Pitchers are being babies: An analysis of the home run surge

Leocadie Haguma, Rosalia Hernandez, Jacob Rojas,

8/23/2018

Abstract

The goal of this paper is to analyze the different components of the surge in home runs of the Major League Baseball that began in the 2015 season. The number of home runs has increased from 4,909 in 2015 to 5,610 in 2016 and 6,110 in 2017. The new high record of home runs achieved in 2017 beat the all time record from the year 2000, reportedly during the peak of the "steroid epidemic" in the MLB. This trend in home runs has sparked many, both from within the MLB and outside, to question why and how this is happening. Many pitchers have voiced their concerns through social media regarding the physical properties of the ball being different. Others, outside the MLB, have tried to investigate this phenomenon to seek conclusions as to why there has been a sudden drastic increase. Most notably, the committee of scientists hired directly by the MLB to analyze the surge in home runs. The task at hand is to attempt to approach this problem from a perspective not yet explored: the players or the culture of the game.

Introduction

The History of the MLB

In 1876, eight teams assembled to call themselves the National League (NL). These were the first teams of the Major League Baseball (MLB) and the birth of what we know today as baseball. Many records of baseball games and baseball teams prior to this year exist. Most notably, the first official baseball game was recorded on June 19th 1846. As one of America's oldest sports, the MLB has made many changes to the game in the past 150. In the paper, "Going Going Gone" by the David Getz (Getz (2009)), the dynamics of pitchers and batters has been influenced by changes to the rules.

According to Getz, baseball in the early 1900s was notoriously known as a pitchers game. The game was widely focused on strategic batting technique rather than powerful hits. During this time, it was commonly frowned upon to aim for a home run. George (Babe) Ruth fundamentally changed the game of baseball after his introduction to baseball in 1914. His technique focused on hitting home runs rather than strategic batting practices. With the increasing popularity of Babe Ruth, more batters followed suit. They began to sacrifice strategy for increase in power. Getz points out that during this time, the number of home runs spiked, but so did the number of strikeouts.

New rule changes allowed batters to hit more home runs. The increasing popularity of the game increased fan attendance, and influenced team owners to add seating in the outfields. In addition, Getz emphasized that more rule changes were added to the game

in order to facilitate home runs. These rule changes included the addition of the cork inside the center of the baseball in 1910 which made it easier to hit a home run (Getz (2009)). These changes, according to Getz, allowed the MLB at the time to combat decreasing attendance by promoting home runs and increasing interest. In his paper, Getz states there has been many instances in baseball's history where the MLB has changed the rules in favor of the home runs because it drew the crowds and spiked up ticket sales.

There have been many decreases followed by sharp increases in overall home runs throughout the course of baseball history. Getz theorizes that, in order to increase decreasing ticket sales, the MLB has looked towards increasing home runs. This trend of overall home run totals can be seen from the 1900s to 2017 in Figure 1. During the early 2010's, there is a decline in home runs followed by a steep increase. This is the time phrase this paper will focus on. Our method will be to analyze the "Who, What, Where, When, and Why" of the home run phenomenon.

Figure Description: In the early 1900's baseball was predominantly a game of "small balls," where hitting home runs was commonly frowned upon. In 1910, George Herman (Babe) Ruth began his baseball career in the MLB. Throughout the next two decades Babe Ruth changed a fundamental aspect of baseball through encouraging an excitement for home runs. Home run rates continue to increase every year after Babe Ruth's new game methodology, which allowed the number of players hitting home runs to increase. The dip that occurs around 1943 is during World War II.[1] After World War II, home runs begin to increase again, more so immediately after 1960. This is when team owners began to implement bleacher seating in the outfields to accommodate for rising fan attendance. According to Getz, this change in stadiums allowed players to more easily hit home runs which caused them to happen more frequently and among more players. The home run increase in the 60's was short lived because as soon as home runs began to decrease, team owners noticed a fall in revenue as well. This challenge influenced new rules to be added to the MLB beginning in 1969 in favor of batters. Some rules include lowering the pitcher's mound by 5 inches, as well as the strike zone being lessened. During June and July of 1981, the player in the MLB went on a players strike which drastically reduced home run totals for the year. In 1994, there was another strike, which led to the cancellation of the World Series for that year.

Figure Description: is a graph composed of the total home runs per team between the 2014 through the 2017 season. As seen on the graph, most of the teams had the biggest increase in home run production during the 2017. This graph is used to see if significant increase in home runs during that time was consistent throughout all the MLB teams. According to the data graph below, we can make the the assertion that all the teams in the MLB contributed to the significant increase in total home run produced independently during the 2015 through the 2018 season.

[1]The United States involvement in World War II began after Pearl Harbor on December 7th, 1941 until the end of World War II in 1945.

