Use of Data to analyze Base Running Statistics

Eamon Deffner*
Frederick Carlson*
edeffner@uvm.edu
frederick.carlson@uvm.edu
University of Vermont
Burlington, Vermont, USA

Figure 1. Seattle Mariners at Spring Training, 2010.

Abstract

How much does base running speed affect offensive statistics? This paper mines data from BaseballSavant on sprint speed and batting averages to answer this question [2]. We use the data science capabilities of anaconda python to answer this question. Using linear regression and clustering, we attempted to find the value of speed. We found that there is little to no advantage to single base hitting due to speed, but we did find that there is an advantage to speed in slugging, and that on average, faster players are able to add to their On Base Plus Slugging (OPS) because of speed. From our clustering, we found that players involved in these hustle plays more often are not necessarily more successful on these plays.

Keywords: baseball, OPS, linear regression, clustering

1 Introduction

Baseball is a sport in which a player's skills are evaluated based on how much they can contribute to either producing and/or preventing runs. Since the publication of *Moneyball* [6] in 2003, the study of statistics in baseball (sabermetrics) has become an important field for the front offices of baseball teams. Being efficient in their choices of players who produce and prevent the most runs will save the team money and win more games. After the rise of sabermetrics, certain statistics cmae to the forefront as correlating well with runs scored for a team. Two of them on the offensive side of the game were On Base Percentage (OBP), which forms one part of On Base Plus Slugging (OPS). OPS correlates well with run scoring, with an R^2 value of 0.9.

Because of statistical analyses, speed is no longer viewed as an important factor in a player's offensive ability. However, speed may play a role in determining a player's ability to get on base, which is one of the statistics that correlates best with run scoring. This paper examines the importance of speed in baseball, specifically the importance of baserunning on a player's own hit. Our project looks to determine a relationship between speed and run scoring, and find if speed does contribute significantly to OBP, Slugging Percentage (SLG), or OPS.

2 Related Work

The rise of sabermetrics, or advanced baseball statistics, has led to multiple studies on the importance of baserunning. In [5], 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. To better measure all kinds of baserunning, the authors created a statistic called Equivalent Base Runs to measure how each baserunning event from a player affected run scoring. They did this by determining how each baserunning event added to run expectation. This was then compared to the average player in the league. From their findings, they determined that for all his speed, Rickey Henderson only added 5 wins to his team over the course of his

*Both authors contributed equally to this research.

career. In [4], 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. They used simulation and linear modeling to predict the average contribution in a nine inning game each player could have. It was found that even the best baserunners can only contribute about 1 win more per season to their teams. In [3], a similar analysis was done using simulation and linear analysis, 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 has been overpaid in years past. 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, speed to first base may have a greater impact on the game than other forms of baserunning.

3 Methodology

3.1 Data Cleaning and Exploratory Data Analysis

Data was scraped from BaseballSavant.com, which has data from every Major League Baseball game. The years 2017-2021 were downloaded as the training set, with 2022 as the test set, with a minimum of 50 hits in play and 50 ground balls. Each player's data had to be scraped in three sections: first, the data for batting average on ground balls; second, for data of slugging on non-home run hits; and third, on sprint speed for both the training and test sets. Each player's data was compiled, merging columns player names by player name. Originally, we had 1012 observations for the data from 2017-2021 (training set), but this was trimmed to 412, as not all players met both the criterion and had enough playing time to have their sprint speed numbers recorded. For the 2022 data (test set), we had 547 observations, but for the same reason, trimmed it down to 260. Age was ignored in our cleaning, as the age would change substantially from 2017-2021, so was left out of our modeling. The statistic of Standardize Slugging was created to standardize slugging above expected to zero. Initial plots were created for batting average above expected on ground balls based on speed, as well as Standardized Slugging based on speed. It was found that there was not much of a relationship for either graph, as slugging had a correlation of 0.464 with sprint speed, while BA - xBA had only a 0.190 r-value with sprint speed.

