JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Final Project

Course Section: 605.662.81
Fall, 2019

Prepared by

James Chatterton

12/9/2019

Abstract

A record number of home runs were hit in 2019 during the MLB regular season. Rob Arthur measured the drag coefficient on the baseball early in the season noting it was well below previous seasons (2). This reduced drag led to the amount of home runs. With the recent public data available from MLB's Statcast, it is easy to obtain data to measure aspects of batted balls, especially those hit for home runs. Taking exit velocity and launch angle data along with the estimated hit distance I determined Arthur's findings of reduced drag leading to more home runs, specifically with home runs from exit velocities and launch angles that do not often result in home runs.

Introduction

During 2019, a record number of home runs were hit in Major League Baseball's regular season. As the season began, writer for Baseball Prospectus, Rob Arthur, detailed how the drag coefficient on the baseball was significantly less than previous seasons (2). He expected home runs to increase as they did in 2017 if the drag remained similar throughout the year. Come October, we can look back at the full season and see that home runs did spike, and many records were set. With batted ball data dating back to the 2015 season, I can compare batted ball distances for home runs and other fly balls to confirm that similar hit balls travelled farther in 2019 than other years. This analysis will help confirm Arthur's analysis that reduced drag helped the baseballs travel longer distances in the air.

Background

Baseball has always been a game of statistics. Counting how many home runs a player hit in a season has been a staple in fans' enjoyment and excitement of watching and following baseball. In 2015, the MLB made public a tracking system called Statcast that implements radar and high-resolution cameras to capture data from many aspects of all events on the baseball field (1). Fielders' routes to the ball flying in the air can be tracked and measured to determine if that player makes an efficient path to catch the ball. The spin of the baseball out of a pitcher's hand can be measured to see how and why a pitcher may have an effective pitch.

Parts of the Statcast dataset was made public, meaning fans and analysts outside of the MLB or its teams, could analyze the data for independent research. Daren Willman developed a website called Baseball Savant to intake the public data available and create charts and other visualizations to create enjoyable content for baseball fans (5). Willman and Savant have been absorbed into the MLB as the official public portal to the Statcast dataset with data on every single batted ball dating back to the 2015 season.

There are two specific pieces of data from all batted balls that have been used to detail how well a ball is hit and the likelihood of the batted ball's outcome. These stats are launch angle and exit velocity. Rob Arthur details how these two statistics combined together give us a great amount of information into what will happen after a batter makes contact with the ball (4). The launch angle is the, "vertical direction of the ball coming off the bat; a launch angle of zero degrees would be a flat line, with positive numbers indicating an upward ball flight and negative ones indicating a ball driven into the ground. (4)" The exit velocity is, "the speed at which a ball leaves the bat. (4)" Balls hit with a specific exit velocity and launch angle can be bucketed into different outcomes, like ground balls, line drives, and fly balls. A batter will be able to hit a home run with the perfect combination of exit velocity and launch angle. The harder the exit velocity the better but the launch angle must fit a certain window to have a trajectory that will result in a batted ball distance far enough to fly over the outfield fence.

This batted ball distance is where the possibility of home runs come into play. Two weeks into the 2019 season, Arthur found that the drag coefficient of the baseball was far lower than the drag coefficient seen in the 2018 season (2). This finding did support the record number of home runs per game the early 2019 season was displaying, higher even than many parts of the record setting 2017 season. At the end of the season, Arthur looked back at the year of drag coefficients while researching the coefficient of the playoff baseballs and found that the number stayed about the same with a slight decline (3). What a lower drag coefficient should mean for batted ball distance is that balls hit with similar exit velocity and launch angle combinations in 2019 should travel farther than those hit in other seasons.

Approach

The goal of this analysis is to provide evidence that the ball traveled farther in 2019 than the balls in previous years, resulting in more home runs. This analysis will help support the determination Arthur made that the drag on the baseball in 2019 was far less than that of previous years. A lower drag should mean that balls hit similarly should have traveled further in 2019 than they did in previous seasons. The dataset obtained from Baseball Savant contains every batted ball from 2015 to 2019 with the exit velocity, launch angle, estimated batted ball distance, and the outcome of the play.

