Predictors of a Successful Movie – A Case Study of IMDB's Top 1000 Movies

Malachy Percy-Campbell

Department of Statistics

University of Connecticut

April 30, 2023

Abstract

This statistical paper aims to investigate the factors that contribute to the success of movies. Using a data set of IMDB's top 1000 movies, we analyzed various variables, including genre, runtime, and critical ratings, to identify the most significant predictors of box office success. Our analysis utilized multiple regression models to identify the key factors that contribute to the success of a movie. Our findings indicate that the most significant predictor of box office success was runtime. We also found that critical ratings play an insignificant role in predicting box office success. Our study provides valuable insights for filmmakers and movie studios to help them make informed decisions about the production and release of movies.

Keywords: Box Office, Film Industry, Movies, Regression, Revenue

1 Introduction

The film industry is a vast and complex network of companies and individuals involved in the creation, production, distribution, and exhibition of films. The importance of the film industry can be briefly summarized in several reasons. Firstly, it is a significant source of entertainment for people around the world. Films provide an escape from reality and offer a means of exploring different cultures, experiences, and perspectives. Secondly, the film industry generates significant revenue for the global economy. And thirdly, films can have a profound impact on society, influencing our beliefs, attitudes, and behaviors. They can spark conversations, promote social change, and raise awareness of important issues. At its core, the film industry is focused on creating, financing, and producing motion pictures that entertain, educate, and inspire audiences all around the world.

In addition to its creative side, the film industry is also a significant economic force, generating billions of dollars in revenue each year (Motion Picture Associations, 2023). The industry's unbelievable financial achievements led me to question which factors in the production process impact the economic success of a movie. The success of a movie ultimately leads to the success of the film industry; this is crucial for the economy as it generates revenue, creates jobs, supports related industries, encourages innovation, and shapes cultural perceptions and values through outreach.

There have been plenty of studies surrounding the question: which factors contribute to the success of a movie? I have seen many factors discussed such as: producer, production house, director, cast, runtime of the movie, the script, time of release, marketing schemes, number of screening days, production budget, IMDB rating, etc., the list goes on and on (University of Technology: Sydney, 2017). Time of release, budget, casting, and ratings were identified as factors most likely to influence a movie's success among the studies we have looked into (Sood, 2017).

These studies are very well rounded, using correct methods and coming up with valid conclusions, which continue to push industry standards. The problem with this topic is how vast the predictors are; there are so many factors it is impossible for one to build a complete model. I hope to offer a unique approach by including movie production as well as social factors to my model. In the end, I choose to observe genre (thirteen genres including: action, adventure, animation, biography, comedy, crime, drama, family, film-noir, horror, mystery, thriller, and western), gross revenue (in USD), Metascore, and runtime (in minutes) as my variables of interest. This distinct cocktail of variables will offer a fresh approach that may have been overlooked in other studies.

The flow of this paper is as followed: Introduction (1), Data Description (2), Methods (3), Application (4), and Conclusion (5). In the introduction section, I briefly explain the importance of the film industry and existing studies. In the data description section, I cover the data I will be testing and what, if any, adjustments are made to the data. In the methods section, I describe what statistical methods / tests will be used in this case study and why they will be used. In the application section, I discuss the results of the testing done. And finally, in the conclusion section I finalize the paper by interpreting the results of the study and drawing back to how this study can be useful to the film industry.

2 Data Description

The data used in this case study was sourced through Kaggle, an online free database used to promote data analysis projects. I found an intriguing data set created by *Harshit Shankhdhar,* who created a sampling frame out of IMDB's top 1000 movies and television shows (Shankhdhar, 2021).

Before we begin the analysis, it is important to understand IMDB, what it is, and why I chose this sampled data. IMDB (Internet Movie Database) is an industry standard online database that provides information related to movies. IMDB rating is a numerical rating

system used to evaluate the quality of movies listed on the platform. IMDB ratings are calculated using a complex algorithm that considers various factors such as the number of votes, the rating given by each voter, and the age of the votes. The algorithm ensures that the rating is constantly updated and reflects the most accurate representation of the overall opinion of the audience. The rating system on IMDB is based on a scale of 1 to 10, with 1 being the lowest and 10 being the highest. In general, a movie with a high rating (7 or above) is good, while a movie with a low rating (6 or below) is below average (IMDB , 2023).

Similar to IMDB rating, Metascore is a numeric rating system used to assess how well received a movie has been to critics. It is calculated by aggregating reviews from a variety of reputable critics, assigning each review a score on a scale of 0-100, and then taking a weighted average of these scores. The Metascore is frequently recognized as a trustworthy measure of a film's overall quality and is designed to give a snapshot of critical opinion. It is frequently used by moviegoers to help them choose which movies to watch, as well as by studios and distributors to sell and promote their movies (Metacritic , 2023).

The original data set contained 1000 observations with 16 variables. After narrowing down the specific variables I chose to study [Genre, Gross Revenue (denoted as Gross in models), Metascore, and Runtime] and cleaning / filtering for blank and "NA" cells, I was left with 750 observations and 5 variables. From there *Model 1* was created. *Table 1* shows a brief summary of statistics used in *Model 1*.

