**Date:** May 14, 2019

**To:** Shawn Janzen, Founder, Janzen Consulting Group

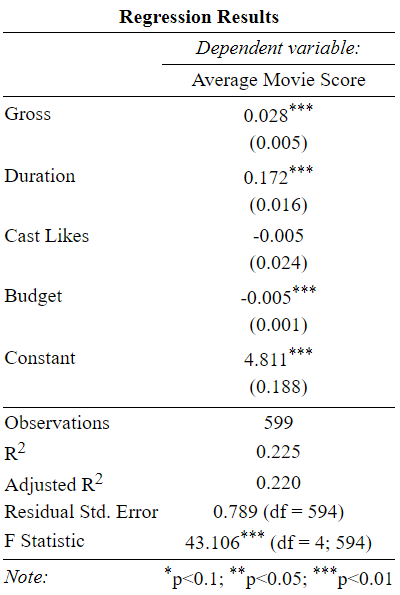
**From:** Andrew Carroll, Janzen Consulting Group

**RE:** Multiple Linear Regression Analysis of Movie’s IMDb Score

For a movie studio to be successful, it is important that they pick and choose movies that they feel will be popular with audiences and do well at the box office. It is critical that these movie studios understand how certain factors will affect the public’s perception of the movie and in turn, the movie’s ratings. For our analysis, the target population is all movies that have been made. I hypothesized that a movie’s budget, gross, duration, and cast Facebook likes does not necessarily indicate that a movie will have a better IMDB rating. During my testing, I found that this may indeed be the case.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Quantitative Summary of Model Variables** | | | | | | |
|  | Freq | Min | Max | Med. | Mean | SD |
| IMDb Score | 599 | 3.40 | 9.20 | 6.70 | 6.74 | 0.89 |
| Budget (in millions of USD) | 599 | 0.04 | 300.00 | 40.00 | 61.22 | 55.93 |
| Gross (in tens of millions of USD) | 599 | 0.01 | 53.33 | 5.54 | 8.48 | 8.85 |
| Cast Likes (in tens of thousands of likes) | 599 | 2.00 | 9.25 | 2.98 | 3.49 | 1.43 |
| Duration (in tens of minutes) | 599 | 8.10 | 24.00 | 11.30 | 11.65 | 2.22 |

In order to find an answer for my hypothesis, I used a sample from the IMDB movie data set. In this sample, there were 599 independent observations of movies that were used in my model. For this test, my independent variables were the budget, gross, duration, and cast Facebook likes. The budget variable describes the amount of money that the studio spent on the film in USD. This variable had to be recoded because some of the values were not in USD. Since converting all the different currencies would take far too long to do, in this analysis I only included movies that were made in the USA. The gross variable describes the total amount of money that a movie made. This was recoded to only include movies that made at least $100,000 and no more than 600,000,000 in order to normalize the variable’s distribution. The cast likes variable measures the combined number of likes each member of the cast has on Facebook. The variable was recoded to have at least 20,000 total cast likes and no more than 100,000 to remove any influential outliers in the data set. The duration variable measures the length of a movie. Movies shorter than 80 minutes were removed since they are not considered feature length films by the Screen Actors Guild. One movie longer than 240 minutes was removed since it was an influential outlier. Finally, the IMDb Score variable rates movies on a score from one to ten. While IMDb has not officially explained how their score is calculated, we do know that it is the weighted average review score of all user submitted reviews.

Before beginning my test, I checked the assumptions and diagnostic plots for a Multiple OLS Regression. Originally, there were many outliers in my data. Upon further investigation, some of the movie’s budgets appeared to be incorrectly entered the dataset and had budgets of over $1 billion. Therefore, I capped the budget at $400 million since that is how much *Pirates of the Caribbean: On Stranger Tides,* the most expensive movie ever made, cost to make. After this recode and my aforementioned recodes, there were no more outliers in my data. Finally, the only assumptions and diagnostics my model may have not passed were having normally distributed errors. However, my QQ plot was close enough to being normal that I felt comfortable enough to not do anymore recoding to my variables and continue my test. In order to test my hypothesis, I conducted a Multiple OLS Regression test on all the variables in my model at an alpha level of .05. My entire model was statistically significant (p < .05) and the only variable in my model that was not statistically significant was the cast likes variable (p = 0.84). That being said, the budget, gross, and duration were all statistically significant. We are able to reject the null hypothesis at an alpha level of p = 0.05, as there are some statistically significant effects on IMDb score from the independent variables. That being said, I am not comfortable generalizing my results to the entire population since the model only includes 599 movies after it has been heavily recoded.

My model’s equation output was = 4.811 - 0.005(Budget) + 0.028(Gross) – 0.005(Cast Likes) + 0.172(Duration). For each additional $1,000,000 in the budget variable, the IMDb score will decrease by 0.005. Each additional $10,000,000 in the gross variable will result in the IMDb score increasing by 0.028. An additional 10,000 likes in the cast likes variable will decrease the IMDb score by 0.005. And finally, for each additional 10 minutes of film length in the duration variable, the IMDb score will increase by 0.172. I believe that the only variable coefficients that are not meaningful are the budget and cast likes variable coefficients since they are incredibly small and a movie would need to have a huge budget or a lot of likes for those variables to have any substantive effect on the score. Additionally my model moderately explained the proportion of variance in IMDb score, R2 =.22, F(4, 594) = 43.11 , p < .05.

Despite my results, my study still had its’ limitations. One such is that I was limited to only using movies made in the United States. While most movies are made in the United States, foreign films were unrepresented in this model. Had foreign films been included, we may have gotten significantly different results than those that were outputted.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generatedA close up of a map

Description automatically generatedA close up of a map

Description automatically generated