Science Meets Sports

Science Meets Sports:

When Statistics Are More Than Numbers

Edited by

Christophe Ley and Yves Dominicy

Cambridge
Scholars
Publishing

Science Meets Sports: When Statistics Are More Than Numbers

Edited by Christophe Ley and Yves Dominicy

This book first published 2020

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-5856-8
ISBN (13): 978-1-5275-5856-4

We dedicate this book to our families and to the memory of Jacques Dominicy. Christophe Ley especially dedicates this book to his wife Nadine.

Contents

Preface

Aim of the Book

The objective of this book is to present the field of sports statistics to two very distinct target audiences. On the one hand the academicians, mainly statisticians, in order to raise their interest in this growing field, and on the other hand sports fans who, even without advanced mathematical knowledge, will be able to understand the data analysis part and gain new insights into their favourite sports. The book thus offers a unique perspective on this attractive topic by combining sports analytics, data visualisation and advanced statistical procedures to extract new findings from sports data such as improved rankings or prediction methods.

Football, tennis, basketball, track and field, baseball–every sport aficionado should find his/her interest in this book. Thanks to cutting-edge data analysis tools, the present book will provide the reader with completely new insights into his/her favourite sport and this in an engaging and user-friendly way.

Context of the Present Book

The world of sports is currently undergoing a fundamental change thanks to the upcoming trend of sports analytics. Recent advances in data collection techniques have enabled the collection of large, sometimes even massive, amounts of data in all aspects of sports, such as tactics, technique, health complaints and injuries, spatio-temporal whereabouts (e.g., tracking data from GPS), but also marketing and betting. Data are by now regularly collected in almost every sport, ranging from traditional Olympic disciplines to professional football, basketball, and handball, to name but a few. Moreover, massive data from individual recreational athletes, such as runners or cyclists, is available. It is by far not only professional and commercially successful sports clubs that aim to analyse data, even recreational athletes and amateur clubs make use of a variety of sensors to monitor their training and performances.

This global rush towards using advanced statistics and machine learning (or, in modern terms, Data Science) methods in sports is due, in large parts, to the success of the Oakland Athletics baseball team in the 2002 season. Prior to that season, they had hired new players in a till then atypical way, namely by not relying on scouts' experience but rather on sabermetrics, the technical term for empirical/statistical analysis of baseball. This particular story of general manager Billy Beane relying on the use of analytics to assemble a competitive team despite Oakland's small budget has been written up in the famous book *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis in 2003, which was released as movie in 2011 under the title *Moneyball.* The success of the Oakland Athletics team has inspired other teams in baseball and soon after in several other sports. Since then, sports analytics as a field has seen a phenomenal development, having led *inter alia* to the developments of new journals such as the Journal of Sports Analytics whose first edition appeared in 2015.

The present book inscribes itself in this context and aims at further contributing to this stimulating research area thanks to its unique feature of targeting academicians and sports fans.

Content of the Present Book

Various popular sports will be described in this book from a scientific and data-driven perspective. The book covers baseball (Chapter 1 by Albert), basketball (Chapter 2 by Manisera, Sandri and Zuccolotto), both baseball and basketball from the perspective of measurement theory in sports (Chapter 3 by Miller), football (Chapter 4 by Brefeld, Lasek and Mair and Chapter 5 by Groll, Schauberger and Van Eetvelde), running (Chapter 6 by Theisen, Nielsen and Malisoux), net games in general (Chapter 7 by Lames) and tennis in particular (Chapter 8 by Kovalchik and Chapter 9 by Koning and Boot, who focus on betting aspects). Finally, Chapter 10 by Goossens, Yi and Van Bulck covers fairness trade-offs in time-tabling, which is a relevant topic for several sports.

Acknowledgement

We wish to thank all contributors to the present book, which, we hope, will please the reader. We also thank Yvonne Fromme for her professional proof-reading of the entire book. All remaining mistakes are ours.

Chapter 1

The Home Run Explosion

Jim Albert
Bowling Green State University

Abstract

In the game of baseball, many of the runs scored are contributed by home runs. There has been a dramatic increase in the rate of home run hitting in recent seasons, prompting a scientific study to better understand the reasons for the home run increase. Using the new Statcast data, one records the launch velocity and launch angle for every batted ball. Using data from the 2015 through 2019 seasons, this chapter explores the relationship between the launch conditions and home run rates. By use of a generalised additive model, we gain some understanding about the reasons behind the home run increase.