Table 1: Descriptive Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Runtime	750	123	26	71	104	136	238
Genre	750						
... Action	130	17%					
... Adventure	58	8%					
... Animation	64	9%					
... Biography	73	10%					
... Comedy	109	15%					
... Crime	81	11%					
... Drama	209	28%					
... Family	2	0%					
... Film-Noir	2	0%					
... Horror	10	1%					
... Mystery	7	1%					
... Thriller	1	0%					
... Western	4	1%					
Meta_score	750	77	12	28	70	86	100
Gross	750	74952069	113328043	1305	5014812	98091571	936662225

After analyzing the data used to build *Model 1*, I chose to remove family, film noir, horror, mystery, thriller, and western as dummy varaibles within the genre predictor. These factors only accounted for about 3% of the genre predictor but they proved to be influential observations. I concluded we would yield fairer results with those factors removed. Therefore, our final data set contains 718 observations and 5 variables. *Table 2* shows a brief summary of statistics used in *Model 2* and onwards.

Table 2: Descriptive Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Runtime	718	123	26	71	104	137	238
Genre	718						
... Action	127	18%					
... Adventure	57	8%					
... Animation	64	9%					
... Biography	73	10%					
... Comedy	108	15%					
... Crime	80	11%					
... Drama	209	29%					
... Family	0	0%					
... Film-Noir	0	0%					
... Horror	0	0%					
... Mystery	0	0%					
... Thriller	0	0%					
... Western	0	0%					
Meta_score	718	78	12	28	70	86	100
Gross	718	76172055	114175234	3600	5204332	100420716	936662225

3 Methods

After constructing my original data set with 750 observations and 5 variables, I created the multiple linear regression model, *Model 1*:

$$\begin{aligned} Gross = \beta_0 + \beta_1(Runtime) + \beta_2(Adventure) + \beta_3(Animation) + \beta_4(Biography) \\ +\beta_5(Comedy) + \beta_6(Crime) + \beta_7(Drama) + \beta_8(Family) \\ +\beta_9(Film - noir) + \beta_{10}(Horror) + \beta_{11}(Mystery) \\ +\beta_{12}(Thriller) + \beta_{13}(Western) + \beta_{14}(Metascore) + E. \end{aligned} \tag{1}$$

This model includes Gross Revenue as the dependent variable and Genre (as a categorical predictor with twelve dummy variables), Metascore, and Runtime as parameters to be

estimated.

After establishing the initial model, my next step was to verify the assumptions of the linear model. This step is crucial in any statistical analysis and SHOULD NOT be skipped. Violating any assumption can lead to biased or inefficient estimates of the regression coefficients, as well as incorrect statistical inference. Let's address the assumptions for a multiple linear regression model:

1. Linearity: The relationship between the dependent variable and each independent variable are linear.

2. Independence: The observations are independent of each other. This means that the value of the dependent variable for one observation is not related to the value of the dependent variable for any other observation. (We will be assuming our data satisfies Independence in this study)

3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables. In other words, the spread of the residuals is the same for all values of the independent variables.

4. Normality: The errors are normally distributed. This means that the distribution of the residuals is symmetric and bell-shaped.

5. No multicollinearity: The independent variables are not highly correlated with each other. This means that there is no perfect linear relationship between any two independent variables.

6. No influential outliers: There are no extreme values in the data that have a disproportionate effect on the regression results.

To verify these assumptions, I ran a plot diagnostics test to receive five plots, as well as running a Shapiro-Wilk test.

The first plot analyzed is the residual vs. fitted plot. This graph plots the residuals (the differences between the observed values and the predicted values) on the vertical axis and the fitted values (the predicted values of the dependent variable based on the independent

variables) on the horizontal axis. This plot is used to evaluate the linearity, normality, and homoscedasticity assumptions.

The second plot analyzed is the Scale-Location plot. Also known as a spread-location plot or square root of standardized residuals plot, this graphical method is used to assess the homoscedasticity assumption.

The third plot analyzed is the normal Q-Q (Quantile-Quantile) plot. This plot is used to assess whether a set of data comes from a normal distribution. If the data is normally distributed, the points in the plot should fall approximately along a straight line.

The fourth and fifth plots analyzed are the Residuals vs. Leverage and Cook's Distance plot, respectively. These plots are used to identify influential observations and outliers, which can have a significant impact on the estimated regression coefficients and affect conclusions.

Finally, the Shapiro-Wilk test is a statistical test used to determine, numerically, whether a data set follows a normal distribution.

After assessing the assumptions of our first model it was very clear we could not perform a valid statistical analysis; even after constructing our final data set with 718 observations and 5 variables. Luckily, there are a few techniques we can do without adding and dropping variables that will help validate our model and form conclusions. My first thought was to apply a squared term to the model to allow for a non-linear relationship between the predictor and the response variable, but when applied the fit of the model was quite off and did not help corroborate the assumptions. My second attempt to validate the model was to apply a log transformation on the response variable. The log transformation is useful when the relationship between the predictor variable and the outcome variable is not linear, additionally, the log transformation can help to normalize the distribution of the variables. This method proved successful, and our new model, *Model 2*, was created:

$$Log(Gross) = \beta_0 + \beta_1(Runtime) + \beta_2(Adventure) + \beta_3(Animation) + \beta_4(Biography) + \beta_5(Comedy) + \beta_6(Crime) + \beta_7(Drama) + \beta_8(Metascore) + E. \quad (2)$$

This new model includes log($Gross$) as the dependent variable and Genre (as a categorical predictor with seven dummy variables), Metascore, and Runtime as parameters to be estimated.

