Stat 535 Project: Marvel vs. DC

Matt Oehler

December 17, 2017

Abstract

In response to the never-ending debate concerning which comic book company is superior (Marvel or DC). This study takes an data-driven approach to objectively answer the question: Which comic book company produces more popular characters. Using data for 64 Marvel/DC superhero films from 1978 to 2016, the fan-based popularity of movies was modeled and analyzed.

1 Introduction

Ever since their origin, Marvel and DC have been rivaling comic book producers. As a result, there is a constant debate amongst fans as to which comic book company produces the best line of superheroes. This ever-raging debate that takes place primarily in the Twittersphere/ blogosphere is rampant with opinionated kerfuffles and predjudice discords. Rather than add to the abundance of opinions that are already out there, we decided to take a data-driven approach to see if we can more objectively determine which comic book company produces more popular superhero characters.

First we gathered data on 64 different Marvel/DC movies from as early as 1978 to as late as 2016. We used Marvel and DC movie data, to assess how a movie's studio, associated comic book company, production budget, and Rotten Tomato critic ratings relate to its fan-popularity as measured by IMDb scores. We chose to collect these variables because we believed that they would all be potentially significant in determining movie quality. High-quality movies typically receive better fan-response, which is a good measure of the popularity of the comic book character. The data will be looked at in more depth in following sections. The intent of this study is not focused on swaying people from one side of the debate to another, but rather to provide an objective baseline with which people can supplement their arguments in the endless debate of comic book superiority. Other audiences, such as the movie making industry may also be interested in seeing what variables have an impact on the popularity of a movie. The following sections will discuss the origin and description of the data, the setup and diagnostics of the statistical model, the analysis and associated discussion, and the results of our data-driven approach to addressing the question: Which comic book company, Marvel or DC, has the most popular characters?

2 Data

In order to begin this study, we needed to gather the data that we wanted to model. We used the 'XML' package in R, an open-source statistical programming language, to make the data set. For the most part, all movie titles were entered in manually for the sake of consistency when compiling everything together. We used information from superheronation.com, the-numbers.com, and imdb.com to gather the movie data for Rotten Tomato critic ratings, production budget, and IMDb scores respectively. Table 1 shows a snippet of what the resulting data set looks like.

The variables we used are as follows:
Response Variable:

- IMDb – A quantitative variable of fan-popularity measured on a scale of 1-10.

Explanatory Variables:

- Budget – A quantitative variable to indicate the production budget of each film listed in millions of dollars.

Movie Title	Studio	Comic	Tomato	Budget	IMDb
Ant-Man	BV	Marvel	80	130.00	7.30
Avengers	BV	Marvel	93	225.00	8.10
Avengers: Age of Ultron	BV	Marvel	75	330.60	7.40
Batman	WB	DC	71	35.00	7.60
Batman and Robin	WB	DC	12	125.00	3.70
Batman Begins	WB	DC	85	150.00	8.30

Table 1: Snippet of the movie data used in this study

- Tomato – A quantitative variable of professional critics' rating from Rotten Tomato measured on a scale of 1-100.
- Studio – A categorical variable indicating which studio produced the movie. There are 8 different studios in the dataset including: Warner Brothers, 20th Century Fox, Buena Vista, Sony, Lionsgate, Paramount, New Line, and Universal.
- Comic – A categorical variable indicating which comic book company the movie characters are from.

We are interested in the popularity of each company's superhero characters, so we figured that IMDb scores (which are given by fans) would be a good measure of popularity. For the explanatory variables, we chose to include the production budget since it is the primary dictator of the amount of quality action scenes and special effects that can be included in a movie. Since superhero movies are overwhelmingly dependent on action scenes, it follows that a movie's production budget would influence how well it is received by fans. Additionally we included Rotten Tomato ratings (which are provided by professionals) since they are less prone to bias than other non-professional movie review measures would be. These professional ratings provide an objective assessment of the individual superhero movies. As for the two categorical variables, comic book company is essential since it is the foundation of the question of interest, and since DC movies are primarily produced by Warner Brothers (because of exclusive rights) we decided it would be good to include a variable that would allow us to adjust for the difference between studios.

