The Rated-R team

presents

Contents

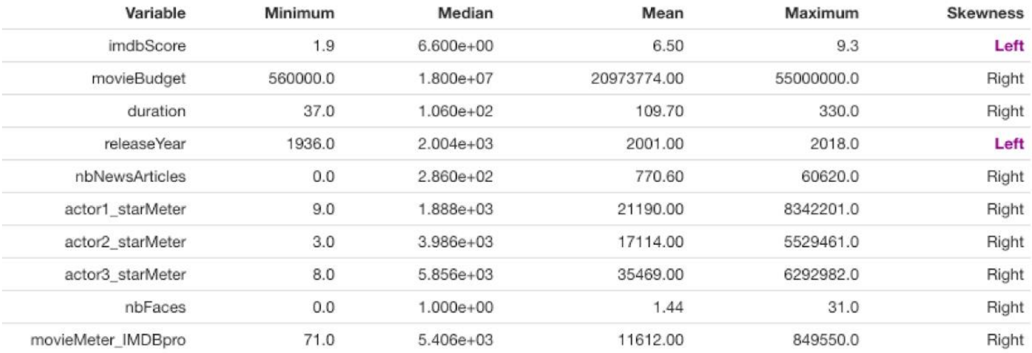
Introduction

With the rapid development of the Internet, film scoring websites have become important platforms for the expression of film reviews and ratings. The most prominent and influential of such websites is IMDB, Internet Movie Database. Data in IMDB has information about films, directors, actors, budget, duration, language, genre, IMDB rating, and much more. IMDB rating has become one of the most popular film scoring metrics, as a result, many people use IMDB to find good movies to watch. However, movies that have not been released don’t have any rating and this poses a problem for many people because they rely on IMDB ratings as a means of evaluating the merit of a movie. It is important to note that a large number of variables contribute to how high or low the IMDB rating of a particular film ends up being. The objective of this project is to create a statistical model that will predict the rating of twelve upcoming movies: Falling for Christmas, Black Panther: Wakanda Forever, Spirited, Paradise City, Poker Face, ¡Que viva México!, Slumberlan, Blue's Big City Adventure, The Menu, The Fabelmans, Devotion, Strange World. A great variety of variables were examined such as director, budget, actors, year, and duration to name a few. Each variable was rigorously analyzed to identify which set of predictors has a significant relationship with IMDB rating. Once such predictors were identified, the team built a statistical model with the objective of increasing the accuracy of IMDB rating predictions. The next sections will elaborate on the data description, the methodology used to build the model, and the results of the final model.

Data Description

Quantitative Variables

The quantitative variables were explored individually to understand their distributions. The findings are summarized in the table below:



Variable	Minimum	Median	Mean	Maximum	Skewness
imdbScore	1.9	6.600e+00	6.50	9.3	**Left**
movieBudget	560000.0	1.800e+07	20973774.00	55000000.0	Right
duration	37.0	1.060e+02	109.70	330.0	Right
releaseYear	1936.0	2.004e+03	2001.00	2018.0	**Left**
nbNewsArticles	0.0	2.860e+02	770.60	60620.0	Right
actor1_starMeter	9.0	1.888e+03	21190.00	8342201.0	Right
actor2_starMeter	3.0	3.986e+03	17114.00	5529461.0	Right
actor3_starMeter	8.0	5.856e+03	35469.00	6292982.0	Right
nbFaces	0.0	1.000e+00	1.44	31.0	Right
movieMeter_IMDBpro	71.0	5.406e+03	11612.00	849550.0	Right

Table 1: Descriptive analytics of Quantitative Variables

Categorical Variables

There are 18 categorical variables in the dataset. For the purpose of analyzing the dataset, we will not include descriptive variables: movieTitle, movieID, imdbLink and plotKeyword. We also decided not to include aspectRatio, as the movies examined are released in cinema and not on TV screen. Additionally, maturityRating was also eliminated as this variable only serves to restrict the audience of the movie, and those who are restricted by this variable would not be able to provide rating for the movie on IMDB.

The remaining categorical variables were also studied one at a time to obtain a better understanding of the data.

The first step was to dummify all the categorical variables. Four exceptions were made for the ReleaseMonth, the Language, the Production Company and the Director. The details of those variables can be seen in the Appendix A.

For the Month of release, we categorized the month into 4 groups for each season. We then created the variable predictor **Season** such as: March-April-May in Spring, June-July-August in Summer, September-October-November in Fall, December-January-February in Winter. The new variable was then dummified.

For the languages, we noticed that most of the observations were categorized as English movies. We then decided to dummify the languages by creating a new predictor called **English**. When the movie is in English we’ll have a 1, when the movie is in another language we get a 0.

For the Production Company, we decided to reduce the number of unique values by filtering out those that have more than 96 observations. The Production Companies that have less than 96 observations in this dataset were put into the general category Other.

We now have the predictor **ProductionCompany** with four selections : Universal Pictures, Paramount Pictures, Columbia Pictures Corporation and Other. The new variable was then dummified.

For the Director we decided to apply the same technique we used for the Production Company. We only kept the Director that appeared in more than ten observations. The others were put in the self-titled category.

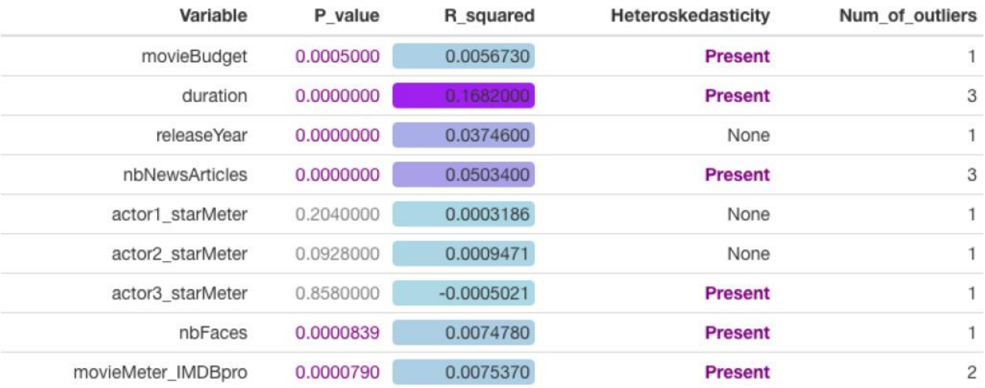
We now have the predictor **Director** with six selections : Woody Allen, Steven Spielberg, Clint Eastwood, Spike Lee, Steven Soderbergh and Other. The new variable was then dummified.

