Character Popularity in Superhero Movies: A Linear Analysis

Kaylee Hodgson

Brigham Young University

Abstract. This paper addresses the debate regarding which comic book company, DC or Marvel, makes superhero movies that have more popular characters. I provide a linear regression analysis to answer this question, along with other subanalyses. The findings indicate that there is not a significant difference in character popularity based on the comic book company that made the movie. However, I find other interesting results that are significant to the model. This research provides a better framework for the debates regarding the success of the two companies to continue, and provides interesting information regarding how DC and Marvel could improve their fan ratings on character popularity in future movies.

Keywords: linear regression, ANOVA, reparameterized model, superhero movies, comic books

1 Introduction

Movies based on Marvel and DC comic books have been popular and widely viewed for years. Professional critics, fans, journalists, and bloggers passionately debate over which comic book company has been more successful and which characters they prefer. Preferences are different depending on the source, and conclusions are based on many different factors. In this paper, I attempt to more systematically analyze which comic book company has had more successful characters in the movies based off of their comic books, by running a linear regression analysis. With this analysis, I intend to create a better framework for the debates regarding these two comic companies by providing an objective analysis of the success of the movies. This is not meant to be conclusive, but rather to ignite a more informed debate around the topic.

The response variable of interest in this analysis is the popularity of characters in the movies based on comic books, which is measured by fan ratings on IMDb. This variable represents well what is being measured in this analysis. IMDb is commonly used by movie fans to review movies, and these fan ratings indicate a movie's level of popularity. This variable is compared between Marvel and DC movies to determine which company has had more popular characters. The comic variable is used to answer the main question of interest in this analysis. Because the differences in popularity ratings are likely not only caused by the difference between the two comic companies, I adjust my analysis for other variables that likely influence the popularity of a movie. I include the movie's production budget, professional critical rating, and production studio. Controlling for these other variables will more accurately answer the research question.

The production budget is included because the budget of a movie can either enhance or detract from the success of the movie's characters. The budget of a comic book movie is particularly important because the movie's quality is largely based on the special effects. Movies with lower budgets may have less popular characters only because the movie's poor quality made fans less likely to like the characters of the movie. Professional critical ratings are included as a more objective rating of the movie. This is a measure of the quality of the movie. The studio is also included because Marvel and DC do not have exclusive rights to all of the studios they use, so there is some overlap. Some studios turn out more successful movies than others, which may also affect the ratings of a movie's characters. All of these variables are included in the analysis to control for other factors that may affect fan's ratings of a movie. They measure different aspects of the quality of a movie, which makes our fan ratings variable a more accurate measure of the popularity of the movie's superhero character.

The dataset includes 64 superhero movies from Marvel and DC. The movies were made between 1978 and 2016. Each movie has a value for the fan popularity score, the comic book company, and each of the other explanatory variables. The results in the analysis will better inform future debates regarding the success of superhero movie characters. While it will likely not change fan's opinions, both sides can begin to consider why one comic book company has found more success than the other and can begin to consider how to improve on that. This analysis is also meant to be informative for the comic companies in providing feedback on how their characters have been received by the fans, and what could be done to improve that.

The paper will proceed as follows: In section 2, I define and describe the data that is used in the analyses. In section 3, I provide summary statistics for all of the variables. In section 4 I introduce the linear model used to evaluate the questions, then in section 5, I preform data diagnostics and consider the abnormalities and assumptions in the statistical model. In section 6, I run the regression analysis to test the significance of the explanatory variables, then in section 7, I discuss the results of the regression analyses. Finally, in section 8, I make concluding remarks regarding the results.

2 Data

The data for this analysis was collected from multiple sources and compiled into the dataset. Most of the data was scraped from the web, using the **XML** package in R, and compiled into an R script. The variables were then cleaned up, ensuring that variable values for the same movie always matched up by adjusting the movie names if needed before merging. The data was then compiled and organized alphabetically by movie.

