**Analyzing Marvel Heros and Possible Predictors for Their Deaths**

Karn Jongnarangsin

**1 Introduction**

Since its inception in 1939, Marvel Comics has produced an overwhelming number of heroes, anti-heroes, side-kicks, and storylines. As superhero stories often involve battles of epic proportion, it is not uncommon that some characters eventually perish. However, death in Marvel comics is, more often than not, temporary and the deceased are brought back in some way, shape, or form. This in turn raises a few questions: are there variables that may predict if a character’s death is due, and can these variables also predict whether a deceased character will be brought back? These questions were answered by analyzing data sets compiled by fivethirtyeight regarding the topic of Avengers’ deaths [1].

**2 Related Work**

In fivethirtyeight’s analysis of Avenger deaths in the comics, data pertaining to Marvel heroes who were part of the Avengers were compiled from the Marvel fan wiki at the time of publication. From this data, they were able to find statistics pertaining to the number of character deaths and associated resurrections. Out of the 173 Avengers observed in this study, the data showed that 69 Avengers or roughly 40 percent had died at least once in their career.

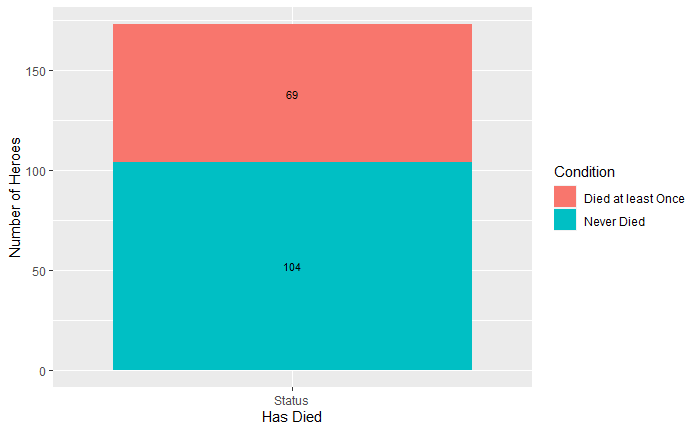


Figure 1: Comparing numbers of Avengers who have never died and those who have died at least once

Furthermore, the number of deaths, including subsequent deaths following resurrections, totaled to 89. Out of those 89 deaths, 57 deaths were followed by a resurrection. Delving further into those statistics, fivethirtyeight found that the proportion of heroes who were resurrected after dying for the first time was greater than the proportion of heroes who were resurrected after dying a second time. These proportions were 66% and 50% respectively. Repeat analysis had results consistent with the original findings and it was also noted that claims towards being resurrected after a third, fourth, or even fifth death are difficult to make as the sample size for these deaths proved to be significantly smaller.

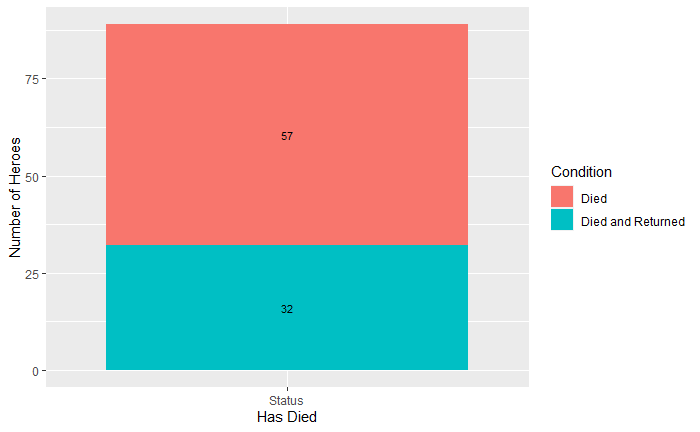


Figure 2: The total number of deaths and the proportion of those that were followed by a return