The Official Investigation

The MLB hired a committee of 10 scientists to study the recent home run surge in an effort to officially investigate the causes of the surge. The committee's chair was Alan Nathan, professor Emeritus of Physics at the University of Illinois at Urbana-Champaign. Other committee members include professors from Caltech, Stanford, MIT, USC, and other prestigious institutions. On May 24, 2018 their report titled "Report of the Committee Studying Home Run Rates in Major League Baseball" was released as an official MLB report. They go into great detail about their research and findings.

Data and Methodologies

Although the main source of data they used in their analysis was statcast data, they also referenced and compared their results to other sources of data, including a dataset provided by the MLB, data from pitchf/x, HITf/x, retrosheet, and baseball reference. They noted that the data obtained directly from the MLB only contained 86-88 percent of batted balls and 96 to 98 percent of home runs for the years 2015 to 2017.

In addition to this, they also performed their own lab tests to measure additional baseball properties not included in Rawlings baseball parameters as well as tests to identify factors that were causing baseballs to have a lower drag coefficient.

Their Research

The committee first identified the basic events that lead to a home run: the ball has to be hit at the right angle and must travel at a certain velocity with sufficient height and distance. They determined that the launch angle, exit velocity and spray angle are the key factors that determine the chances of a home run and that almost all home runs happen in the "red zone." The red zone refers to the specific region that optimizes the chances of hitting a homerun. The red zone, as they described, is between 90 and 115 mph exit velocity and a 15 to 45 degree launch angle. The committee was interested in this because they wanted to know if the batters were hitting the ball stronger than before or if they have been changing their batting techniques. They were interested in analyzing changes in the flights of batted balls as well as if there were concentrations of home runs within a particular group of players, within certain ballparks, or if there had been significant changes in pitching strategy. Their analysis concluded that there had been no significant changes to any of the above theories.

The committee then switched their focus to begin exploring characteristics of batted balls that are home runs. They found that that proportion of balls hit in the "red zone" had not changed as much as the proportion of balls that are hit in the "red zone" and are home runs. The major takeaway from this was that there were more home runs because there had been a change in the probability of making a home run for specific spray angles, launch angles and exit velocities. This finding led the committee to question if the drag properties of the baseball itself had been identical for all years.

One possible explanation for a change in drag properties that was explored was the temperature. The committee studied the temperature effects on home runs and found that, although home run rates go up in warmer temperatures, the rate of increase is consistent for the 2015 to 2017 seasons, including a similar result when looking at that constant temperature at ballparks that are closed stadiums. They concluded that the changes in temperature could not explain the home run surge.

Another possible explanation for a change in drag properties was that the ball itself had changed.

Figure 5 shows home runs hit by rookies from 2012 to 2017. The home runs of each year are totalled from unique groups of players whose debut seasons are on the x-axis. An increase in home runs is observed from the 2014 to 2015 season where total home runs jumps from 207 to 360 home runs, a 74% increase. The rookie classes after 2015 show a slight decrease in home run production, though still much higher than seasons prior to 2015 where the jump occurred.

Figure 6 compares the height and weights of rookie classes from the same six seasons of 2012 to 2017. Each player is represented by a bubble which is sized and colored according to the amount of home runs the player scored during their debut season. A slight trend of home runs in areas where height and weight are above average can be seen. The graph also contains lines representing average height and weight of each rookie class. Height of entering rookies during these years stays relatively constant around 73.5 inches while weight shows an 8 pound drop from years before 2014 to 2017.

Figure 7 is a composition of all the pitches thrown during the 2015 - 2017 season respectively. The graph illustrates the 18 different pitches, although many very similar. The graphs breaks down all the pitches in quantity and speed. As seen in figure, most pitches are thrown between 90 to 100 miles per hour (mph). In addition to that, most pitches that are thrown are characteristically fastball. There are not many drastic trends that can be made out in this is graph. The pitching speeds are consistent in this short time frame.

Literature Review

In many of the papers published regarding the surge in home runs, the focus has mostly been in analyzing the properties of the ball itself. In the report by the committee of scientists hired by the MLB, it was concluded that the overall mass of the baseballs had changed about 0.5 grams, and the drag coefficient had also changed. The report's analysis of the phenomenon focused on the aspects of the game that could be affecting the increase in home runs. In many of their figures, the increase in home runs was concluded to be comparable amongst all the players. The report put most of its focus on the properties of the ball that could have attributed to this increase. Although there were changes to the ball from the first half to the second half of the 2015 season, all changes were analyzed individually.

The committee concluded that these individual changes to the ball could not explain the surge in home runs after the first half of the 2015 season. While this is true, the report lacked to connect these insignificant changes to determine if they were significant when compounded. In addition to this, the Committee of Scientist did not analyze different aspects of the game besides the baseball that could have been contributors to the increase in home runs.

