Anthony Brown
Stats 406
4/26/20

The Effect of a 15+ Game Hitting Streak on a Baseball Player's Teammates' Offensive Performance in 2019

Introduction:

A long hitting streak is a rare occurrence in baseball. A hitting streak is defined as a stretch of consecutive baseball games wherein a batter records at least one hit in each game. The longest such streak in MLB history was a 56-game hitting streak by Joe DiMaggio in 1944. Since then, there have only been 30 streaks of 30 or more games, the longest of which was 44 games by Pete Rose in 1978. Most of the time, a hitting streak means a big offensive improvement in the player who has the streak, but what is the impact on that players' teammates? The question being asked in this study is whether or not a players' hitting streak can improve the offensive performance of his teammates.

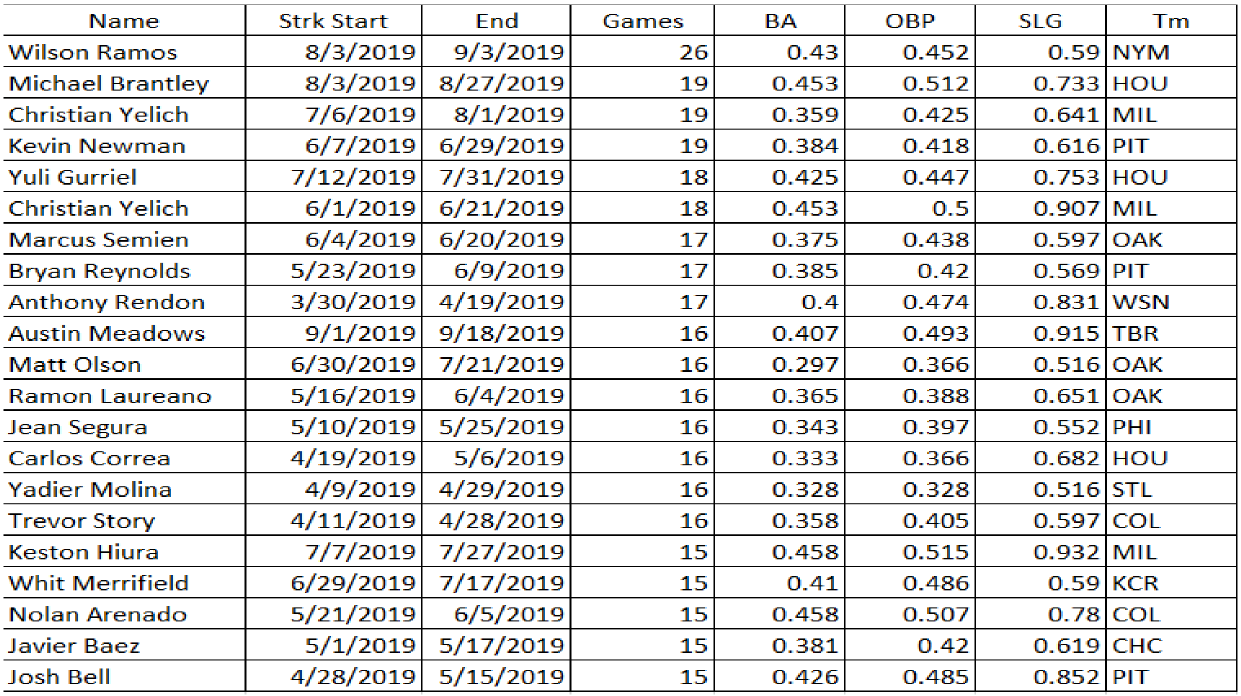
Offensive performance is a bit of a touchy subject in modern baseball discussions. The debate over using old-school statistics vs. new-age sabermetrics in measuring output is hotly contested to this day with no clear winner. For this project, I have selected to use the following statistics to measure offensive performance: batting average, on-base percentage, and slugging percentage. Batting average is calculated by taking the total number of hits a player has and dividing it by the total number of at-bats and is a measure of roughly how well a player hits the ball. On-base percentage adds the number of hits to the number of walks and hit-by-pitches for a player and divides it by the number of plate appearances, which is at bats plus walks, hit-by-pitches, and sacrifices, and measures how well a player gets on base. Slugging percentage is calculated by taking the number of total bases a player gets (Single = 1 total base, Double = 2 total bases, Triple = 3 total bases, Home Run = 4 total bases) and divides it by the number of at bats and is used to measure roughly the amount of power a player has. I am using these three statistics because they are easy for non-baseball fans to understand and, combined, effectively measure a player's offensive performance.