Relationships between variables

A simple linear regression was run for each predictor against the target variable (imdbScore) to understand the relationship between them. We noted the Adjusted R-squared and the p-value to understand the significance of that predictor. If the p-value is lower than 0.05, we consider that the predictor in question is significant.

We also conducted a heteroskedasticity test on each predictor as well as an outlier test. The initial findings are summarized below.

Quantitative Variables

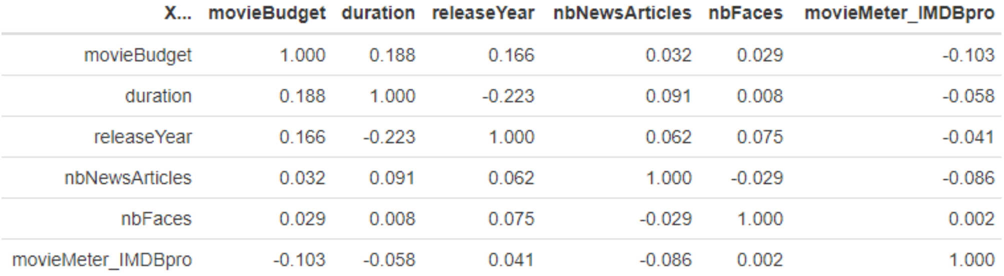


Variable	P_value	R_squared	Heteroskedasticity	Num_of_outliers
movieBudget	0.0005000	0.0056730	Present	1
duration	0.0000000	0.1682000	Present	3
releaseYear	0.0000000	0.0374600	None	1
nbNewsArticles	0.0000000	0.0503400	Present	3
actor1_starMeter	0.2040000	0.0003186	None	1
actor2_starMeter	0.0928000	0.0009471	None	1
actor3_starMeter	0.8580000	-0.0005021	Present	1
nbFaces	0.0000839	0.0074780	Present	1
movieMeter_IMDBpro	0.0000790	0.0075370	Present	2

Table 2: Summary of linear regression results with each quantitative variable
*NOTE: For variables **duration**, **releaseYear** and **nbNewsArticles**, the p value is <2e-16, which is why it is represented as 0.0000000 in the table.*

In terms of quantitative variables, the ones that are significant regarding our target variable are: **movieBudget, duration, releaseYear,nbNewsArticles, nbFaces, movieMeter_IMDBpro.**

The actor star meters (actor1_starMeter, actor2_starMeter, actor3_starMeter) were considered as not significant at all (their p-values are > 0.05). Finally, we also led a test on collinearity for the quantitative variables. The results displayed in the correlation matrix below show that the predictors are not collinear.

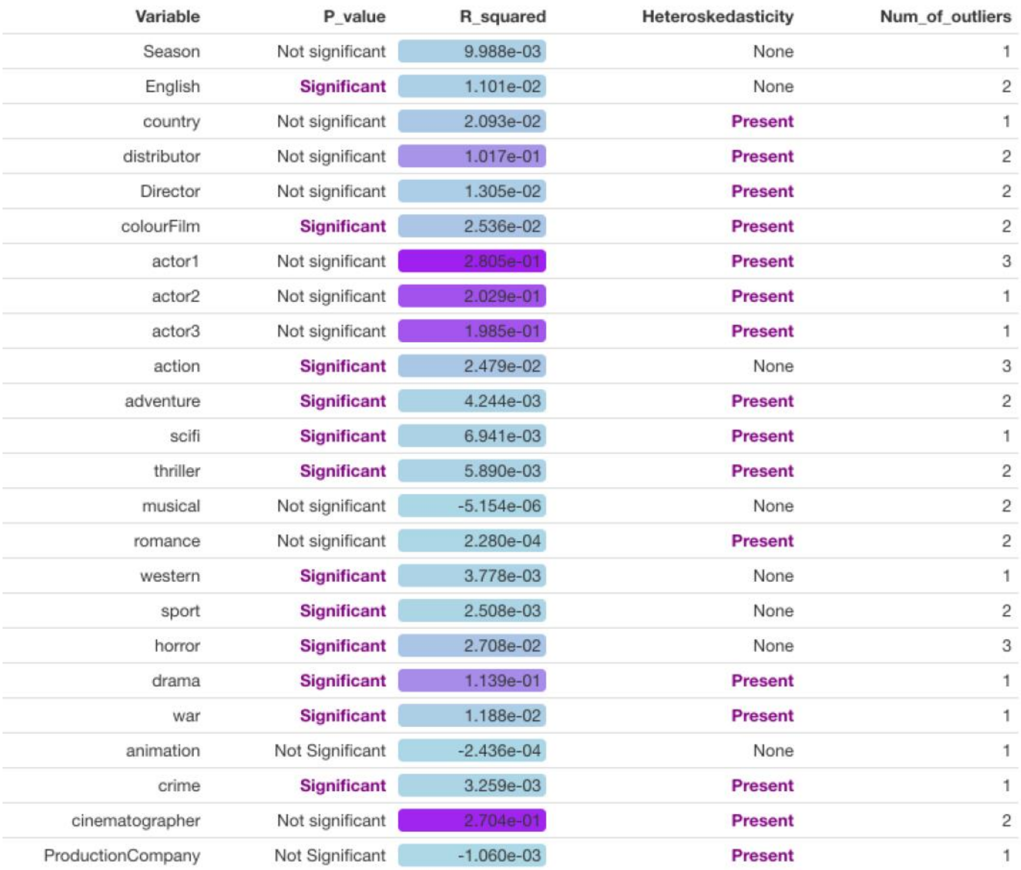


X...	movieBudget	duration	releaseYear	nbNewsArticles	nbFaces	movieMeter_IMDBpro
movieBudget	1.000	0.188	0.166	0.032	0.029	-0.103
duration	0.188	1.000	-0.223	0.091	0.008	-0.058
releaseYear	0.166	-0.223	1.000	0.062	0.075	-0.041
nbNewsArticles	0.032	0.091	0.062	1.000	-0.029	-0.086
nbFaces	0.029	0.008	0.075	-0.029	1.000	0.002
movieMeter_IMDBpro	-0.103	-0.058	0.041	-0.086	0.002	1.000

Table 3: Correlation Matrix of Quantitative Variables

Categorical Variables

We run a simple linear regression on each categorical variable against the target variable. For the variables with a lot of categories, we looked at the p-value for all the categories. If the majority was not significant, we considered that variable as not significant. This can be seen in the following table under ‘most p-value’. The details of the p-value can be seen in the Appendix B (for the variables with a lot of different observations like distributors, cinematographer, actor and country, please refer to the code).