Deep analysis on the ball was done by fivethirtyeight.com to verify the results of the committee and test other parameters not tested. MLB refused to give fivethirtyeight.com the data they needed to verify the results of the committee. Instead, fivethirtyeight.com purchased balls that were certified used MLB balls, as well as a set of new balls used

during the 2014 to 2017 seasons. One of their methods to analyze the balls was to put them through a Computed Tomography(CT) Scan. This revealed that there was a change in the density of the cork used in the pill between the 2014/2015 balls and the 2016/2017 balls. Further analysis done by Kent State and the University of Southern California (USC), revealed that in addition to the decrease in the density of the cork, the chemical composition of the rubber lining surrounding the cork had changed. This change was discovered to make the ball bouncier (Arthur and Dix (2018)).

fivethirtyeight.com concluded that the physical changes to the ball could account for about a 25% increase in home runs. The overall home run surge was a 35-45% increase in home runs from the first half the 2015 season to the 2017 season. Furthermore, fivethirtyeight.com believed that the remaining 10-20% increase in home runs could be explained by a recent change in player batting philosophy. This means that batters may have been changing their swing in favor of an upward swing.

Data collection

First we began by scraping data from the MLB website which included all 30 major league teams back to 2005 and their batting statistics(home runs, at bats, etc.). While this data was readily availiable to use for out analysis. Though this data was avaliable for the public, it was designed to be examined by some who were looking quickly at player information. The problem was that we were looking for player information that ranged several years. In addition to this, the type of data desired for further analysis of specific variable was so large that obtaining without a computerized and automated scraping tool/function would too much difficulty and time. Though hard to locate, we were able to obtain a package in r called rstudio that we could use for this type of data collect.

Baseballr

The baseballr package is a data scraping package made by Bill Petti, a Global Data Strategist who consults with and advises multiple major league baseball teams in the areas of player analysis, evaluations, and in-game strategy . The package was used to scrape initial team data from baseballsavant.com, who allows its users to browse the MLB's statcast database. This tool allowed the us to efficiently collect large amounts of baseball data with variable that were able to compare against home run production. This initial data frame contained player data for an entire season. While this was a good start, the model we wanted to create needed play by play data.

The baseballr package in r was design to render player data for several games quickily. For example, if one wanted to know the game scores of the games played in one weekend, they can simply enter the desire time frame they wanted and using a baseballr function, could obtain this information. In addition to this, player were also able to locate the picters, batters, and umpiers of everygame played in the MLB since the 1890s. The tool seemed very usefull in our quest to collect the desired data, but it did cause us some difficulty with obtaining the information for our analysis.

Initilaly, installing the baseballr package was difficult. This was due the recent updates of rstudio. RStudio had recently updated but most of it's packages had not updated. Baseballr was one of those packages without an update. After days of serching for the recent baseballr installment of rstudio, we were able to find an installment from

github[2] by Bill Petti (Petti (2018)). Even after having access to this baseball package, it took a while to understand how to use all function to obtain the type of informatin we desired. Most notibly, because the package is designed to only scrape a couple of games, it was difficult for us to collect data for serveral year. While this problem was not was never solved, were able to bypass the usage of baseballr package for most of our player demographic data with the use of the Lehman database [^3] (ADD LEHMAN DATBASE INFO).

Lahman Database

The Lahman database was designed to collect player demographic data such as date of birth, height, weight, position, ect. These variable were important for our analysis of the effects player demographics on homerun production. These variables collected from the Lahman database established the fouctions for our initial investigation. With this database we were able to see how what kinds of athletes were in the MLB. In addition to this, we were able to quantify team averages that we later compared against the MLB website for accuracy. While this data information served as an initial set for analyzing the surge in homeruns, we needed more play by play do measure quanties such as pitch speed, and exit velocies like to compare agianst the Committee's findings.

Baseball Savant

Like the Lehman database, baseball savant is a source for large MLB data. What sets the to databases apart is access to statcast data. Statcast is a combination of instruments such as pitchfx and hitfx, which collect play by play data. These instraments were install in all 30 MLB baseball stadiums in 2015 in order to quantify the ablilities of players. We used this data to analysize what factors affected the probability of hitting a homerun the most.

This database which we called "pitching.data" only contained 2015 to 2017 information. Though it was incredibly large, we still wanted prior 2015 data in order to run a logistic regression before the surge. Due to this we had a desire to obtain 2010 through 2014 pitch-by-pitch data. The same baseballr function, called "scrape_statcast_savant_batter_all" did not work to obtain play-by-play data prior to 2015. Due to this, obtaining prior 2015 play-by-play data became very difficult. Obtaining 2011 to 2014 data included a combination of many functions inside the baseballr package.