The following paper consists of five sections: Data, Methods, Simulations, Analysis, and Discussion. The sample of baseball players which will be used in this project will be introduced in the Data section. For Methods, I will be introducing the concept of bootstrapping and confidence intervals which will be used later on. In Simulations, I will test out the strength of the bootstrap confidence interval, which will be applied to the observed data from the sample in Analysis. Lastly, the Discussion section will be used to summarize the results from Analysis and provide further context into this topic. I hypothesize that a hitting streak of 15+ games positively affects the offensive performance of the player's teammates.

Data:

For this study, the population is all MLB players in 2019. In order to measure offensive performance, the statistics being used will be batting average (BA), on-base percentage (OBP), and slugging percentage (SLG). The sample of the population that I will be taking consists of players who played with a hitter that had a 15+ game hitting streak in the 2019 MLB season. The sample will also only consist of players who averaged at least two plate appearances per game during the streak in order to remove small sample size bias, which comes out to a total sample of 170 players. Then, the control sample will be the same group of players except it will comprise of the players' stats over the rest of the season. The aforementioned statistics for the control group will be compared to the other group. All of the data mentioned above was collected from the Play Index at BaseballReference.com ("Play Index Home").

Table 1: 2019 Hitting Streaks >= 15 Games

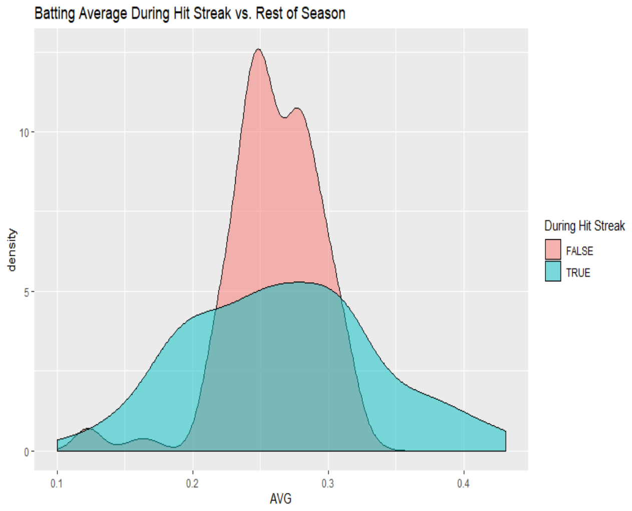


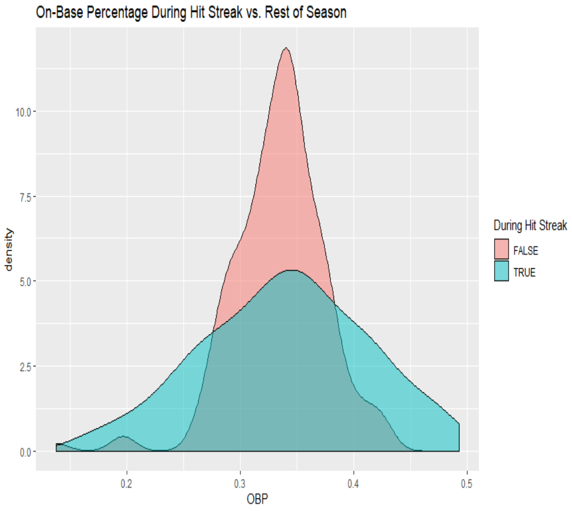
Name	Strk Start	End	Games	BA	OBP	SLG	Tm
Wilson Ramos	8/3/2019	9/3/2019	26	0.43	0.452	0.59	NYM
Michael Brantley	8/3/2019	8/27/2019	19	0.453	0.512	0.733	HOU
Christian Yelich	7/6/2019	8/1/2019	19	0.359	0.425	0.641	MIL
Kevin Newman	6/7/2019	6/29/2019	19	0.384	0.418	0.616	PIT
Yuli Gurriel	7/12/2019	7/31/2019	18	0.425	0.447	0.753	HOU
Christian Yelich	6/1/2019	6/21/2019	18	0.453	0.5	0.907	MIL
Marcus Semien	6/4/2019	6/20/2019	17	0.375	0.438	0.597	OAK
Bryan Reynolds	5/23/2019	6/9/2019	17	0.385	0.42	0.569	PIT
Anthony Rendon	3/30/2019	4/19/2019	17	0.4	0.474	0.831	WSN
Austin Meadows	9/1/2019	9/18/2019	16	0.407	0.493	0.915	TBR