1.1 Introduction

1.1.1 Game of Baseball

Baseball is a bat-and-ball game played by two opposing teams. A game of baseball consists of a series of innings, with each inning consisting of two half-innings. In the top half-inning, the visiting team is batting, and the home team is on the field, and the roles of the two teams are switched for the bottom half-inning.

During a particular half-inning, the pitcher on the fielding team will throw a ball which a player on the opposing team, the batter, will attempt to hit. The intention of the batting team is to have runners advance through four bases (first base, second base, third base, and home plate) to score runs. Each batter will advance to a base or get out, and the half-inning concludes when three outs are recorded. In a professional baseball game, nine innings are played, and the team that scores the most cumulative runs is the winner.

When a new batter comes to bat in a “plate appearance”, the pitcher will throw a sequence of pitches. If the batter does not swing at a pitch, then the umpire will call a “strike” or a “ball” depending on the location of the thrown ball. If the batter swings and misses, or if a batter hits the ball in foul territory, a strike is recorded. There are several ways that the plate appearance can conclude. If three strikes are recorded where the last pitch is a called or swinging strike, the batter strikes out. If four balls are recorded or if the batter is hit by the pitch, then the batter can advance to first base. The plate appearance can also end when the batter hits the ball “in-play”. When a ball is put in-play, there can be an out (achieved usually by a grounder thrown to first base or a pop-up or fly ball that is caught by a fielder) or a base hit. There are several types of hits (single, double, triple, and home run) distinguished by the number of bases reached by the batter on the hit. The most dramatic hit is the home run, typically achieved when the batter hits a ball a good distance so that it lands over the outfield fence. In this case, the batter, and all runners currently on base will score runs for the batting team. A grand slam home run is a home run hit when all three bases are occupied with runners.

In a typical professional baseball game, a few hundred pitches are thrown, and most of the outcomes of these pitches are strikes or balls, and a relatively small number of balls are put in-play.

1.1.2 A Plate Appearance

Outcomes

A plate appearance (PA) is the basic confrontation between a batter and a pitcher. In a PA there are three basic events that can occur (ignoring other events such as a hit-on-pitch and catcher interference that are unlikely to occur.) The batter can strike out, he can receive a base on balls (called a walk), or he can put the ball in-play. Figure 1 displays the rates of these three events from the 1960 season to the current season (2019). One can see from the figure that there are clear patterns in these rates. The rate of striking out has shown a steady increase in recent years, and the rate of putting a ball in play has steadily decreased. The rate of walking has vacillated over this period of baseball but has shown some increase in recent seasons.

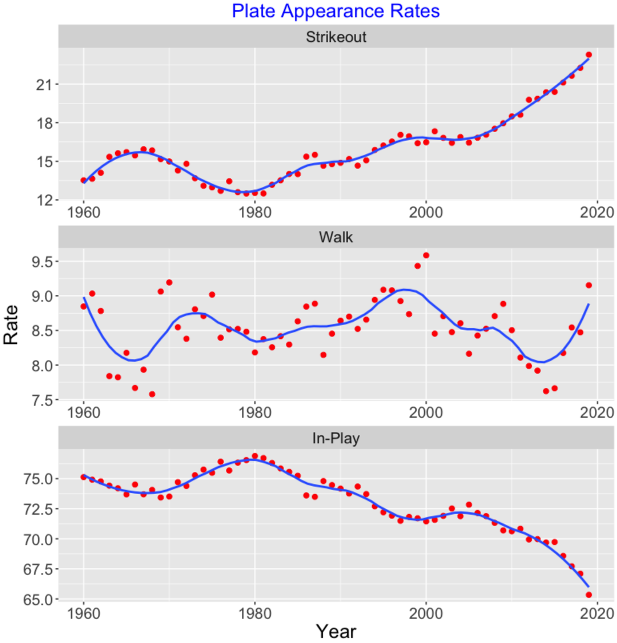

Figure 1.1: Historical pattern of three rates during a plate appearance.

