Hayden Lamprecht

Professor Keith

Econ 2810:001

12/3/24

MLB Hall of Fame

The question surrounding my research for this project is to find if there is a correlation between a baseball player's stats and their hall of fame probability. This is a very broad question and needs to be reorganized to allow for the use of data to answer it. To do this, I am going to set up hall of fame probability as the dependent variable, and the two statistics I will use as independent variables are wins above replacement and all-star game appearances.

First of all, wins above replacement, more commonly known as WAR, is currently the leading stat in Major League Baseball when it comes to measuring a player's overall value. The reason WAR has become one of the more important statistics in baseball is because it is calculated for both pitchers and batters. Whereas many other statistics are independent to what position a player plays. For example, batters do not have pitching stats and pitchers do not have batting stats. This obviously makes it difficult to compare certain players' value to their team, which is where wins above replacement comes into play. The calculation of WAR is essentially runs contributed via batting and baserunning and runs saved via fielding divided by their team's total runs with some league average adjustments. For pitchers it is the number of wins they have contributed based on individual statistics and total innings pitched. The second independent variable in this model is all-star game appearances, which is somewhat self-explanatory. Each season there is an all-star game, where the top players for each position get voted in. This

obviously measures a player's value to their team, position, and even the league. To quantify the dependent variable, hall of fame probability, I used the number of years between retirement and hall of fame induction. This obviously rules out a couple demographics. We will only be using players who are currently retired and in the hall of fame. The current qualifications to be in the hall of fame are to have played ten seasons in the MLB, been retired for five years, and receive 75% of the votes. Prior to 1954 the qualification was to only have been retired for 1 year, and even earlier there was no qualification for retirement years. Because of this, in my data there were players who were below the ten seasons played mark, and the five years retired mark. They were removed from the sample to cause more accuracy in the model. Beyond that, the all-star game did not exist until 1933, so I also removed any players whose career began before 1933, because some would have zero all-star appearances.

My data was taken from baseballreference.com, which is the top website for finding any current or all-time baseball statistics. The data I pulled was the top WARs of all time, as well as a second set of data for hall of fame players and combined the data in one set. After this, I deleted any players who were not yet retired or in the hall of fame, as well as deleted the players who did not meet the current hall of fame qualifications or made their debuts prior to 1933. After this I was left with one hundred and thirty observations. This set had the year each player retired and was inducted into the hall of fame, so from there I simply took the difference. It also included each players all-star appearances, which was very convenient. Before calculating descriptive statistics or running the regression, I decided it was best to use average WAR per season instead of total WAR because certain players play twenty seasons or more while others may only have ten to fifteen. To calculate this in the data I divided total war by number of seasons played to get each player's average WAR per season. The data that I have for this research question is time

series data because these are player's stats that are measured each season and compiled over time. The first player in this data set began their MLB career in 1933 and the most recent retired in 2018. The descriptive statistics showed WAR per season has a mean of 3.65 with a standard error of 0.096. The variable all-star game appearances has a range of 2 to 25 and a mean of 9.68 with a standard error of 0.39. The dependent variable years to hall of fame has a range of 6 to 59 with a mean of 14.38 and a standard error of 1.09.

Before running the regression, it is important to indicate my model and my expectations for the model. The model is as follows:

Years to HOF = B_0 - B_1 WAR/Season - B_2 All-Star Appearances + e

My expectations are for both signs to be negative because as each dependent variable increases, you would expect years to HOF to go down. For example, a player who has a higher WAR/Season and more all-star appearances will get into the hall of fame quicker than a player who has lower of both stats. The regression results aligned with my expectations. Both variable coefficients are negative. The model yielded an adjusted r squared of 0.1015; meaning that 10.15% of the variation can be explained in this data. This model also has an F stat of 8.28 with a very low significance, meaning this model is significant. The regression model is as follows:

Years to HOF = 29.06 – 3.12 (WAR/Season) – 0.34 (All-Star Appearances) + e

This states that a one unit increase in WAR/Season will decrease years to hall of fame by 3.12 years and one additional all-star appearance will decrease years to hall of fame by 0.34 years.

To conclude this research project, the results follow my expectations of both variables being negative. It should be noted that WAR/Season is more influential, and all-star appearances are not significant, with a p-value of 0.185. Seeing as the model has a low R squared, I assumed

that I ran into multicollinearity. I went ahead and tested using a correlation matrix in excel to test the correlation between WAR/Season and all-star appearances. The correlation matrix provided a calculation of 0.3799, which indicates that the independent variables are not strongly related to each other. I also calculated the VIF, which came up as 1.13, which also does not indicate any multicollinearity. More than likely, the reason for the low r squared is the model not being a strong predictor, not having enough observations, or not having enough variables. To fix this I could replace all-star appearances with some other variable or add a third or fourth to see what kind of results I may get.