Matt Olson	6/30/2019	7/21/2019	16	0.297	0.366	0.516	OAK
Ramon Laureano	5/16/2019	6/4/2019	16	0.365	0.388	0.651	OAK
Jean Segura	5/10/2019	5/25/2019	16	0.343	0.397	0.552	PHI
Carlos Correa	4/19/2019	5/6/2019	16	0.333	0.366	0.682	HOU
Yadier Molina	4/9/2019	4/29/2019	16	0.328	0.328	0.516	STL
Trevor Story	4/11/2019	4/28/2019	16	0.358	0.405	0.597	COL
Keston Hiura	7/7/2019	7/27/2019	15	0.458	0.515	0.932	MIL
Whit Merrifield	6/29/2019	7/17/2019	15	0.41	0.486	0.59	KCR
Nolan Arenado	5/21/2019	6/5/2019	15	0.458	0.507	0.78	COL
Javier Baez	5/1/2019	5/17/2019	15	0.381	0.42	0.619	CHC
Josh Bell	4/28/2019	5/15/2019	15	0.426	0.485	0.852	PIT

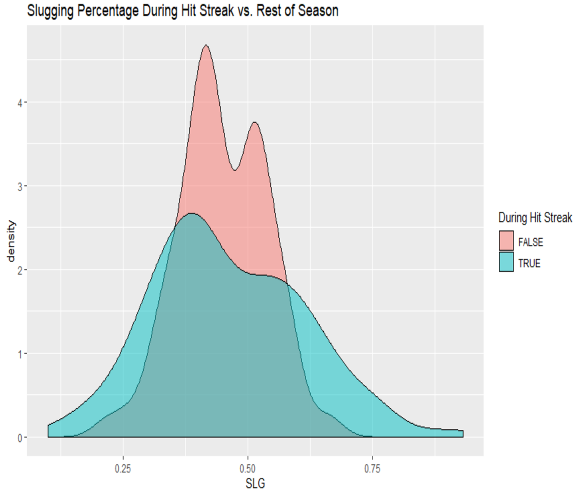
Figure 1: This table contains the length of the hitting streak, the player who had the streak, the team they played for, the start and end dates of the streak, as well as the players' offensive performance during the streak.

An examination of the density plots placed below for each offensive statistic shows that the stats during the streak appear to have a similar mean compared to the stats over the rest of the season. The main difference between the two time periods is that for the period during the hitting streak, there is a much higher variance than there is over the course of the rest of the season.

Anthony Brown
Stats 406
4/26/20







Method:

For this study, I will be analyzing the data using the bootstrap method (Yen). In the bootstrap method, a sample of size "n" is taken from a larger population. Then, from that sample, a bootstrap sample of the same size is taken with replacement, and this process is then replicated "B" number of times. For each bootstrap sample, I will be taking the mean of the difference between the three measures of offensive performance during the streak and the three measures of offensive performance over the rest of the season for each player in the bootstrap sample, giving me “B” number of estimates for the mean. Finally, using the “boot” package in R, I will compute a 95% confidence interval for the mean for each measure of offensive performance, which will determine whether or not a hitting streak does affect a player’s teammates. For the bootstrap to work, we must assume that all samples are independent and identically distributed, and that we know the size of the population. A further description of bootstrapping can be found in Maria Rizzo’s textbook, *Statistical Computing with R* (Rizzo).