The response variable of interest in this analysis, the popularity of characters in the movies based on comic books (IMDb), is measured by Fan Popularity, which is based on the fan ratings on IMDb's website. Movies that are more popular receive higher fan ratings, while movies that are less popular are rated lower. Because superhero movies are all about the superhero character, this measure is a good indication of the popularity of the character in the movie. As discussed in the introduction, I evaluate multiple variables in an attempt to explain differences in these fan ratings.

The explanatory variables in this analysis are as follows:

- Comic: This variable is a binary explanatory variable, which labels a movie as either Marvel or DC. This is the main variable of interest in the analysis. The information for this variable was manually coded in for each movie.
- Budget: The budget variable measures the production budget (in $ million) for a movie. While this is not the main variable of interest, I predict that it significantly affects a movie's fan ratings. This will therefore need to be adjusted for in the analysis. The data for this variable was pulled from "The Numbers: Where Data and the Movie Business Meet" website. The Production Budget on the website is written in single dollar amounts, but was adjusted into terms of $ million for the purpose of this analysis. So, for example, the movie budget for Superman Returns was changed from $ 232,000,000 to 232.0 ($ million) in this dataset.
- Tomato: The Tomato variable measures the professional critics ratings from Rotten Tomatoes. According to Rotten Tomatoes, the Tomatometer (the metric for the ratings) is "based on the published opinions of hundreds of film and television critics [and] is a trusted measurement of movie and TV programming quality for millions of moviegoers." The value is based on the percentage of positive reviews given to a movie by these prifessional critics. The data for this analysis was actually pulled from Superhero Nation, which reports the Rotten Tomatoes score for superhero movies. This variable is also not the main variable in the analysis, but I predict that this is also significant to a movie's fan ratings, and will adjust for that in the analysis.
- Studio: This variable is measured because Marvel and DC often do not have exclusive rights to studios, so this variable also needs to be adjusted for in the analysis. This was also coded in manually for each movie. The variable abbreviations have the following meanings, which will be used in the paper from here on: WB = Warner Brothers, Fox = 20th Century Fox, BV = Buena Vista (Disney/Marvel), Sony = Sony, Lions = Lionsgate, Par. = Paramount, NL = New Line, Uni. = Universal.

3 Summary Statistics

The following tables display the descriptive statistics for the data, given for the entire dataset, then only Marvel and only DC movies. Table 1 gives the descriptive statistics for both of the categorical variables: Comic and Studio.

These provide both the count and the proportion of the factors in these variables. Notice that Marvel has over double the movies in this dataset than DC has, so proportion is a more standardized way to evaluate the differences in studios based on the comic company. Most studios are used by one comic book company or another, however, two of the studios have produced movies for both companies: Lions and Sony. Table 2 displays the descriptive statistics for the quantitative variables: IMDb ratings, Tomato ratings, and movie budget.

Table 1. Descriptive Statistics: Categorical Variables

		Comic Company	**Studio Name**							
			WB	BV	Fox	Lions	NL	Par.	Sony	Uni.
Overall	Frequency	64	18	10	14	3	3	4	9	3
	Proportion	1	1	1	1	1	1	1	1	1
Marvel	Freuqency	43	0	10	14	2	3	4	7	3
	Proportion	0.67	0	1	1	0.67	1	1	0.78	1
DC	Frequency	21	18	0	0	1	0	0	2	0
	Proportion	0.33	1	0	0	0.33	0	0	0.22	0

Table 2. Descriptive Statistics: Quantitative Variables

IMDb Fan Ratings						
	Mean	Min	Q1	Median	Q3	Max
Overall	6.62	3.30	3.70	6.95	8.34	9.00
Marvel	6.83	4.30	4.33	7.00	8.09	8.10
DC	6.19	3.30	3.50	6.50	8.70	9.00

