Impact on the Review Rating for Horror Films 2012-2017

Ethan A. Kelley
West Virginia University
Introduction to Econometrics

December 2020

Abstract

This paper measures the impact on IMDB review ratings for horror films from 2012- 2017. Applying multivariate regression techniques using R to measure how specific variables, such as run time and genre, in R studio. All to observe the impact such variables have on a films' review rating.

1 Introduction

It's that time of year, the leaves are falling, there's a chill in the air, and the shallow glimmer of Jack O lanterns light up the streets. It's Halloween, the time of year where we all love a good fright. Many people love scary themed entertainment during this time of year especially horror films, including myself. There are a lot of horror films out there making it difficult to pick. So most people generally rely on review scores. This begs the question, what factors contribute to a films' review rating; especially a high rating? In this paper I will be analyzing specific variables that can contribute to an increase in the review score of horror films.

2 Data Collection and Analysis

The data set I will be using is a created data set from the website kaggle.com titled IMDB Horror Movie Data Set [2012 onwards]. This dataset contains a list of Horror movies released from 2012 to the present day. The dataset does not only include US horror films it also includes films from other countries such as Canada, the UK, the Philippines, and many others. The dataset also lists variables for movie rating, movie run time, plot, cast, release date, review rating, and the language the movie is in. For this project I will be focusing on how the other determinants in the dataset affect the review rating for movies in the US.

The IMDB data set allows me to address my question given the number of listed films present in the data set. Providing an ample amount of data to determine what variables influence a film review score and to what extent. Some limitations are the fact that there are non-US films listed in the data set and the budgets of these films are listed in their native currency which makes it difficult to use the initial Budget variable in

the data frame. Another limitation is the large number of missing values present in certain variables. This can make it difficult to use these variables in the analysis; some hardly have any values listed such as the movie rating variable although given that the data set consists of horror related films it would to assume that these movies are rated between PG-13 to Not Rated. Another issue is the issue that a film's favor-ability is largely subjective in nature. This is difficult to account for however, I can still reasonably gauge the impact the determent's may have on the review rating variable with what is provided in the dataset.

In this data set will be focusing on the present variables, Genre, Run Time, Release Date, and Budget. Now, since the data set also includes non-US films; for this analysis I'll be focusing on films in the US. Here, I will need to subset the original IMBD data set and create a new data frame for only US films. I will need to create new variables from each of the previously mentioned variables specifically for US only observations. From here, I will need to clean the variables Run Time, Genre, Release Date, and Budget. After cleaning the data I will then proceed to perform a multivariate regression analysis on determent's with respect to the review rating variable.

Starting with the Run Time variable, each observation is listed with "min" for the total number of minutes the move lasts. Here I will need to remove the "min" units from each observation to not cause any issues later on in the analysis. After removing "min" from Movie.Run.Time, I created two new columns for dummy variables. The first column would return a 1 if the movie run time is greater than or equal to 60 minutes. The second column will return a 1 if movie run time is less than 60 minutes.

Next, I will need to edit the "Release Date" variable to keep just the year by removing the days, months, and the dashes between the month, day and year using the "gsub" command in R. Removing both the dashes and the month and year will provide greater ease during the rest of the cleaning process and when running the regression. From here, I created six binary variables using ifelse statements and created new columns for the specified year. The first column would return a 1 if the Release date is equal to 2012. The second column would return a 1 if release data is equal to 2013. The third column would return a 1 if the movie release is equal to 2014. The fourth column would return a 1 if the release date is equal to 2015. The fifth column would return a 1 if the release date is equal to 2017. For the Genre variable, I will need to create a new column containing observations for specific genres included in the data set. Some genres observed in the data set include, "Comedy", "Thriller", "Drama", "SciFi", "Action", and "Adventure".

I will be selecting six genres to use in this analysis; given that this dataset consists of all horror films creating new columns from three non-horror genres will not pose an issue as all films are listed as in the horror genre. I will select the "Horror", "SciFi", "Comedy", "Action", "Adventure", "Drama", and "Thriller" genres and create a new column using the "grepel" function in R; which will return a value of 1 if the film is listed as either one of the three genres or 0 otherwise. Lastly, I have created a new column for the budget

variable removing all non-US dollar observations from the column. Despite creating a new data frame for US only films, the filming location could still be in another country such as Hong Kong, India or the Philippines resulting in a Budget listing in the countries native currency. To limit the observations in the Budget variable to US only observations I will be using the "gsub' function again to remove the "." and the dollar sign for budget observations in dollars. Preventing any NA values from occurring in the dataset.

