Comic Book Regression

Jackson Curtis

Brigham Young University

Abstract. This project applies the concepts of linear models to a dataset about movies based off comic books. Techniques are employed to estimate the popularity of the different comic book universes. Hypotheses about the effect of variables such as franchise, studio, budget, and professional rating are tested.

1 Introduction

In our research we are interested in discovering whether Marvel or DC comic book characters are more popular with fans. Many business ventures could be interested in the answer to this question. For example, if toy makers are going to license toys from one comic book franchise or the other, knowing which is more popular would be extremely helpful as popularity would be a key driver of sales.

In recent times superhero popularity is largely driven by movies. Movie studios have been producing movies based on comic books at a break neck pace recently due to the major successes of movies such as The Dark Knight trilogy and The Avengers. In order to answer the question of popularity, we will look at how fans respond to movies in the superhero genre.

To do this we have collected data about superhero movies released over the last few decades. To answer the question of which characters are more popular, we will look at fan popularity based on ratings from the Internet Movie Database. However, because not all movies are similar in quality, we will use several other control variables to explain movie quality, and then analyze the effect of franchise on fan response. Some of the things we will control for are budget, critic rating, and studio.

This analysis will help people make decisions on the future of these franchises. What movies to produce, what video games to create, and what advertisements to make can be answered by better understanding superhero popularity.

2 Data

The dataset was culled from several different online sources. Each row in our dataset contains information about a superhero movie released in the past three decades. Occasionally movie names had to be reformatted in order to merge smoothly with data from a different website. Below we will walk through each variable and how it was collected and formatted.

Fan Popularity

Fan popularity was scraped from a IMDb web page containing average user ratings for superhero movies submitted by IMDb users. Fans rate movies on a scale of one to ten, and the ratings are averaged over all submissions. More recent movies were not available on the web page, so they were collected manually from IMDb and added to the dataset.

Comic

The variable comic identifies whether it came from the DC or Marvel comic book franchise. It was entered manually.

Budget

The variable budget reflects what is known about how much the studio spent producing the movie. The data is collected from The Numbers website, but they note that budget figures are not always reliable, as studios sometimes have interest in inflating or deflating the true costs. After collection the data was formatted to remove dollar signs and commas.

Tomato Meter

The tomato meter rating is an aggregation of critic consensus as calculated by RottenTomatoes.com. We scraped the data from superheronation.com where they collect many superhero movies in one convenient list. We supplemented this list with additional superhero movies that were not included.

Studio

The studio variable identified which movie studio produced the movie. This data was entered manually.

3 Summary Statistics

We begin with some exploratory data analysis about the individual variables. Table 1 shows that our data is not balanced across levels of comic franchise or studio. We have more information about Marvel movies, and we have many studios who have only a few superhero movies.

For our numeric variables we note that critic rating can take on values between 0 and 100, but the range of our dataset has a range of 8 to 94. Similarly, Fan popularity ranges from 0 to 10, but our movies have values of 3.3 to 9. Summaries of the spread are contained in Table 3.

Table 2 summarizes our quantitative variables broken up by comic franchise.

	DC	Marvel	Proportion
Warner Brothers	18	0	0.28
Buena Vista	0	10	0.16
Fox	0	14	0.22
Lionsgate	1	2	0.05
New Line	0	3	0.05
Paramount	0	4	0.06
Sony	2	7	0.14
Universal	0	3	0.05
Proportion	0.33	0.67	1

Table 1: Counts for studio and franchise for movies in dataset

Franchise	Budget	Tomato	Fan Popularity
DC	125.71	49.19	6.19
Marvel	141.08	63.93	6.83

Table 2: Averages by comic book franchise

	Critic	Budget	Fan Popularity
Min.	8.00	17.0	3.30
1st Qu.	30.00	72.5	5.78
Median	66.00	137.8	6.95
Mean	59.09	136.0	6.62
3rd Qu.	84.25	185.2	7.50
Max.	94.00	330.6	9.00

Table 3: Summary statistics for quantitative variables in the model (Budget in millions of USD)

4 Choosing a Model

We will answer the question of which universe has more popular characters by looking at fan ratings of movies in which the characters appear, but we will control for our other explanatory variables which influence movie popularity.

To answer these questions, we want to model how fan popularity is effected by our explanatory variables studio, franchise, critic rating, and budget, so we can make inference on the franchise effect. In order to do this we need to define the relationship between the explanatory variables and the response. A good place to start is to ensure that the effects of the explanatory variables are linear with respect to our parameters.