Tomato Professional Ratings						
	Mean	Min	Q1	Median	Q3	Max
Overall	59.09	8.00	9.57	66.00	94.00	94.00
Marvel	63.93	9.00	10.20	73.00	93.00	94.00
DC	49.19	8.00	9.00	44.00	94.00	94.00

Budget ($ million)						
	Mean	Min	Q1	Median	Q3	Max
Overall	136.04	17.00	28.00	137.75	265.22	330.60
Marvel	141.08	28.00	28.25	140.00	257.60	330.6
DC	125.71	17.00	26.00	110.00	262.50	275.00

4 Statistical Model

In this analysis, I evaluate how fan ratings are affected by the comic company that makes the movie, the professional ratings on Rotten Tomatoes, a movie's production budget, and the studio the movie is produced in. To find the effects and determine the significance of the these variables on fan ratings, I create a linear regression model. I chose to include an interaction in my model between comic company and budget. I suspect that one reason one comic company may perform better than another is because one may make higher-budgeted movies, which fans respond positively to in terms of ratings. I test this as one of the research questions in my analysis. I model the data using the reparameterized model where $Y = X\boldsymbol{\beta} + \epsilon$, where Y is a vector of all of the values of IMDb fan ratings. The X matrix and the $\boldsymbol{\beta}$ vector of the model are defined as follows:

$$\mathbf{X} = \begin{bmatrix} J & 0 & 0 & 0 & 0 & 0 & 0 & 0 & X_B & X_T & 0 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & 0 & 0 & X_B & X_T & J & X_B \\ J & J & 0 & 0 & 0 & 0 & 0 & 0 & X_B & X_T & 0 & 0 \\ J & J & 0 & 0 & 0 & 0 & 0 & 0 & X_B & X_T & J & X_B \\ J & 0 & J & 0 & 0 & 0 & 0 & 0 & X_B & X_T & 0 & 0 \\ J & 0 & J & 0 & 0 & 0 & 0 & 0 & X_B & X_T & J & X_B \\ J & 0 & 0 & J & 0 & 0 & 0 & 0 & X_B & X_T & 0 & 0 \\ J & 0 & 0 & J & 0 & 0 & 0 & 0 & X_B & X_T & J & X_B \\ J & 0 & 0 & 0 & J & 0 & 0 & 0 & X_B & X_T & 0 & 0 \\ J & 0 & 0 & 0 & J & 0 & 0 & 0 & X_B & X_T & J & X_B \\ J & 0 & 0 & 0 & 0 & J & 0 & 0 & X_B & X_T & 0 & 0 \\ J & 0 & 0 & 0 & 0 & J & 0 & 0 & X_B & X_T & J & X_B \\ J & 0 & 0 & 0 & 0 & 0 & J & 0 & X_B & X_T & 0 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & J & 0 & X_B & X_T & J & X_B \\ J & 0 & 0 & 0 & 0 & 0 & 0 & J & X_B & X_T & 0 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & 0 & J & X_B & X_T & J & X_B \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_{BV} \\ \beta_{Fox} \\ \beta_{Lions} \\ \beta_{NL} \\ \beta_{Par} \\ \beta_{Sony} \\ \beta_{Uni} \\ \beta_B \\ \beta_T \\ \beta_M \\ \beta_{B:M} \end{bmatrix}$$