Next, I would create two new columns using the observations under the budget variable using ifelse statements. One column would return a value of 1 if Budget1 is greater than or equal to 1,000,000. The last column would return a 1 if Budget1 is less than 1,000,000. The last adjustment to the data I made was to create a new data frame containing only the variables i'm using in this analysis. This was done for greater and organization and ease when creating the summary statistics before moving forward.

For this analysis, I will be running a standard regression with the variable "Review.Rating" as my dependent variable and the independent variables for this analysis are the genres Horror, Comedy, Scifi, Thriller, Adventure, and Action. The budget variables "Budget great equal mil", and "Budget less mil". The release year variables, year2013, year2014, year2015, year2016, and year2017. The run time variables, more equal 60, less60. In this regression I have left out the variables Horror, year2012, less60, and "Budget less mil", out of the regression, using them as reference groups. I would expect to see to what degree the variables included in the regression impact a movies review rating. Initially, I would expect the variable "Budget greater equal mil" and perhaps the variable "more equal 60" to have a larger impact on a films review rating then the other variables listed. I would infer "Budget greater equal mil" to have a larger impact than say the run time variable, as a larger budget would be able to encompass other factors such as, specific directors, cast members, writers, and other crew members for the film. This would also include filming location and expensive props, if the movie requires it. Improving the quality of the film which could lead to a more successful and higher rated film. Although I would not say that the budget variable is causally related per say; but there may be a high correlation between a larger budget and higher movie reviews.

The possibility of a low budget film receiving a higher rating is possible and can be seen with films such as Night of the Living Dead (1968) Evil Dead (1981) and the original Halloween (1978) being two well-known examples in the horror Genre. In the case of the run time variable, the variable may have a significant effect on Review.Rating, although I'd expect it to level out and then decrease after a certain length of time with diminishing marginal returns with every additional minute added to a flims' run time. For the other variables, there may be some interesting findings for the specific genres or years. Although for the genre variables, the impact each genre would have on Review.Rating I would expect to be minor as certain genres, like Scifi, may appeal to a very niche audience especially when combined with elements of horror. Unless perhaps certain subgenres became quite popular during certain years. This could mean that

certain genre variables may be correlated with the year variables. Making use of the vif function will be useful when checking for multicollinearity. Listed below are the summary statistics for the final data frame containing the variables specific to this analysis.

2.1 Summary Statistics

Table 1: Table 1: horrorus final Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Drama	2,092	0.155	0.362	0	0	0	1
Comedy	2,092	0.153	0.360	0	0	0	1
Scifi	2,092	0.097	0.296	0	0	0	1
Thriller	2,092	0.438	0.496	0	0	1	1
Adventure	2,092	0.031	0.174	0	0	0	1
Action	2,092	0.100	0.301	0	0	0	1
Budget_great_equal_mil	739	0.388	0.488	0.000	0.000	1.000	1.000
Budget_less_mil	739	0.612	0.488	0.000	0.000	1.000	1.000
year2013	2,092	0.126	0.332	0	0	0	1
year2014	2,092	0.152	0.359	0	0	0	1
year2015	2,092	0.181	0.385	0	0	0	1
year2016	2,092	0.200	0.400	0	0	0	1
Horror	2,092	1.000	0.022	0	1	1	1
year2012	2,092	0.089	0.285	0	0	0	1
year2017	2,092	0.251	0.434	0	0	1	1
more_equal_60	1,773	0.987	0.113	0.000	1.000	1.000	1.000
less60	1,773	0.013	0.113	0.000	0.000	0.000	1.000
Review.Rating	1,915	5.019	1.489	1.000	4.000	6.000	9.800

3 Regression

For the analysis, I ran a traditional multivariable linear regression, with Review.Rating as my dependent variable and the created variables for the subgenres, the variables for budgets equal to and greater than 1,000,000 dollars, the variables for years 2013-2017, and the variables for run times equal to and greater than 60 minutes. Constructing a linear model such as; Review.Rating = Drama + Comedy + Scifi + Action + Thriller + Adventure + Budget great equal mil + year2013 + year2014 + year2015 + year2016 + year2017 + more equal 60. Once I ran the regression of my model in Rstudio, I used the stargazer package to tabulate my results. Below is the table for the results of my regression.

