Box Office Success V.S. Release Date Analysis

Statistical Data Analysis

Xi He

M.S. Engineering Science and Applied Math

McCormick School of Engineering, Northwestern University

December 1st, 2017

Introduction

Although numerous different factors may influence the box office of a movie, this project concentrates on studying the relationship between a movie's box office and its release time. More specifically, this project relates a sequel's gross to the time gap between the sequel and its precedent. This study would focus on Marvel Cinematic Universe movies in order to minimize other features' influences. The goal of this project is to determine the time gap that makes a successful sequel, and to build a probability density function that predicts a sequel's success rate, given information of the previous movie.

Method

A. Data Collection

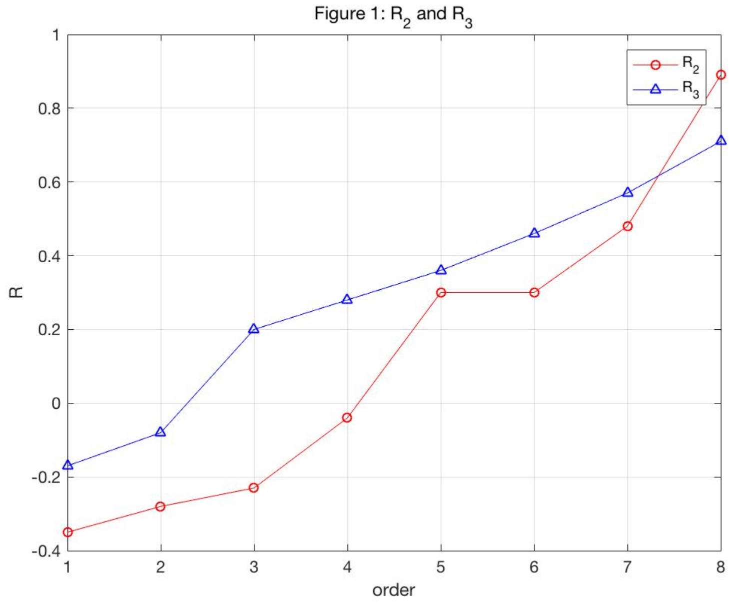
Eight Marvel Comics movie series, including 30 movies, have been analyzed in this project. Data has been collected from IMDb's Box Office Mojo online database. The dataset includes each movie's title, production studio, total gross (*G*), opening weekend gross (*g*), and release date.

B. Variable Creation

In order to rule out the influences from features such as ratings and reviews, opening weekend gross is chosen as the main variable instead of total gross. A sequel's success rate (R) is defined as: $R = \frac{g_{sequel} - g_{previous}}{g_{previoius}}$ to quantify how successful a sequel is compared to the previous movie. If f R is positive, the sequel is marked as a "success"; otherwise, it is marked as a "failure". Gaps (*Y*) between sequels and their precedents in the series are drawn from the differences of release years, without the exact dates for the sake of simplicity. Consequently, there exists 2 one-year, 8 two-year, 9 three-year, 1 four-year, and 2 five-year gaps. Considering the sample size, only sequels with two-year and three-year gaps will be analyzed. The movie series *Guardian of the Galaxy* with a three-year-gap sequel has been selected randomly as a testing sample.

C. Analysis and Results

Taking two subsets of success rate: $\{R_2 \in R : Y = 2\}$ and $\{R_3 \in R : Y = 3\}$ for 2-year gap and 3-year gap sequels. Each subset is ranked in ascending order and plotted in the same figure:



According to Figure 1, it is clear that given the same order number, every single point in R_3 is greater than point in R_2 except for the last one. It is easy to calculate from these 16 samples that:

$$P(sucess|Y=2)=0.5,\ P(sucess|Y=3)=0.75$$

Thus, it is reasonable to make a hypothesis that the length of gap affects the success rate of sequel. Success-failure result forms a binomial distribution. A significance test is used to check whether R_2 and R_3 come from the same distribution.

1. Significance Test

First, making null hypothesis that H_0: R_2 and R_3 come from the same distribution. The random variate $Z \approx 1.03279$ can be found by calculation. If the null hypothesis were correct, it should be normally distributed with $\mu = 0, \sigma = 1$. Since $Z > \sigma$, the null hypothesis is rejected to a 68% confidence level.

2. Sign Test

Next, test the success rates of R_2 and R_3 separately. Since it is too vague to assume that the success rate forms a Gaussian distribution, non-parametric statistics method is used without making any assumption about the data distribution. Sign Test would be the best choice to see whether the ratio of getting success is the same as getting fail.