Home Run Rates

The focus of this study is on the rate of home runs per batted ball. Figure 2 displays the home run rate (expressed as a percentage) for the seasons 1960 through the most recent season 2019. From 1960 through 1980 one has seen a decrease in the rate of home run hitting, followed by an increase through 2000, and then a gradual decrease until the 2014 season. There has been a dramatic increase in home run hitting the past five seasons, and the 2019 rate of 5.4% home runs per batted ball is an all-time high.

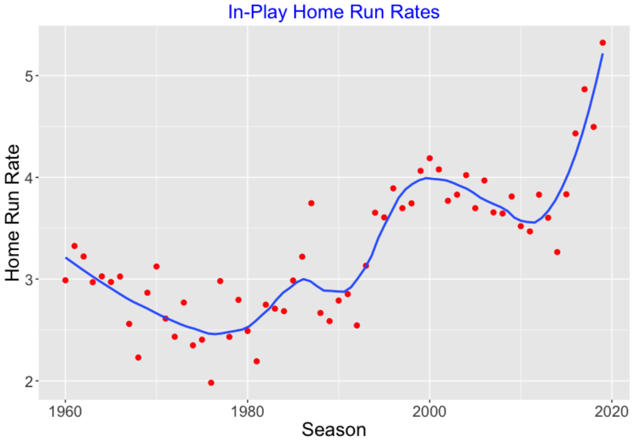

Figure 1.2: Rate of home runs per batted ball (expressed as a percentage) for the seasons 1960 through 2019.

1.1.3 MLB and the Home Run Committee Report

Major League Baseball (MLB) has been concerned about the increase in home run rates. Since home runs are currently prevalent, teams may think of home runs as a primary means of run scoring and fill their batting line-ups with players who are proficient in hitting home runs. An alternative way of scoring runs is based on putting runners on base through base hits or walks and advancing the runners by stolen bases or hits. Due to the home run increase, teams may be less interested in using these “small-ball” methods to score runs. Indeed, one observes, on average, 0.92 stolen bases in the 2019 season which is the smallest average in the past 50 seasons.

In the fall of 2017, a scientific committee was charged by the Office of the Commissioner of Baseball to "give the full benefit of their knowledge and expertise and to conduct primary and secondary research in order to identify the potential causes of the increase in the rate at which home runs were hit in 2015, 2016, and 2017."

This committee explored several possible explanations for the home run increase.

- **The batters.** It is possible that the batters are hitting balls in a different way that would contribute to the rise in home runs. Perhaps they are hitting balls harder or at a more suitable launch angle or spray angle that would result in more home runs.
- **The pitchers.** Pitchers throw different types of pitches and we have observed a general tendency of the pitch speeds of pitchers to increase in recent years. Perhaps the changes in pitch types and/or pitch speeds are causing the home run increase.
- **The weather.** Baseball is played in a six-month season from April through October and there is a great variation in the weather in the ballpark. It has been documented that it is less likely to hit a home run in cold weather. Perhaps weather changes over recent seasons have contributed to the home run increase.
- **The ballpark.** Every ballpark has a unique shape and ballparks differ in terms of the distance from home plate to the outfield fences. Also, the weather conditions differ among the 30 ballparks. For example, the altitude of Coors Field is 5280 feet and the light air contributes to changes in the movement of a baseball. Perhaps ballpark effects are contributing to the home run changes.
- **The ball.** The composition of the manufactured baseball plays an important role in how the ball moves through the air. It is possible that there have been subtle changes in the ball in recent seasons that have contributed to the home run increase.

The committee explored changes in the launch conditions (the exit velocity, the launch angle, and spray angle) of batted balls over the 2015 to 2017 period. They did not believe that the changes in launch conditions were the primary cause for the increase in home run hitting over this period. Instead, they found that the increase in home runs was primarily due to better "carry" of the baseball for given values of the launch conditions. Furthermore, the committee found that the better carry of the baseball was not due to changes in the weather condition, but instead due to changes in the aerodynamic properties of the baseball. Although the committee believed that changes in the baseball were the main culprit, it was unclear what aspects of the manufactured baseball would lead to a decrease in the drag coefficient and an increase in the ball's carry.

In this chapter we explore this increase in home run hitting given data from the 2015 through 2019 baseball seasons.