After verifying *Model 2* assumptions, we are ready to run the required statistical analysis. To reiterate, we are performing a multiple regression test. To execute this test, we used the statistical software R (RStudio) to run the model and obtain the necessary $F - statistics$ and corresponding $p - values$ to draw conclusions. We ran 4 tests altogether and created a new model for each test.

Test 1 assesses the overall significance of a linear regression model that includes all predictor variables, i.e., the full model, *Model 2*.

$$H_{01} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \, vs. \, H_{a1} : At \, Least \, One \, Correlation \, Coefficient \neq 0,$$

$$Log(Gross) = \beta_0 + \beta_1(Runtime) + \beta_2(Adventure) + \beta_3(Animation) + \beta_4(Biography) + \beta_5(Comedy) + \beta_6(Crime) + \beta_7(Drama) + \beta_8(Metascore) + E.$$

Test 2 assesses the significance of the reduced model which includes a subset of the predictor variables including Runtime and Genre, i.e.,

$$H_{02} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0 \, vs. \, H_{a2} : At \, Least \, One \, Correlation \, Coefficient \neq 0,$$

$$\begin{aligned} Log(Gross) = \beta_0 + \beta_1(Runtime) + \beta_2(Adventure) + \beta_3(Animation) + \beta_4(Biography) \\ +\beta_5(Comedy) + \beta_6(Crime) + \beta_7(Drama) + E. \end{aligned} \tag{3}$$

Test 3 assesses the significance of the reduced model which includes a subset of the predictor variables including Runtime and Metascore, i.e.,

$$H_{03} : \beta_1 = \beta_8 = 0 \; vs. \; H_{a3} : At \; Least \; One \; Correlation \; Coefficient \neq 0,$$

$$Log(Gross) = \beta_0 + \beta_1 Runtime + \beta_8(Metascore) + E. \tag{4}$$

Test 4 assesses the significance of the reduced model which includes a subset of the predictor variables including Genre and Metascore, i.e.,

$$H_{04} : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \, vs. \, H_{a4} : At \, Least \, One \, Correlation \, Coefficient \neq 0,$$

$$\begin{aligned} Log(Gross) = \beta_0 + \beta_2(Adventure) + \beta_3(Animation) + \beta_4(Biography) + \beta_5(Comedy) \\ +\beta_6(Crime) + \beta_7(Drama) + \beta_8 Metascore + E. \end{aligned} \tag{5}$$

4 Application

Below follows our assessment of *Model 1* assumptions:

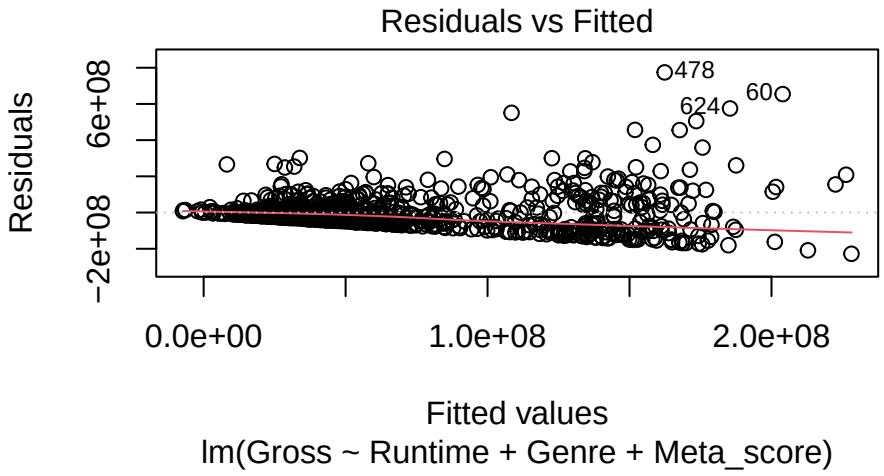

Figure 1: *Model 1* Residual vs. Fitted Plot

In Figure 1, we can see no discernible pattern at first glance. However, you can notice the spread of the residuals increase as the fitted values increase, which could indicate a problem with homoscedasticity. We also notice the residual vs. fitted line (red) shows a slight dip as the fitted values increase; we want this line to be as horizontal as possible. Therefore, we can assume the linearity assumption is violated.