For sequels with a 2-year gap, null hypothesis is that H_0: success and failure are at the same rate. Marking success as (+) and failure as (-), the following table is produced:

R_2	-0.35	-0.28	-0.23	-0.04	0.3	0.3	0.48	0.89
Success	-	-	-	-	+	+	+	+

The parameter is $sum = \sum_{i=0}^{4} P(i,8) \approx 0.6367$. At the 68% confidence level, the null hypothesis is accepted.

For sequels with a 3-year gap, making the same null hypothesis, the following table is produced:

R_3	-0.17	-0.08	0.2	0.28	0.36	0.46	0.57	0.71
Success	-	-	+	+	+	+	+	+

The parameter $sum = \sum_{i=0}^{6} P(i,8) \approx 0.9649$. At a 68% confidence level, the null hypothesis is rejected. In fact, even at a 99% confidence level, the null hypothesis is still rejected.

As the result, it is reasonable to say that 2-year gap and 3-year gap do make a difference in a sequel's success rate. A sequel with 3-year gap has more chance to succeed than the one with 2-year gap.

One goal of this study is to make predictions about success rates given information of the previous movie. Looking at the data, it seems that a sequel is more likely to fail if its precedent in the series has a high total gross (*G*) roughly above $250,000,000. Making a hypothesis saying that a sequel's success rate is related to its previous movie's total gross $(G_2, R_2), (G_3, R_3)$ are plotted in Figure 2:

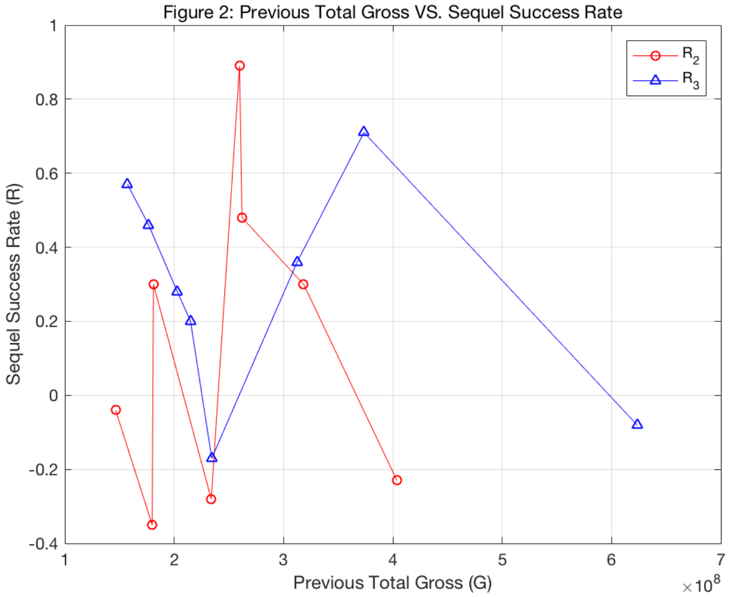


Figure 2: Previous Total Gross VS. Sequel Success Rate

It is hard to find a clear relation between the total gross and the success rate from Figure 2. The correlation coefficient method is needed for further analysis.

3. Correlation Coefficient

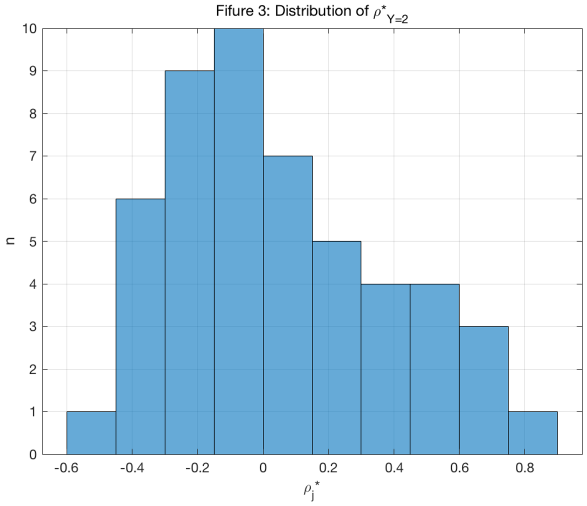
For the 2-year gap, the correlation coefficient is $\rho_{Y=2} = \frac{\sigma^2_{R_2G_2}}{\sigma_{R_2}\sigma_{G_2}} \approx 0.0486$