In order to scrape the same statcast data from baseballsavant.com. When we obtained this data, we noticed some of the variables we relied on in our model were not measured in their entirety for some years. For example, release speed was not consistently measured by statcast prior to 2011. In addition to this, launch speed was also not recorded in its entirety prior to 2015. This missing data could not be accounted for in our databases because it does not exist. As a result, we plan on only using years that contained consistent information throughout. For example, release speed analysis will only be performed after 2010 and launch speed analysis will only be performed after 2015.

[2]GitHub Inc. is a web-based hosting service for version control using Git. It offers all of the distributed version control and source code management functionality of Git as well as adding its own features

Obtaining this play-by-play data was not easy. The initial data frame obtained using baseballr that contained only team information dating back to 1890s, contained about 2 thousand observations. This was a very manageable data frame to work with. The play by play data frame initially contained about 3 million observations before it was filtered for the desired variables listed below.

pitch_type" release_speed" release_pos_x" release_pos_z" player_name" batter" pitcher"
events" zone" stand" p_throws"
home_team" away_team" balls" strikes"
game_year" pfx_x" pfx_z" plate_x"
plate_z" inning" vx0" vy0"
vz0" ax" ay" az"
hit_distance_sc" launch_speed" launch_angle" effective_speed"
release_spin_rate" release_extension" pitch_number" pitch_name"

The utilization of the baseball savant data ws initially difficult. While there is a search page on the websit to obtain our desired data, learning to use the search engine was not trivial. In addition to this, the design of the search engine was for very specific player information inside a small time farame window. As stated earlier, our desired time window was several year. With the implimentaion of the baseballr package, we were able to scrape this data from the baseball savant websit. This data was play by play data from the 2010 to the 2017 season.

Data cleaning

For any given game, Statcast produces about 2 TB of data. This is a large data set to work from and not all the varibles measured by statcast would be informative in our analysis. Due to this, we had to clean most of this data. In addition to cleaning, we also had to join databases together to obtain measurements of the players not recored in either database. For example, the age of the player was not a variabe recored by either lehman or baseball savant. By joining the two databases, we were able to obtain the birth date of the player in addition to the seaons. Though some practicle calculations, we were able to produce the age of any given player on any given seaons. This was crucial on our analysis of the distribution of ages amongst the player.

Though we believe the data is as accurate as can be, we have encountered some inconsistencies. Most notably, during an analysis of launch speeds, there was an observation which seemed anomalous in comparison to the rest of the data. A particular home run by Bryce Harper, a Philadelphia Phillies slugger, was especially peculiar. During the 2015 game between Philadelphia and Washington at Nationals Park, Harper hit a home run in the second inning against pitcher Sean O'Sullivan. The pitch speed was clocked at 91.9 mph yet the launch speed was recorded at an astoundingly low 36.6 mph. This value immediately stuck out in our regression model as an outlier. The slowest home runs tend to be in the 80 mph range and most struggle to make it over the fences. Harper's home run would be physically impossible at 36.6 mph and watching a recording of the home run makes it clear that the launch speed is an input error. We contacted Daren Willman, the creator of Baseballsavant, via email about our concerns with Haper's 36.6 mph home run. We have not yet received a response from Daren Wilman, nor baseballsavant.

Hypothesis Testing

Before exploring this problem further, we wanted to test weather the increase in home runs from the 2015 to the 2017 seaons was significnat. Using the Lehman database above, we were able to obtain yearly totals of homeruns divided into teams. Using this information we wanted to formulate questions for our hypotheis test. More specifically, we want to test whether the increase in home runs during the 2016 and 2017 seasons is significant compared to the total home runs 2015 season. In order to perform our hypothesis test, we used the comparisons test to analyze p - values to determine the significace of the three designated years: 2015, 2016, and 2017. From this we were able to identify our null and alternative hypthosis.

The null hypothesis

$$H_0 :$$

There is no change in the home run production during the three seasons.

Alternative hypothesis

$$H_0 :$$

There is an increase in the home run production during the three seasons.

Using these hypothesiss, we established a "true-mean" value which was determined to be the average home runs from the 2014 season(before the surge). Whiile utilizing the formate above we obtained three p-values corresponding to the individuale seasons, which gave the probability of the null hypothesis being correct. For each of three seasons we recieved corresponding t-values and p-values at a 95% cofidence interval. The t-values were obtained from the t-distribution, where the degrees of freedom were said to be 29. The t-values and the p-values showed conlusive evidence in favor of the alternative hypotheses.

We decided to rerun this test again. But now with 2,000 entried corresponding to each player in the MLB. Our initilay test only had a sample size of 30 teams. This was not a big enough date set to account for sensitivity. After recalculating our t-vlaues and p-values for all three years, we came to same conclusion: reject the null hypothesis in favaor of the alternative hypothesis. As a result, we were able to conclude that the infact the surge in homeruns between the 2014 through the 2017 seaons was statiscally significant.

