To: Bob Ross of Janzen Consulting Group

From: Ali Namin

Date: April 23, 2019

Subject: Do the budget, duration, IMDB score, and number of users who voted on IMDB for a movie influence gross profit?

Project 4

**Introduction:**

The movie industry is massive within the US, therefore those in charge of its financials must ensure that they are getting their money’s worth. Nowadays, large companies must analyze if their movies budget is going to be profitable or loss-making, which influences what movies end up being made. Using a multiple linear regression model, we can see if multiple factors have an influence on a specific outcome, in this case gross profits of a movie.

**Research:**

The target population is all movies within the dataset. My hypothesis is that the variables budget, duration, IMDB score, and number of users who voted on IMDB will have a positive effect on a movies gross profit. This analysis will provide useful information for New Line Cinemas in regard to future marketing and business tactics. After conducting the analysis, I can conclude that my hypothesis was correct, and these factors do in fact have a positive correlation on gross movie profits.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Quantitative Summary** | | | | | | |
|  | Freq | Min | Max | Med. | Mean | SD |
| Movie Budget | 3,509 | 218 | 300,000,000 | 20,000,000 | 35,755,986 | 42,688,841 |
| Duration | 3,801 | 7 | 330 | 103 | 106.4 | 23.7 |
| IMDB Score | 3,807 | 1.6 | 9.3 | 6.5 | 6.4 | 1.1 |
| Number of Voters | 3,807 | 5 | 1,689,764 | 39,786 | 91,053 | 147,270 |
| Gross sales | 3,235 | 703 | 760,505,847 | 32,178,777 | 55,214,607 | 71,733,124 |

For my research, I decided to use the IMDB Movie Data specifically for movies within the United States in order to narrow down our dataset. For the independent variable, I chose budget, duration, IMDB score, and number of users who voted on IMDB for a movie which are interval/ratio values. Budget is included in the IV’s because it is a major factor in the production quality of a movie with a sample size (n) of 3,509, min. of $218, max of $300,000,000, mean of

20,000,000, and a std. dev of 42,688,841. Duration of a movie can determine if a movie is too long, too short, or just the right length. In minutes, duration had a sample size (n) of 3801, min. of 7, max of 330, mean of 103, and a std. dev. of 23.7. IMDB score reflects a user’s feedback to a movie with a sample size (n) of 3,807, min. of 1.6, max of 9.3, mean of 6.4, and a std. dev. of 1.1. The number of users who voted on IMDB for a movie are taken into account when determining how accurate those scores are reflected a sample size (n) of 3,807, min. of 5, max of 1,689,764, mean of 91,053 and a std. dev. of 147,270. The dependent variable I chose was gross sales which is also interval/ratio because it is also a dollar value with a sample size (n) of 3,235, min. of $703, max of $760,505,847, mean of 55,214,607 and a std. dev. of 71,733,124. Sample size (n) was determined by subtracting NA’s from original sample size. Summary statistics for all variables can be found in Table 1.

Based off these summary statistics, we can determine that gross profit is the only variable that is significant so one must be cautious when considering if the other variables are significant. From the summary stats for our subset, we can see that there are a different number of missing values per IV, and we got rid of these by encoding our values in R to drop any missing values and exclude them from potential calculations. I also took a look at our subsets by looking at our assumptions table and finding that although there are some outliers, none of them are significant outliers. There is also a gap in our dataset as we cannot assume that our results extend to the population (every movie) because I am analyzing only American movies.

**Tests and Results:**

I chose an alpha level of 0.05 to accommodate any possible errors in the dataset.

I completed the multiple OLS regression analysis with the regression equation of Ŷ= 7,072,000 + 0.7828 (budget) -114,000 (duration) + 1,417,000 (IMDBscore) +207.3(voters). Based on our assumptions table, we can see that there is a linear relationship between the independent and dependent variables and that errors between observed and predicted values should be normally distributed. We can also assume that there is no multicollinearity in the data which would cause our independent variables to be too closely related.