These aspects of a batted ball event can be compared to show how 2019 resulted in more home runs given a similar exit velocity and launch angle. I will employ different visualizations, mostly scatterplots, to make these comparisons.

One approach will be to show histograms of the estimated batted ball distance on all home runs hit across each season. This method does not employ exit velocity and launch angle but should show a relationship between distance and the season.

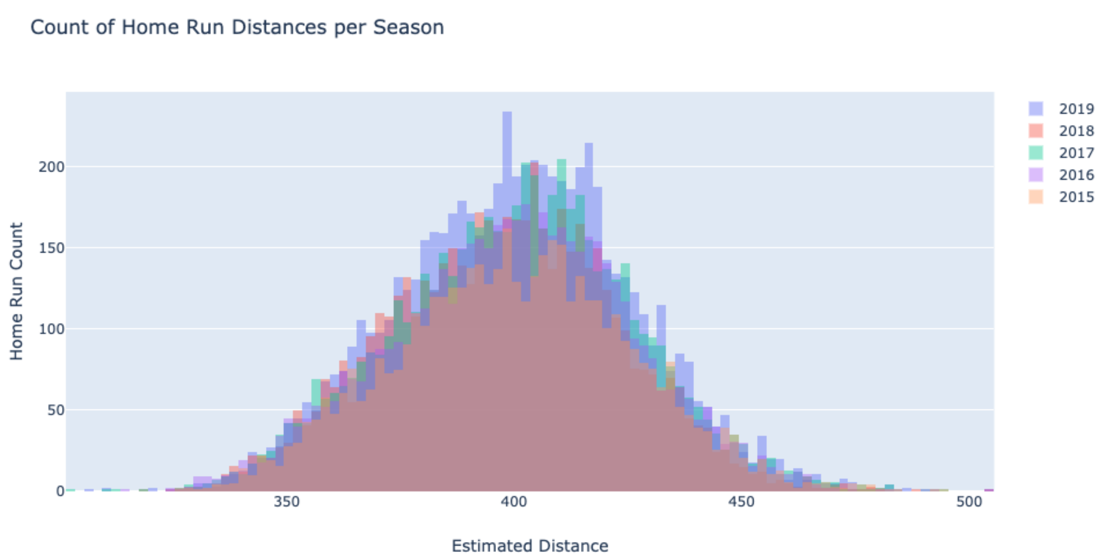
Another approach will be to make a scatterplot of home runs across each season with exit velocity the x-axis, launch angle the y-axis, and year the color of each point. The expectation is that this will show a wider range of launch angles and exit velocities resulting in home runs for 2019 than other seasons.

Additionally, another method can be used to bucket exit velocity and launch angle groups together to compare estimated batted ball distance. Similar exit velocities and launch angles should have farther distances in 2019 than previous seasons. The seasons can be compared with boxplots and violin plots for each bucket.