3.1 Regression Results

Table 2: Table 1: ols reg horrorus Results

	Dependent variable:
	Review.Rating
Drama	0.149
	(0.171)
Comedy	0.246
	(0.170)
Scifi	0.123
	(0.217)
Action	0.063
	(0.199)
Thriller	0.203
	(0.132)
Adventure	0.115
	(0.287)
Budget_great_equal_mil	0.112
	(0.128)
year2013	0.053
	(0.231)
year2014	−0.026
	(0.229)
year2015	0.260
	(0.226)
year2016	0.512**
	(0.225)
year2017	0.660***
	(0.242)
more_equal_60	−1.196**
	(0.540)
Constant	5.726***
	(0.553)
Observations	592
R^2	0.048
Adjusted R^2	0.027
Residual Std. Error	1.494 (df = 578)
F Statistic	2.263*** (df = 13; 578)
Note:	*p<0.1; **p<0.05; ***p<0.01

Now that I have the results, I can rewrite the linear model as such. Review.Rating = 5.726 + 0.149(Drama) + 0.246(Comedy) + 0.123(Scifi) + 0.063(Action) + 0.203(Thriller) + 0.115(Adventure) + 0.112(Budget great equal mil) + 0.053(year2013) + (-0.026)(year2014) + 0.260(year2015) + 0.512**(year2016) + 0.660***(year2017) + (-1.196**)(more equal 60) With the new linear model, I can begin to interpret the results from our regression.

For the variable Drama, if a film is under the category of Drama then this is associated with a 0.149 increase in the film's review rating. For Comedy, if a film is a Comedy then it is associated with a .246 increase in the films review rating. For Scifi, if a film is in the Scifi genre then it is associated with a .123 increase in the films review rating. For Action, if a film is in the Action genre then it is associated with a .063 increase in the films review rating. For Thriller, if a film is in the Thriller genre then it is associated with a .203 increase in the films review rating. For Adventure, if a film is in the Adventure genre then it is associated with a .115 increase in the films review rating. For Budget great equal mil, if a film has a budget greater than or equal to 1,000,000 dollars then it is associated with a .112 increase in the film's review rating. For year2013, if a film was released in the year 2013 then this is associated with a .053 increase in a films review rating. For year2014, if a film is released in the year 2014 then it is associated with a -.026 decrease in a films review rating.

For year2015, if the film was released in the year 2015 then it is associated with a .260 increase in a films review rating. For year2016, if a film was released in the year 2016 then it is associated with a .512 increase in a films review rating and is statistically significant at a p value of 0.05. So the null hypothesis, that year2016 has no impact on the review rating, is rejected. For year2017, if the film was released in the year 2017 then it is associated with a .660 increase in a films review rating and is statistically significant at a p value of 0.01. So the null hypothesis, that year2017 has no impact on a films' review rating, is rejected as well. For the variable more equal 60, If a film's runtime is greater than or equal to 60 minutes then this is associated with a -1.196 decrease in a film's review rating and is statistically significant at a p value of 0.01. So the null hypothesis of more equal 60 having no impact on a films review rating is rejected. To check for signs of multicollinearity I check the variance inflation factor using the vif function in R studio. None of the variables included in the regression had a variance inflation factor of 10. The highest vif factor was 2.10 from the year2016 variable. Below are charts from the initial regression for Residuals vs Fitted, Normal Q-Q, Scale Location, and the Residuals vs Leverage