5 Data Diagnostics

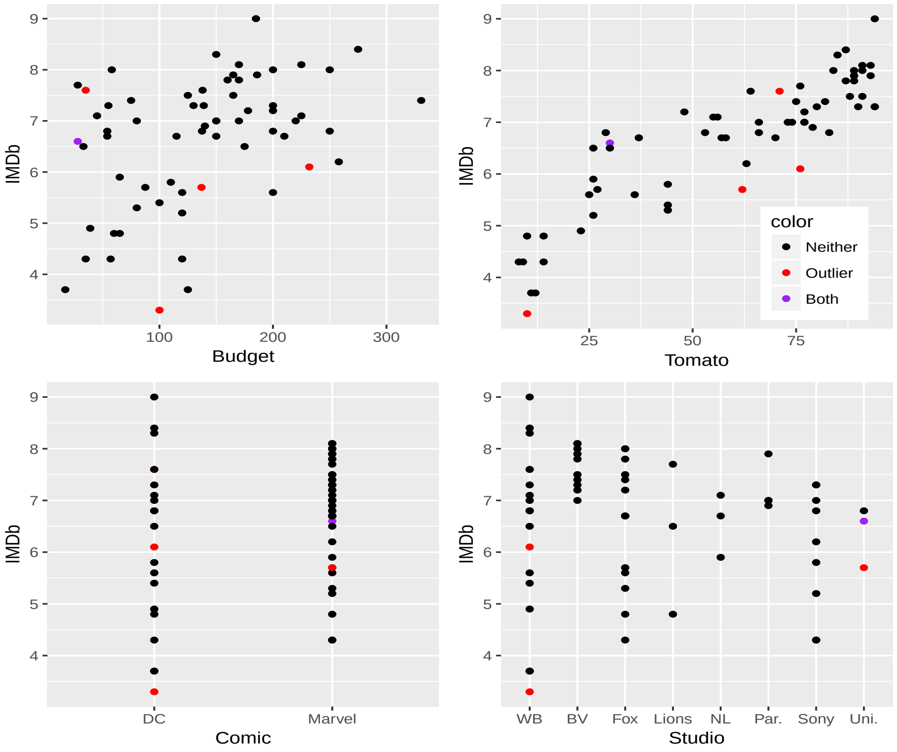
Diagnostic checks are run on the data to verify the validity of the model chosen in the previous section. Leverage and Cook's distance tests are employed to identify influential points in the dataset. Only one movie is identified, with high Cook's distance: Kick *** 2. This is likely because the movie's professional rating (30) was much lower than the average professional rating (59.09), but its IMDb rating was an average score. This tells us that the fan ratings were surprisingly high for the movie given the low professional rating it received. I also test for outliers in the data, and find that there are multiple outliers to consider. Table 3 displays the movies that were found to be outliers, and their values for the explanatory and response variables. Most of these outlier observations are fairly popular superhero movies, which many fans care a lot about. Therefore, exclusion of any of these observations should be very carefully considered.

Table 3. Outliers

Movie Title	Studio	Comic	Tomato	Budget	IMDb
Batman	WB	DC	71	35	7.6
Catwoman	WB	DC	10	100	3.3
Hulk	Uni.	Marvel	62	137	5.7
Kick *** 2	Uni.	Marvel	30	28	6.6
Superman Returns	WB	DC	76	232	6.1

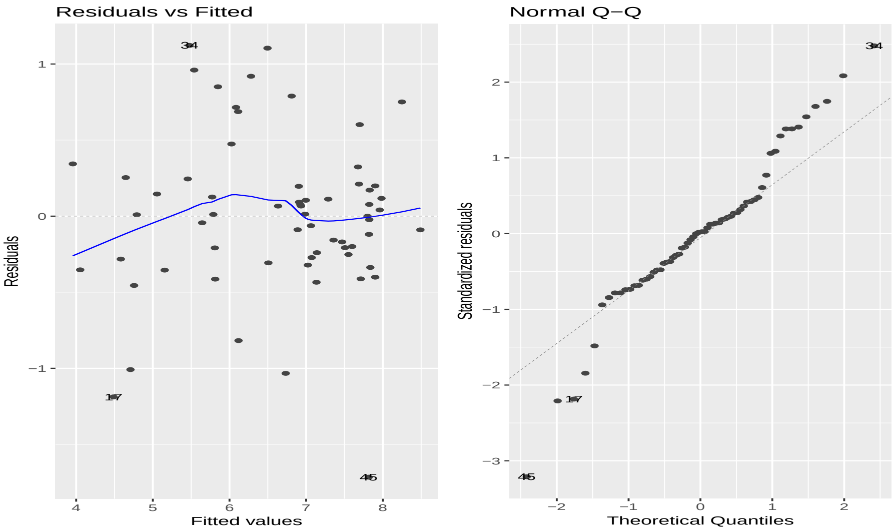
Figure 1 plots the outlier and influential points against the four different explanatory variables. The only movie that is both an outlier and an influential point is Kick *** 2. I determine that this point should be kept in the model because the test values are not high enough to justify excluding the movie. Its benefit of the unique information it offers offsets the risk of a possible slight decrease in the certainty of the predictions in the model. The rest of the outlier observations will also remain in the model because, given the high interest in these particular movies, I do not find sufficient reason to exclude them.