3.2 Linear Regression Models for Added Hustle OPS

While our initial results showed that speed did not generate an addition to batting average, this did not tell the whole story, as it is unclear how slugging and batting average on our two types of "hustle plays" worked together to impact a player's On Base Plus Slugging (OPS). To do this, we created a feature called Added Hustle OPS, which was BA − xBA ∗ Ground Ball Rate + Standard Slugging ∗ Hit in Play Rate. Added Hustle OPS was the addition to OPS that could be found on the two types of plays analyzed. We then created a linear regression model for Added Hustle OPS using the features of sprint speed, home plate to first base, and number of plate appearances. After viewing the feature importances and p-values, it was determined that home plate to first base was the only practically and statistically significant predictor of Added Hustle OPS (coeff = -0.0645, p-value 0.000). After removing the other features, the model had a mean absolute error of 0.0166, with a coefficient value of -0.0515 for time from home to first.

3.3 PCA and Clustering of Speed

To determine if certain groups of players were more likely to be faster than others based on number of hustle plays, clustering was performed. This was done by dimensionality reduction using PCA on percentage of hits in play and percentage of ground balls. Percentage of hits in play and ground balls were determined to have a strong linear relationship. Before pefrorming PCA, we used the StandardScaler package in scikitlearn to standardize home plate to first base time, ground ball percentage, and hits in play percentage. The amount of clusters was determined by an inertia curve and the difference from the origin of the graph. It was determined using the curve that four was the correct number of clusters. Results are shown in figure 6.

4 Results

Results for our initial models on batting average above expected based on speed and standardized slugging based on speed revealed that there is not a strong direct relationship between these statistics and speed (R^2 = 3.09% for Batting Average above expected, R^2 = 14.8% for Standardized Slugging). After creating the model for Added Hustle OPS, our model with all three features had a R^2 = 24.4%. However, feature pruning was necessary, as sprint speed had a p-value in this model that was not statistically significant (p = 0.168) using an alpha value of 0.05. We determined that while number of plate appearances was statistically significant in this model (p-value = 0.005), it was not practically useful as a predictor of Added Hustle OPS, given that the number of plate appearances for a player in a five year period could be very different than in a one year period. Only time from home plate to first (along with a constant) was statistically and practically significant, with a p-value of 0.000. Feature importances before pruning are shown in Figure 7.

After pruning the model, we found an R^2 = 22.8%. This means that less than a quarter of the variation of Added Hustle OPS comes from the variation in sprint speed. However, there is some form of a linear relationship between time to first base and Added Hustle OPS (p-value = 0.000). The coefficient in the model for time to first base was -0.0515,

meaning that a player being a second faster to first base would increase their OPS on average by 0.052, which would be a large increase in OPS for many players. However, this is an impossible feat for many, but it does show that faster players benefit from their speed over slower ones. A visual representation of this relationship is shown in Figure 5.

As shown in Figure 6, after reducing the dimensions of Percent Ground Balls and Percent Hits in Play, we sought to find if players who had a higher percentage of these "Hustle Plays" were faster on average, and if there were particular groups of players who were involved in these plays more often. The results from the clustering show that there was not a strong relationship between speed and amount of "Hustle Plays." The clusters seem to be based solely on involvement in these plays, and speed has almost no role in the results of the clustering.

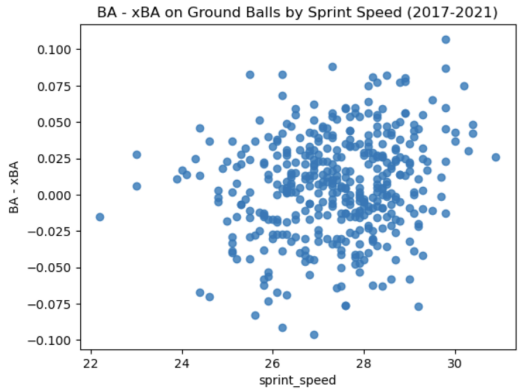

Figure 2. Batting average above expected on ground balls by sprint speed (2017-2021)

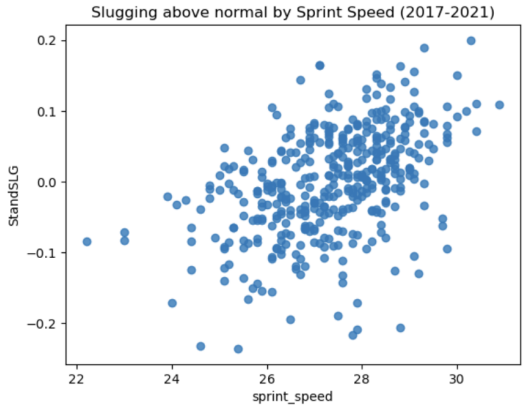

Figure 3. Slugging Percentage above normal on hits in play by sprint speed (2017-2021)