3.2 Regression Plots

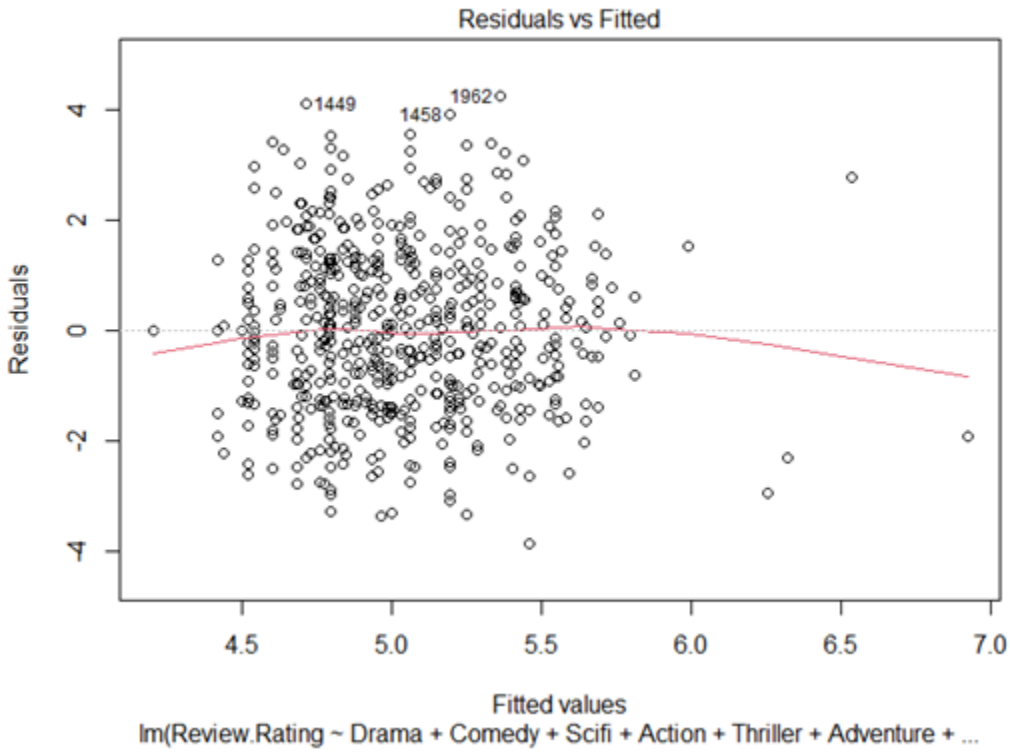

Figure 1: Review Rating Residual Plot

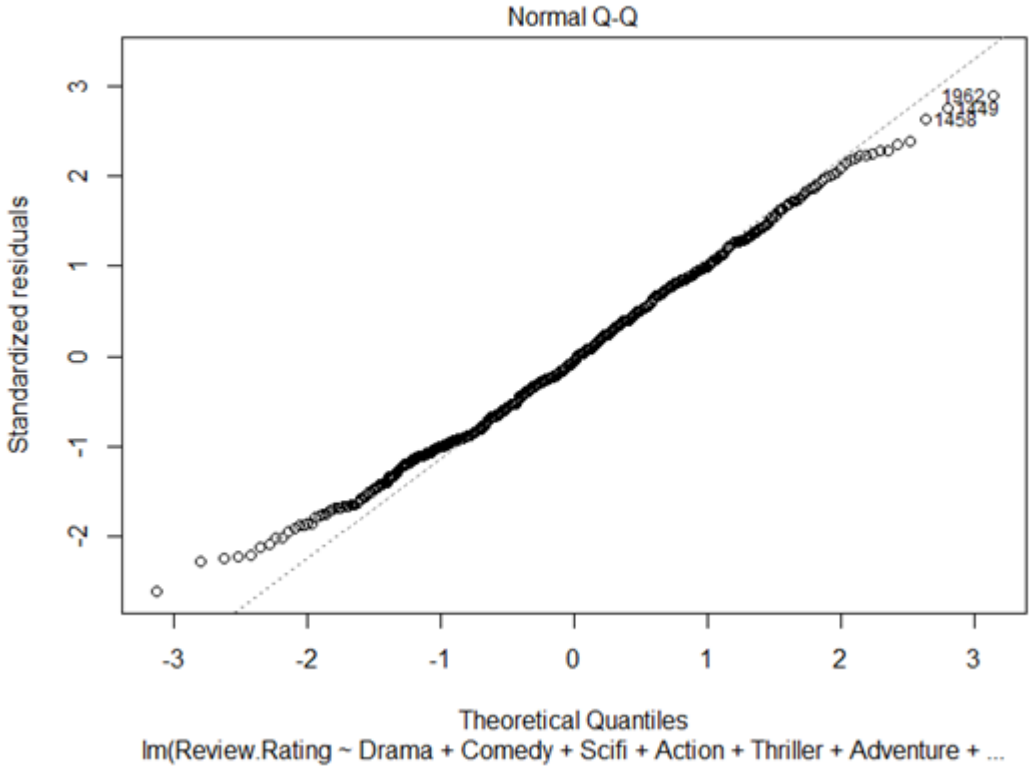

Figure 2: Review Rating Normal Q-Q

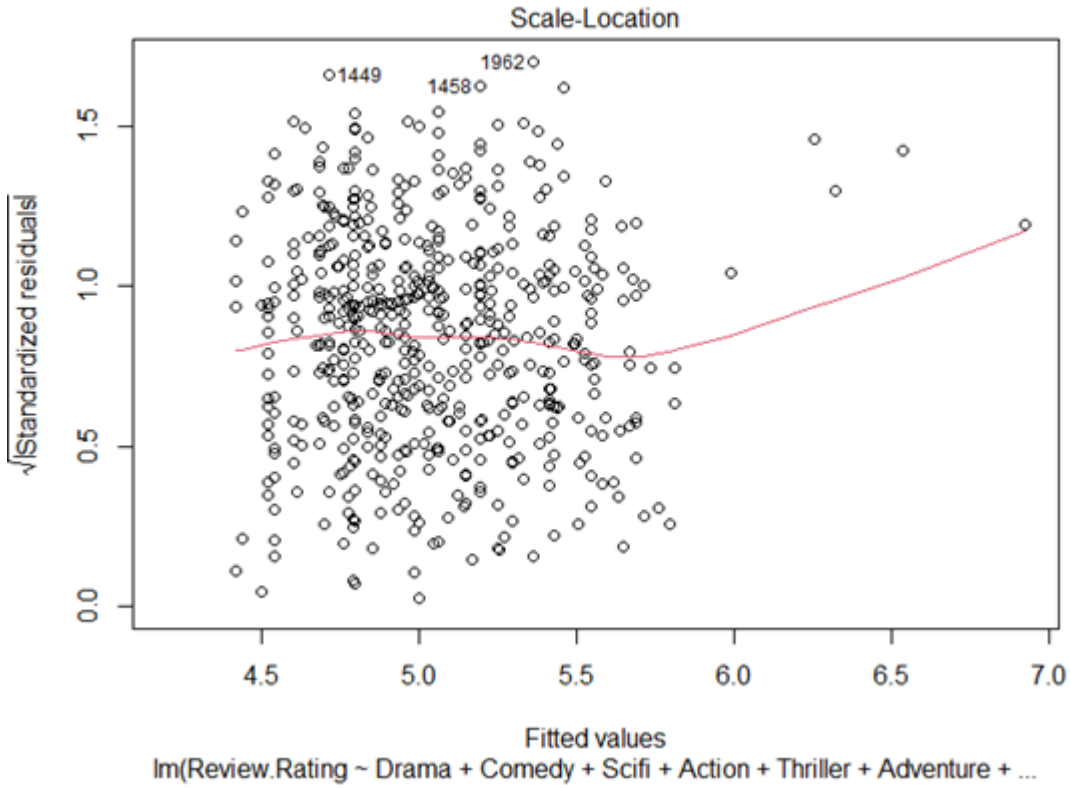

Figure 3: Review Rating Scale Location

In the regression results, the statistically significant variables were year2016, year2017 and more60. For each of these variables I created Scatterplots to display the relationship between each variable and a films review rating below.