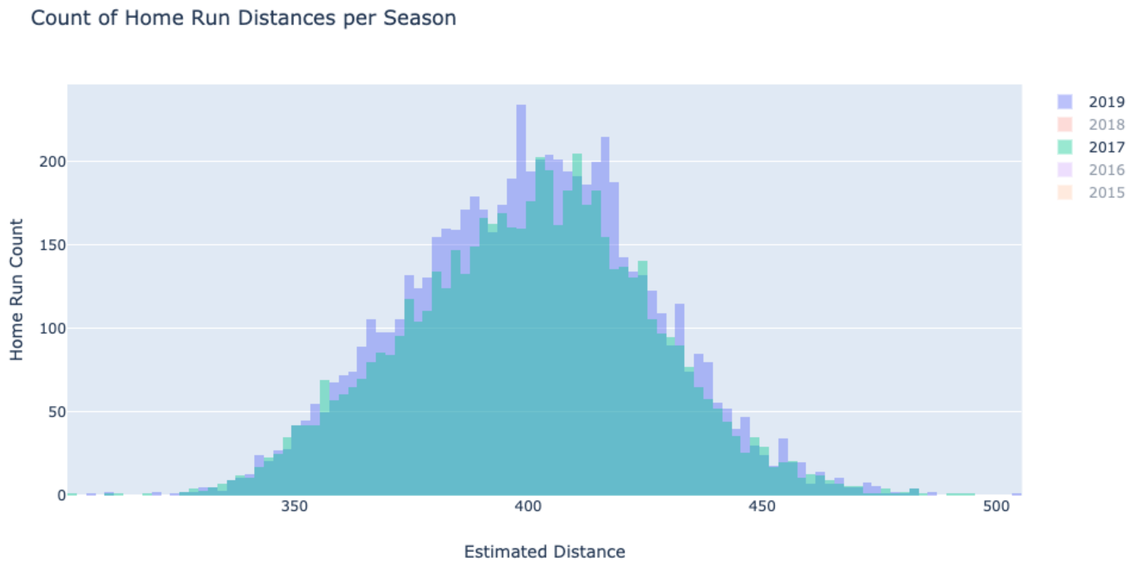
Results

I first had to clean up the data and narrow down the scope of batted balls. I changed the criteria of batted balls to limit just to home runs with a batted ball distance greater than 300 feet.

First, let's establish the volume of home runs hit during the 2019 season versus the other seasons. The following plot is a histogram of each season's home run count given the estimated home run distance overlaid.



This plot shows a distinct volume of home runs over nearly all bins of distance for the 2019 season. Comparing the former record setting 2017, we see a clear lead, especially with home runs hit between 398 and 399 feet with 234 for 2019 and 160 for 2017.



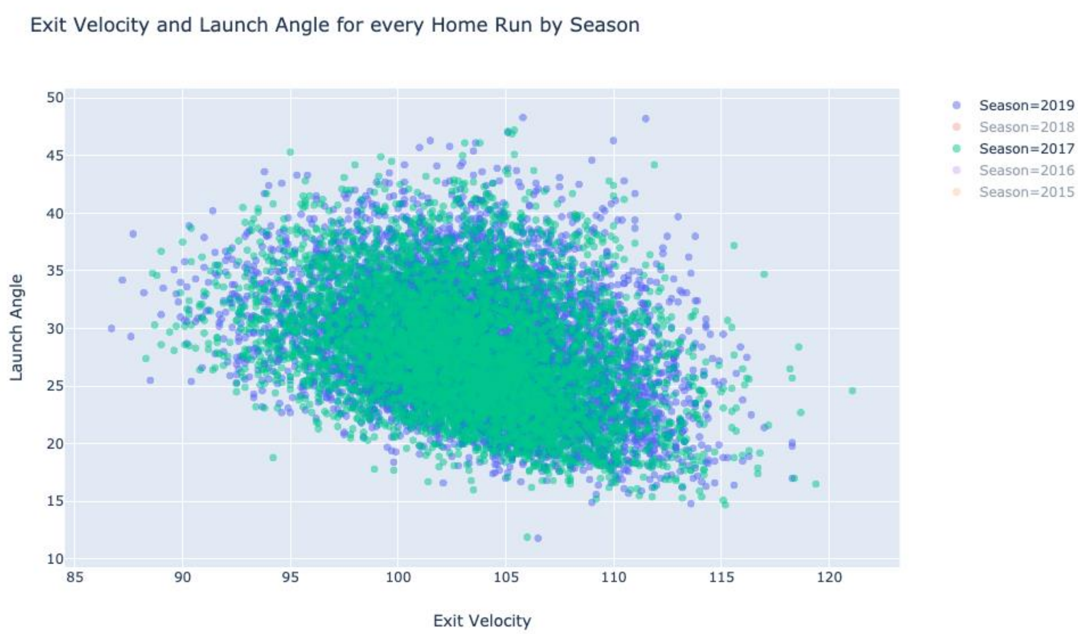
The plots show not just the all-around increase but smaller spikes in the 380 to 390 foot range and the 415 to 420 foot range.