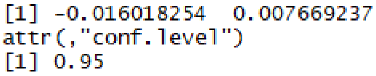
For this project, the null hypothesis will be that a 15+ game hitting streak has no effect on the offensive performance of the player’s teammates. The alternative hypothesis will be that there is an effect on a team’s offensive performance during the hitting streak of one of its members.

Simulations:

For the simulation, since the distribution of both the streak statistics and the rest of season statistics is roughly normal, I will be creating a sample of normally distributed random variables and using a t-test to find the confidence interval. This is to compare how the t-test confidence interval would be compared to the interval determined by the bootstrapping. Using the means and standard deviations for batting average during the hitting streak and over the rest of the season, I created 170 normally distributed random variables for each the streak sample and the control sample, which matches the size of the samples of observed values.

Anthony Brown
Stats 406
4/26/20

After that, I used the t.test function in R to find the confidence interval for the difference in means between the streak sample and the rest of the season sample. As you can see in the result below, the confidence interval for the effect of a hitting streak on teammates' batting average at the 95% confidence level is (-0.016, 0.007). This means that the average change in batting average during a teammate's hitting streak is between a 16 point drop and a 7 point gain. Since the confidence interval includes zero, we are unable to reject the null hypothesis that there is no change in offensive performance during a teammate's hitting streak.

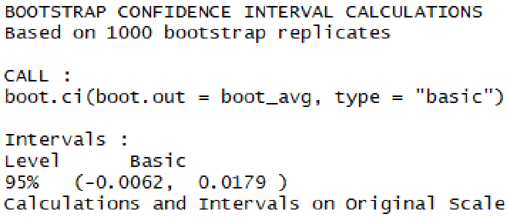


```
[1] -0.016018254  0.007669237
attr(,"conf.level")
[1] 0.95
```

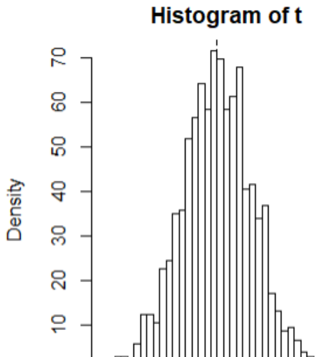
Analysis:

1. Batting Average

As seen in the picture on the left, the confidence interval for mean difference in batting average at the 95% level is (-0.0062, 0.0179). This means that the average change in a player's batting average during a teammate's hitting streak is somewhere between a six point drop and an eighteen point gain in batting average, with a mean gain of six points as seen on the histogram to the bottom right. Due to the confidence interval containing zero, we fail to reject the null hypothesis that there is no effect from a hitting streak on a teammate's batting average.



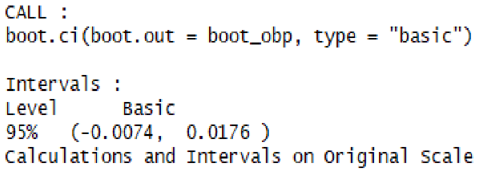
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_avg, type = "basic")

Intervals :
Level      Basic
95%   (-0.0062,  0.0179 )
Calculations and Intervals on Original Scale
```



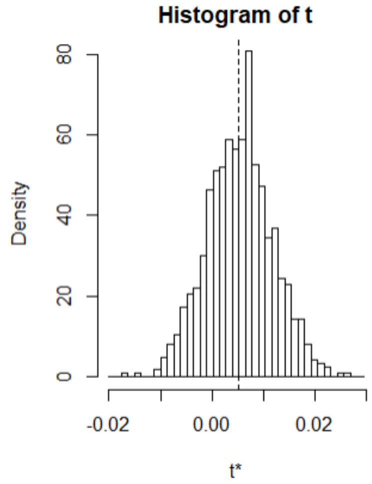
Anthony Brown
Stats 406
4/26/20

2. On-Base Percentage

As seen in the picture on the bottom left, the confidence interval for mean difference in on-base percentage during a teammate's hitting streak at the 95% level is (-0.0074, 0.0176). This means that, on average, the difference in a player's on-base percentage during a teammate's streak is between a seven point drop and a seventeen point gain, with a mean gain of five points as seen on the histogram to the bottom right. However, this also fails to reject the null hypothesis that there is no effect, as the confidence interval contains zero.

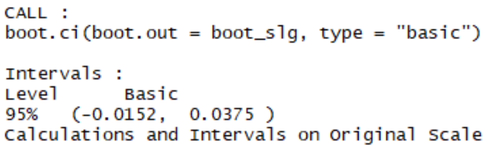