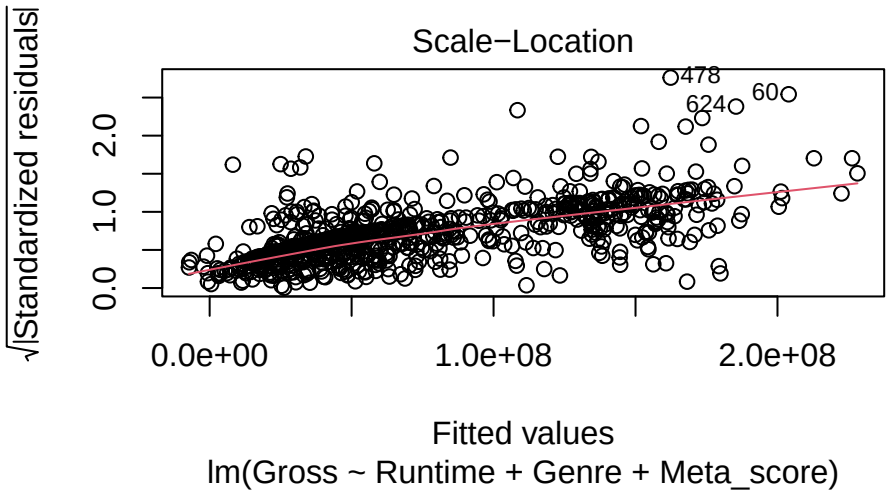

Figure 2: *Model 1* Scale-Location Plot

In Figure 2, we can see the Scale-Location plot shows a bit more spread as the fitted values increase, which is the same observation in the residuals vs. fitted plot. This can be a case of heteroscedasticity, but this plot shows a lot more overall spread than the residuals vs. fitted plot so further assessment may be needed.

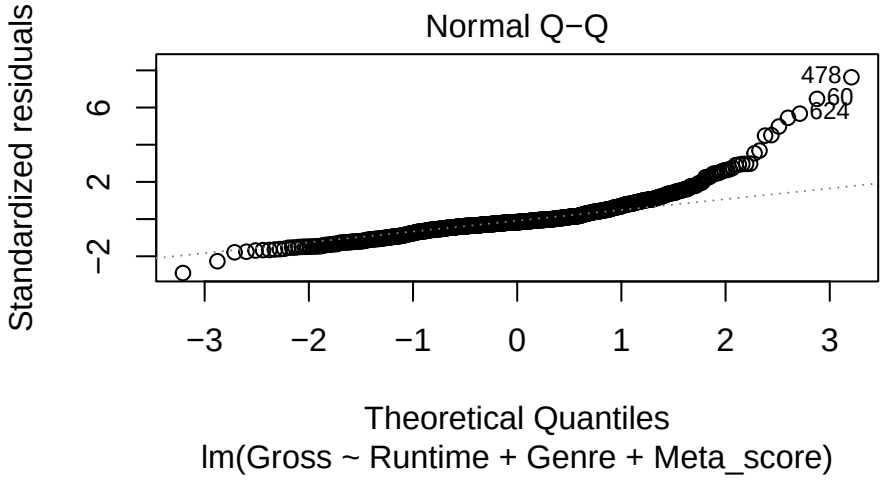

Figure 3: *Model 1* Q-Q Plot

In Figure 3, we can see the data points appear to follow a straight line up until 1.5-2 theoretical quantiles. Since the data curve trends upward, we can assume the data has heavier tails and more spread than what we would expect from a data set that follows a normal distribution. We can assume the normality assumption is violated, but to further access this assumption I performed a Shapiro-Wilk test in R with its outputs as followed:

Shapiro-Wilk Test

$$W = 0.83461,\ p - value < 2.2e - 16,$$

At a 5% alpha level, we can reject the null hypothesis and conclude our data set does not follow a normal distribution.