Experiment setup

The problem being posed in this document is why are there significantly more home runs during the 2015-2017 seasons compared to the rest of baseball. This problem can be approached in many different ways, and many have found unique ways to attack the problem as summarized in the literature review section. We chose to examine the problem from the following perspectives: the players, the ball, and the game. We chose these because they have been briefly examined but never examined relative to each other. They are also the basic components of how to make a home run in baseball. We asked ourselves how these components could have affected the increase in home runs individually and together.

As seen in the literature review all these perspectives have been taken into account. This paper aims to investigate more about the culture of the MLB and how these factors come into play. Our approaches will focus on historical trends that could help

us forecast what is being seen now. This is not the first time the MLB has seen a significant increase in home runs. From the historical perspective as told by our data we can link social events and cultural shifts within the MLB to the falls and rises of home run production.

The Players

Player data was collected from many different sources, but ultimately the data seen in this research study was recorded and documented by Statcast technology and was scraped from the Lahman database and baseballreference.com using the baseballr package in R. The questions we want to investigate are whether the players are significantly impacting the increase in home runs. Are the players getting stronger? Are they getting bigger? Are the players' age an important factor? Are rookie players making a bigger impact than before? All of these are the questions we believe could help us answer our main question: Are the players contributing the most impact to the increase in home runs?

Are the players getting stronger?

To measure this variable, we used a height and weight proxy to establish relative strength between all the players in the MLB from 2010 to 2017. The reasoning for this was to see if there was a correlation in an increase in height and weight ratio to the increase in home run production. Moreover, we wanted to see if a certain range of height/weight ratios was producing the most home runs and if there had been any kind of shift in recent years in correlation to the increase in home runs.

We first collected all the heights in inches of all the players in every year from the beginning of the MLB (since 1871) using the Lahman Database. We then decided to filter the data to the desired time frame, which was between 2010 to 2017. Within the smaller database we were able to graph the heights vs the total home runs as seen in figure 3(?).

We saw that there is a wider disbursements of home runs in 2017 than compared to 2010 and the change is gradual over the time frame. This reinforces the fact that there are no small concentration of heights that are making the home runs. Home runs are being made by players of a wide variety of heights. We would also like to note that heights were recorded only once per player in the Lahman Database. We don't believe heights will drastically change from when it is measured to throughout the rest of a players career.

In this database we were also able to collect the weights in pounds of each of the players. We graphed the weight of each of the players against the total home runs per season within the same time frame as the heights. We believe that a players' muscle composition can be proxied using their weights. We did not see any significant changes in the weights of the players with respect to the home runs. As noted above, weights were also only measured once, and not updated per season.

Historically, pitchers tend to be taller and weigh more than their batters counterparts. Furthermore, in the AL, pitchers are not allowed to bat, while in the NL pitchers are allowed to bat. Due to the inconsistency of roles within the divisions of the MLB, we decided to filter out the pitchers from the database in an attempt to get a more accurate reading of batters who contribute the majority of home runs. As shown in figure above.

Are the players' age and important factor?

We wanted to see whether age was a significant factor in home run production by graphing age against home runs and looking at the age distribution of each season from 2014 to 2017. Were the older players making the most home runs because their careers are longer and they have more experience? Were the younger players making more home runs because they have a different player philosophy that has been shaped by optimized home run statistics?

Our graphs show that there is a higher frequency of players aged 24 to 28 in 2017 that are making more home runs. In addition to this, there seems to be a trend of more home runs being produced by younger players than in the past. We built the graphs from 2014 to 2017 and saw few trends. To make our graphs more consistent with our height and weight graphs, we wanted to extend the year back to 2010 so that we could see eight years of age and home run trends, instead of just four.

While looking at the eight year distribution of home runs and age, there was still no significant trends as can be seen in the figure below. We suspect that age does not have a significant impact on home run production based on these graphs. Further review using a logistic regression will be necessary to confirm our observation.

Are rookie players making a bigger impact than before?

We wanted to see if the incoming rookies were changing, physically, over this period of interest and affecting home run rates. We did this by looking at their home runs with respect to heights and weights, which was our strength proxy as mentioned above. We will define rookies as players who debuted their first game in that season. It was important to look at the incoming class of players in order to determine whether the spike was internal or external. By this, we wanted to see whether this spike was caused by players who had been in the league for a while or by players who were new to the league.

To look at if rookies were making a bigger impact than before, we had to combine data from two tables, both from the Lahman Database. The debut date of players was in one table that focused on player specifics (like height, weight, age, etc), and the amount of home runs was in another that focused on player performance (like home runs, times at bat, team, etc) . We filtered through players whose first MLB game was played on or after the 2012 season. This allowed us to capture rookie classes from this season onward.