```
CALL :
boot.ci(boot.out = boot_obp, type = "basic")

Intervals :
Level      Basic
95%   (-0.0074,  0.0176 )
Calculations and Intervals on Original Scale
```

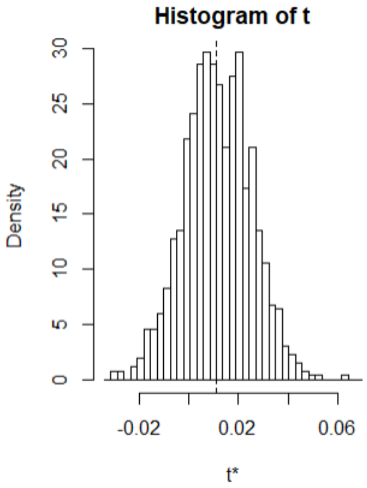


3. Slugging Percentage

As seen in the picture on the bottom left, the confidenc in slugging percentage during a teammate's hitting streak at the 95% level is (-0.0152, 0.0375). This means that, on average, the difference in a player's slugging percentage during a teammate's streak is between a fifteen point drop and a thirty-seven point gain, with a mean gain of eleven points as seen on the histogram to the bottom right. Once again, this fails to reject the null hypothesis that there is no effect, as the confidence interval contains zero.



```
CALL :
boot.ci(boot.out = boot_slg, type = "basic")

Intervals :
Level      Basic
95%   (-0.0152,  0.0375 )
Calculations and Intervals on Original Scale
```



Discussion:

As seen in the bootstrapping tests done in the Analysis section, we failed to reject the null hypothesis at the 0.05 level for all three offensive performance statistics. What this means is that although we cannot conclusively say that a hitting streak has zero effect on teammates' offensive performances, we can't say that there is an effect either. The reason I believe prevented us from rejecting the null hypothesis is the sample size of the games. While 15+ games isn't a small stretch in baseball by any means, it is small enough to cause a large variance in the statistics. Baseball statistics tend to vary wildly towards the beginning of the season, before finally evening out after a month or two. By only using 15+ game hitting streaks, we are risking having a bad stretch of games from one player throwing off the entire project.

Anthony Brown
Stats 406
4/26/20

In the future, we can try to take into account several more variables for this project. For example, the opposing teams that are played could have an effect on offensive performance, as bad teams are more likely to give up a lot of hits than good teams are. Also, the time of year the streak occurs in could have an effect. Some players struggle at the beginning of the season, due in part to cold weather or just shaking off the rust from the offseason. This could cause offensive performance during hitting streaks at the beginning of the season to be worse than they would during the middle of the season. We could also try restricting the hitting streak size to streaks of 20+ or 25+ in order to increase the sample size of games for the sample to decrease the variance.

Anthony Brown
Stats 406
4/26/20

Works Cited

Bock, Joel R, et al. “Hitting Is Contagious in Baseball: Evidence from Long Hitting Streaks.” *PloS One*, Public Library of Science, 12 Dec. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3520861/.

“Play Index Home.” *Baseball*, Baseball Reference, www.baseball-reference.com/play-index/.

Rizzo, Maria L. *Statistical Computing with R*. CRC Press, Taylor and Francis Group, 2019.

Yen, Lorna. “An Introduction to the Bootstrap Method.” *Medium*, Towards Data Science, 28 Jan. 2019, www.towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60.