Before modeling the data, we looked at basic summary statistics to get a better glimpse of how the data are distributed and any unusual data characteristics. To do this we have included the 5 number summary and means of all quantitative variables in table 2, and also the frequency and proportion of the categorical variables in tables 3, 4, and 5.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
IMDb	3.30	5.77	6.95	6.62	7.50	9.00
Tomato	8.00	30.00	66.00	59.09	84.25	94.00
Budget	17.00	72.50	137.75	136.04	185.25	330.60

Table 2: 5-number summary and means of quantitative variables

	DC	Marvel
Frequency	21	43
Proportion	0.33	0.67

Table 3: Frequency and proportion of comic book company

Looking at table 2, we can see that, for the data we have collected, superhero movie budgets range from $17 million to $330 million dollars. It is also important to note, as shown in table 5, that several of the studios have produced significantly less movies than other studios, and so we don't have an equal spread of data points for each of the studios.

	WB	Sony	Lions	Fox	BV	Par.	Uni.	NL
Frequency	18	9	3	14	10	4	3	3
Proportion	0.28	0.14	0.05	0.22	0.16	0.06	0.05	0.05

Table 4: Frequency and proportion of studios

	DC	Marvel
WB	18	0
Sony	2	7
Lions	1	2
Fox	0	14
BV	0	10
Par.	0	4
Uni.	0	3
NL	0	3

Table 5: Comic frequency for each studio

3 Statistical Model

We chose to model the data using all of the variables additively with the exception of comic and budget, which we also included an interaction term for. The reason we chose to include this interaction is because it seems realistic that movies with certain comic book characters might be able to get more out of their budget than movies with other characters. This interaction plot is shown later in figure 1, and it confirms our belief of this interactive relationship. The model we used for this analysis is defined in equation 1.

$$\begin{aligned} Y = \beta_0 + \beta_1(Sony) + \beta_2(LionsGate) + \beta_3(Fox) + \beta_4(BuenaVista) + \\ \beta_5(Paramount + \beta_6(Universal) + \beta_7(NewLine) + \beta_8(Marvel) + \\ \beta_9(Tomato) + \beta_{10}(Budget) + \beta_{11}(Marvel)(Budget) \end{aligned} \quad (1)$$

In this dummy variable regression model, the intercept, β_0, is the base case which represents the average IMDb score of a DC movie produced by Warner Brothers (assuming that all else was set to 0). We chose to use DC and Warner Brothers as the factor levels for the base case since Warner Brothers has exclusive rights for almost all of the DC movies. This makes it so we can look at the other studios ($\beta_2 - \beta_9$) in relation to Warner Brothers. Likewise, we can look at the effect of Marvel (β_8) in terms of how it compares to DC.

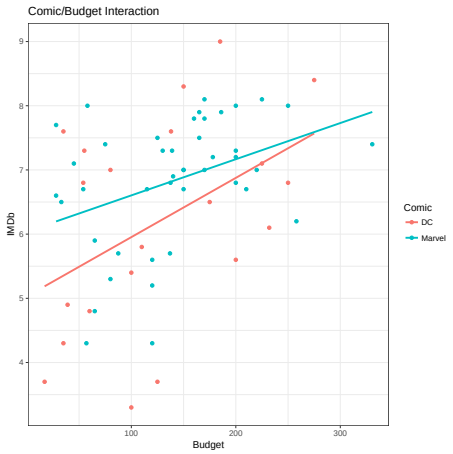

Figure 1: Plot of the interaction between production budget and comic.

4 Data Diagnostics

In this section we justify the selected model, assess the assumptions that need to be met, and inspect influential observations. These diagnostics confirm that we can proceed with the analysis using the model that is listed and defined in the previous section.

For the independence assumption we can be confident that the IMDb score of one movie doesn't affect the IMDb score of another movie, which means that are data are independent. This assumption might not hold in the case of movie sequels, but we determined that there aren't enough of those to get us worried, so we maintained this assumption.

The next assumptions pertaining to the equal variance and distribution of the residuals can be assessed using the plots in figure 2. According to standard linear model assumptions, the variance of the residuals needs to be constant throughout the observations, they also need to be normally distributed. Figure 2 shows that the variance is approximately equal throughout the observations, and that the residuals are approximately normally distributed. These assumptions aren't perfectly met by any means, but given that they hold as well as they do for only 64 observations, it is safe to maintain these model assumptions.