Variable	P_value	R_squared	Heteroskedasticity	Num_of_outliers
Season	Not significant	9.988e-03	None	1
English	**Significant**	1.101e-02	None	2
country	Not significant	2.093e-02	**Present**	1
distributor	Not significant	1.017e-01	**Present**	2
Director	Not significant	1.305e-02	**Present**	2
colourFilm	**Significant**	2.536e-02	**Present**	2
actor1	Not significant	2.805e-01	**Present**	3
actor2	Not significant	2.029e-01	**Present**	1
actor3	Not significant	1.985e-01	**Present**	1
action	**Significant**	2.479e-02	None	3
adventure	**Significant**	4.244e-03	**Present**	2
scifi	**Significant**	6.941e-03	**Present**	1
thriller	**Significant**	5.890e-03	**Present**	2
musical	Not significant	-5.154e-06	None	2
romance	Not significant	2.280e-04	**Present**	2
western	**Significant**	3.778e-03	None	1
sport	**Significant**	2.508e-03	None	2
horror	**Significant**	2.708e-02	None	3
drama	**Significant**	1.139e-01	**Present**	1
war	**Significant**	1.188e-02	**Present**	1
animation	Not Significant	-2.436e-04	None	1
crime	**Significant**	3.259e-03	**Present**	1
cinematographer	Not significant	2.704e-01	**Present**	2
ProductionCompany	Not Significant	-1.060e-03	**Present**	1

Table 4: Summary of linear regression results with each chosen categorical variable

In terms of categorical variables, the ones that are significant regarding our target variable are: **English, colourFilm, action, adventure, scifi, thriller, sport, horror, drama, war and crime.**

Model Selection

As mentioned in the previous section, we decided to build our first model (Model 1) with all the predictors that have individually significant relationships with the dependent variable imdbScore. However, for all variables that indicate genre, we decided to only include those that have more than 100 observations. Hence, 'western', 'sport' and 'war' variables are eliminated. The summary of Model 1 is in Regression Table 1, Appendix C.

After testing Model 1, we decided to remove 'thriller', 'adventure', 'director' and 'productionCompany' variables due to their insignificance (each have p-value that is larger than 0.05). At this phase, we also carried out the outlier test and eliminated 7 outliers to increase the MSE of this model (see Appendix D).

We then ran a new model (Model 2) with the new updated dataset along with the new set of predictors. The non-linearity test reveals that duration, nbNewsArticles and movieMeter_IMDBpro are currently significantly not linears (see Appendix E). In order to fix the non-linearity, we decided to use spline and polynomial regression. To avoid overfitting, we chose to keep the number of knots between 2 and 4 and the degree of polynomial between 2 and 5. For each number of knots, we run a loop of degree of polynomial to determine the model with the lowest MSE and would then compare those models to determine the one with the lowest MSE overall. After choosing our final model (Model 3), we will then run the test for heteroskedasticity and correct to finalize the model.

Results

Our final prediction model is:

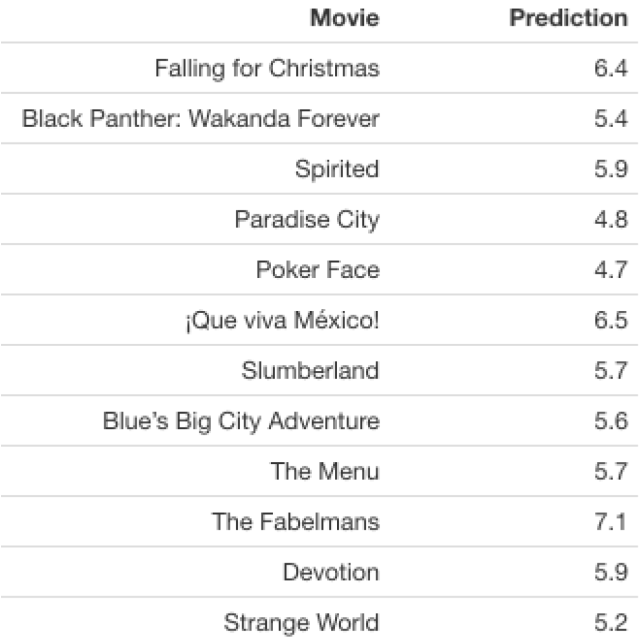
```
    lm(imdbScore~movieBudget + releaseYear +
                    bs(duration,knots = c(c1,c2,c3), degree  =  4) +
                    bs(nbNewsArticles, knots = c(d1,d2,d3), degree = 2) +
                    nbFaces +
                    bs(movieMeter_IMDBpro,knots = c(f1,f2,f3), degree = 2) +
                    colourFilm + action + horror + drama + crime +
          English)
```

**(c1,c2,c3); (d1,d2,d3); (f1,f2,f3) are the 0.25, 0.5 and 0.75 knots for duration, nbNewsArticles, movieMeter_IMDBpro respectively.*

The summary of the regression model after correcting for heteroskedasticity is found in the appendix.

The K-fold cross-validation test (with K=10) for this model gives an mse score of 0.62. The model has an adjusted R-squared score of 47.31%.

Our **final prediction** for the test data is:



Movie	Prediction
Falling for Christmas	6.4
Black Panther: Wakanda Forever	5.4
Spirited	5.9
Paradise City	4.8
Poker Face	4.7
¡Que viva México!	6.5
Slumberland	5.7
Blue’s Big City Adventure	5.6
The Menu	5.7
The Fabelmans	7.1
Devotion	5.9
Strange World	5.2

Table 5: Movie rating predictions

Interpretation of the predictors

The coefficient estimate for movieBudget is -9.5274e-09. This means that there is a negative correlation between the budget spent on a movie and the rating it receives on IMDB (i.e., movies with higher budgets are receiving lower ratings all other things being equal). A standard error of 1.3553e-09 suggests that the data adheres tightly to the regression line. Since the absolute value of this estimate is very small, it is reported in the coefficient summary table as -0.0000. However, this does not mean that it is irrelevant. Since the values for movieBudget are very large, it results in an extremely small coefficient estimate.

