Analysis of rating scores on IMDb.

The focus of this project was the analysis of the movie ratings given by users of a movie database IMDb. Hypothesis tests were performed to discover the types of relationships between variables “gross”, “budget”, “votes”, “runtime” and the variable “score”.

The relationship of votes and score, runtime and score were found to be close to moderately significant positive linear. The relationships of gross and score, budget and score appeared to be very weak positive linear. Scatterplots and Least Squares Fit were used to support the analysis.

Hypothesis of scores of horror movies being rated lower vs other genres of movies was tested with the help of CDFs, PMFs, and a chi-square test. The difference between the scores of these two variables was concluded to be statistically significant, with PMFs and CDFs illustrating the lower scores for horror movies genres being more prevalent than those of the other genres.

Creating histograms that would look good was quite a challenge. I couldn’t get a few of the histograms to show the graph so I ended up using matplotlib.pyplot as a workaround.

I experienced some difficulty understanding the type of relationship between two variables (budget and score) due to Pearson’s and Spearman’s coefficients resulting in values with opposite signs. I believe the relationship is very weak positive linear which coincides with the results of Pearson’s correlation which was found to be statistically significant with p-value less than .05. The scatterplot, Least Squares Fit and simple linear regression model provided support for the choice of the positive value of the correlation coefficient. Squaring the value in the model resulted in a negative number of R squared which indicated that the model with a polynomial relationship would not fir the data well, thus making the choice more in favor of the linear relationship.

The warning of possible multicollinearity appeared for all of the models except for one. I had to investigate which variables could possibly cause that because multicollinearity can make interpreting results problematic. The analysis performed was problematic because of the changes in the sign of coefficients of the variables that have higher VIF, and have correlation close to high. Decision to interpret results for the models with the variables staying stable overall was made. I am still not fully confident it was the right call but I was able to get a model that didn’t include variables that have a moderate or high correlation between each other, and variables that didn’t show a stable behavior.

The project has some limitations. Even though it provides results for effects of variables budget, gross, runtime and votes, it doesn’t explain other factors that could be influencing the outcome “scores”. Considering performing the analyzes of other variables from the data set can allow for discovery of additional factors influencing the outcome variable “score”. I believe that variables like “company” (the production company), “actors” and “director” could potentially have some effect on the rating scores of movies and provide more information about the factors influencing the rating.

Missing values significantly reduced the amount of observations available thus reducing the possibility of accounting for all of the information for the variables. Looking for the sources containing the information that was missing could potentially solve this issue.

Additional analysis of all of the assumptions that multiple linear regression model should meet for valid results can be considered to double check the validity of the model and its results. I am not sure I did enough and might have missed something since I only looked at the assumption of no multicollinearity.

The time frame for the movies used in the data set is 1986-2016. Adding more recent information can provide results that are more current.

Overall, the project let me explore various concepts of statistical analysis and learn even more by trying to overcome the challenges of the analysis.

References

Kaggle. *Movies.* Retrieved from <https://www.kaggle.com/danielgrijalvas/movies>

Minitab. *Enough Is Enough! Handling Multicollinearity in Regression Analysis.* Retrieved from<https://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis>

Frost J., *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions.* Retrieved from <https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

Downey, A.B. October 2014. Think Stats.