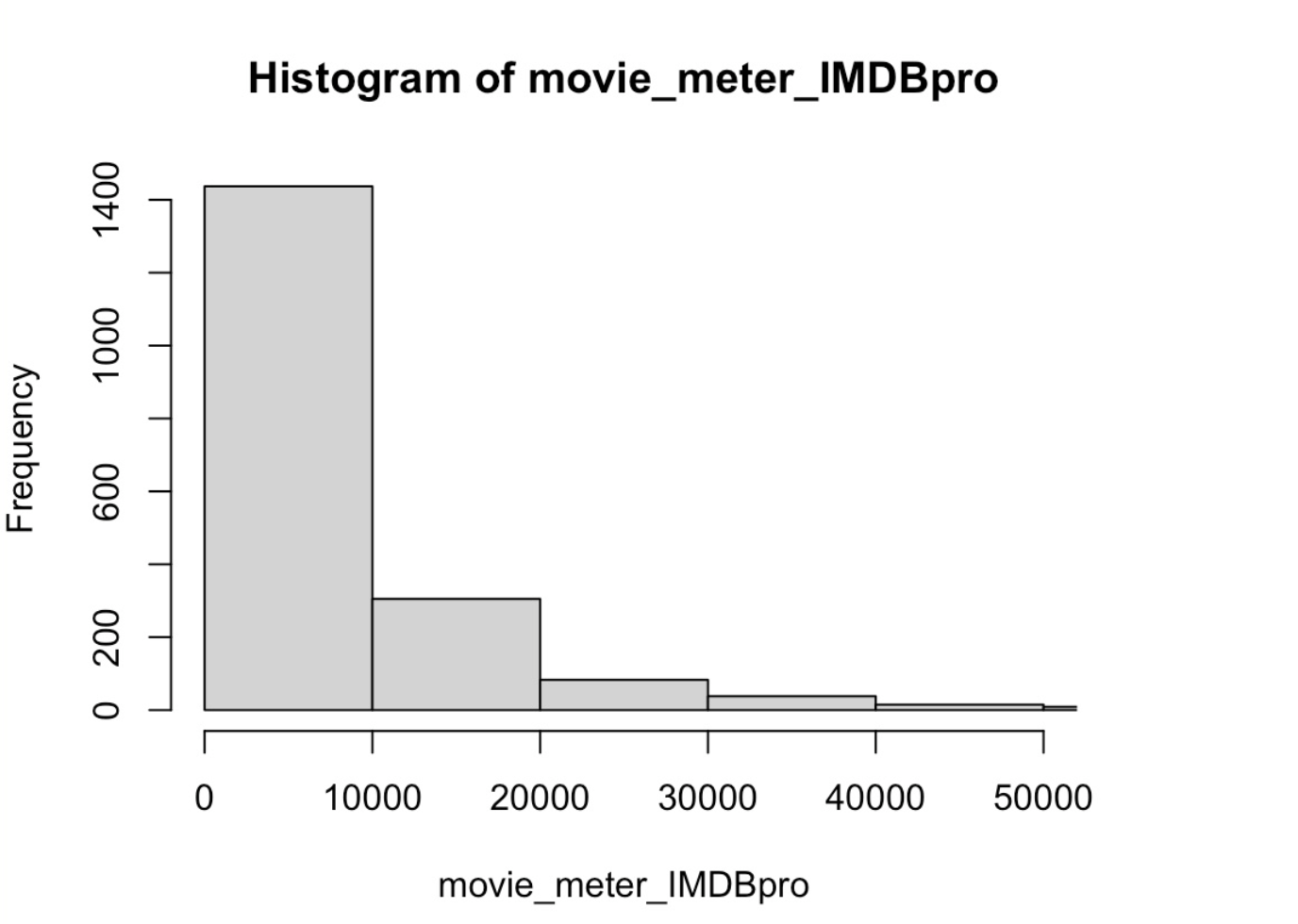


Figure 6: Normalized movie\_meter\_IMDBpro (Post-Log Transformation)