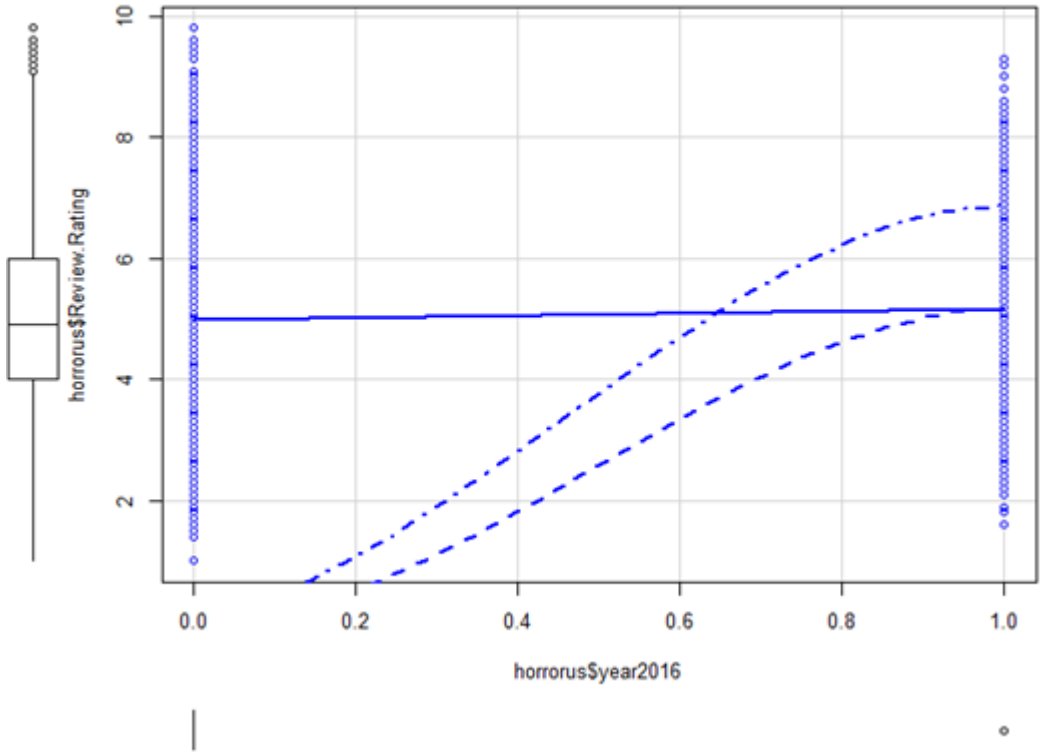

Figure 4: 2016 Review Rating

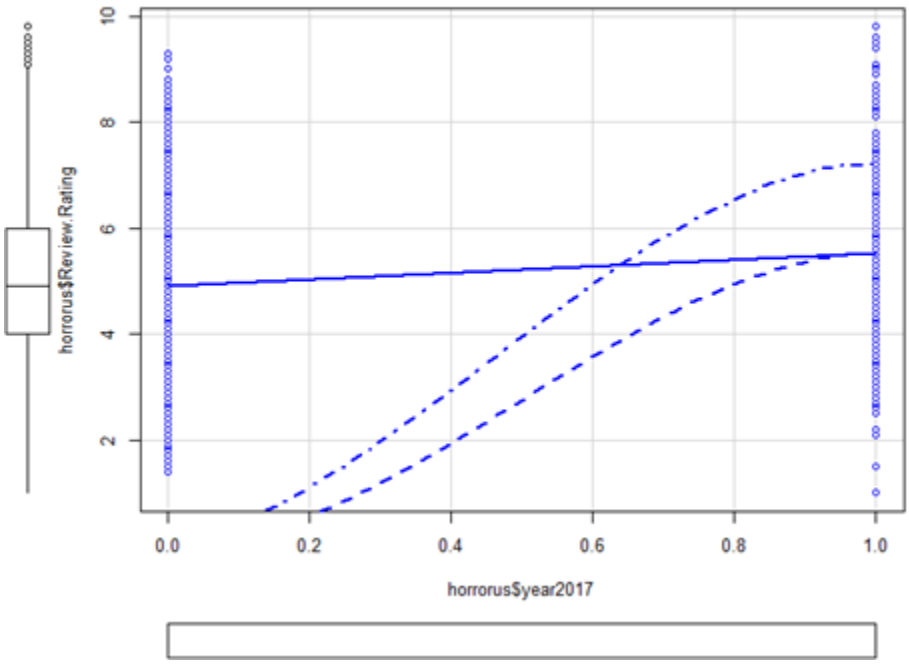

Figure 5: 2017 Review Rating

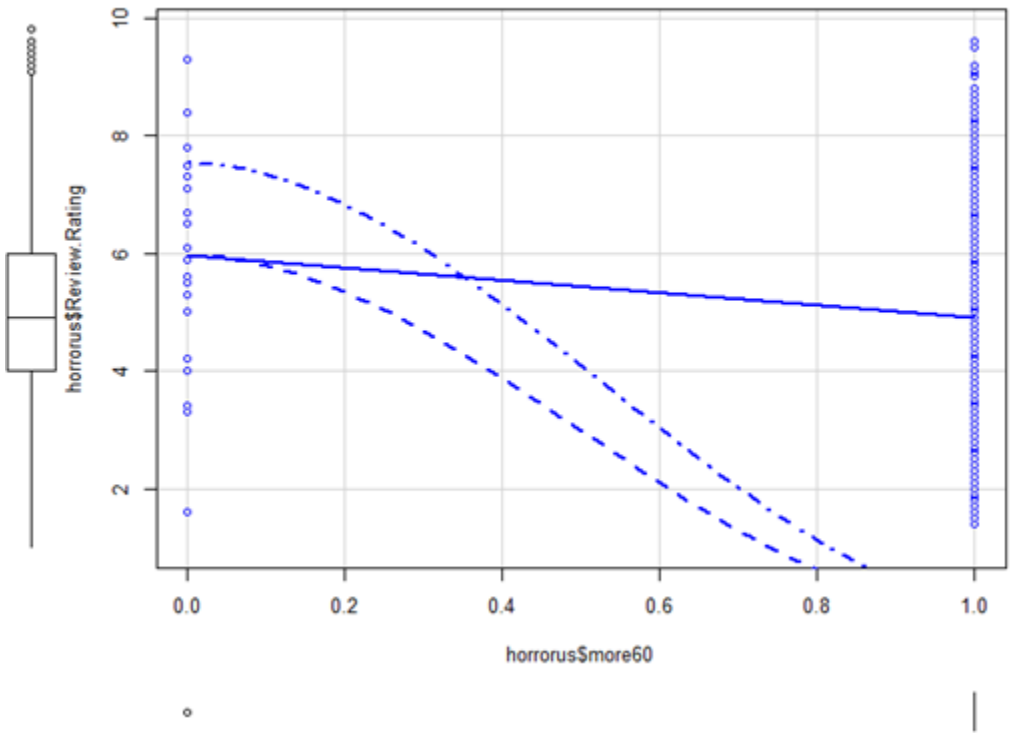

Figure 6: Budget Greater Then 60 Million

The results of this analysis were rather surprising in some areas and returned results that were expected. The more equal 60 variable returned a negative correlation with review rating. This was expected likely due to diminishing marginal returns as the length of a film increases. Although what came as a surprise was that the results for this variable was statistically significant. Despite that not being my initial expected outcome, the statistical significance does make sense. As with each additional minute added to a films' run time the amount of entertainment a person could receive would decrease. Displaying a closer relationship between run time and review rating than what would other wise be explained by mere chance.

Audiences may be less willing to sit through films with a run time of over 90 minutes. In certain cases, maybe but, not very frequently. Next, the budget more equal mil variable returned an associated .112 increase in the review rating but was not statistically significant. This was not initially expected as I anticipated that given a higher budget this would improve production value and increase the chances of a higher rating. Perhaps it is not statistically significant given the creative aspect that is tied to making a horror film. Even if a film possesses a lower budget if the film is creative and well made it can still be quite successful. We can see this in the three films I mentioned previously, Night of the living dead, Evil Dead, and Halloween.