1.1.4 Statcast Data

Baseball is remarkable for the amount of data collected on each game. Ever since the beginning of professional baseball in the 19th century, box score data was collected containing the number of at-bats, hits, and runs for each player for every game. (An at-bat is a plate appearance which does not result in a base-on-balls or a hit-by-pitch.) Due to the grassroots efforts of Retrosheet, play-by-play computerised records for every game have been collected, and entire play-by-play records for entire seasons are available at `retrosheet.org`. Starting in 2006, Major League Baseball began to install cameras in every baseball stadium to record information about each pitch. The PitchFX system, created and maintained by Sportsvision, provides information about the speed, movement, and location of every pitch. This data is publicly available and R packages such as pitchRx (Sievert, 2014) allow a person to easily download PitchFX data for a particular group of games.

Statcast (Statcast, 2019) represents the new generation of baseball data. This system was started in 2015 and collects the movements of each player on the baseball field. For a specific player one observes spatial-temporal data–specifically, his location on the field over a fine grid of time values during each inning. In addition, many variables are recorded for each batted ball that is put in-play. One collects the following variables:

- the launch speed, the speed off the bat, measured in miles per hour;
- the launch angle, the angle, measured in degrees, which the ball leaves off from the bat relative to the horizon;
- the spray angle, the horizontal angle of the batted ball relative to home plate.

The complete Statcast dataset is presently only available to the professional teams. However, the batted ball measurements are currently available through the Baseball Savant website at baseballsavant.mlb.com. The `baseballr` package of Petti (2019) provides functions for downloading this selected Statcast data over a time period of interest.

For the work of this chapter, Statcast data were collected for all batted balls during the five seasons 2015 through 2019. For each batted ball one collects the launch speed, launch angle, and spray angle. In addition, the data includes the indicator variable `HR` which is equal to one if a home run was observed and equal to zero otherwise.

1.1.5 Plan of the Chapter

The general goal of this chapter is to gain insight about the increase in home run hitting by examining the relationship between the batted ball launch measurements and home run rates. Section 2 takes an exploratory approach where one identifies the launch angle and launch speed measurements that lead to home runs, and one looks at the rate of home run hitting in this region of measurements. We explore how specific rates vary across months of a season and between the 2015 and 2019 seasons. Section 3 uses a modelling approach to see how the probability of a home run depends on launch conditions, month, and season. This modelling approach allows us to predict the home run count in the 2019 season based on the "ball carry" characteristics of previous seasons. In Section 4 we summarise the main findings and discuss related research about the home run increase.

1.2 Empirical Perspective

1.2.1 Launch Conditions: The RED Zone

Following the general strategy in the MLB Home Run Report of Albert et al. (2018), we focus on the values of the launch conditions (launch speed and launch angle) that tend to produce home runs. Figure 3 displays a contour graph of (launch angle, launch speed) values for all home runs hit during the 2019 season and a rectangle is drawn which contains the launch conditions for 78% of the home runs hit for this season. This rectangle is defined by launch angles between 20 and 35 degrees and launch speeds between 98.5 and 108.5 mph. For the remainder of this chapter, we will refer to this region of launch conditions of batted balls as the "RED" region.

We are interested in values of launch conditions among all batted balls that are favourable for hitting home runs. In addition, among all of these "favourable" batted balls, it is of interest to see how many are home runs. This discussion motivates the consideration of the following two rates.

- R_{RED} = fraction of batted balls with launch angle and launch speed measurements in the RED region.

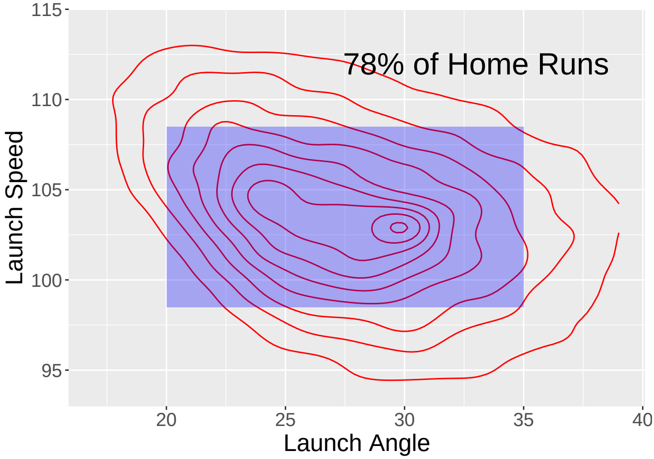

Figure 1.3: Density estimate of the launch angle and launch speed measurements for the home runs hit during the 2019 season. The rectangle contains the values of launch angle and launch speed for which 78 percent of the home runs occur.