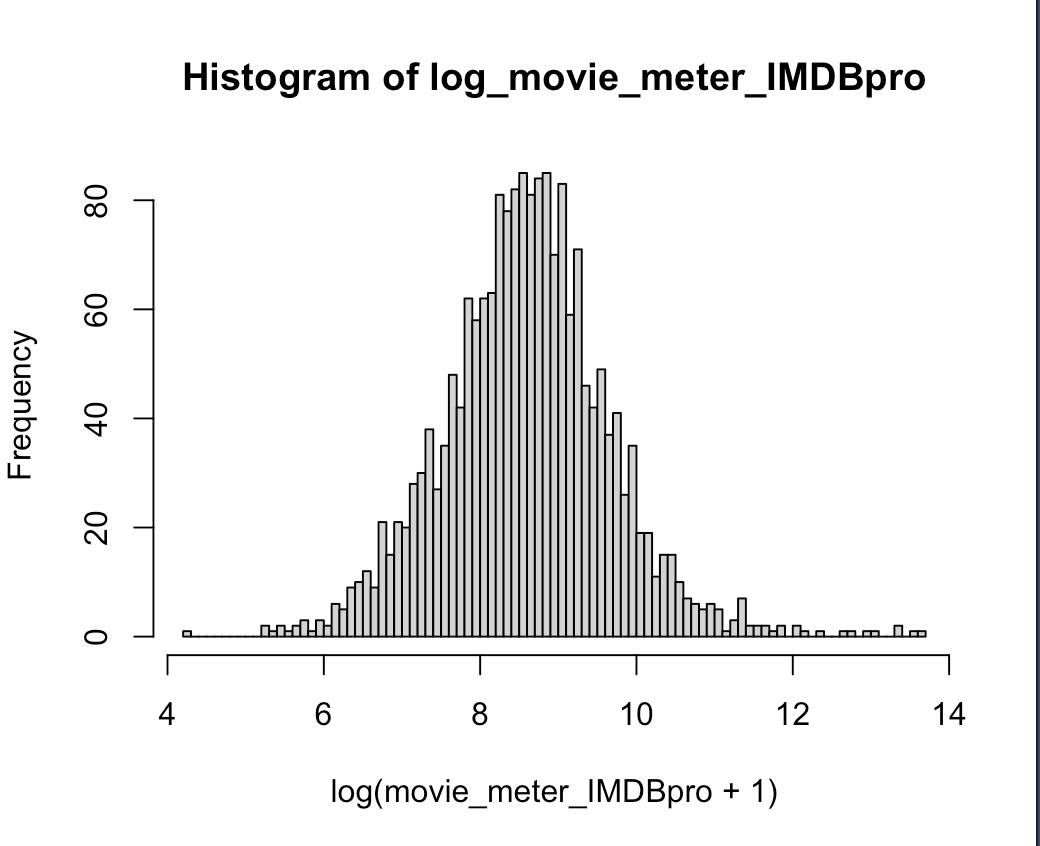
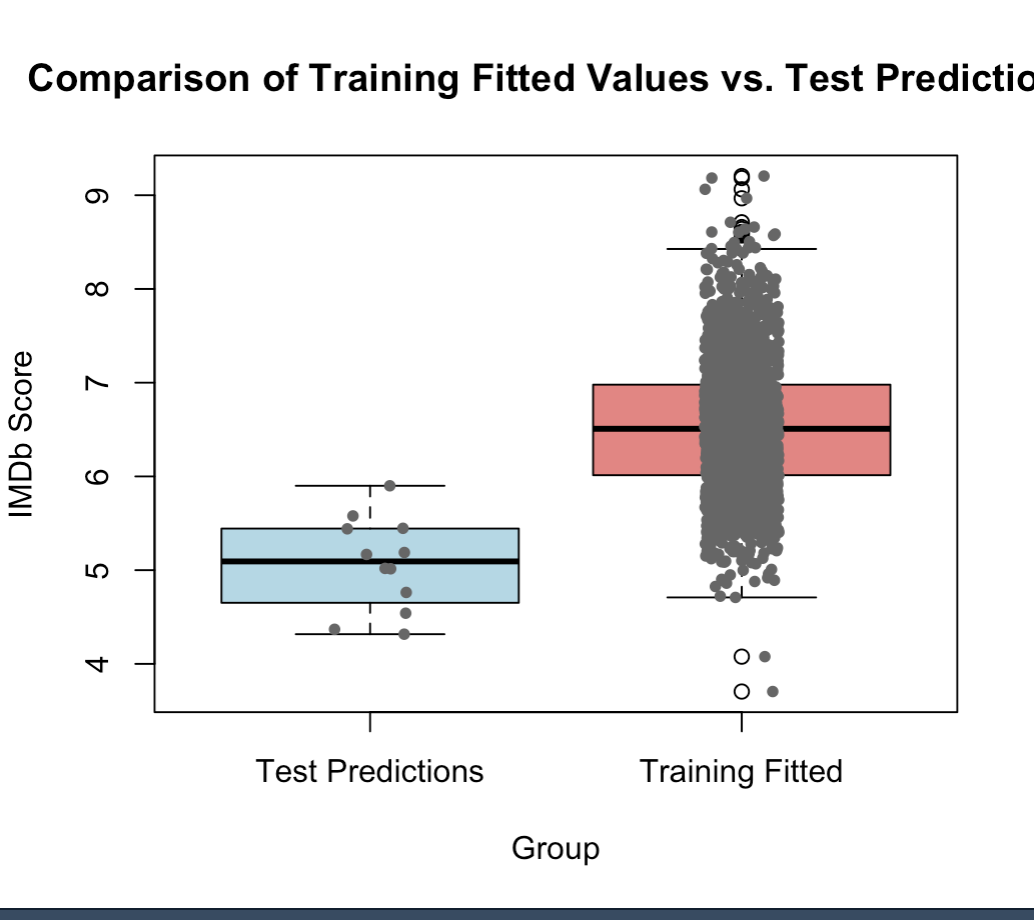