A film with a large budget can flop as well especially if the project was deemed risky for the film studio to begin with but still took a risk. Some risk factors may be an inexperienced director managing a big budget film, studio interference during production, and a poorly written script being a few examples. The next two variables, year2016 and year2017, both return statistically significant results to my surprise. The statistical significance of these two variables is hard to say but perhaps this might be explained by popular horror film franchises being released during both years. The Conjuring and Annabelle franchise comes to mind as examples of well-known horror franchises that had films released during both years. It's also possible that both years may have been simply successful years for horror films in general. The subgenre variables did not return statistically significant values and neither did the other three year variables. The results from this analysis could be useful for determining the factors for a horror film with a high review rating. Though there may still be some other factors that can go into a movie review rating that may not be included in my initial regression. At the very least the work from this analysis can be built upon for exploring other factors not considered in this analysis.

4 Conclusion

In the end, the factors that go into a horror films review rating vary in several ways. Though looking at specific variables can provide a sufficient gauge for determining what aspects impact a horror film's review

rating. Some improvements to this analysis are to look at not just the release year but the release month and see how the review rating for films was impacted on a more seasonal level. The data set as well possessed some issues due to missing values which may have impacted the analysis but, with added data these issues can be easily addressed. This analysis despite it's flaws has laid a foundation down which can easily be built on to when new data is added. Providing a more clearer picture on which determinants impact a films review rating.

5 Citations

PromptCloud. "IMDB Horror Movie Dataset [2012 Onwards]," October 31, 2017.
https://www.kaggle.com/PromptCloudHQ/imdb-horror-movie-dataset.