The coefficient estimate for releaseYear is -1.4711e-02. This means that there is a negative correlation between the year of release and the rating a movie receives on IMDB (i.e., older movies are receiving higher ratings than newer movies all other things being equal). A standard error of 1.6897e-03 suggests that the data adheres tightly to the regression line.

The coefficient estimate for nbFaces is -3.9701e-02. This means that there is a negative correlation between the number of faces in the main poster of a movie and the rating it receives on IMDB (i.e., movies that have many faces in the main poster are receiving lower ratings compared to movies with fewer faces, all other things being equal). A standard error of 9.6833e-03 suggests that the data adheres tightly to the regression line.

The coefficient estimate for colourFilmColor is -4.2195e-01. This means that all other things being equal, on average a coloured movie is receiving a lower rating (specifically by 4.2195e-01) when compared to a black and white movie. A standard error of 6.8358e-02 suggests that the data adheres tightly to the regression line.

The coefficient estimate for action is -3.0591e-01. This means that all other things being equal, on average an action movie is receiving a lower rating (specifically by -3.0591e-01) when compared to other genres of movies (not including horror, drama, and crime). A standard error of 5.3010e-02 suggests that the data adheres tightly to the regression line.

The coefficient estimate for horror is -4.5808e-01. This means that all other things being equal, on average a horror movie is receiving a lower rating (specifically by -4.5808e-01) when compared to other genres of movies (not including action, drama, and crime). A standard error of 6.5092e-02 suggests that the data adheres tightly to the regression line.

The coefficient estimate for drama is 3.8190e-01. This means that all other things being equal, on average a drama movie is receiving a higher rating (specifically by 3.8190e-01) when compared to other genres of movies (not including action, horror, and crime). A standard error of 4.5149e-02 suggests that the data adheres tightly to the regression line.

The coefficient estimate for crime is 1.7138e-01. This means that all other things being equal, on average a crime-based movie is receiving a higher rating (specifically by 1.7138e-01) when compared to other genres of movies (not including action, horror, and drama). A standard error of 4.2363e-02 suggests that the data adheres tightly to the regression line.

The coefficient estimate for English is -7.7225e-01. This means that all other things being equal, on average a movie with English as the dominant language is receiving a lower rating (specifically by -7.7225e-01) when compared to movies where other languages are dominant. A standard error of 1.2232e-01 suggests that the data adheres tightly to the regression line.

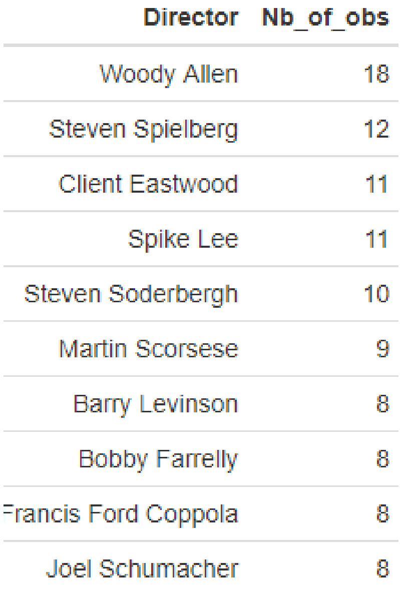
Since each predictor above has a p-value of less than 0.001, it means that their coefficient estimates are statistically significant at the 0.1% level. Thus, there is strong evidence against the null hypothesis for each predictor (which claims that there is no correlation between the predictor and imdbScore). Since duration, nbNewsArticles, and movieMeter_IMDBpro are polynomial splines in the regression model, we lose the ability to interpret their coefficients.

An immediate area of improvement for this analysis would be to go back after getting the regression results and further transform the data, so that it becomes easier to visualize the impact of the coefficient estimates. For example, transforming movieBudget so that it represents a movie's budget in terms of millions of dollars would greatly increase the absolute value of its coefficient estimate, which would make it easier to comprehend.

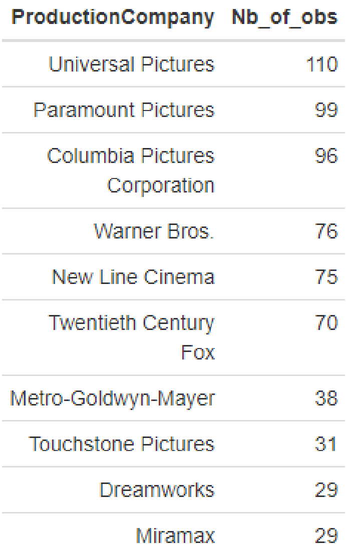
Appendices

Appendix A : Data Description

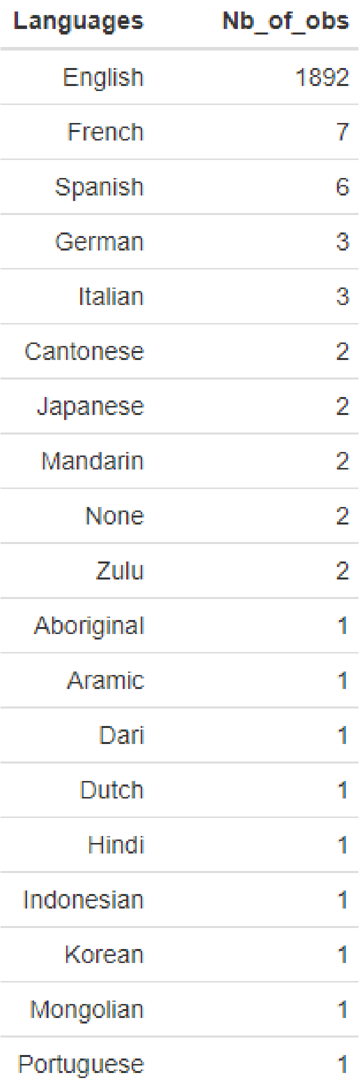
Number of Movies Per Language   Number of Movies Per Production   Number of Movies Per Director



Director	Nb_of_obs
Woody Allen	18
Steven Spielberg	12
Client Eastwood	11
Spike Lee	11
Steven Soderbergh	10
Martin Scorsese	9
Barry Levinson	8
Bobby Farrelly	8
Francis Ford Coppola	8
Joel Schumacher	8