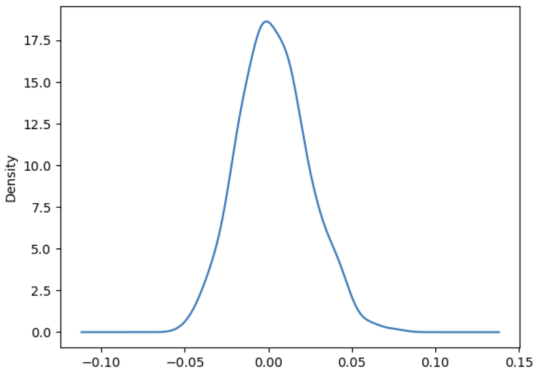

Figure 4. Density curve of Added Hustle OPS

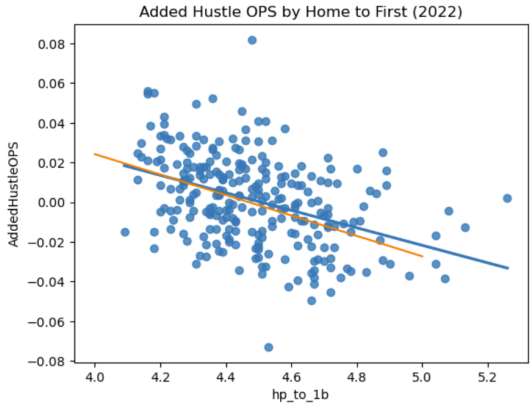

Figure 5. Added OPS from 2022 with trendline (blue). In addition, the trendline from the 2017-2021 data is included (orange).

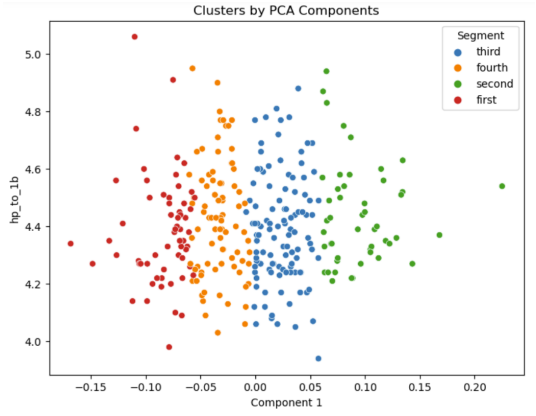

Figure 6. K-means clustering on home plate to first base after PCA

5 Discussion

Although the relationship is not necessarily a strong one between Added Hustle OPS and speed, faster players do

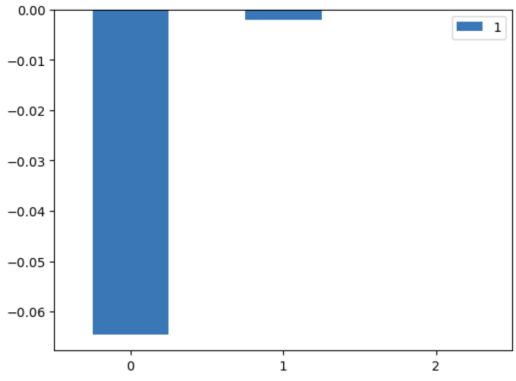

Figure 7. Feature importances before pruning of features. It was determined that only feature 0 (home to first time) was statistically and practically significant