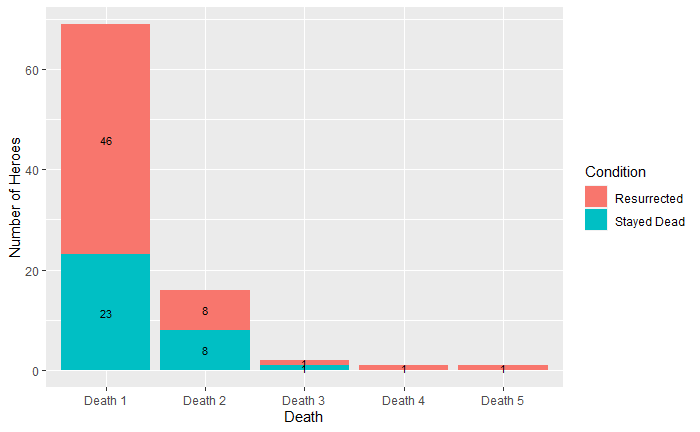


Figure 3: Subsequent deaths and resurrections

**3 Dataset**

Continuation of analysis on Marvel Avenger deaths was performed using the data that was originally compiled by fivethirtyeight. This dataset includes 173 observations, one per hero, and each data entry contains 24 variables. These variables include the URL to their associated wiki page, their name or alias, the number of appearances, whether or not they are a current avenger, the date at which probationary status as an Avenger may have been given, the month and day for when they became an Avenger, the year, the number of years since joining, whether or not they are a full Avenger, 5 separate variables corresponding to 1-5 deaths, and 5 more variables corresponding to 1-5 resurrections.

**4 Methods**

Although fivethirtyeight provided many statistics towards the topic of Marvel Avengers’ deaths, they do not provide much in the way of finding relationships between the number of deaths and other observable variables. To further expand on the analysis performed by fivethirtyeight, the questions sought to be answered were if the variables in this dataset- excluding the URL, name/alias, and dates- could be used to predict the number of deaths that a character would have and also to predict the number of resurrections, likewise.

As these each question entails multiple variables affecting a single response variable, the models were created from multivariate linear regressions, ridge regressions, and lasso regressions. These models allowed for determining which variables held the most significance and demonstrated how isolating these variables could create improved prediction models. In total, six models were created with each of the three types of regressions used for each question. For each of these models, the data was randomly split into training and testing sets for model validation.

For the linear models for each of the two questions, these models served as a baseline that the subsequent models could be compared with. These models included the full range of variables used to predict the response variable. Performing the regression was as simple as utilizing R’s ‘lm’ function, specifying the response variable, which differed for each question, and predictor variables in the function call, and specifying that the training data subset be used. Afterwards, predictions using this model were made for the testing data and the testing R2 and MSE were calculated to be used for comparison to the subsequent models.

The ridge regression models created penalties for collinearity, and the model saw greater variable isolation than that of the linear model. For each question, the model was created with a lambda tuning parameter selected by using cross validation with a threshold of 1 \* 10-12 and using 10 folds, an alpha of 0 for ridge regression, the corresponding response variable, the same set of predictor variables used in the linear model, and using data from the training subset. From this point, the final models for each question were then also used to make predictions for the testing data, with testing R2 and MSE calculated.

Finally, the lasso regression models performed shrinkage to create a more sparse model and so the model saw even greater variable isolation than that of the ridge regression. Creation of the models was similar to that of the ridge model, with parameters being the same with the exception of alpha being 1 for lasso regression instead of 0 for ridge regression. As was the procedure for the linear and ridge regression models, this model was also used to make predictions for the testing data, with the testing R2 and MSE being calculated afterwards.

**5 Results**

Starting with the models that predict the number of times a character will die, the linear model saw the number of appearances, the years since joining, and being a Full Avenger as significant predictors for the model, with a significance of < 0.01. Adjusted R2 was -0.03663622 and the test MSE was 0.8367952.

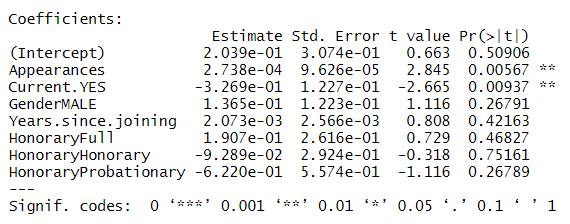


Figure 4: Linear regression coefficients for Deaths

For the ridge model, the best lambda was found to be 0.3764936. A notable difference in coefficients is seen in the “years since joining” predictor having a much smaller effect, as it may have collinearity with other variables. Adjusted R2 was -0.03931852 and the test MSE was 0.6049598.

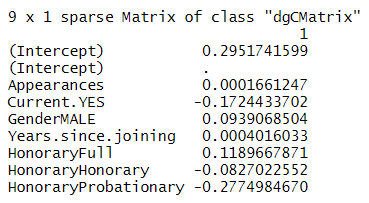


Figure 5: Ridge regression coefficients for Deaths

The lasso model for predicting the number of deaths had a best lambda of 0.03053856. From this model, it can be seen that only the coefficient for “years since joining” has been zeroed out. Adjusted R2 was -0.03256476 and the test MSE was 0.5972518.