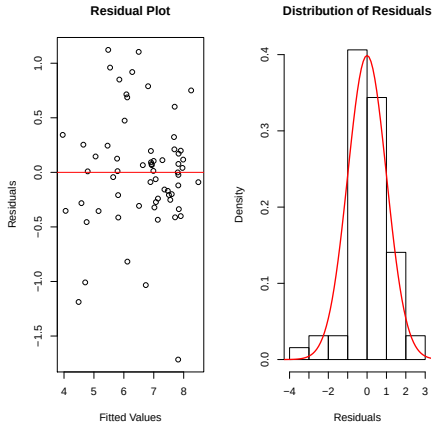

Figure 2: Plot of variance and distribution of residuals

To avoid having outliers influence the results of the analysis, we decided that it would useful to inspect any influential observations. To do this we flagged any movies that violated the rule of thumb for leverage, Cook's Distance, and R-studentized residuals. Figure 3 shows how many points are above or outside the rule of thumb threshold for each of the methods just mentioned, and table 6 lists the title of all influential movies with the methods that deemed them influential.

It can be seen in figure 4 that the influential points don't appear to be major outliers and don't

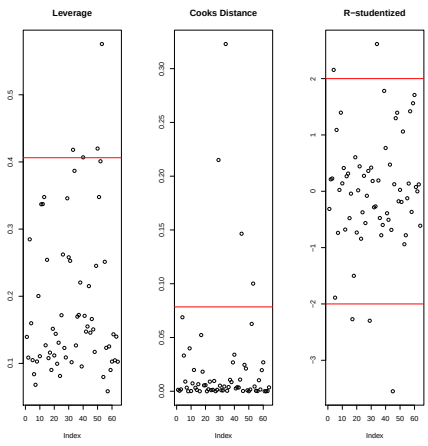

Figure 3: Plot of influential points by rule of thumb

Influential Movies	Leverage	Cook's Distance	R-studentized
Kick ***	Yes	No	No
Supergirl	Yes	No	No
The Green Hornet	Yes	No	No
The Spirit	Yes	Yes	No
Hulk	No	Yes	Yes
Kick *** 2	No	Yes	Yes
Superman Returns	No	Yes	Yes
Batman	No	No	Yes
Catwoman	No	No	Yes

Table 6: Table of which and how observations are influential

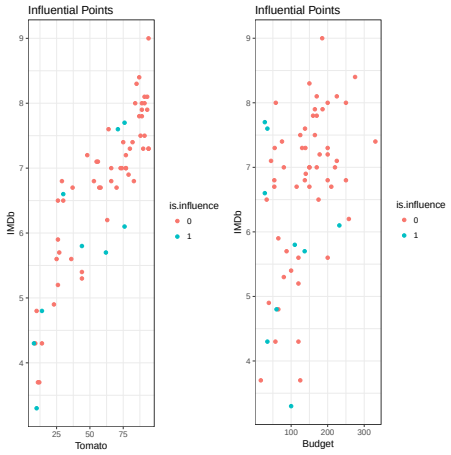

Figure 4: Plots of influential points vs Tomato and Budget

seem to follow any particular pattern (other than being close to the edge of the observations). Because of this, we concluded that these observations aren't negatively influential, and we decided to keep them in the model.

5 Results of Analysis

Fitting the model shown in equation 1 yields the β estimates shown in table 7. The 95% confidence intervals for these estimates are shown in table 8.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5268	0.3039	11.60	0.0000
StudioSony	-0.0821	0.4133	-0.20	0.8433
StudioLions	0.7392	0.4797	1.54	0.1294
StudioFox	0.2820	0.4627	0.61	0.5449
StudioBV	0.2260	0.4793	0.47	0.6392
StudioPar.	-0.0651	0.5191	-0.13	0.9007
StudioUni.	0.1913	0.5511	0.35	0.7299
StudioNL	0.6415	0.5643	1.14	0.2608
Tomato	0.0390	0.0031	12.60	0.0000
ComicMarvel	0.5875	0.4682	1.25	0.2152
Budget	0.0057	0.0017	3.31	0.0017
ComicMarvel:Budget	-0.0056	0.0024	-2.31	0.0251

Table 7: Estimated model coefficients

	2.5 %	97.5 %
(Intercept)	2.92	4.14
StudioSony	-0.91	0.75
StudioLions	-0.22	1.70
StudioFox	-0.65	1.21
StudioBV	-0.74	1.19
StudioPar.	-1.11	0.98
StudioUni.	-0.91	1.30
StudioNL	-0.49	1.77
Tomato	0.03	0.05
ComicMarvel	-0.35	1.53
Budget	0.00	0.01
ComicMarvel:Budget	-0.01	-0.00