One area of interest is the effect of budget. Budget has an unbounded upper limit (studios could spend any amount of money on a film), but fan rating is bounded at an upper bound of ten, so we should be concerned that the relationship cannot be linear indefinitely. In addition, from an economic perspective we would expect diminishing returns as more money fails to increase the quality of the finished product. Indeed Figure 1 shows that while Budget appears to have a strong positive effect for low budgets, the effect weakens after about $150 million. To linearize the relationship, we will model the square root of budget, as this does a good job of reflecting the phenomenon of diminishing returns.

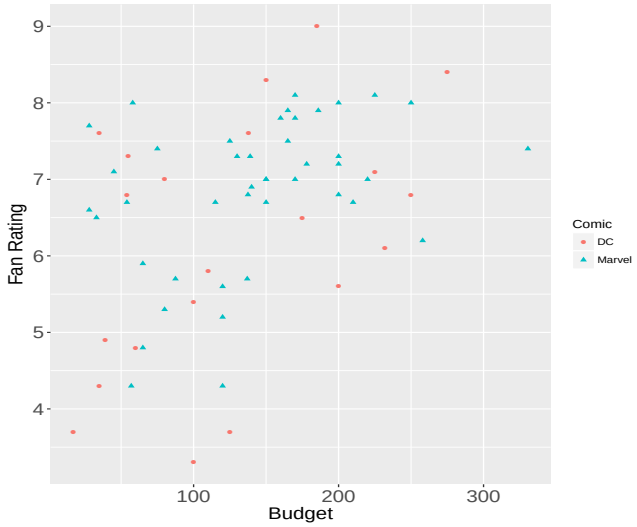

Fig. 1: Budget suffers from diminishing returns

The relationship between critic and fan ratings is quite linear, so we do not have the same concerns as we did with budget. From there, we might be interested in what interactions might be significant to consider. We are limited in the interactions we can consider by what is estimable. For example, our data is too

sparse to estimate effects of a studio by franchise interaction, but we can consider interactions between franchise and our quantitative variables. One that we might be interested in from what we know about superhero movies is a budget and franchise interaction. In recent years, Marvel movies have become famous as big summer blockbusters, and money has poured into the special effects and casting the biggest stars. DC got a later start to building their cinematic universe, so it seems reasonable to expect that the effect of budget might differ by franchise. We can look at an added variable plot to justify including this interaction in our model.

Figure 2 gives enough evidence for an interaction that we will include it in our model so that we can run hypothesis tests on it. We do not have a good reason to suspect other interactions, nor do we see anything in the data to justify it, so we will use only one interaction.

Our model is:

$$y_{ijk} = \beta_i + \beta_j + \beta_{10}\sqrt{Budget} + \beta_{11}Critic + \beta_{12}\sqrt{Bud} * I(i = 2) + \epsilon \qquad (1)$$

$$\epsilon \sim N(0, \sigma^2)$$

Where y_{ijk} is the fan popularity for movie k for the ith level of franchise and the jth level of studio. There are two levels of franchise, Marvel and DC. The seven βs correspond to the eight movie studios, with the effect for the first studio (Warner Brothers) being built into the intercepts of β_1 and β_2. The indicator function next to β_{12} indicates the interaction that will only apply if the movie is in the Marvel universe. ϵ represents our normally distributed errors.

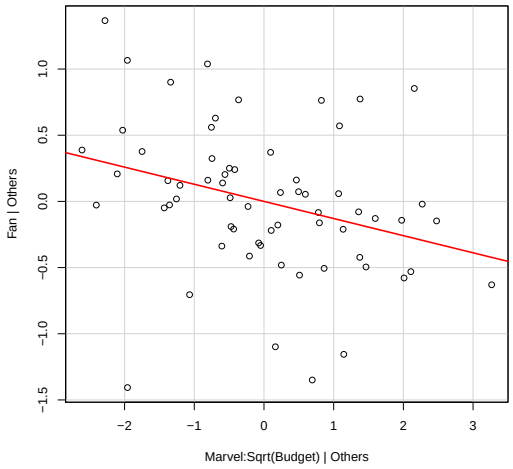

Fig. 2: The added variable plot suggests evidence of an interaction

	$\hat{\beta}$	Standard error	t statistic	p-value
DC	2.9910	0.444	6.73	0.0000
Marvel	4.2632	0.628	6.79	0.0000
Bueno Vista	0.2402	0.485	0.50	0.6222
Fox	0.2616	0.465	0.56	0.5760
Lionsgate	0.6671	0.485	1.37	0.1752
New Line	0.5757	0.566	1.02	0.3138
Paramount	-0.0665	0.525	-0.13	0.8997
Sony	-0.0837	0.417	-0.20	0.8415
Universal	0.1535	0.552	0.28	0.7822
SqrtBudget	0.1177	0.037	3.18	0.0025
Tomato	0.0392	0.003	12.52	0.0000
Marvel:SqrtBud	-0.1294	0.055	-2.36	0.0218

Table 4: Estimates for $\hat{\beta}_1$ to $\hat{\beta}_{12}$ with hypothesis test $H_0 : \beta_j = 0$

5 Data Diagnostics

Now that we have a model we can diagnose our model and look for problems. One thing we want to look for are influential observations. Leverage is a measure of how much an observation contributes to the prediction of itself. Unsurprisingly, Figure 3 shows that movies from studios with few movies in the dataset have the highest leverage. Noticing this, we will proceed with the note that predictions involving movies from these studios will likely have larger uncertainty than our model suggests.