As can be seen in figure 8(?) during the 2012 to 2014 season, rookies made very little home runs. During that time, they peaked at around 225 home runs in the 2013 season. Its lowest was under 200 in the 2012 season. Moreover, the biggest spike in rookie total home run production was in 2015 when they scored over 350 home runs. This is a 74% increase from 2014. Its slight decline followed, but is still significantly higher than before the spike in 2015. This reinforces the initial overall home run spike in 2015, and showed that it affected rookies as well as everyone else in the league.

To look at if the kind of rookie that entered the league was different, we used the same strength proxy as before to measure whether the new players entering the league were generally heavier and taller than in previous seasons. We found that the average height was consistent throughout all the seasons displayed. In addition, the average weight of the newer rookie classes lowered by about 8 pounds. To measure if this makes a difference in terms of home run production, we added a third component to the graphs

which measured home run totals against height and weight. Our goal was to see if there was a concentration of a large number of home runs within certain height and weight areas. We found that home run's are not concentrated among a specific height and weight range, and seem to be about evenly distributed among all the regions where height and weight are correlated. In the last graph of 2017, there are higher home run rates among the taller and heavier players. We will run a logistic regression to verify our results.

The Game

Changes in player philosophy can be seen throughout the history of the MLB as far back as the introduction of Babe Ruth. These changes have historically been met with rule changes from MLB executives. Most notably, during the early 1960's, teams notoriously changed mound heights in order to give their pitches an advantage. This practice introduced a new form of strategic play that was available due to the current mathematical concepts. These inconsistencies in parks were deemed unfair to the players by the MLB. By 1969 the MLB established a new rule mandating that all mounds be the same height. After this period the MLB has continuously changed the height of the mounds in favor of home runs. As stated by Getz, the increase in home runs continues to be a profitable play for the MLB. This is because fans are intrigued by home runs.

Are there more points/hits per game?

When was the last time the mounds were changed?

Have the lengths and heights of the fences changed?

The Ball

When speaking about the recent increase in home run production many say the ball is "juiced." This term stems from the notion that the ball has been made lighter and/or bouncier. In turn, the behavior of the ball would favor home runs because this would make the ball bounce off the bat faster and stay in the air longer. The notion of a "juiced" ball is not unique to just this recent surge. Many players within the MLB have spoken out about their speculations that the ball has been juiced. Most notably, Justin Verlander, Houston Astros pitcher, tweeted "All I'm saying is I don't care if balls are juiced (seriously). We're all using the same ball so it's a fair field. My issue is I don't like being lied to. I knew something was different. Century old records are being broken and numbers are skewed." Our investigation into the dynamics of the ball include looking into the physical changes of the ball: mass, seam height, and pill.

Tweet from Justin Verlander

Manufacturing of the Baseball

The manufacturing of the baseball begins with the construction of the pill. The pill is the center of the baseball. First, a cork is lined with two rubber materials. The composition of these rubber materials starts with a black rubber lining surrounding

the cork followed by a pink rubber lining surrounding the initial black lining. The pill is very essential to the baseball. The manufacturer of the pill is different from the manufacturer of the entire baseball. The manufacturer of the baseball is Rawlings Sporting Goods, located in Costa Rica while the manufacturer of the pill is Hultec, located in Fort Worth, Texas.

After the pill is molded by Hultec, it is transported to Rawlings Sporting Goods, and the construction of the whole baseball begins. The initial step that Rawlings performs is to cover the pill with an adhesive so that layers of yarn may be wound. The first layer of yarn is a wool fabric. This winding process is performed by a computerized machine that is specialized for this task. An additional two layers of the wool yarn are wound around the pill. In addition to winding the multiple layers of yarn, the machine also weighs the ball at each step to verify that it is kept within MLB standards. All these different kinds of yarn give the ball a resilient factor such that the ball can be hit multiples times and keep its shape. The last layer of yarn is thin so that it smooths out the baseball into a defined sphere. Adhesive is then placed on in preparation for the leather hide.

The leather hide is manufactured by Rawlings owned tannery operation, located in Tullahoma, Tennessee. The cow hides are acquired mostly from a Cargill(ADD FOOTNOTE) beef plant in Wyalusing, Pennsylvania. Two leather hides cut out in figure eights are wrapped around the ball. A Rawlings employee then is tasked with sewing the two leather hides together. The sewing technique involves using two needles with waxed red thread. After this process the balls are rolled for 15 seconds to level any raised stitches, then the balls are inspected and, if they are good enough, are pressed with the official logos of the MLB.