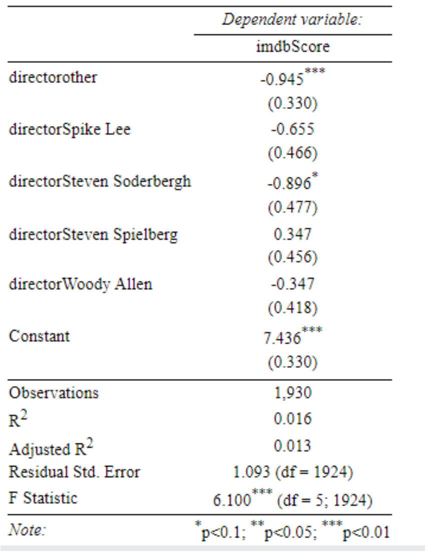
ProductionCompany	Nb_of_obs
Universal Pictures	110
Paramount Pictures	99
Columbia Pictures Corporation	96
Warner Bros.	76
New Line Cinema	75
Twentieth Century Fox	70
Metro-Goldwyn-Mayer	38
Touchstone Pictures	31
Dreamworks	29
Miramax	29



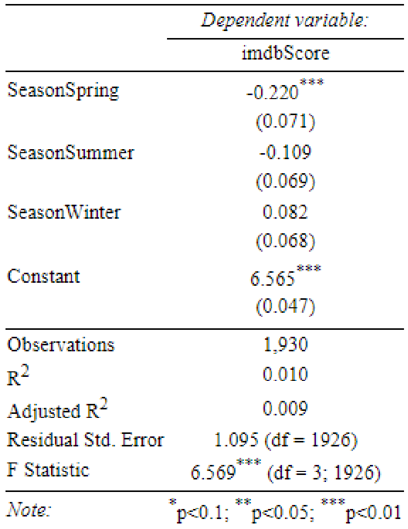
Languages	Nb_of_obs
English	1892
French	7
Spanish	6
German	3
Italian	3
Cantonese	2
Japanese	2
Mandarin	2
None	2
Zulu	2
Aboriginal	1
Aramic	1
Dari	1
Dutch	1
Hindi	1
Indonesian	1
Korean	1
Mongolian	1
Portuguese	1

Appendix B: P-Value of the categorical variables

Regression Table: imdbScore~Director

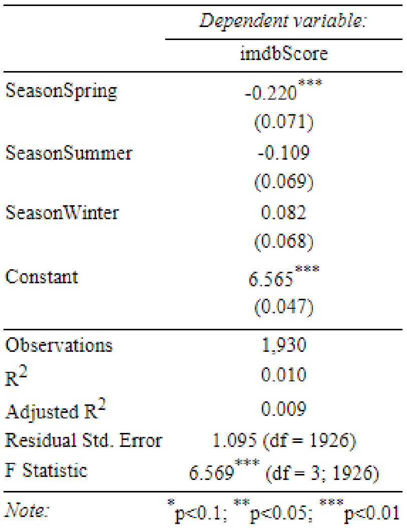


	Dependent variable:
	imdbScore
directorother	-0.945^{***}
	(0.330)
directorSpike Lee	-0.655
	(0.466)
directorSteven Soderbergh	-0.896^{*}
	(0.477)
directorSteven Spielberg	0.347
	(0.456)
directorWoody Allen	-0.347
	(0.418)
Constant	7.436^{***}
	(0.330)
Observations	1,930
R^2	0.016
Adjusted R^2	0.013
Residual Std. Error	1.093 (df = 1924)
F Statistic	6.100^{***} (df = 5; 1924)
Note:	*p<0.1; **p<0.05; ***p<0.01

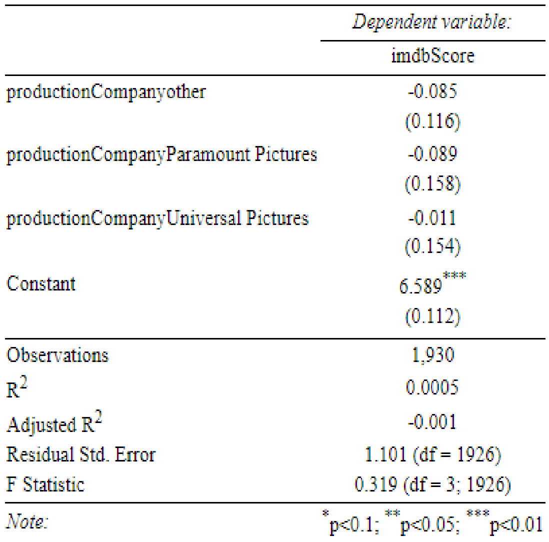


	Dependent variable:
	imdbScore
SeasonSpring	-0.220^{***}
	(0.071)
SeasonSummer	-0.109
	(0.069)
SeasonWinter	0.082
	(0.068)
Constant	6.565^{***}
	(0.047)
Observations	1,930
R^2	0.010
Adjusted R^2	0.009
Residual Std. Error	1.095 (df = 1926)
F Statistic	6.569^{***} (df = 3; 1926)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Table:imdbScore~Season



	Dependent variable:
	imdbScore
SeasonSpring	-0.220^{***}
	(0.071)
SeasonSummer	-0.109
	(0.069)
SeasonWinter	0.082
	(0.068)
Constant	6.565^{***}
	(0.047)
Observations	1,930
R^2	0.010
Adjusted R^2	0.009
Residual Std. Error	1.095 (df = 1926)
F Statistic	6.569^{***} (df = 3; 1926)
Note:	*p<0.1; **p<0.05; ***p<0.01