The second plot is a scatter plot of all those home runs with exit velocity as the x-axis and launch angle as the y-axis. Each point is differentiated with a different color by season.

The main season we are concerned with is 2019, which is the blue behind all the other dots. What we are looking for here is a wider spread of exit velocities and launch angles for the 2019 season. This would mean that home runs are more possible at the extreme values.

First, the orange 2015 season is concentrated heavily in the center while the other seasons are more spread. Removing the other seasons to compare just 2015 and 2019 shows a higher concentration of home runs in 2019 on the edges of the data like 100 to 105 MPH exit velocities and 35 to 40 degree launch angles.

Additionally, another season that Arthur had discussed with a juiced ball was the 2017 season. 2019 broke the home run records that 2017 had set. Comparing just these two seasons shows the following.



Again, there are still extreme values where the 2017 season has some home runs, but the edges of the data look to be more concentrated with the 2019 home runs.

Lastly, I bucketed ranges of exit velocities and launch angles together to evaluate their specific estimated batted ball distances and compare them across seasons. The goal here is to show how the ball may travel farther given similar exit velocities and launch angles in 2019 given the reduced drag.

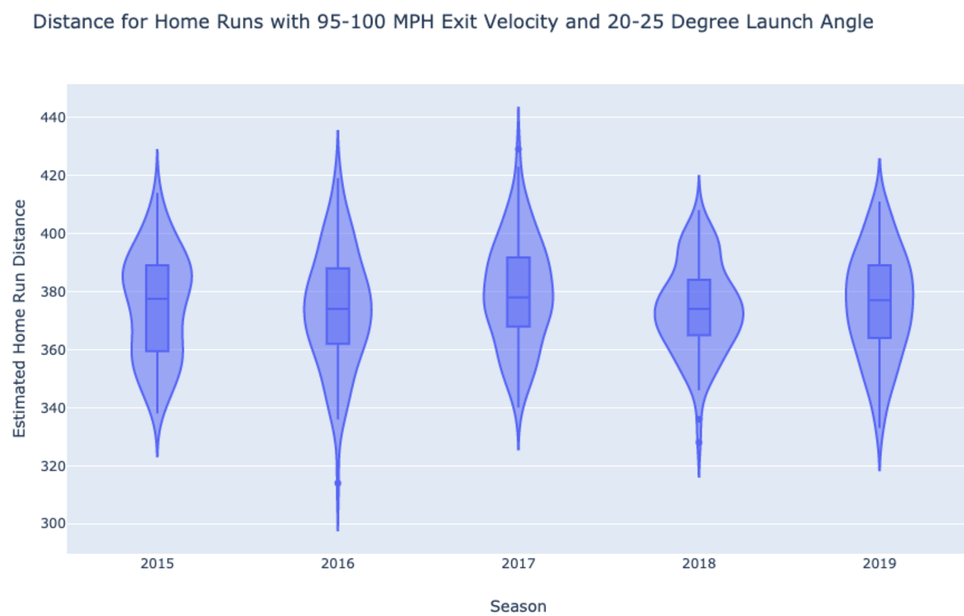
I created 14 buckets based on the high volume of home runs in sections of five for both exit velocity and launch angle. For launch angles of 20 to 25 and 25 to 30 degrees, there are four buckets for exit velocity; 95 to 100, 100 to 105, 110 to 115, and 115 to 120 MPH. For 30 to 35 and 35 to 40 degrees, there are three buckets; 95 to 100, 100 to 105, and 110 to 115.

The results here are where things get interesting. There is no consistent difference when comparing the box plots for each bin. Sometimes 2019 has a higher median and interquartile range but many times the box plots are essentially the same, even with 2015.