Fig. 1. Outliers and Influential Points



I also verify the independence of the explanatory variables in the model. The variance inflation factor (VIF) is calculated for each of the explanatory variables. The general rule of thumb is that if the VIF is greater than 10, that is an indication of a multicollinearity issue in the model. However, the VIF for the variables this model are all lower than that value, so we are not concerned with the collinearity of the explanatory variables. I lastly verify the normality and linearity assumptions of the model. The normal Q-Q plot in Figure 2 does not raise concerns regarding the normality assumptions of the response variables in the model. While there are a few points that give the normal Q-Q plot heavy tails, they are not enough of a concern to remove from the model. Additionally, the residuals vs. fitted plot in Figure 2 shows that the model has approximately constant variance.

Fig. 2. Diagnostic Plots for Constant Variance and Normality



I conclude that this is a well-fitted model and that none of the observations should be removed. While there may be a couple of small deviances from the model assumptions, they are not significant enough to remove any data or change the design.

6 Analysis

The coefficients of the reparameterized model can be found in Table 4.

Table 4. Model Summary

Coefficients	Estimate	Standard Error	t value	p-value
β_0	3.53	0.304	11.604	0.000
$\beta_{StudioBV}$	0.226	0.479	0.472	0.639
$\beta_{StudioFox}$	0.282	0.463	0.609	0.545
$\beta_{StudioLions}$	0.739	0.480	1.541	0.129
$\beta_{StudioNL}$	0.641	0.564	1.137	0.261
$\beta_{StudioPar.}$	-0.065	0.519	-0.125	0.901
$\beta_{StudioSony}$	-0.082	0.413	-0.199	0.843
$\beta_{StudioUni.}$	0.191	0.551	0.347	0.730
β_{Budget}	0.006	0.002	3.306	0.002
β_{Tomato}	0.039	0.003	12.595	0.000
$\beta_{ComicMarvel}$	0.587	0.468	1.255	0.215
$\beta_{Budget:ComicMarvel}$	-0.006	0.002	-2.307	0.025

At an $\alpha = 0.05$, Budget, Tomato ratings, and the interaction between Budget and Comic are significant to the model. The estimate for the coefficient corresponding with Budget indicates that, all else held constant, and assuming that the Comic is DC and the studio is WB, as the Budget increases by $1 million, the fan's ratings increase by 0.006 points. The interaction tells us that the effect of budget on fan ratings significantly changes when the comic is Marvel instead of DC. This coefficient indicates that when the movie is based on a Marvel comic, the fan rating changes by 0.006 less than it would have changed for DC given a $1 million increase in budget. So, we can say that the average change in fan ratings given a $1 million increase in budget for Marvel movies is 0 (Budget+Budget:ComicMarvel=0.006-0.006). The professional Tomato ratings also significantly affect the model. All other variables held constant, and assuming that Comic is DC and studio is WB, as a Tomato rating increases by 1, the fan ratings will on average increase by 0.039.