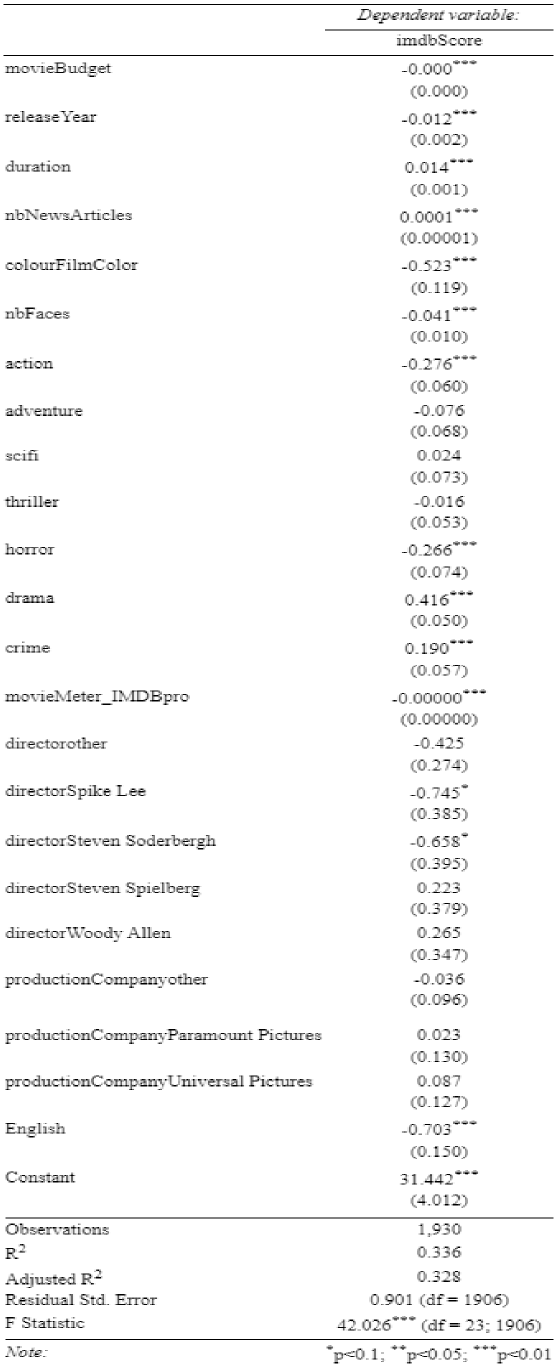


	Dependent variable:
	imdbScore
productionCompanyother	-0.085
	(0.116)
productionCompanyParamount Pictures	-0.089
	(0.158)
productionCompanyUniversal Pictures	-0.011
	(0.154)
Constant	6.589^{***}
	(0.112)
Observations	1,930
R^2	0.0005
Adjusted R^2	-0.001
Residual Std. Error	1.101 (df = 1926)
F Statistic	0.319 (df = 3; 1926)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Table: imdbScore~Production Company