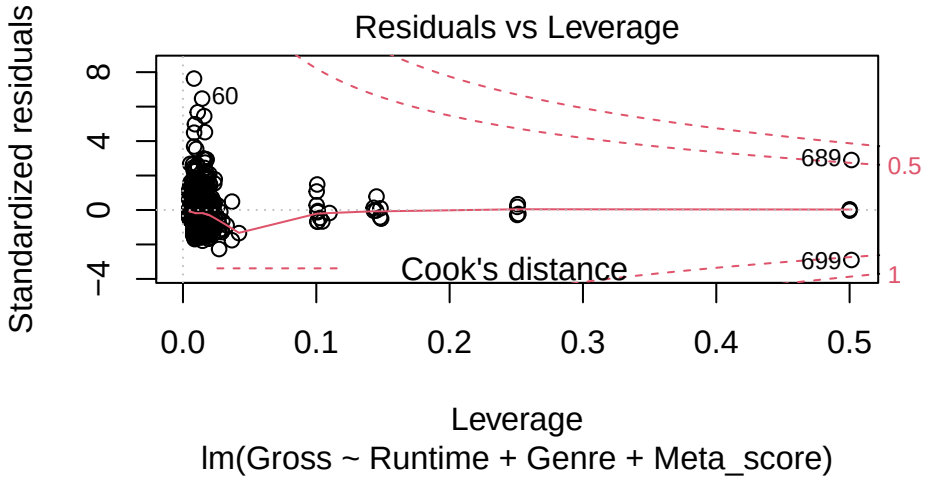

Figure 4: *Model 1* Residual vs. Leverage Plot

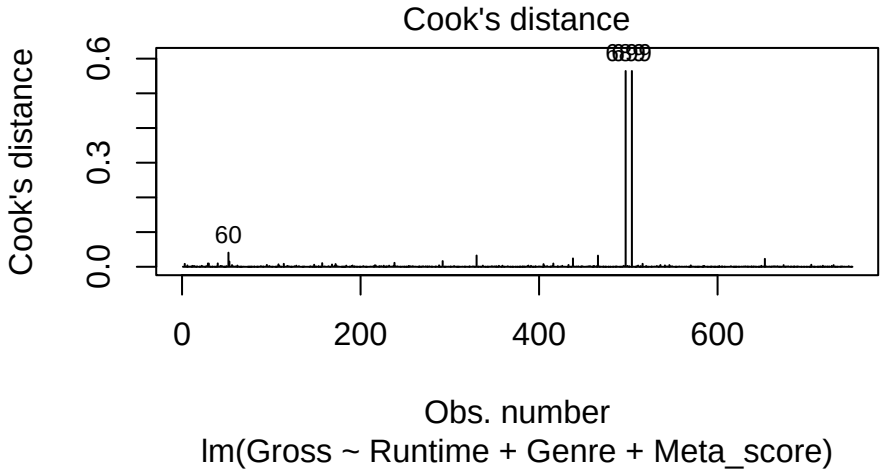

Figure 5: *Model 1* Cook's Distance Plot

In Figures 4 and 5, we can see clear outliers and influential observations. When performing the test in R (RStudio), observation 506 could not be read since it had a Cook's Distance value higher than 1. We can see observations 689 and 699 are clear outliers. I included a specific Cook's Distance plot to illustrate the outliers and influential observations further.

As we briefly discussed earlier, *Model 1* assumptions are badly violated. If we were to further conduct testing our results would yield statistically invalid. Therefore, we created *Model 2*. Below is our assessment of assumptions for *Model 2*:

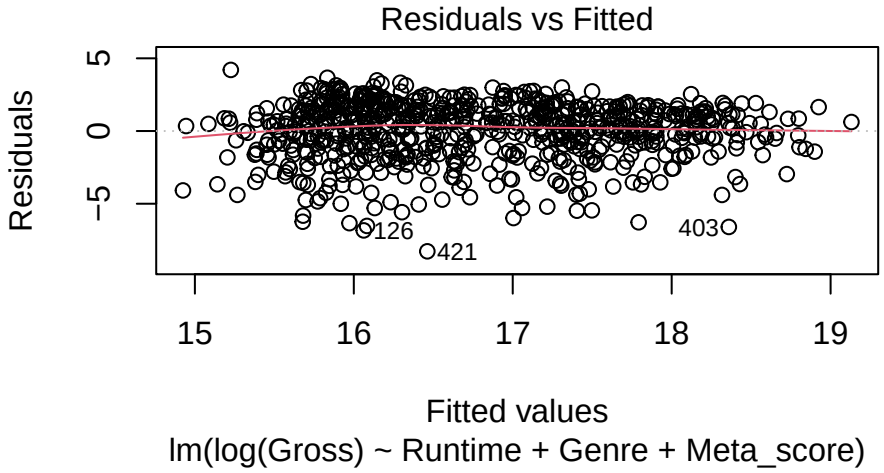

Figure 6: *Model 2* Residual vs. Fitted Plot

In Figure 6, we can see a clear improvement from Figure 1. There seems to be no recognizable pattern and the residuals are randomly scattered about the 0 line, which is exactly what we were looking for. This validates our model, and the variance of random errors are consistent. We can assume the linearity assumption is valid.

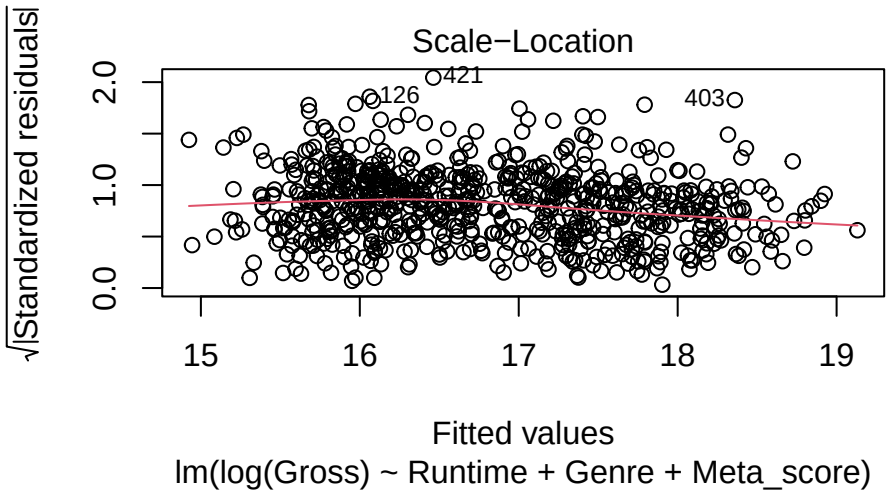

Figure 7: *Model 2* Scale-Location Plot

In Figure 7, we can see the scale-location plot, which looks a lot better than Figure 2. Here we can see a consistent spread across fitted values. The standardized residual vs. fitted line (red) appears to be much more horizontal than in Figure 2, indicating our assumption of homoscedasticity is valid.