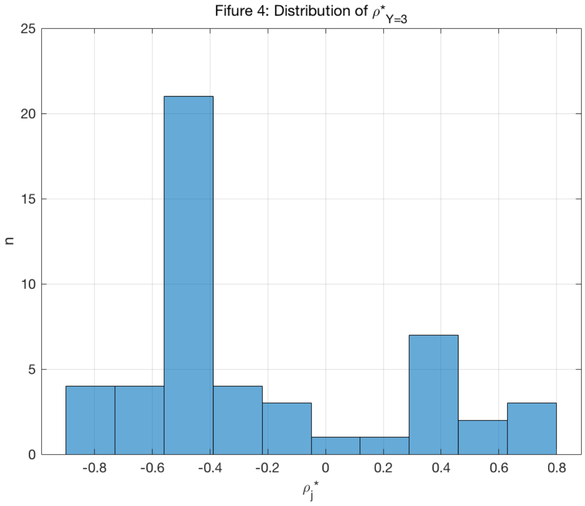
For the 3-year gap, the correlation coefficient is $\rho_{Y=3} = \frac{\sigma^2_{R_3G_3}}{\sigma_{R_3}\sigma_{G_3}} \approx -0.2943$

ρ is bounded by $-1 \leq \rho \leq 1$ and $\rho \approx 0$ indicates weak or no correlation. The Bootstrap Method is then used to determine the confident interval.

4. Bootstrap Method

In each sample set $\{G_2, R_2\}$ and $\{G_3, R_3\}$, a subset of 6 random paired-values with replacement is taken and the values' $\rho_j *$ are calculated. This process is repeated for 50 times and the mean $\overline{\rho *}$ and standard deviation $\sigma_{\rho*}$ are found. Figure 3 and Figure 4 show the distribution for $\rho_{Y=2} *$ and $\rho_{Y=3} *$:


Fifure 3: Distribution of $\rho^*_{Y=2}$


Fifure 4: Distribution of $\rho^*_{Y=3}$

Both distributions can be roughly regarded as a Gaussian distribution. The standard deviation of $\rho_{Y=2}*$ is $\sigma_{\rho*_2} \approx 0.4466$, and the standard deviation of $\rho_{Y=3}*$ is $\sigma_{\rho*_2} \approx 0.4538$. Thus, in a 68% confidence level, both of $\rho_{Y=2}$ and $\rho_{Y=3}$ will neither approach -1 nor 1. In other words, at a 68% confidence level, there is weak or no correlation between a sequel's success rate and its previous' total gross, regardless of the gap length.

The result is different from what has been expected. Other factors related to the success rate need to be found. Regarding the current dataset, a new hypothesis about the relationship between a sequel's production studio and its success rate is generated.

5. Hypothesis Test

According to the data set, there are 5 movies made by *Sony Studio*, 8 made by *20th Century Fox* and 7 made by *Buena Vista Studio*, out of 22 sequels. Taking these three main studios as the variable of this hypothesis test, their relations are tested in pairs. The table below shows the success rate for each production studio:

									μ	σ^2
Sony	-0.23	0.71	-0.59	0.48	0.28				0.1294	0.2266
Fox	0.57	0.20	-0.17	-0.35	-0.04	0.71	-0.28	0.34	0.1237	0.1371
BV	0.46	0.89	0.36	0.30	0.43	-0.08	0.55		0.4166	0.0715

The parameter is defined as $Z \equiv \frac{\overline{R_{studio\ 1}}-\overline{R_{studio\ 2}}}{\sqrt{\sigma^2_{studio\ 1}-\sigma^{\ 2}_{studio\ 2}}}$.

(i) Compare *Sony* with *20th Century Fox*, the null hypothesis that

H_0: R_{Sony} and R_{Fox} have no difference

and $Z_1 \approx 0.009464 < 0.05$ is found, thus accepting the null hypothesis.

(ii) Comparing *Buena Vista* with *20th Century Fox*:

H_0: R_{BV} and R_{Fox} have no difference

Computing $Z_2 \approx 0.6412 > 0.05$, the null hypothesis is rejected.

(iii) Comparing *Buena Vista* with *Sony*:

H_0: R_{BV} and R_{Sony} have no difference

Computing $Z_3 \approx 0.5260 > 0.05$, the null hypothesis is also rejected.

As the result, *Sony Studio* and *20th Century Fox* do not make huge differences in their sequels' success rates. Nonetheless, *Buena Vista Studio* is more likely to make sequels with higher success rates compared to the other two production studios.

Conclusion

Both the significance test and the sign test confirm the initial prediction on the relation between a sequel's success rate and release time gap. According to these tests, sequels with a three-year gap are more likely to success compared to sequels with a two-year gap. It is reasonable enough to suggest that a movie series should release its sequel in 3 years rather than 2 years. However, the correlation coefficient test contradicts with the initial prediction, showing that there is no strong relationship between a sequel's success rate and its previous movie's total gross. This contradiction makes it impossible to build an efficient model that predicts a sequel's success rate based on its previous movie's information. Features other than total gross may relate stronger to the success rate. The last hypothesis test proves that, among the three production studios, a sequel from *Buena Vista Studio* is more likely to have higher success rate. Therefore, multiple features must be investigated in a further analysis in order to build an accurate prediction model on success rate.