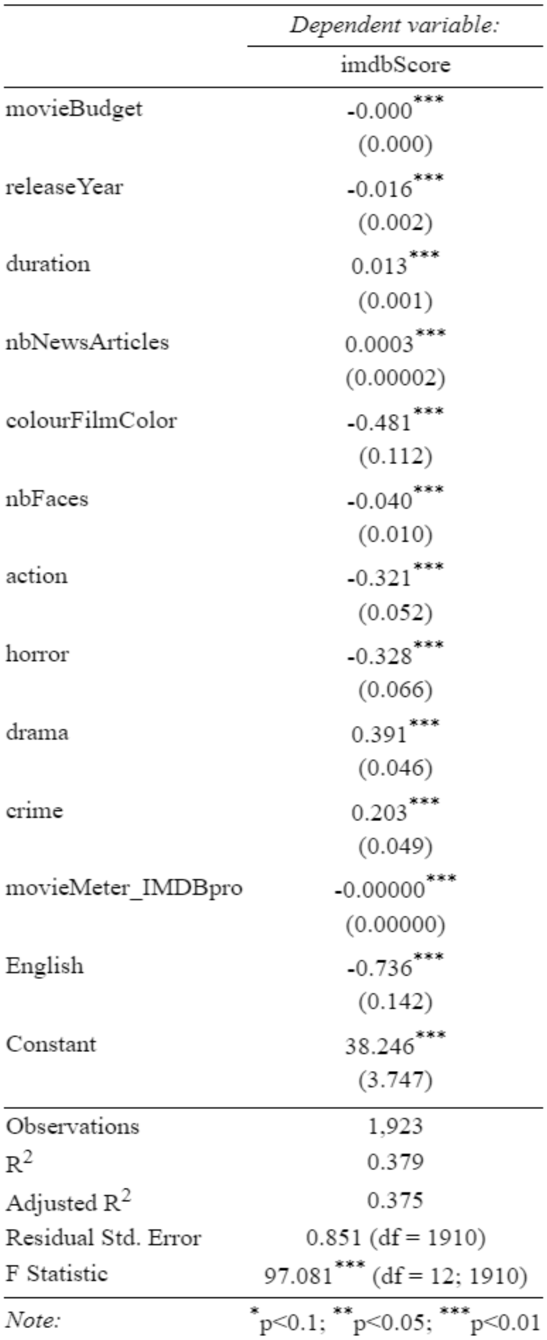
Appendix C: Regression Output

Regression Table 1: Regression Output of Model 1



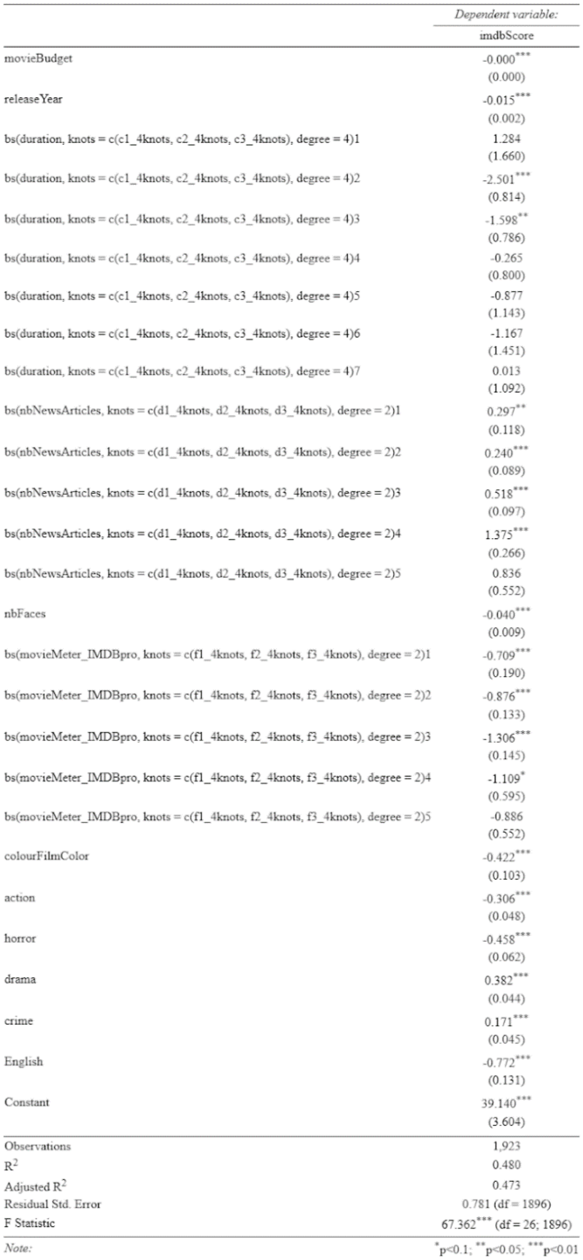
	Dependent variable:
	imdbScore
movieBudget	-0.000*** (0.000)
releaseYear	-0.012*** (0.002)
duration	0.014*** (0.001)
nbNewsArticles	0.0001*** (0.00001)
colourFilmColor	-0.523*** (0.119)
nbFaces	-0.041*** (0.010)
action	-0.276*** (0.060)
adventure	-0.076 (0.068)
scifi	0.024 (0.073)
thriller	-0.016 (0.053)
horror	-0.266*** (0.074)
drama	0.416*** (0.050)
crime	0.190*** (0.057)
movieMeter_IMDBpro	-0.00000*** (0.00000)
directorother	-0.425 (0.274)
directorSpike Lee	-0.745* (0.385)
directorSteven Soderbergh	-0.658* (0.395)
directorSteven Spielberg	0.223 (0.379)
directorWoody Allen	0.265 (0.347)
productionCompanyother	-0.036 (0.096)
productionCompanyParamount Pictures	0.023 (0.130)
productionCompanyUniversal Pictures	0.087 (0.127)
English	-0.703*** (0.150)
Constant	31.442*** (4.012)
Observations	1,930
R^2	0.336
Adjusted R^2	0.328
Residual Std. Error	0.901 (df = 1906)
F Statistic	42.026*** (df = 23; 1906)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Table 2: Regression Output of Model 2



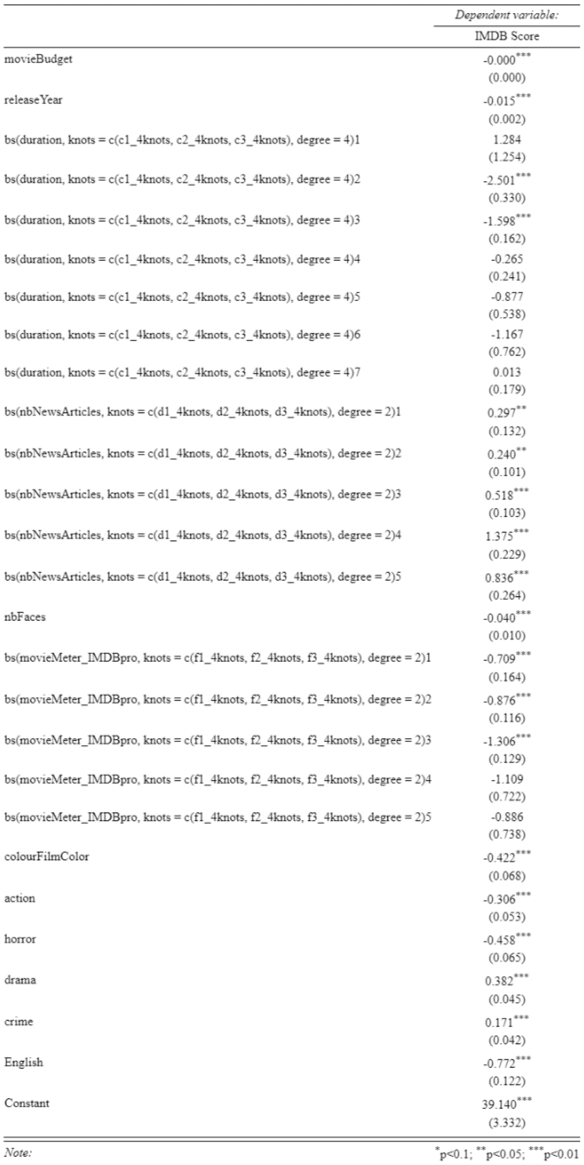
	Dependent variable:
	imdbScore
movieBudget	-0.000*** (0.000)
releaseYear	-0.016*** (0.002)
duration	0.013*** (0.001)
nbNewsArticles	0.0003*** (0.00002)
colourFilmColor	-0.481*** (0.112)
nbFaces	-0.040*** (0.010)
action	-0.321*** (0.052)
horror	-0.328*** (0.066)
drama	0.391*** (0.046)
crime	0.203*** (0.049)
movieMeter_IMDBpro	-0.00000*** (0.00000)
English	-0.736*** (0.142)
Constant	38.246*** (3.747)
Observations	1,923
R^2	0.379
Adjusted R^2	0.375
Residual Std. Error	0.851 (df = 1910)
F Statistic	97.081*** (df = 12; 1910)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression Table 3: Regression output of Model 3



	Dependent variable:
	imdbScore
movieBudget	-0.000*** (0.000)
releaseYear	-0.015*** (0.002)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)1	1.284 (1.660)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)2	-2.501*** (0.814)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)3	-1.598** (0.786)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)4	-0.265 (0.800)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)5	-0.877 (1.143)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)6	-1.167 (1.451)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)7	0.013 (1.092)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)1	0.297** (0.118)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)2	0.240*** (0.089)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)3	0.518*** (0.097)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)4	1.375*** (0.266)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)5	0.836 (0.552)
nbFaces	-0.040*** (0.009)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)1	-0.709*** (0.190)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)2	-0.876*** (0.133)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)3	-1.306*** (0.145)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)4	-1.109* (0.595)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)5	-0.886 (0.552)
colourFilmColor	-0.422*** (0.103)
action	-0.306*** (0.048)
horror	-0.458*** (0.062)
drama	0.382*** (0.044)
crime	0.171*** (0.045)
English	-0.772*** (0.131)
Constant	39.140*** (3.604)
Observations	1,923
R^2	0.480
Adjusted R^2	0.473
Residual Std. Error	0.781 (df = 1896)
F Statistic	67.362*** (df = 26; 1896)
Note:	*p<0.1; **p<0.05; ***p<0.01