7 Discussion

In order to answer the different research questions in this study, I employ the Wald test, which uses an F statistic and p-value to determine significance in hypothesis testing. The F statistic is calculated as:

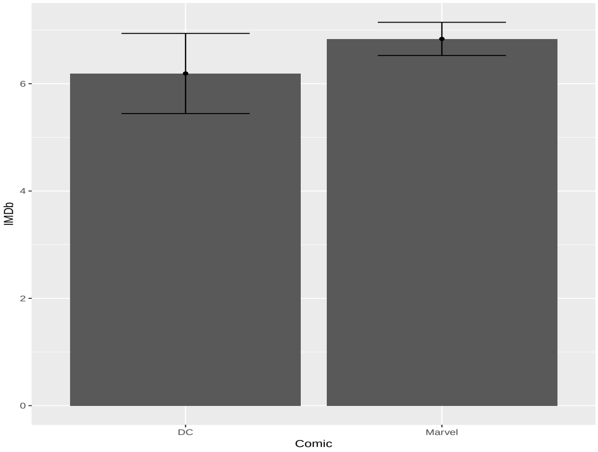
$$F = \frac{(C\hat{\theta})'C(W'W)^{-1}C'(C\hat{\theta})}{qs^2}, \tag{1}$$

where q is the number of columns in the contrast (C) matrix. Note that the coefficient vector in this function is not the same as the vector defined in the previous section, and that the W matrix is not the same as the X matrix in the model. These are from the cell means model. Therefore, in order to test my research questions, I convert the reparameterized coefficients listed in the previous section into the parameters for the cell means model. I perform this conversion using the "M" matrix, so that $C\theta = CM\beta$, where θ is the parameter vector for the cell means model, and β is the parameter vector for the reparameterized model. This "M" matrix is also used so that $C(W'W)^{-1}C' = CM(X'X)^{-1}C'M$. The M matrix is implemented throughout this section to answer the research questions.

For all of the analyses that use the Wald test with the contrast matrix, the answers are verified using the full-and-reduced models method. This method, while its theoretical basis is different, gives equivalent output as the Wald test. This method takes in as arguments a full model with all of the variables and a reduced model, which excludes the variable whose significance is being tested. The method returns an F statistic and p-value, which are the same as the Wald test output. This test is used solely for the purpose of checking the results of the Wald test to ensure accuracy in the analysis

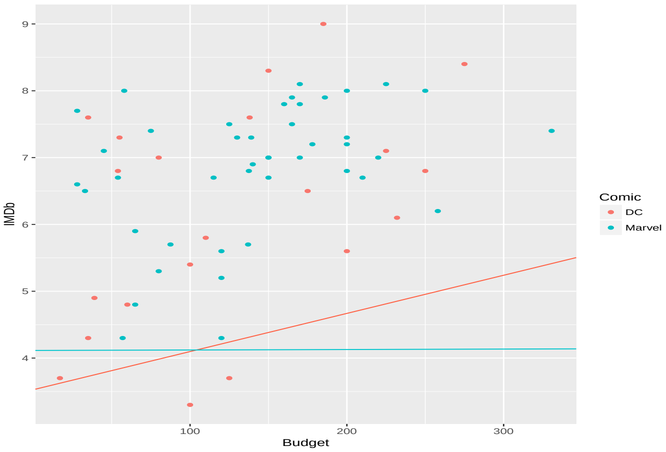
The main question of interest in this analysis is whether one comic company's movies has more popular characters than the other. Employing the Wald test, I evaluate the main effect of the Comic variable, and obtain an F statistic of 2.67, with a p-value of 0.08. With these values, I fail to conclude that one Comic book company performs significantly better than another. This is visually demonstrated in the mean and error bars plot in Figure 3, where the confidence intervals all overlap, indicating that there is no clear difference in IMDb ratings between the two comic companies. This result is interesting given the constant debate regarding which comic company performs better. This ongoing debate seems to be justified, given that this analysis cannot conclude that one has significantly more popular characters than the other. However, there may be other interesting conclusions to make regarding what affects fan popularity.