Figure 7: Initial test predictions were systematically lower than training fitted values



*After generating initial predictions for the 12 upcoming movies, we noticed the predicted IMDb scores were unusually low compared to the training fitted values (mean scores: 5.06 vs. 6.53). To confirm this, we plotted a boxplot comparing training and test predictions. Further investigation showed that the* ***number of news articles*** *variable had much lower values in the test data, likely introducing bias. Removing this predictor improved generalizability.*

Figure 8: Residuals vs. Fitted; no clear patterns, suggesting our final model is robust

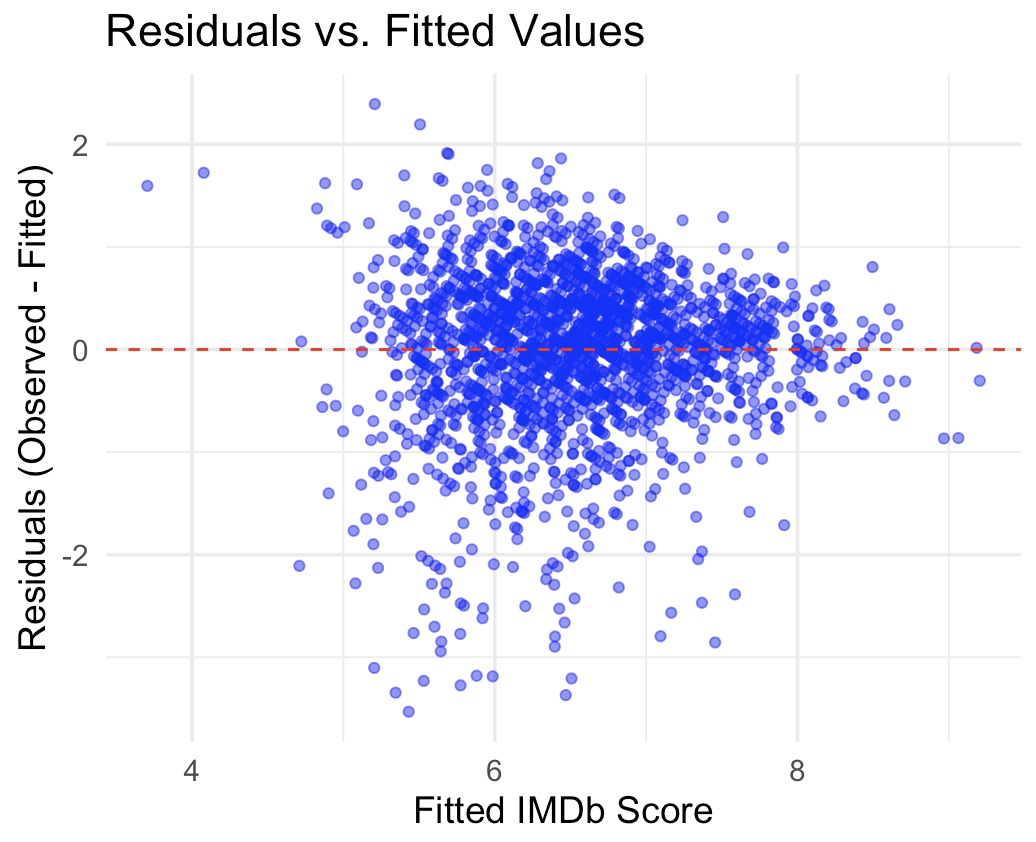


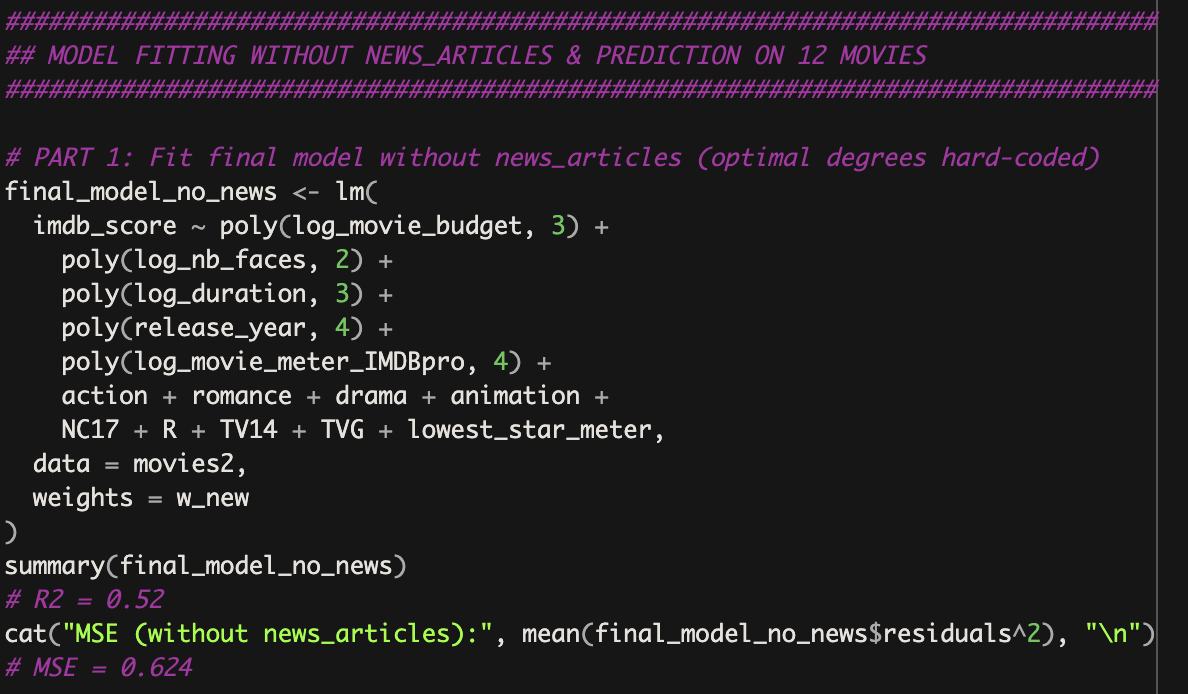
Figure 9: Final model implementation for IMDb score predictions