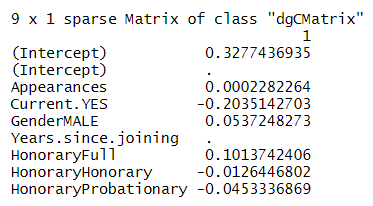


Figure 6: Lasso regression coefficients for Deaths

In regards to the prediction models for the number of times heroes return from the dead, the linear model saw the number of appearances, years since joining, and status as a full Avenger as significant predictors with significance of < 0.01. Adjusted R2 was -0.005140219 and the test MSE was 0.602403.

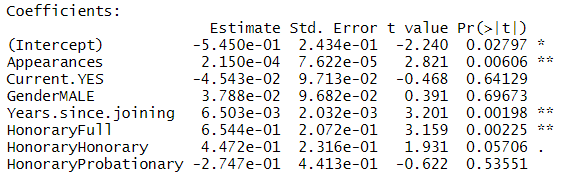


Figure 7: Linear regression coefficients for Returns

The ridge model for predicting the number of returns had a best lambda of 0.1232847. In comparison to the linear model, there was an observable reduction in contribution for the variables of being a full avenger as well as the number of years since joining. Adjusted R2 was -0.01473043 and the test MSE was 0.6046697.

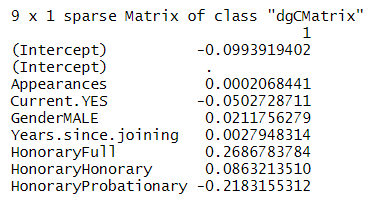


Figure 8: Ridge regression coefficients for Returns

Finally, the lasso model for predicting the number of returns had a best lambda of 0.03053856. This model had zeroed out the coefficients for gender, having honorary status as an Avenger, as well as having probationary status as an Avenger. Adjusted R2 was -0.01948241 and the test MSE was 0.6085919.

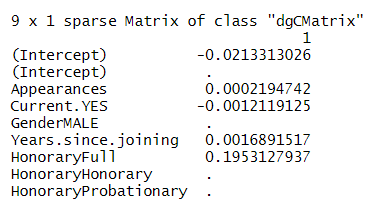


Figure 9: Lasso regression coefficients for Returns

**6 Discussion and Conclusions**

With the models presented, it appears that the regressions did not seem to be a good fit with the data. When looking at the adjusted R2 values for the linear, ridge, and lasso regression models for predicting number of deaths, the values were -0.03663622, -0.03931852, and -0.03256476 respectively. The adjusted R2 values for the linear, ridge, and lasso regression models for predicting number of returns did not prove to be any more compelling, with values of -0.005140219, -0.01473043, and -0.01948241 respectively. As all adjusted R2 values are negative, this means that the inclusion of multiple variables for each of the regressions did not help the models any more than by chance. The similar values between linear, ridge, and lasso are interesting as that means even after isolating variables, the prediction models did not improve correlation to the underlying data.

In terms of models being improved when compared to another, prediction accuracy did appear to improve slightly for deaths. The test MSE for the linear, ridge, and lasso models for predicting number of deaths were 0.8367952, 0.6049598, and 0.5972518 respectively. This shows that the ridge regression had an observable decrease in test MSE in comparison to the linear model, and lasso only saw a marginal decrease when comparing to the ridge. However, differences in test MSE for the linear, ridge, and lasso models for predicting number of returns was not as distinct. The test MSE for the linear, ridge, and lasso regression models for predicting the number of returns were 0.602403, 0.6046697, and 0.6085919 respectively and this shows that the variable isolation did not have a great effect towards minimizing testing error.

In terms of inferences that can be made from these models, the performance of the prediction models show that the variables observed in this analysis are not correlated to whether a character is more likely to die as well as whether they are likely to be revived afterwards. This would imply that these variables do not have distinguishable predictive value towards the number of deaths and number of resurrections.

For future studies, investigation of different variables may prove to be more insightful towards whether heroes are more likely to be killed off and/or resurrected. Variables of interest include real-world statistics, such as character popularity ratings, sales figures for the comics that these characters appear in, number of different writers responsible for a character, etc., as well as fictional statistics, such as various power ratings.

**References**

1. Hickey, Walt. “Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building.” *FiveThirtyEight*, FiveThirtyEight, 12 May 2015, [fivethirtyeight.com/features/avengers-death-comics-age-of-ultron/.](https://fivethirtyeight.com/features/avengers-death-comics-age-of-ultron/)  
   Data repository: <https://github.com/fivethirtyeight/data/tree/master/avengers>