- R_{HR} = fraction of balls in this RED region that are home runs.

Both rates are informative about the process of home run hitting. The R_{RED} rate is helpful for understanding possible changes in batting style over seasons. Since players are becoming more familiar with launch angles, it is possible that they will adjust their swing to produce batted balls with suitable launch angles leading to home runs. The R_{HR} rate is helpful for understanding the carry effect of the baseball. If the ball is made in such a way that will change the drag or resistance to the air, this change would result in increased carry and a change in the values of R_{HR}.

1.2.2 Changes in Rates of Batted Balls in the RED Region

Figure 4 graphs the rate of batted balls in the RED region, R_{RED}, for each month of the seasons 2015 through 2019. First, one observes an interesting pattern for the 2015 season–the rate was low in the first three months of the 2015 season but dramatically increased in the second half of the season. Comparing seasons, one observes a steady increase in "favourable" batted balls from 2015 through 2019. For example, this RED rate was in the 5 to 5.5 percent range in the 2016-2017 seasons, increasing to 5.5-6.0 in the 2018 season and 6-6.5 in the 2019 season.

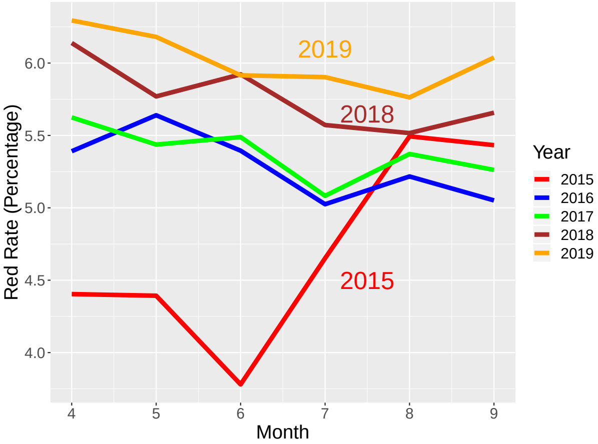

Figure 1.4: Rate of RED zone batted balls for each month of the seasons 2015 through 2019.

Although there is strong evidence for an increase in "home run favourable" batted balls over seasons, there are different populations of hitters for the different seasons. For example, there are rookie players in the 2019 season who did not play in previous seasons and veteran players in earlier seasons who may have retired and did not play in the 2019 season. To control for the changing groups of players, one can focus on particular players who played in both the 2015 and 2019 seasons and see how the favourable batted rates have changed for these players.

We focused on the players who had at least 200 batted balls in 2015 and 100 batted balls in 2019. Figure 5 displays a scatterplot of the RED region rates for these players together with a smoothing curve found using the loess (locally estimated scatterplot smoothing) procedure (Cleveland, 1979). One can compute that 75% of these players had a higher batted ball rate in the RED region in 2019. This indicates that players are indeed changing their swinging style to produce harder hit batted balls with good launch angles. Also, it is interesting that this increase in RED batted ball rates appears to be largest for players with moderate RED rates. The conclusion is that this change in batted ball rates is occurring for players with low, moderate, and high slugging abilities.

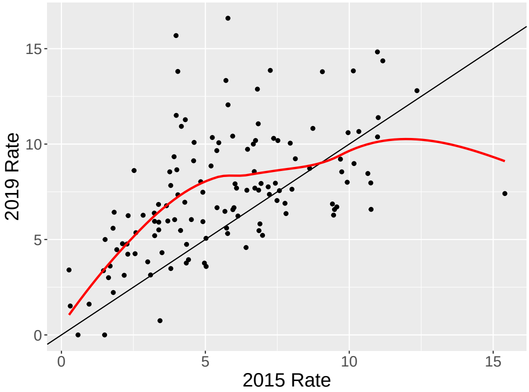

Figure 1.5: Scatterplot of RED region rates for players with large number of batted balls in the 2015 and 2019 seasons.

1.2.3 Changes in Home Run Rates for Batted Balls in the RED Region

Next, we focus on the percentage of batted balls in the RED region that are home runs. Given that a batter has hit a ball with suitable values of launch angle and launch velocity, what is the chance that it will be a home run? This question deals with the characteristic of the baseball to have sufficient carry for a home run.