The regression coefficients for budget (.7828), duration (-114,000), IMDB score (1,417,000), number of voters (207.3). The resulting p-value for my model is so small that it is negligible and means that my chosen alpha of 0.05 is greater than our p-value, so my model is statistically significant (p<0.05). From the given regression coefficients, we can conclude that for every one unit increase in budget (USD), we can see a $0.7828 increase in gross sales (p<0.05). For every one unit increase in duration (minutes), we can see a $114,000 decrease in gross sales (p<0.05). For every one unit increase in IMDB score, we can see a $1,417,000 increase in gross sales (p<0.05). For every one unit increase in number of voters, we can see a $207.3 increase in gross sales.

Our analysis also provided us with an adjusted r-squared value of 0.5869 or 58.7% of variance in the dependent variables can be explained by the independent variables, which shows that our sample has moderate substantive significance. This substantive significance translates to the IV’s regression coefficients being meaningful.

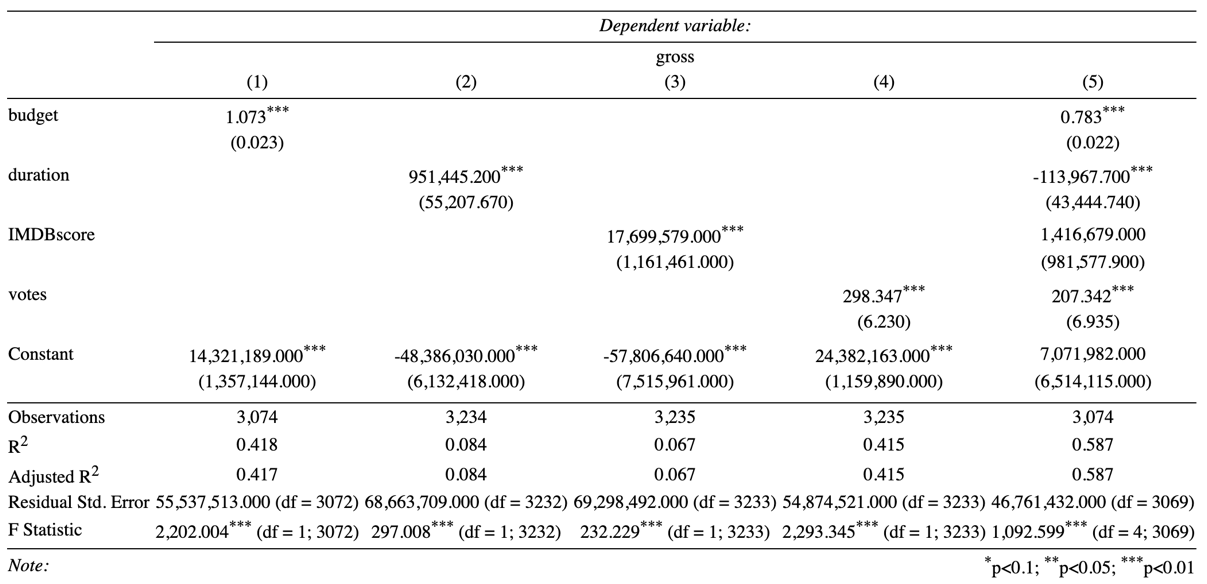
**Conclusion:**

Based off of my analysis, I can determine that there is a positive correlation between our dependent variable and independent variables and that budget, duration, IMDB scores, and number of IMDB voters is significant and substantive. Therefore, New Line Cinema should promote more movies with larger budgets, high IMDB scores and high IMDB voters.

When conducting our analysis of the IMDB Movies dataset, I had to consider some of the weaknesses and limitations of the dataset and possibly our analysis. We had to look out for missing values that could throw off our data and be interpreted as a 0 instead of an NA value that can be discarded.

An additional weakness of the dataset were outliers did have an impact on our analysis, however it wasn’t significant enough for us to take it out of the dataset entirely instead of having to scrub the from our subset. This decision could have skewed our data in the direction

Stargazer:



Assumptions:

