Zach Palmer
Dr. Nelson
STAT 1341 – Sports Analytics
7 November 2021
~2200 Words (sorry)

Midterm Paper

Introduction

Question: What impact does aging have on an athlete in a given sport?

I believe that the impact of aging on athletes in various sports is relevant as it offers valuable insights to coaches, GM's, managers, etc. in terms of roster decisions. Similar to how we discussed the impact of Moneyball, and the roster building strategies mentioned therein, being able to assess how much you can expect out of aging players and having a better understanding of how players will potentially regress in their later years allows teams' management to make smarter, more informed decisions. Additionally, knowing a typical player's trajectory in terms of performance is vital in roster building because of the "championship window" that you will sometimes hear people colloquially refer to. Attempting to align the individual peaks of your key players is a crucial aspect of roster building and answering this question will likely offer useful information in that department. Finally, I think that another interesting and valuable insight that researching this question can offer is how the effects of aging may differ amongst the various positions withing a given sport. For instance, does aging affect a striker in soccer more than defender? If it does not, then is this idea simply a false misconception, or could there be another cause in the difference between players' longevity—perhaps their particular playing style? Regarding any research I know of that has been done in this area, I know a significant amount of research has been done on the impact of aging on baseball players, with many articles discussing said research on FanGraphs (I included links to many of the articles I looked at on the very last page of the report).

Data Collection

I obtained the final datasets I used in my analysis of aging in baseball from the FanGraphs website. I used the leaderboards feature and specified that I wanted the statistics for qualified players from the 2006 season to the 2021 season. I also further customized the leaderboard by adding a few more statistics I knew or thought I might need (e.g. age, batting runs, etc.) using the tool at the bottom of the page. I repeated this process for both batting and pitching stats and then exported the resulting leaderboards to generate the two raw datasets I used for my analysis. In addition to cleaning the dataset for incomplete observations in the traditional fashion, due to the unique nature of the question I am attempting to answer—the impact of aging—I only kept the subsets of each dataset that would provide useful insights into how aging affects players. Any observation—a player's statistics for a single season—that does not have a sibling observation, that is to say another observation representing that same player's statistics for the previous or following year, is essentially useless in our analysis of the impact of aging. We are not concerned with how well a player of a given age typically does, but rather with how aging impacts their performance. Therefore, we only want to keep observations that represent one of two consecutive seasons by the same player so we can investigate how their performance changed as they grew one year older. Finally, for both position players and pitchers I removed those columns that I wouldn't be investigating using the dplyr select() function to reduce the amount of clutter and unnecessary information in both data frames.

Descriptive Statistics

Before I dive into my examination of the descriptive statistics I included in my analysis of aging in baseball, I should note that my case is somewhat of a unique one, as I am not following the traditional format of creating a model that uses certain statistics to predict a specific response. Since I am investigating the impact of aging, what I am doing is looking at how certain statistics change on average for players as they grow a year older. Moreover, I am looking at how this change differs depending on the ages they are transitioning between—are they young and simply getting closer to their "prime" or are they old and getting a year closer to potentially retiring. With this in mind, in my study of descriptive statistics I simply generated summary statistics and visualizations for the statistics for which I would be creating aging curves in the following inferential section.

For position players, I decided to look at how the following statistics changed as players grew older: wOBA, batting runs, baserunning runs, defensive runs, and WAR. The thought process behind my choices was the following: wOBA gives us a general idea of a player's batting ability as a whole, WAR gives us a general idea of the total value they offer to their team, and the three other statistics may potentially offer insights into how aging could affect different skills at different rates. To start with, I generated a "table" of summary statistics, that is to say I created a one row data frame using summarise() that contains the mean, median, and standard deviation of each of the statistics I chose to investigate. The table does not really tell us much, but it provides us with a general idea of the center and spread of each of the statistics as well as what values they could typically be expected to take. The scatterplots I created of player age versus each of the statistics demonstrate that each of the variables has a fairly uniform spread/variability regardless of age, although at the extreme age values we do see some differences, likely amounting to the small sample sizes. Regarding the histograms, I notice two distinct groupings. The histograms for wOBA, as well as baserunning and defensive runs, appear to be normally distributed, whereas the histograms for both batting runs and WAR seem to be right skewed, with the skew being slightly more pronounced for batting runs.

For pitchers, I decided to look at how the following statistics changed as players grew older: vFA (avg. 4-seam velocity), WHIP, xFIP, and WAR. My rationale for these choices was as follows: vFA provides us with an idea of a pitcher's overall throwing power, WHIP provides us insight into the number of opportunities (walks and hits) they give the opposition per inning, xFIP provides information about the number of runs we would expect them to give up per inning, and WAR provides us with a good sense of the total value they offer to their team. Just as with the position players, I created a one row data frame using summarise() that contains the mean, median, and standard deviation of each of the statistics I chose to investigate for the pitchers. It does not provide many insights but is useful in offering a general synopsis of the statistics. Similar to the scatterplots for position players, we see about the same variability in the statistics regardless of age except at the extreme values, where the difference in variability/spread is especially apparent for the pitchers. Additionality, for vFA I would note that we see a slight but noticeable general downward trend as pitchers age. With the histograms, we see that WHIP and xFIP are normally distributed, WAR appears slightly right skewed, and vFA is left skewed.