Figure 6 displays these home run rates over different months and seasons. There is a clear weather effect. Generally, home run rates are smallest in the cold weather month of April and home run rates are larger for the warmer months of June, July, and August.

The pattern of change of home run rates in the RED region over seasons is more complicated. In the 2015 season, the home run rates sharply decreased in the second half of the season. From the second half of the 2015 season through the 2017 season, there was a steady increase in home run rates. This indicates that there was a systematic change in the characteristics of the baseball that led to less drag and an increase in home run rates.

In the last two seasons, we see a different pattern in these home run rates. In 2018, the home run rates in the RED region dramatically decreased and the rates in 2019 resemble the rates in the 2016 season.

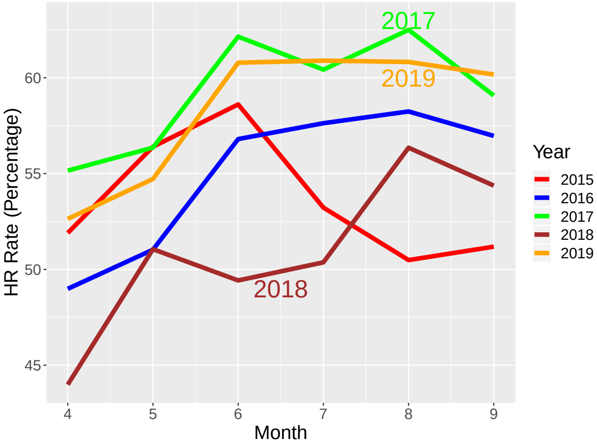

Figure 1.6: Home run rates of batted balls for different months and seasons from 2015 to 2019.

This indicates an increase in drag characteristics of the baseball for 2018 and a decrease in drag in 2019, but not to the level of the 2017 season.

1.3 Modelling Perspective

1.3.1 Introduction

In the empirical approach the RED home run rate was helpful in learning about the likelihood of a home run given suitable values of the launch angle and launch velocity. An alternative approach is to use a statistical model to better understand how the likelihood of a home run depends on the launch conditions. Specifically, one is interested in modelling the probability that a batted ball is a home run based on the launch conditions and season and month effects.

1.3.2 Generalised Additive Model

Let p denote the probability a batted ball is a home run. Suppose we consider the use of the generalised additive model (GAM)

$$\log\left(\frac{p}{1-p}\right) = s(LA, LS) + Season + Month + Season * Month.$$

In this model the term $s(LA, LS)$ denotes a smooth function of the launch angle (LA) and launch speed (LS) and $Season$ and $Month$ denote categorical effects to the season and month, respectively. It is possible that the month effect depends on the season, so this model includes an interaction effect of season by month.

To demonstrate the usefulness of the nonparametric function $s(LA, LS)$, Figure 7 displays contours of the GAM fitted probability a batted ball is a home run for a region of values of launch angle and launch speed. The contour levels are drawn at fitted probability values of 0.1, 0.3, 0.5, 0.7, and 0.9. In Figure 7, one sees that batted balls hit higher than 100 mph with a launch angle between 25 and 35 degrees tend to be home runs. In addition, note that a higher launch angle compensates for a lower launch speed. For example, the probability of a home run for a batted ball at 30 degrees and 102 mph is approximately equal to the probability of a home run for a batted ball at 20 degrees and 110 mph.

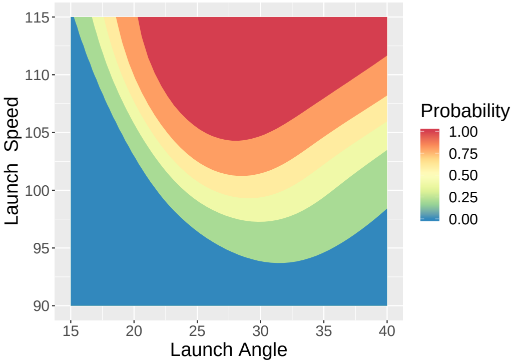

Figure 1.7: Contour graph of probability of home run as a function of the launch angle and launch speed.

