Linear Regression Analysis on Video Game Movie Review and Sales Data

Aaron Hum

1. Introduction to Problems

Video game movies represent to a lot of consumers a “so-bad-it’s-good” experience. Meaning that even though the movies are of bad quality and generally disliked critically, this actually drives people to the theatres to see the movie. To test this theory, I have a dataset of 30 years worth of video game movies(42 total entries) listing their title, release date, worldwide box office sales (in US dollars), Rotten Tomatoes score (percentage), Metacritic score (out of 100), film distributor, and publisher. I will test the relationship between the review site scores (Metacritic and Rotten Tomatoes) to box office sales.

Data Source: https://www.kaggle.com/bcruise/film-adaptations-of-video-games

2. Models and Numerical Methods

Models used:

Linear Regression, Multiple Linear Regression

Numerical methods:

The equation will measure Y, the box office sales of a movie in US dollars, in the equation:

Y = box office sales of a movie in US dollars

X_1= Metacritic Score, X_2= Rotten Tomato Score

B_0 = Intercept, B_n = Slope of the regression line for each variable, e = error

$Y = B_0 + B_1X_1 + B_2X_2 + e$

2b. Python Libraries I Plan To Use

---

numpy: Numerical computations library in Python

pandas: Data manipulation in Python

matplotlib: Data visualization in Python

seaborn: Statistical data visualization in Python

2c. More Variables / Models

In the original model, I plan to just use the numerical data. However, we also get several non-numerical data in this dataset including the movie distributors (Sony Pictures, 20th Century Fox, etc.) and original game publishers (Capcom, Nintendo, etc.). I believe separating the data by these variables could yield interesting results and comparisons to each other.

For example, I expect movies made about Nintendo games to generate more sales than movies about Electronic Arts games, or movies distributed by Disney, may have higher box office sales than Sony.

I’m curious to see what effect on the models this data might have.

3. Expectations

I expect Metacritic scores and Rotten Tomatoes scores will have a negative correlation to box office sales, due to my hypothesised “so-bad-it’s-good” effect.

I expect the model will have a lot of error, due to the wide variance among video game movie box office sales and scores, and also, due to the relatively small sample size (42), there could be a lot of error as well.