Figure 10: Our final model’s coefficients showing high statistical significance

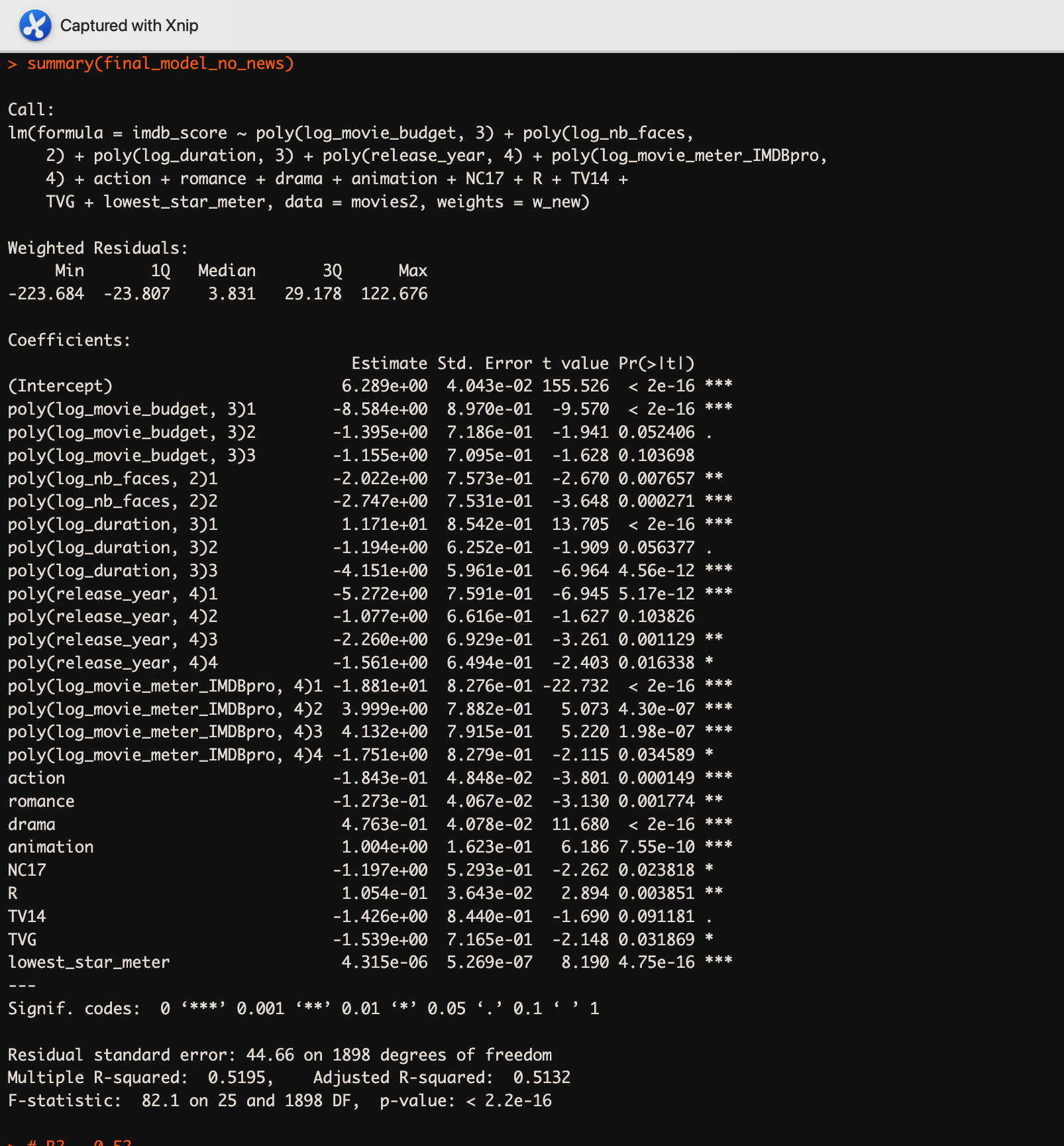
****

Figure 11: Test for Heteroskedasticity before adjustment

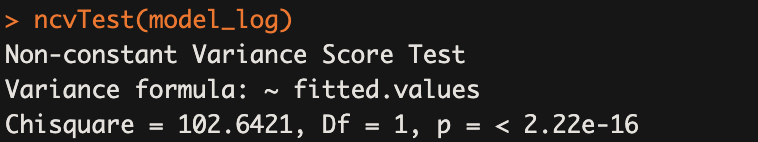
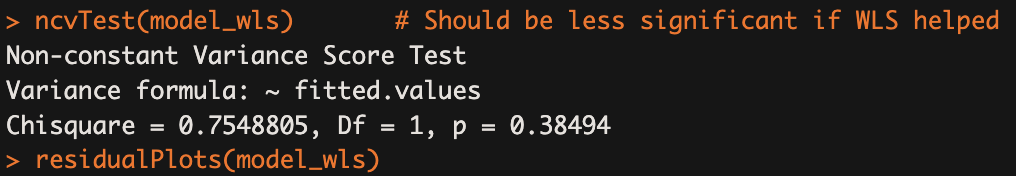


Figure 12: Test for Heteroskedasticity after weighted adjustment - issue resolved



**Bibliography**

[1]. Yes, Remakes Do Suck -- And the Tomatometer Proves It. https://editorial.rottentomatoes.com/article/yes-remakes-do-suck-and-the-tomatometer-proves-it/. Accessed 9 Mar. 2025.