Inferential Statistics

To answer my research question about the affect of aging on baseball players, I created aging curves for each of the statistics for both position players and pitchers. I accomplished this by iterating over each of the data frames and calculating the difference between the statistic values for a given player across consecutive seasons. I then took the average of those differences for each age range/interval. Using these average differences, I created plots of the average change in each statistic at each age range/interval to visualize the affect of aging on both position players and pitchers. I also printed out the sample sizes for each of the age range "bins".

For position players, we see that, generally speaking, players appear to improve very little or not all once they enter the league. The aging curves for wOBA, batting runs, and WAR all demonstrate small gains or no change in each of the statistics from approximately ages 20/21 to 26/27 before we observe a steady increase in the rate at which the statistics decline as players age past 26/27 years old. The three graphs are remarkable similar, although this makes sense given that batting runs is derived from wOBA through RAA, and batting runs is a component of WAR. Despite what I would've expected, the aging curve for defensive runs seems to suggest that aging has little to no impact on a player's defensive abilities and/or contributions as all of the values hover around zero. The curve for baserunning runs shows a steady, gradual decline in baserunning ability of about 0.5 runs per year, which is essentially what I expected as players generally slowly lose their speed as they age.

The aging curves for pitchers are more interesting/unusual (perhaps because of the smaller sample sizes). I should first note that my analysis of the curves will ignore age ranges after 37-38 because of their incredibly low sample sizes (I probably should've removed them). With average 4-seam velocity, what's fascinating is that it actually seems like there is little to no change in velocity until pitchers reach around age 30 at which point we start to see their throwing power slowly fall off. WHIP seems to remain fairly constant throughout a pitcher's career as all of the values hover around 0 and fluctuate between being positive and negative. With xFIP we actually see a slow gradual improvement until a pitcher reaches around age 28, where thereafter their xFIP appears to only gradually increase until they retire with what looks like an average increase of about 0.1 a year until age 38 where the data becomes incredibly volatile. We see about the same trend with pitcher's WAR as we do with position players, WAR stays about the same and actually increases somewhat until around age 26 at which point we see a decrease in WAR at gradually increasing rates until retirement (a weird anomaly at 34-35 with N = 20).

Discussion/Conclusion

What I have learned is potentially valuable to managers and front offices of baseball teams as they can use these aging curves to project the future performance of players they currently have rostered or players they are considering adding in order to determine whether a given player is worth the price they are asking for. For example, consider you are thinking about acquiring a 4 WAR (last season) player who is 28 years old. Based on the WAR aging curves, regardless of whether he is a position player or a pitcher, as he ages he will lose about 0.5 WAR on average per year. Thus, if you are signing him to a 3-year contract you can project him to offer your team WAR values of 3.5, 3, and 2.5 over those 3 seasons for a total of 9 WAR. Now all you have to do is decide how much one win is worth to you and you have a baseline approximation of his value. Moreover, the aging curves can offer useful insights to team managers/management. for instance, that they don't necessarily need to worry about a position player's defensive capabilities decreasing, or that you typically shouldn't expect a player to improve much, if at all, once they make it to the major leagues.

In a broad sense, yes the results of my investigation into the impact of aging in baseball are generalizable to other sports. What I mean by this is that, while the exact rate at which the average player declines or reaches their peak may differ from sport to sport, the overarching idea of understanding the typical progression of a players' abilities and their eventual regression is essential to projecting players' future performance and thus roster building.

If I was to mention the limitations of my project, the first thing that comes to mind is the sample size. Preferably, I would've been able to have a larger sample size and I considered lowering the requirements to be including in the dataset; however, if we ignore the extreme values I am fairly happy with the number of observations I have for the age "bins" from ages 21 to 37 for position players and ages 22 to 35 for pitchers.

Regarding ideas for future research, these mostly stem from areas of interest that I wanted to delve further into, but simply ran out of time to. I would be fascinated to see how aging curves differ by sport and to use the insights provided by these potential differences to draw conclusions about the various natures of different sports. Perhaps in sports like basketball, hockey, or soccer, where players enter the top tiers of competition at such young ages, we would see very different aging curves as the average player may need more time to mature and reach their full potential. Beyond just looking at different sports, I think it would be worthwhile to break down the question further and to study the impact of aging on smaller subsets of players with similar playstyles or strengths. It is a common belief amongst sports fans that players who rely on their superior athletic ability and physicality tend to regress earlier and at a faster rate than their more skilled and "intelligent" counterparts. Do claims such as these actually have statistical backing? I think it would be very interesting to find out.

Resources/Inspiration

The Beginner's Guide to Aging Curves: https://library.fangraphs.com/the-beginners-guide-to-aging-curves/

Are Aging Curves Changing?: https://blogs.fangraphs.com/hitters-no-longer-peak-only-decline/

How Do Baseball Players Age?: Investigating the Age-27 Theory: https://www.baseballprospectus.com/news/article/9933/how-do-baseball-players-age-investigating-the-age-27-theory/