Table 8: 95% confidence intervals for the model coefficients

6 Discussion

The results of the fit model will now be utilized to test several relationships of interest that pertain to our research question: Which comic book company produces more popular characters? The first relationship we'll look at is the difference between marvel and DC comics. At first glance, one might think that since the estimated coefficient for ComicMarvel is positive the conclusion is that Marvel produces more popular comic book characters. However, this is not the case (at least according to the results of this study). As shown in table 7, the calculated p-value for the effect of Marvel (compared to DC) is 0.21. Since this is significantly higher than $\alpha = 0.05$ we concluded that there isn't a significant difference between Marvel and DC comics. We also conducted a full and reduced model comparison to look at the overall effect of comic in the model (i.e. we compared the original model to a model without the variable for comic or the interaction between comic and budget). This resulted in a p-value of 0.78, which led us to conclude that comic does not have a significant effect. Table 5 provides a visual illustration of the non-significance of comic book company. It shows that the 95% confidence interval for the effect of Marvel compared to DC contains 0, implying that there isn't a significant effect. This result prevents us from coming to a definitive conclusion in regards to one comic book company being more popular than the other. One thing that we did notice however, is that the interaction between Comic and Budget is significant with a resulting p-value of 0.025. This the estimated coefficient for this interaction is -0.0056. This means that (holding all else constant) as budget increases the average increase in IMDb score is greater for DC movies than it is for Marvel movies. This significance doesn't carry over to say that overall DC is better than Marvel, but it is important to note that this implies that in terms of fan popularity, DC movies appear to better utilize their budget.

We also examined the relationship between fan popularity and production budget. Budget is one variable that, according to the selected model, proved to be significant in its relationship with IMDb score. We found that holding all else constant, for each additional $1 million increase in production budget, fan ratings (IMDb) increase on average by 0.0057. This makes sense intuitively because movies with higher budgets are able to have more high-quality action scenes, which leads to higher fan popularity regardless of comic book company.

The next relationship we inspected was the difference in effect between movie studios on fan popularity. Table 7 shows that none of the levels of studio resulted in a p-value less than 0.05. This means that none of the movie studios have a significant effect in comparison to Warner Brothers (the base case) on fan-popularity. We affirmed this by doing a full and reduced model comparison of the original model with and without the studio variable which resulted in a p-value of 0.51 implying that none of the studios have a significant effect on movie popularity. This non-significance is expressed visually in figure 5.

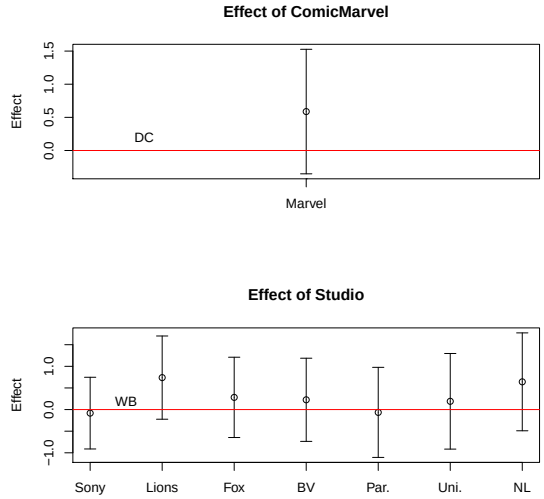

Figure 5: Plot of estimated effect and 95%confidence interval for Comic and Studio

7 Summary

After modeling the data we gathered and testing several relationships, we have reached several conclusions pertaining to the original research question: Which comic book company produces more popular characters? Our results showed that neither company produces characters that have significantly more popular movies than the other. Consequently we concluded that neither company produces more popular characters. We did, however, come to the conclusion that DC movies appear to get more bang for the buck when it comes to movie popularity based on production budget, as the interaction between comic and budget was significant. Lastly, we concluded that movie studios do not have a significant effect on the popularity of super hero movies.

Since this study did not come to a conclusive results for one comic book company being more popular than the other, the Marvel vs DC debate will continue on without much disturbance. Further steps could be taken to augment this study and perhaps reach conclusive results. These steps include gathering more movies for the data set, or possibly incorporating additional variables such as comic book sales. Whatever steps are taken, it would be beneficial in the field of the comic book company debate to have objective, data-driven results by which fans could supplement their existing opinions.

8 References

- http://www.imdb.com/list/ls051507615/?start=1&view=compact&sort=listorian:asc&defaults=1&scb=0.74955
- http://www.the-numbers.com/movie/budgets/all
- http://www.superheronation.com/2011/08/22/rotten-tomatoes-ratings-for-superhero-movies/
- https://cran.r-project.org/web/packages/XML/index.html