The Pill

The pill is a construction of a cork center and a rubber lining outside. The pill is what contributes the most to the "bounciness" of the ball. In relation to the entire baseball, the pill is the most dense and holds the most weight. The manufacturer of the baseball is Rawlings Sporting Goods, located in Costa Rica while the manufacturer of the pill is Hultec, located in Fort Worth, Texas. Changes to the pill have been identified by researchers at fivethirtyeight.com. Their experiment involved implementing the Computerized Tomography(CT) scan. In addition to this, the team was also able to implement thermogravimetric analysis(TGA) in order to determine the chemical composition of the cork.

The researchers were able to collect official baseballs used by MLB teams prior to the 2015 All Star Games and after. These balls were bought on eBay and were authenticated through MLB's authenticator program. In addition to this, one unused ball was purchased from Rawlings Sporting Goods which is the official manufacturer of baseballs used in the MLB. These balls were separated into two groups: pre-2015 All-Stars, and post-2015 All-Stars. All these balls underwent the same analysis as said above.

The first test the researchers at fivethirtyeight.com performed on the balls was the CT scan. This test was performed by Dr. Meng Law, Dr. Jay Acharya and Darryl Hwang at the Keck School of Medicine at the University of Southern California (USC). The CT scan revealed that baseballs in the same group had insignificant differences. Furthermore, baseballs compared against other groups showed significant differences.

Post-2015 balls appeared to be less dense when tested through the CT scan. Most notably the outer pink rubber layer of the pill appeared to be 40% more dense than the pre-2015 balls.

Moreover, to identify these differences in the density of the rubber lining, these balls underwent the thermogravimetric analysis. This test was done at Kent State University's chemistry and biochemistry lab. The findings were very interesting. The balls after the 2015 all-Star game had 7% more polymer and 10% less silicon, which made them more porous and less dense. This difference made the post-2015 baseballs, on average, approximately 0.5 grams less than the pre-2015 baseballs. Fivethirtyeight.com determined that these changes in the cork center could only attribute about 6 inches in flight distance. The changes in the rubber of the pill, compounded together with other simultaneous changes, had a significant effect. These effects include a change in the drag coefficients of the ball, which in turn, made the ball less resistant to the air.

As a result, according to fivethirtyeight.com, these changes could only be attributed to 25% of the increase in home runs during the 2015 to 2017 seasons. This is only a fraction of the total increase in home run production, which was about 46% from 2014 to 2017. Fivethirtyeight.com concluded that the remaining 21% could be accounted for by analyzing player philosophy. By this, the researchers speculated that a new batting trend which involved an upswing has been growing in popularity.

Batting in an upswing motion results in a powerful hit but more often than not, limits the batter in their ability to make contact with the ball. As a result, if more batters were adapting this new trend of an upswing batting motion, then we should expect a higher percentage of strikeouts. In our investigation, we chose to use strikeouts as a proxy for new batting techniques. Our findings reinforced that notion that strikeouts had also been increasing alongside home runs. This is a perspective of identifying the causes of the recent spike in home runs that had not been previously analyzed in most of the other efforts we evaluated. We believe that analyzing the increase in strikeouts was as important as analyzing the increase in home runs. This phenomenon shed some light to recent commentary from MLB pitchers about the recent surge in home runs. This was contradictory to recent claims by pitchers who are saying that people who used to strike out are now hitting more home runs as shown in the figures below.

Image of twweets

Strike outs

We believe that the results from fivethirtyeight.com could also explain the recent surge in strikeouts. In addition to giving batters an advantage, in terms of lower drag coefficients, increase in bounciness, and a lighter ball, these properties could also give pitchers an advantage. As shown in figure(density speed/hr/strikeout) the density of home runs and strikeouts, with regard to release speed, can be observed. The distribution of home runs peaks in the range of 90 to 95 MPH pitches. Strikeouts resemble a bi-modal distribution, with the first peak around 85 MPH pitches, and the second peak closer to 95 MPH pitches. The phenomenon of a double peak in strikeouts describes the existence of two distributions. We believe the first peak in the strikeout distribution is a result of slower pitches. Of these, we suspect pitches such as a change-up and types of curve-balls, those which by nature have a slower speed, would contribute to the first peak of strikeouts. The second peak of strikeouts trends towards higher pitch speeds, which would resemble fast balls. This high concentration

of pitched balls in the strikeout and home run distribution identifies the popularity of a fast ball, even though balls thrown in that range are most commonly home runs, are also most likely to be strikeouts.

Logistic Regression

We chose to use a logistic regression to model the probability of hitting a home run. Our input variables included: release speed, launch speed, weight, height, age, handedness of the pitcher and batter. The interaction of weight and height was examined in a third regression.

