Eamon Deffner

Stat 287: Data Science

Professor Nicholas Cheney

3 March 2023

Project Proposal

Motivation

Baseball is a game in which speed, which was often valued as an important factor in a player’s skillset, is no longer as sought after as an attribute. However, players who can beat out ground balls get on base more, leading to the driving question for this project. What is the contribution towards runs scored of speed from home to first in baseball?

Relevant Work

The rise of sabermetrics, or advanced baseball statistics, has led to multiple studies on the importance of baserunning. In [3], the authors analyze how Rickey Henderson, the most prolific base stealer of all time, would have had his impact on the game changed if he was not fast. The authors then analyzed how and when teams should employ the stolen base, and determined that in close games in the later innings is the best time. In [2], baserunning was analyzed by creating vectors for each baserunning event and comparing how often a player attempts and is successful at each compared to the average player. It was found that even the best baserunners can only contribute about 1 win more per season to their teams. In [1], a similar analysis was done, except using nine person lineups of the same player. They found that while some players need to be aggressive on the bases, others should not because their offensive profile is best used for power. All three of these studies conclude that the value of baserunning is overrated. However, none of them look at the impact of speed to first base, which is overlooked when looking at how speed affects the game. However, it may have the greatest impact on the game.

Data Collection, Cleaning, and Exploration

All data for this project will be extracted from baseballsavant.com using the Statcast Search tool. Using this tool, statistics will be collected on batting average above expected ground balls, slugging above expected for hits (that are not home runs), as well as sprint speed from the statistics part of baseballsavant.com for the years 2017-2022. Each of these will have a minimum number of 50 attempts to eliminate players with too few results. The data will be collected into CSV files, which are easily downloadable from the site. As preliminary data exploration and cleaning, I will concatenate all the data frames into one and use mathematical feature engineering to create columns that are easier to use for my analysis.

Modeling and Analysis

I plan on creating two linear regression models using sprint speed to predict both batting average above expected on ground balls, as well as slugging above expected on hits that are not home runs. All models will be built from 2017-2021 and using 2022 as a test. This data will be split by simply adjusting the features on Baseball Savant. Built-in functions from scikit learn will be used to create the models, and mathematical tools in Pandas will also be used to order the tables and create new statistical columns. Data will be visualized for the models using seaborn, and tables will also be created to show the best and worst performers in each category, as well as their contributions to runs created from these two statistics. In order to make sure the relationships are linear, I will create residual plots for each model created and adjust them based on the shape.

Needs Analysis

Creating the two linear models will form the first step of the project. By doing this, I will find the added contribution to run scoring from each additional foot/second of sprint speed to determine the true value of speed from home to first. Using the models and numpy functions, I will calculate how many runs a player with an “average” offensive profile contributes if his speed is fast, average, or slow, compared to the league average. This will be done by adding an average performing player to the product of how often each event happens and its added contribution to on base percentage or slugging (it is known that on base plus slugging (OPS) has about a 0.9 R^2 with runs scored). If the model appears to have distinct groups for players, I would like to cluster players using their speed and other parts of their profile and be able to predict their actual OPS based on sprint speed, batting average above expected and slugging percentage above expected. This model may be easier to complete without using clustering, but clustering the players would help to eliminate the variance of offensive profiles for players of similar speeds. This would be shown in a graph showing the value of speed to each cluster. This would form the final step in a minimal viable project to determine how speed impacts run scoring. A stretch goal would be to take salary data from single seasons and use it to see how players with an above average speed from home to first are paid relative to those with average speed. I would use bootstrapping and clustering here to predict the difference in salaries of similar players with differing speeds but similar run contributions. This is a loftier goal, and could only come after completing the other steps, but could be accomplished because salary data is easily available.

Timeline

Week 1: Extract data from baseball savant, create data frame and do some EDA

Week 2: Create linear regression models, and confirm linearity of relationships

Week 3: Find players who have the highest addition to their offensive profile from speed, create a model to determine how much of a role how speed, batting average above expected, and slugging above average on hits determine a player’s OPS.

Week 4: Compare how speed affects OPS for average, fast, and slow players. This will be done by building a fake player with a given speed and an offensive profile particular to his speed.

Week 3: Use clustering to run the same regression as in week 3, breaking players into groups, and then see if speed matters in different ways for each group of players, and how this affects their overall offensive contributions.

Week 6: Read in a file with salary values for players from 2017-2022. Cluster players based on OBP, SLG, and speed, and use a linear model to determine how speed plays a role in how players are paid

Works Cited

[1] Baumer, Ben S.; Piette, James; and Null, Brad. 2012. Parsing the Relationship Between Baserunning and Batting Abilities Within Lineups. *Statistical and Data Sciences: Faculty Publications*, Smith College, Northampton, MA. <https://scholarworks.smith.edu/sds_facpubs/40>

[2] Baumer, Ben S; and Terlecky, Peter. 2010. Improved Estimates for the Impact of Baserunning in Baseball. https://www.researchgate.net/publication/265059918\_Improved\_Estimates\_for\_the\_Impact\_of\_Baserunning\_in\_Baseball.

[3] Click, James, et al. 2007. “What if Rickey Henderson had Pete Incaviglia’s Legs.”  
 *Baseball between the Numbers: Why Everything You Know about the Game Is Wrong.* Basic Books, pp. 112-126.