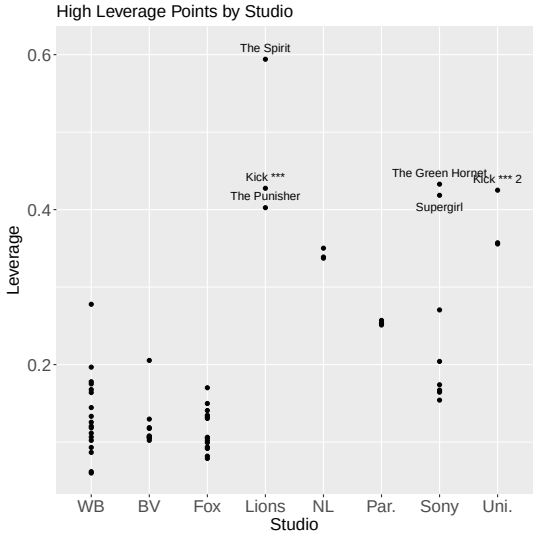

Fig. 3: High leverage points occur when studios have produced few movies

A big concern for our model are outliers. Although our residuals look fairly normally distributed, we have issues with outliers. Superman Returns has a R Studentized residual of -3.36. Because those residuals are t distributed, we would expect a residual that large or larger just over one in a thousand times. For only 64 data points, it is quite an extreme figure. Looking into Superman Returns, it appears the movie has such a large residual because the critics viewed it favorably, but the fans did not respond similarly. Because that appears to be the main cause, and it's not unlikely that that would happen with future movies, we will leave it in our dataset, but note that it could lower our power.

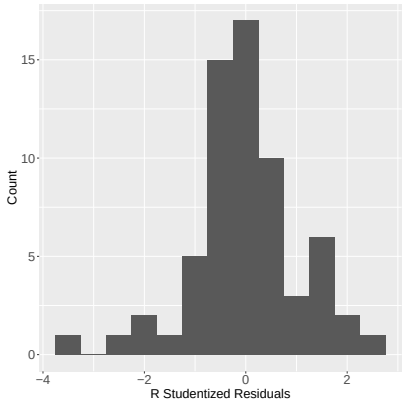

Fig. 4: R Studentized residuals show a few outliers

6 Discussion

After satisfying ourselves that our model is doing what we want, we can answer the original research questions. Our primary research question was whether the comic book franchise had an effect on fan popularity. Our model suggests two ways this could happen: either through the two intercepts (β_1 or β_2) being different, or our slope for budget not being the same for Marvel and DC (β_{12} not being equal to zero). Thus we can test the hypothesis $H_0 : C\beta = 0$ with the C-matrix:

$$C = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

This hypothesis test produces an F-statistic of 2.81 on 2 and 52 degrees of freedom which gives us a p-value of 0.0693. This may initially be surprising when you consider that Table 4 tested H_0: $\beta_{12} = 0$ and returned a p-value of 0.02, but it's important to remember that F-tests of multiple hypothesis at once do a better job of controlling Type I error rate by adjusting for the fact that we are testing multiple things. Figure 5 demonstrates visually that these two effects

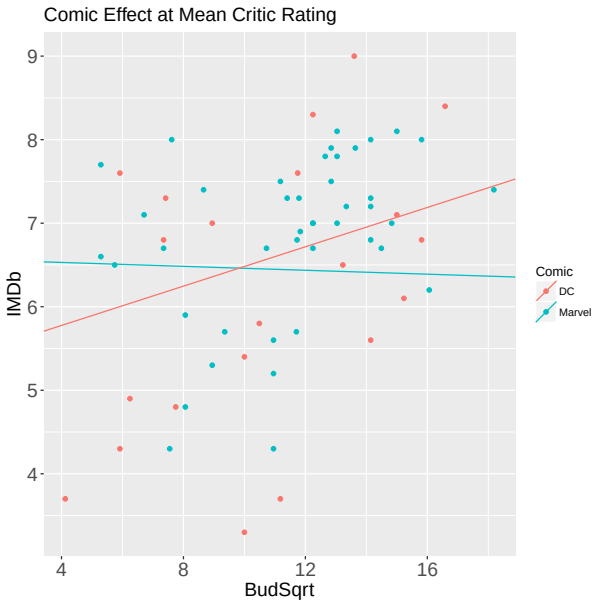

Fig. 5: Our first F-test tests for differences in the two plotted lines

(Marvel with a higher intercept, DC with a higher slope) have a canceling effect on each other when tested simultaneously.