Fig. 3. Main Effect of Comic Company on Fan Ratings



Although the main effect of comic companies is not significant to the model, there may be an interaction effect. I again employ the Wald test to find whether there is a significant interaction between comic company and the movie's budget. The interaction test indicates whether the slopes in the DC and Marvel models that regress IMDb scores on Budget are significantly different. The interaction between these two variables is plotted in Figure 4. The F statistic for this interaction term is 5.32, with a p-value of 0.025, which indicates that the interaction between comic company and a movie's budget is significant to the model. As discussed in the previous section, the coefficient corresponding to that interaction term means that for Marvel movies, the slope of the regression line of Budge on IMDb scores is significantly less than the slope for DC movies. The budget of Marvel movies does not matter as much to their ratings as the budget for DC's movies, which have a positive slope along IMDb ratings.

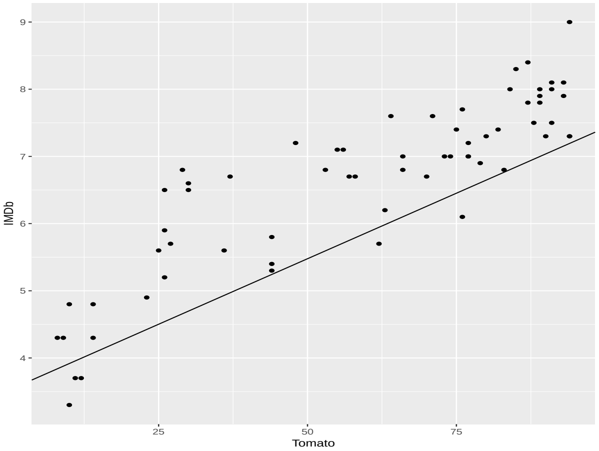
Fig. 4. Interaction Effect Between Comic Company and Budget



While budget has a significant main effect (F=5.47, p-value=0.007), concrete conclusions cannot be made regarding its impact alone because its level of effect is dependent on which comic company we are evaluating. The plot in Figure 4 indicates that budget does not affect average IMDb scores at all for Marvel movies. However, for DC movies, increased production budgets appear to improve the ratings quite significantly. This indicates that it may be more worth the cost for DC to invest more money in its movies than Marvel if their concern is character popularity from the fans.

Although definitive conclusions cannot be made regarding how budget affects the overall model, the professional ratings from Rotten Tomatoes are highly significant, with an F statistic of 158.64, and a p-value ¡ 0.000. The coefficient associated with these professional ratings is positive, so from the model we can conclude that, holding all other variables constant, as professional ratings increase by 1, fan ratings also increase by 0.039 on average. The plot in Figure 5 shows the average increase in IMDb scores for unit changes of professional rating scores, holding all else in the model constant. This indicates that it may be beneficial for both comic book companies to invest more resources into ensuring that their movies are appealing specifically to the professional critic community.

Fig. 5. Average Effect of Tomato on IMDb (Holding all else constant)



This research also evaluates how the studio that the movie was produced in affects the fan ratings of the movie. The mean and error bar plot in Figure 6 shows the distributions of the IMDb scores for different studios. As is indicated by the heavy overlap in the confidence intervals, the Studio variable is not significant on any level to the model (F=0.899, p-value=0.515). Studio is modeled as an additive term, meaning that the coefficient corresponding with each Studio represents the change of the regression line intercept from the baseline, which in this case is the WB studio. Because none of the coefficients are significant, we can conclude that we do not have enough evidence to assume a significant effect of studio on the fan ratings of the movie. Another way to represent the additive effects of the model is plotted in Figure 7, which shows how the y-intercept changes for different studios. These changes are very small, so again it is unsurprising that the results were not significant.