Here are some examples:

The above bin is essentially the best hit balls on the scatterplot where the furthest hit balls fall. 2015 has a higher median with a larger spread of distances despite the same bucket of exit velocity and launch angle.

Here is the smallest and weakest bucket:



This bucket is the lowest exit velocity and launch angle bucket I created for these plots. They all have similar enough structures to not see a difference.

I have two thoughts regarding the results of this comparison. First is how the Statcast estimated batted ball distance is calculated. One concern about this is that it uses previous data to come to the measurement. There are no tape measures pulled out during the game. All that can be done is use the software to calculate the trajectory. MLB's Statcast glossary states that it is determined by finding the parabolic arc of the ball and projecting that to the end of the batted ball path (http://m.mlb.com/glossary/statcast/projected-home-run-distance). This does mean that the distance there is a projection based on the actual flight path of the ball. However, the projection could use previous data from other seasons as well.

My second thought is that the buckets are too wide. From the first box plot example above, 2019 has a projected distance range of 394 to 487. That is almost a 100 feet difference. If I were to narrow the buckets greatly to compare, I am curious if the results would be any different. However, with the measurements taken above, there is no distinct difference in estimated home run distance given the above buckets of exit velocity and launch angle.

Conclusion

In 2017, R.J. Anderson wrote an article for CBS Sports detailing the reception of Statcast's effect on the MLB (1). He brought up further research by Arthur on how Statcast can have difficulty tracking batted balls. There can be plenty of faults in the data which I found through this research. Many estimated home run distances were null in the dataset and were removed. Additionally, some launch angles were incorrect. I found one home run by Avisail Garcia with a negative launch angle in the data set. That would be impossible. I found a video of the home run, and the launch angle was definitely positive (https://www.youtube.com/watch?v=9oQ5EXXoTqQ). These errors make for difficult conclusions especially around the launch angle, exit velocity, and distance data.

It is easy to show the volume difference in the home runs in 2019 than the other seasons. This does prove there is at least some merit to Arthur's research. Reduced drag can lead to more home runs. However, does that mean that fly balls that would not have travelled far enough in earlier seasons would have been home runs in 2019? The scatterplots do confirm this. Fly balls with exit velocity and launch angle buckets that normally would not have been home runs in other seasons were more often home runs in 2019. However, the estimated distances of these home runs in those buckets were not noticeably farther in 2019 than the previous four seasons. Further research with smaller buckets may be necessary to find that difference. Overall, reduced drag did have a result on batted balls in 2019 producing more home runs than any other season including more home runs that would not have been in other seasons.

Bibliography

1. Anderson, R.J. “How Statcast Has Changed MLB and Why Not Everybody Seems All That Happy about It.” *CBSSports.com*, 6 June 2017, www.cbssports.com/mlb/news/how-statcast-has-changed-mlb-and-why-not-everybody-seems-all-that-happy-about-it/.

2. Arthur, Robert. “Moonshot: The Baseball Is Juiced (Again).” *Baseball Prospectus*, 5 Apr. 2019, www.baseballprospectus.com/news/article/48260/moonshot-the-baseball-is-juiced-again/.

3. Arthur, Robert. “Moonshot: The Rocket Ball Has Disappeared In October.” *Baseball Prospectus*, 10 Oct. 2019, www.baseballprospectus.com/news/article/54306/moonshot-the-rocket-ball-has-disappeared-in-october/.

4. Arthur, Robert. “The New Science Of Hitting.” *FiveThirtyEight*, 13 Apr. 2016, fivethirtyeight.com/features/the-new-science-of-hitting/.

5. Natarajan, Senthil. “A Journey Into the Heart of Sports: Data Viz with Daren Willman.” Medium, Nightingale, 5 Aug. 2019, medium.com/nightingale/a-journey-into-the-heart-of-sports-data-viz-with-daren-willman-ce429da3f1dc.