Logistic Regression pre and post 2015 comparisons

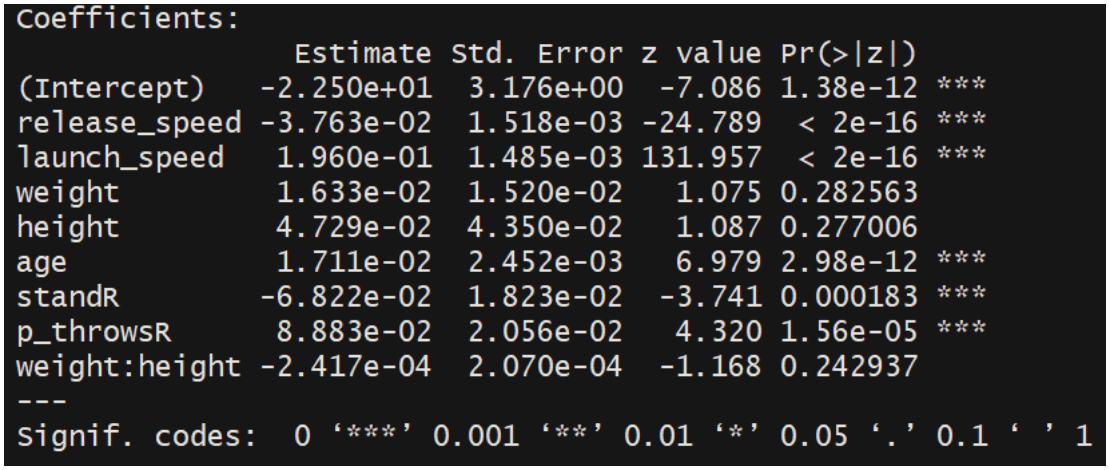
We found that all variables were significant in the pre-2015 regression. Upon running the post-2015 regression, the variable of age lost significance in the model. We believe this is due to all age groups being affected by the surge and therefore lowering the effect of age on hitting a home run.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.8816546  0.3540160 -25.088  < 2e-16 ***
release_speed -0.0063463  0.0012061  -5.262 1.43e-07 ***
weight        0.0094586  0.0005333  17.737  < 2e-16 ***
height        0.0500789  0.0049122  10.195  < 2e-16 ***
age           0.0116417  0.0025300   4.601 4.19e-06 ***
standR        0.0908204  0.0147309   6.165 7.03e-10 ***
p_throwsR     0.0441629  0.0162320   2.721  0.00651 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.4279544  0.3722917 -19.952  < 2e-16 ***
release_speed -0.0058135  0.0013788  -4.216 2.48e-05 ***
weight        0.0090255  0.0005521  16.347  < 2e-16 ***
height        0.0403999  0.0053456   7.558 4.11e-14 ***
age          -0.0038989  0.0022585  -1.726   0.0843 .
standR        0.0703056  0.0167903   4.187 2.82e-05 ***
p_throwsR     0.0792483  0.0189381   4.185 2.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression 2015 through 2017

Data on exit velocities was available in years after 2015. Because of this, we wanted to see the effect of adding exit velocity as a contributing variable to hitting a home run. Upon adding exit velocity, we found that both height, weight and their interaction no longer were significant variables. This is because those variables are being used as a proxy for strength. When exit velocity is present, weight and height are no longer the most accurate variables to capture the strength of the player, velocity is.



```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.250e+01  3.176e+00  -7.086 1.38e-12 ***
release_speed -3.763e-02  1.518e-03 -24.789  < 2e-16 ***
launch_speed   1.960e-01  1.485e-03 131.957  < 2e-16 ***
weight         1.633e-02  1.520e-02   1.075 0.282563
height         4.729e-02  4.350e-02   1.087 0.277006
age            1.711e-02  2.452e-03   6.979 2.98e-12 ***
standR        -6.822e-02  1.823e-02  -3.741 0.000183 ***
p_throwsR      8.883e-02  2.056e-02   4.320 1.56e-05 ***
weight:height -2.417e-04  2.070e-04  -1.168 0.242937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Pre-2015 Logistic Regression

Conclusion

Acknowledgments

This work was supported through NSA grant number H98230-18-1-0009 and NSF grant DMS-1731082, to Dr. Javier Rojo PI, through the RUSIS @ Oregon State University.

Reference

Arthur, Rob, and Tim Dix. 2018. “We X-Rayed Some MLB Baseballs. Here’s What We Found. | FiveThirtyEight.” https://fivethirtyeight.com/features/juiced-baseballs/.

Getz, David. 2009. “Going , Going , Gone !: How the Home Run Has Changed Major League Baseball Going , Going , Gone !: How the Home Run Has Changed Major League” 10 (1).

Petti, Bill. 2018. “Bill Petti - Global Data Strategist - Gallup | LinkedIn.” https://www.linkedin.com/in/billpetti.