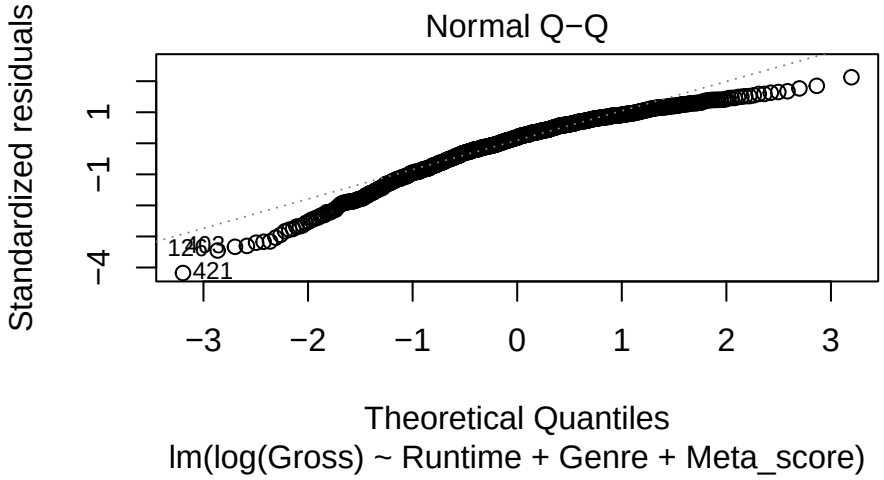

Figure 8: *Model 2* Q-Q Plot

In Figure 8, we can see the data points still do not quite follow a straight line. After applying the $log(y)$ transformation it now appears both tail ends are heavy. If a normal Q-Q plot has heavy tails, it means the points on the plot curve upwards and or downwards at the ends. This suggests the tails of the distribution are heavier than expected under a normal distribution. This indicates our data has extreme values that occur more frequently than would be expected if the data were normally distributed. A heavy-tailed distribution can have a significant impact our statistical inference. Let's check our findings with a histogram of *Model 2*.

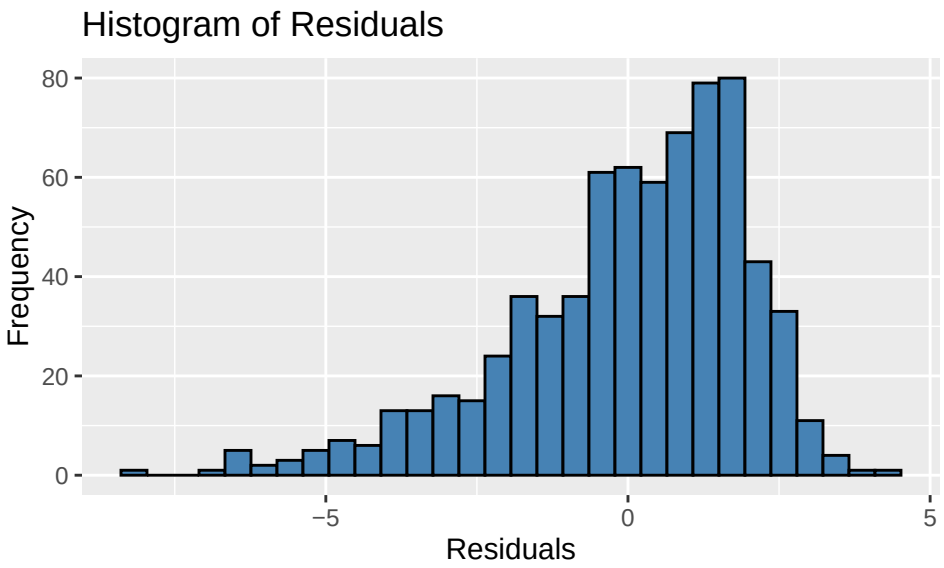

Figure 9: *Model 2* Histogram

In Figure 9, we can see the histogram is skewed to the left. In a skewed left histogram, majority of the data are clustered to the right, and there are relatively fewer observations on the left side. This indicates that the distribution is negatively skewed. A negatively skewed distribution has a mean that is less than the median, and the tail of the distribution is on the left-hand side. This suggests that there are some extreme values on the left side of the distribution that are pulling the mean towards that side. This confirms our suspicions in Figure 7, and the normality assumption is violated.