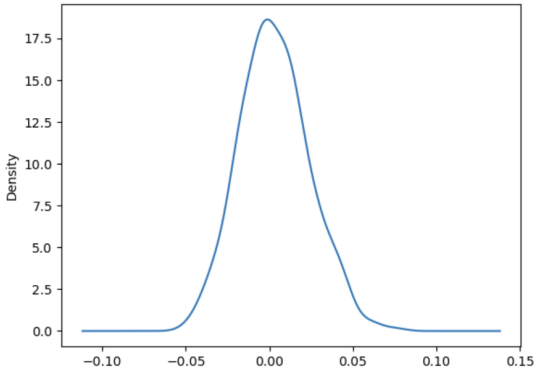

Figure 8. Density curve of Added Hustle OPS

seem to benefit from their speed. However, from the density curve in Figure 4, it is apparent that most players do not have a significant change to their OPS from hustle plays. For faster players, this can be important to their batting profile, as their ability to get on base hinges more on their speed, among other factors, compared to other players who may hit more home runs, and therefore do not need to worry about speed as much. Although the R^2 value of our model was relatively low, we believe there are a number of other factors besides speed that can determine whether or not a batted ball becomes a hit, including location. Some players have an ability to find holes in the defense, meaning they can turn more batted balls with soft contact into hits without using their legs. While the effect of OPS on overall run scoring was not examined in this study, previous studies have shown that OPS is a strong predictor of run scoring. With an average OPS around 0.750, players with even a contribution of 0.040 from hustle plays add an extra 5% to their offensive profile. In our training set, we found 20 of the 412 players had a contribution of 0.040 or higher from Added Hustle OPS.

From the clustering, we have found that players who are involved in hustle plays more often are not necessarily any faster than players. If this were true, in Figure 6 we would see a negative linear relationship between time to first base and Component 1 of our PCA. If there were a relationship here, there would be a cluster of players in the bottom right of the graph who were speedsters who also hit more ground balls and had more hits in play per at bat. However, the verticality of the clusters shows that players are closer together in the K-means clustering are only similar in how often they are involved in these plays, and speed appeared to have little impact in the clustering.

One finding that was surprising from our study that time from home to first had much higher coefficients for predicting Added Hustle OPS than sprint speed did. At the start of this study, we believed that the two may work together, especially on the slugging data. The exclusion of sprint speed in our model for Added Hustle OPS implies that top speed is less of an important factor than being able to reach top speed quickly. This contradicts our initial finding that slugging had more of a relationship with speed than batting average, as top speed would be expected to play a more important role on plays that involve longer runs. As mentioned earlier, in our models to predict Added Hustle OPS, sprint speed was not statistically significant.

One assumption made in this statistical modeling involves the fact that we did not factor hitting ability into our model. This was because we believed that while players may have some ability to find holes in the defense, defensive shifting (which was allowed during the time studied) would have mitigated this ability. Another factor that may have affected our results is the limitations of the linear model. Although the linear model is most common in baseball statistics, and our residuals were found to be relatively random, we would like to employ other tests for prediction.

6 Future Work

There are existing implications for future work in the intersection of Data Science and Baseball Statistics. We believe that one of the most relevant areas of future work will be using Data Science to predict the effects of rule changes in Major League Baseball, how those changes will affect both the strategic and tactical management of the game itself, and the practical recruiting and price points of the talent, skill, and financial management of the player base. Another possibility for future work would be to address the impact that the perceived importance of speed has on the labor market. A question that we did not look at is whether faster players still paid more, even if they do not have better offensive statistics than other players.

Another future derivative work is to see if Data Science insights can be used to analyze the game of cricket. Some work is being done in analyzing cricket, mostly with NumPy

packages [1]. That said, it appears that the work in Sabermetrics and even PyBaseball dwarfs what is happening in cricket. Since cricket's fan base is more extensive worldwide than U.S. Baseball, this may be a profitable and exciting area of future work.

7 Conclusion

For this paper, we investigated the importance of speed as a factor in a player's overall offensive profile. While we found that for most players, speed to first base does not matter as much, but for the fastest players in the league, it can have a profound impact on their other offensive statistics. We believe that these players would be able to have higher batting averages and slugging percentages if they were involved in these plays more often, and our results suggest that players are not necessarily playing to their strengths. Like most statistical research in baseball, there is a variety of possibilities for investigating how speed affects the game, and we hope this study opens the door for more research on the impact of speed out of the batter's box.

8 Acknowledgements

We would like to thank Professor Cheney for teaching this class, and for his help during this process. We would also like to thank Csenge for her help.

References

[1] Case study: Cricket analytics, the game changer!, 2023.
[2] Statcast search, 2023.
[3] Ben Baumer, James Piette, and Brad Null. Parsing the relationship between baserunning and batting abilities within lineups. *Faculty Publications:Smith College*, 2007.
[4] Ben Baumer and Peter Terlecky. Improved estimates for the impact of baserunning in baseball. 2010.
[5] James Click. *Baseball between the Numbers*, chapter Chapter 4-1, pages 112–126. Basic Books, New York, 2007.
[6] Michael Lewis. *Moneyball*. WW Norton, New York, NY, 2004.

Received 7 May 2023