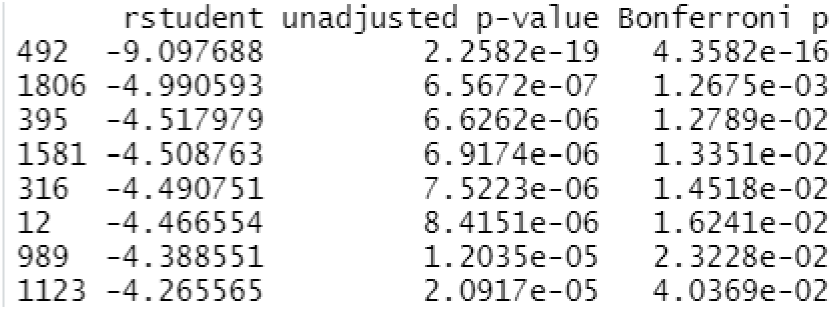
Regression Table 4: Regression output of Model 3 corrected for heteroscedasticity



	Dependent variable:
	IMDB Score
movieBudget	-0.000*** (0.000)
releaseYear	-0.015*** (0.002)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)1	1.284 (1.254)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)2	-2.501*** (0.330)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)3	-1.598*** (0.162)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)4	-0.265 (0.241)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)5	-0.877 (0.538)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)6	-1.167 (0.762)
bs(duration, knots = c(c1_4knots, c2_4knots, c3_4knots), degree = 4)7	0.013 (0.179)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)1	0.297** (0.132)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)2	0.240** (0.101)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)3	0.518*** (0.103)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)4	1.375*** (0.229)
bs(nbNewsArticles, knots = c(d1_4knots, d2_4knots, d3_4knots), degree = 2)5	0.836*** (0.264)
nbFaces	-0.040*** (0.010)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)1	-0.709*** (0.164)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)2	-0.876*** (0.116)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)3	-1.306*** (0.129)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)4	-1.109 (0.722)
bs(movieMeter_IMDBpro, knots = c(f1_4knots, f2_4knots, f3_4knots), degree = 2)5	-0.886 (0.738)
colourFilmColor	-0.422*** (0.068)
action	-0.306*** (0.053)
horror	-0.458*** (0.065)
drama	0.382*** (0.045)
crime	0.171*** (0.042)
English	-0.772*** (0.122)
Constant	39.140*** (3.332)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix D: Outlier Test

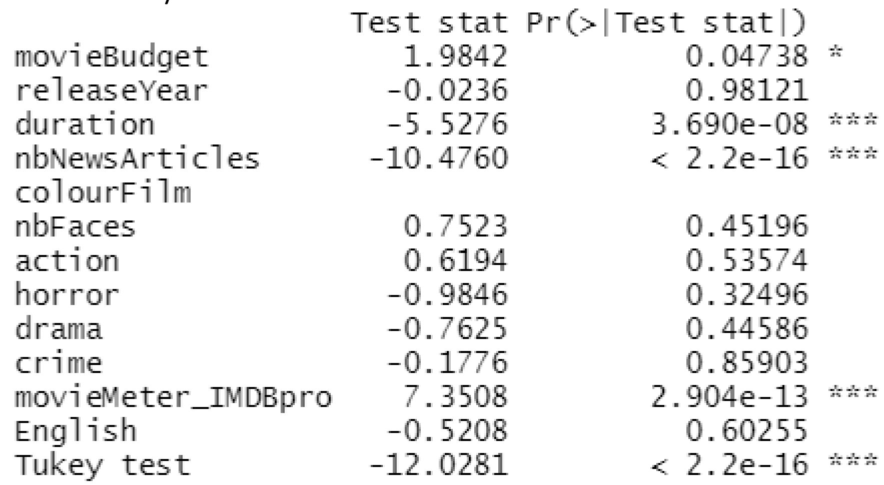
Table 1: Outlier Test for Model 1



```
      rstudent unadjusted p-value Bonferroni p
492  -9.097688         2.2582e-19   4.3582e-16
1806 -4.990593         6.5672e-07   1.2675e-03
395  -4.517979         6.6262e-06   1.2789e-02
1581 -4.508763         6.9174e-06   1.3351e-02
316  -4.490751         7.5223e-06   1.4518e-02
12   -4.466554         8.4151e-06   1.6241e-02
989  -4.388551         1.2035e-05   2.3228e-02
1123 -4.265565         2.0917e-05   4.0369e-02
```

Appendix E: Non-collinearity test for Model 2

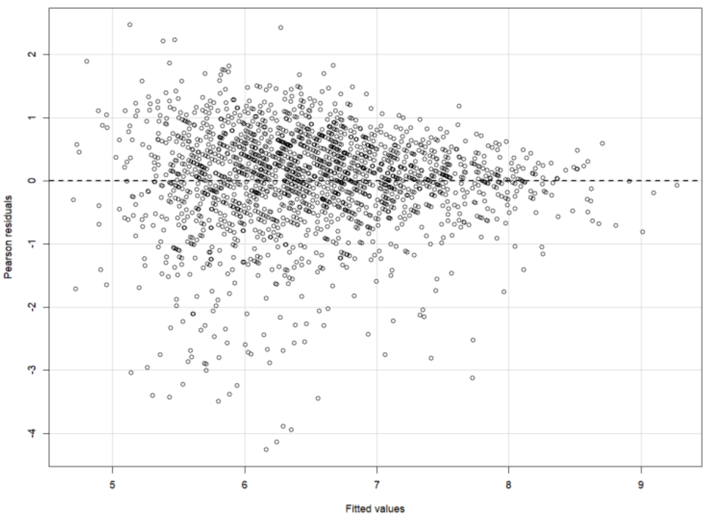
Table 1: Tukey test for Model 2



```
                   Test stat Pr(>|Test stat|)
movieBudget           1.9842          0.04738 *
releaseYear          -0.0236          0.98121
duration             -5.5276        3.690e-08 ***
nbNewsArticles      -10.4760        < 2.2e-16 ***
colourFilm
nbFaces               0.7523          0.45196
action                0.6194          0.53574
horror               -0.9846          0.32496
drama                -0.7625          0.44586
crime                -0.1776          0.85903
movieMeter_IMDBpro    7.3508        2.904e-13 ***
English              -0.5208          0.60255
Tukey test          -12.0281        < 2.2e-16 ***
```

Appendix F: Heteroskedasticity test

Graph 1: Funnel test for Model 2


Pearson residuals
2
1
0
-1
-2
-3
-4
5
6
7
8
9
Fitted values