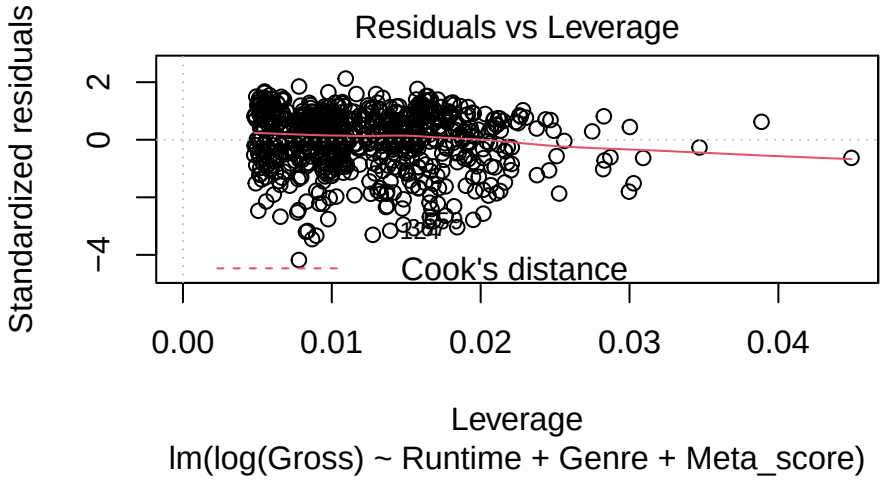

Figure 10: *Model 2* Residual vs. Leverage Plot

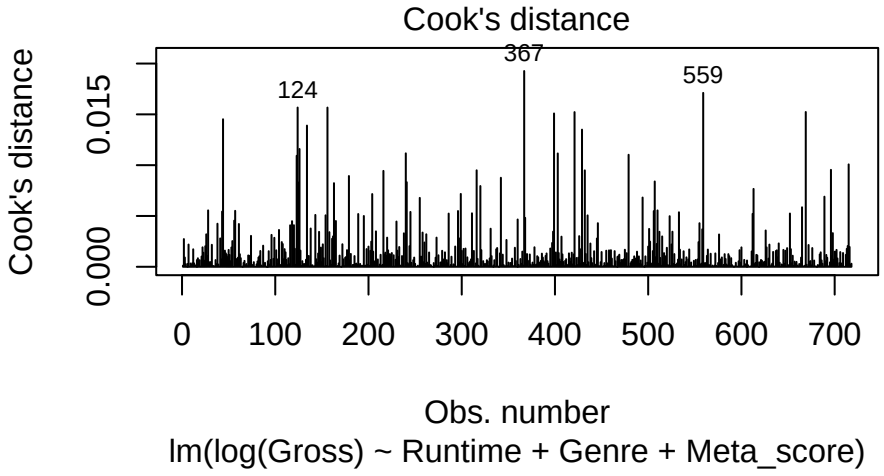

Figure 11: *Model 2* Cook's Distance Plot

In Figures 10 and 11, we can see from both Residual vs. Leverage and Cook's Distance plots, there are no obvious outliers or influential values. For standardized residuals, we have no value that exceeds a Cook's Distance of .05 and for observations, we have no value exceeding .020 on the Cook's Distance scale. We have no values exceeding the Cook's Distance of 1, which is great. We can assume all outliers and influential values were successfully removed from the data set.

Now that we verified our assumptions from *Model 2* we can conduct our testing. Below are the results for *Test 1, 2, 3, and 4*:

Table 3: Linear Regression Testing

	Dependent variable:			
	log(Gross)			
	Test 1	Test 2	Test 3	Test 4
	(1)	(2)	(3)	(4)
Runtime	0.018^{***}	0.018^{***}	0.017^{***}	
	p = 0.000	p = 0.000	p = 0.00000	
Genre: Adventure	-0.720^{**}	-0.757^{**}		-0.643^{**}
	p = 0.024	p = 0.018		p = 0.049
Genre: Animation	-0.259	-0.342		-0.790^{**}
	p = 0.423	p = 0.284		p = 0.013
Genre: Biography	-0.553^{*}	-0.576^{**}		-0.431
	p = 0.060	p = 0.050		p = 0.151
Genre: Comedy	-1.325^{***}	-1.373^{***}		-1.664^{***}
	p = 0.00000	p = 0.00000		p = 0.000
Genre: Crime	-1.892^{***}	-1.925^{***}		-1.889^{***}
	p = 0.000	p = 0.000		p = 0.000
Genre: Drama	-1.882^{***}	-1.933^{***}		-1.919^{***}
	p = 0.000	p = 0.000		p = 0.000
Metascore	-0.010		-0.015^{**}	-0.009
	p = 0.112		p = 0.021	p = 0.149
Constant	16.399^{***}	15.684^{***}	15.839^{***}	18.682^{***}
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	718	718	718	718
R^2	0.171	0.168	0.050	0.132
Adjusted R^2	0.162	0.160	0.047	0.123
Residual Std. Error	1.988 (df = 709)	1.990 (df = 710)	2.119 (df = 715)	2.033 (df = 710)
F Statistic	18.291^{***} (df = 8; 709)	20.498^{***} (df = 7; 710)	18.836^{***} (df = 2; 715)	15.422^{***} (df = 7; 710)
Note:				*p<0.1; **p<0.05; ***p<0.01

To conclude our testing, we found none of the models to be a good fit. We see models 2, 3, 4, and 5 have corresponding R^2 values of .171, .168, .050, and .132, respectively. These are extremely low values, which tell us we have a poor overall fit.

After testing the overall significance of each model, we found very low $p-values$ (close to 0) for each model. This provides strong evidence that an association exists between at least one coefficient in each model and our response variable, the Logarithm of Gross Revenue. Let's break down each model one by one to verify this conclusion.

In *Test 1*, we see that two factors of Genre, Genre: Animation and Genre: Biography, as well as Metascore prove to be insignificant in predicting the Logarithm of Gross Revenue, while Runtime and the four other factors of Genre were significant in predicting the Logarithm of Gross Revenue. This test proves nothing for the categorical predictor, Genre. However, it proves my initial hypothesis wrong, in that Metascore would have the largest association with Gross Revenue, in fact, based on this we can conclude there is no linear association between Metascore and the logarithm of Gross Revenue after accounting for the other predictors in the model. Runtime surprised me by being statistically significant in helping to predict the Logarithm of Gross Revenue.