While the results of this analysis are not conclusive regarding which comic book company has more popular characters, the results provide a significant contribution to the Marvel vs. DC debate. These results actually lead to more interesting and informative conclusions. Given that one comic company is not significantly better than another, we can begin to evaluate what makes their

Fig. 6. Effects of Studio on IMDb Ratings

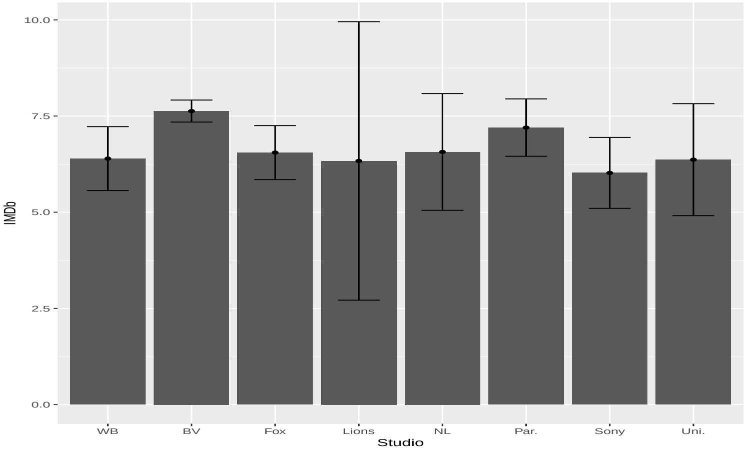
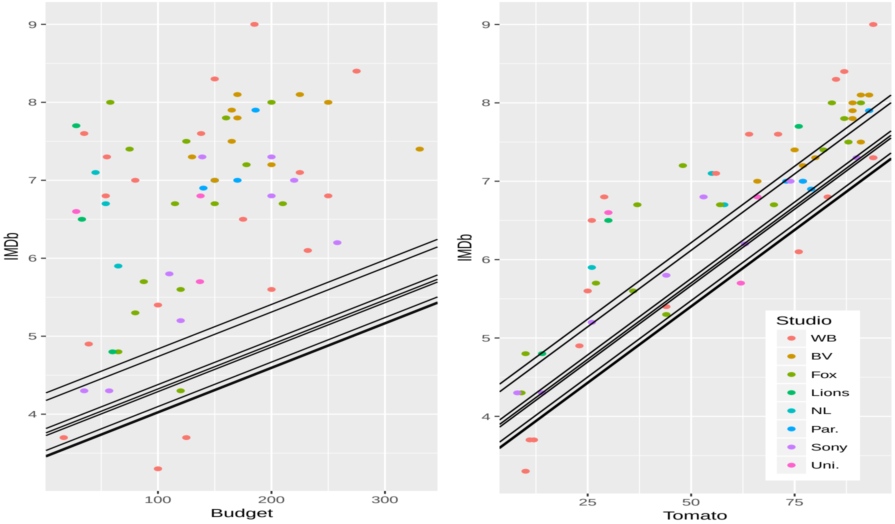


Fig. 7. Additive Effects of Studio on the Model



individual movies' characters more popular. First, I find that their is a significant interaction effect between comic book company and budget, and that higher production budgets are more beneficial to DC than to Marvel in increasing fan ratings. While this result may not drastically change the amount of money that the companies can spend on their movies, this can help inform the companies when there is a marginal decision to be made regarding spending. DC benefits more in terms of fan ratings by investing more in their movies, and alternatively Marvel doesn't improve their average fan ratings by investing more, so that it may not make as much sense for them to add marginal spending to their movie production. I also find that the professional critic ratings are significant to the character popularity of the movies. This is true for both comic companies. Making a movie appealing to professional critics is a useful strategy in increasing character popularity.

8 Summary

The findings of this study are useful in both systematically framing the ongoing debate regarding the two comic companies and in informing the companies on the best ways to improve the popularity of the characters in their movies. Future research should continue this evaluation as these two companies continue to produce more superhero movies. Conclusive results regarding which comic company has had more successful characters may be possible in the future as more movies become available to add to the analysis. Future studies should also evaluate the reasons for some of the significant effects found in this study and should attempt to explain more explicitly what the comic companies should do to improve fan ratings. For example, it would be useful for the companies to understand what qualities of a movie significantly improve professional critics ratings.