We can create confidence intervals to summarize the effect of franchise on popularity. Because we are interested in a confidence interval on the difference between the intercepts and a interval on the difference between slopes, I will use a Bonferroni correction when computing my intervals.

The confidence interval for $\beta_1 - \beta_2$ is (-2.776, 0.231). This means that we expect a movie made by DC instead of Marvel to score between 2.8 points worse or 0.2 points better all else held equal.

The difference between slopes is given by β_{12}. The confidence interval on this parameter is (-0.256, -0.003). This means we expect the slope to be lower for Marvel movies somewhere in the range of 0.256 and 0.003. While the effect is small we do expect that a big budget helps a DC movie more than a Marvel movie.

Table 2 demonstrates that while Marvel had greater fan popularity, its movies also had bigger budgets and were rated higher by critics. Critic rating in particular was a great predictor of fan popularity, and after controlling for it the difference between Marvel and DC was smaller than it originally appeared.

Next we are interested in assessing the effect of an increase in budget on fan popularity. Because we modeled an interaction with budget and franchise, we can report the effect of budget using two confidence intervals, one for Marvel and one for DC. For DC, the confidence interval on β_{10} is (0.043,0.192). The effect of this on the response will vary depending on the magnitude of the budget

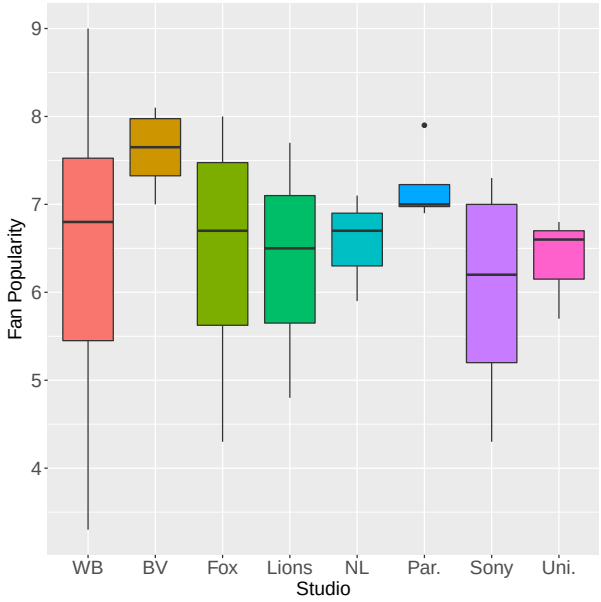

Fig. 6: There are no clear differences between studios

(because of the transformation), but, as an example, if we were to move from the mean budget to one million more than the mean budget (all else equal), we could expect an increase in fan popularity between 0.0019 0.0082.

We can analyze the effect of budget for Marvel movies similarly, but this time it will be a confidence interval on $\beta_{10} + \beta_{12}$. That interval (-0.0953 0.072). From this we cannot conclude definitively whether an increased budget hurts or helps Marvel movies with fan popularity. For example, at the mean budget that interval corresponds to 0.0041 decrease or 0.0031 increase in fan popularity for an additional million spent.

Next we are interested in whether the studio that produced the movie effects fan popularity. This test naturally lends itself to a full and reduced model test, where our reduced model is a model fit without the βs for studio. This test gives us an F-statistic of 0.748 and a p-value of 0.6326. This gives no evidence that studio significantly predicts how well a movie will be received. Figure 6 represents this visually.

7 Summary

In conclusion, our analysis did not show that audiences had clear preferences for one comic book universe over another. Instead, it seems they value high quality movies regardless of where they come from. How critics rated the quality of the movie was the best predictor of fan response.

Our best conclusion about what we can say about the movies is that the Marvel franchise is more consistently putting out high quality movies. DC movies can do quite well when they have the budget to support it, but many DC movies had a smaller budget. Marvel movies had much bigger budgets, and we didn't see any evidence that pumping more money into Marvel movies increased fan rating. They seem to have found a sweet spot for getting the most bang for their buck.

Further research could be done about the differences between studio. Really small sample sizes weakened our ability to discern meaningful differences. An interesting approach would be to look at more than one genre of movie to increase sample size and to estimate the quality

8 References

Superman Returns (2006). Rotten Tomatoes.
https://www.rottentomatoes.com/m/superman_returns/