In *Test 2*, reduced model (3), two factors of Genre, Genre: Animation and Genre: Biography prove to be insignificant in predicting the Logarithm of Gross Revenue, while Runtime and the four other factors of Genre were significant in predicting the Logarithm of Gross Revenue. This is the same exact conclusion we arrived to looking at the full model, this time excluding Metascore. It's intriguing that both Animation and Biography Genres are proven to be insignificant again. Still, no conclusions can be drawn based on this. Removing Metascore from the model changed nothing, proving it has zero significance in the full model.

In *Test 3*, reduced model (4), when excluding Genre, both Runtime and Metascore prove to be significant in predicting the Logarithm of Gross Revenue. This can mean several things. First, it could be Genre was obscuring the relationship between Metascore and the Logarithm of Gross Revenue. Or it could be Genre is collinear with Metascore. To

assess the collinearity, we can run a Variance Inflation Factor (VIF) test to calculate the amount of correlation between each predictor variable and all other predictor variables. The VIF score indicates how much the variance of an estimated regression coefficient is inflated due to multicollinearity. A VIF score of 1 means there is little to no correlation between the predictor variable and the other variables in the model, while a score typically greater than 10 is considered high correlation. After running a VIF test, we can dismiss the multicollinearity thought:

VIF Test

$$\textit{Runtime}\ 1.197592.$$

$$\textit{Genre}\ 1.238598.$$

$$\textit{Metascore}\ 1.035585.$$

Finally, in *Test 4*, reduced model (5), when excluding Runtime, Genre: Biography and Metascore prove to be insignificant when predicting the Logarithm of Gross Revenue. All other factor levels of Genre are significant when predicting for the Logarithm of Gross Revenue. Again, we cannot make any conclusion on Genre. However, Genre: Biography proved to be insignificant in every test, which should be noted.

5 Conclusion

With this study, I aimed to provide useful information to film production companies when creating a budget for a new movie. To achieve this, I focused on three predictor variables: Runtime, Genre, and Metascore. I determined the variable of success (response variable) to be gross box office revenue in USD. Unfortunately, the assumptions in my original model were horribly violated so I had to apply a $log(y)$ transformation to help normalize the distribution of the variables and improve linearity. This worked out relatively well, however there was still a normality violation, which can and will lead to incorrect conclusions by

affecting $p-values$. Nonetheless, I concluded there was overall significance found in the full model and reduced models for at least one coefficient. Meaning, at least one predictor has an established association with predicting the Logarithm of Gross Revenue. Focusing on the full model, to my surprise, the only variable that proved significant in predicting the Logarithm of Gross Revenue was Runtime. Therefore, film production companies should focus their efforts on making quality movies with a longer screen time. Based on this conclusion, we can assume consumers purchase more box office tickets the longer the movie.

There are a few limitations on this current study. For one, I based this study off a linear regression model, which has limitations of its own. To be more specific, a linear regression model assumes linearity and normality, however in the real world these conditions are hardly ever true (like we saw in our study). When these conditions are being violated, they can lead to inaccurate statistics and wrong conclusions. I believe my data set was also a limitation. If I had more time, I would have preferred to collect the data myself. I felt very limited in what predictor variables I could use for this study and, ultimately, I was not able to use the variables I wanted. For example, I wanted to include director as a predictor variable in my model, but due to time constraints I was not able to collect the appropriate data to be able to include the director variable.

There's so much that can be done for future studies. The film industry is vast. There are so many variables unaccounted for, which can help draw more accurate, and useful conclusions. With further analysis, stakeholders can identify trends in audience preferences, marketing strategies, and distribution channels, leading to better-informed decisions that can result in increased profits. Another approach could be identifying and addressing issues related to diversity and representation in the film industry. By analyzing data on the representation of different groups in films, studios can identify areas where they need to improve and work towards creating more inclusive and diverse content. Overall, future studies are crucial to the future of the film industry, as it can help businesses make better decisions, create better content, and work towards a more inclusive and diverse industry.

References

[1] Gutbezahl, J. (2017, June) *5 types of statistical biases to avoid in your analyses*, Business Insights Blog.

[2] Sood, N. (2017, November) *Factors affecting the success of movies- a case study of twin movies*, International Journal of Innovative Science and Research Technology.

[3] Arthur, M. B. (2018, October) *Why the film industry matters – not for entertainment, but for your career*, Forbes.

[4] Motion Picture Association. (2023, January) *Driving economic growth*, Motion Picture Association.

[5] Hlavac, M. (2022, October) *Stargazer: Well-formatted regression and summary statistics tables.*.

[6] IMDB. (2023, April). *IMDB*, IMDB.

[7] Kassambara. (2018, March) *Linear regression assumptions and diagnostics in R: Essentials*, Statistical tools for high-throughput data analysis.

[8] Rtatman. (2017, December) *Removing influential points*, Kaggle.

[9] Shankhdhar, H. (2021, February) *IMDB movies dataset*, Kaggle.

[10] Thieme, C. (2021, June) *Understanding linear regression output in R*, Towards Data Science.

[11] University of Technology, Sydney (2017, November) *Study explores what really makes a movie successful*, PHYS ORG.

[12] Zach. (2022, April) *How to interpret regression coefficients*, Statology.

[13] Metacritic. (2023, April) *How we create the metascore magic*, Metacritic.