1.3.3 Estimating Home Run Probabilities

One use of the GAM model is to estimate the probability a batted ball is a home run at particular launch conditions (launch speed and launch angle) during a particular month and season. This fitted probability is helpful for understanding how the ball carries as a function of launch measurements and specifically how the carry of the ball changes as a function of the month and season.

We focus on particular values of launch angle and launch speed that lead to large home run probabilities. Specifically, consider a batted ball hit at a launch angle of 25 degrees and a launch speed of 102 mph. Figure 8 displays the fitted GAM home run probability at these launch conditions for different months and seasons. There are some interesting takeaways from this graph. First, one notices the weather effect–for each season, the smallest home run probability occurs during April, and the probability increases as one moves from April to August. Second, there are substantial differences between the fitted home run probabilities across seasons. Focusing on the home run probability in August (Month = 8), note that there was an increase in the fitted probability from the 2015 to 2017 seasons, but this probability decreased in the 2018 season. These patterns are consistent with the patterns of rates of home runs in the RED region seen in the empirical approach in Figure 6.

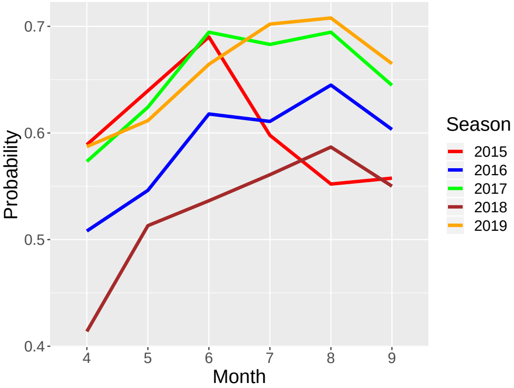

Figure 1.8: Fitted GAM probability of a home run for different months and seasons when the launch angle = 25 degrees and launch speed = 102 mph.

1.3.4 Predicting Home Run Counts

This GAM model can also be used to predict home run counts in future seasons. For example, we observed a surge in home run hitting for the 2019 season. The current home run record was 6105 from the 2017 season and there were 6776 home runs hit in 2019, which broke the season record by 11 percent. Scientists are discussing the reason for this home run surge. Is this surge due to the carry of the ball, or is this surge due to the change in the launch measurements of the hitters?

One can address this question by use of the GAM model. First, we fit this GAM model using all the data from the 2015 through 2018 seasons. Essentially, one is using all the baseballs from the 2015 through 2018 seasons to understand the relationship between launch angle, launch speed, month, and the probability of a home run. Then this fitted GAM model is used to predict the 2019 home run count using the observed 2019 launch condition measurements. If the GAM prediction is close to the actual 2019 home run count, then this tells us that the increase in 2019 home run hitting is due to the changes in launch measurements for the 2019 season. If, instead, our GAM prediction is too small (underestimates the actual 2019 home run count), then that suggests that other inputs, such as the change in the carry of the 2019 balls, are contributing to the 2019 increase.

Using the fitted GAM model, one can obtain a predictive distribution for the 2019 home run rate. Using the launch conditions for each of the batted balls in the 2019 season, one obtains fitted probabilities $\hat{p}$ of a home run for these batted balls. By use of random numbers together with these fitted probabilities, one can predict the total 2019 home run count. By repeating this exercise for 1000 iterations, one obtains a predictive distribution for the home run count.

Figure 9 displays a histogram of the simulated predictions of the 2019 home run count from the GAM model. This figure tells us that the prediction of the 2019 home run count is likely to fall between 6422 and 6593. The actual 2019 home run count of 6776 is represented by a vertical line in the figure.

What does one conclude? One takeaway is that the observed 2019 home run count is larger than the likely range of predicted values. So, this observed home run count is inconsistent with the GAM model based on data from the previous four seasons. This means that one cannot explain the 2019 home run surge solely by the changes in launch conditions in the 2019 hitters. The carry behaviour of the 2015 to 2018 balls (as measured by the GAM fitted model) together with the change in launch measurements in 2019 appear to jointly explain the 2019 home run hitting.

1.4 Conclusions and Problems for Future Study

There has clearly been a dramatic change in home run rates in recent seasons of Major League Baseball, but the reasons for this change are not clear. As described in Section 1, there are many possible explanations for the increase in home runs, such as changes in pitching and batting styles and the composition of the baseball. This chap-

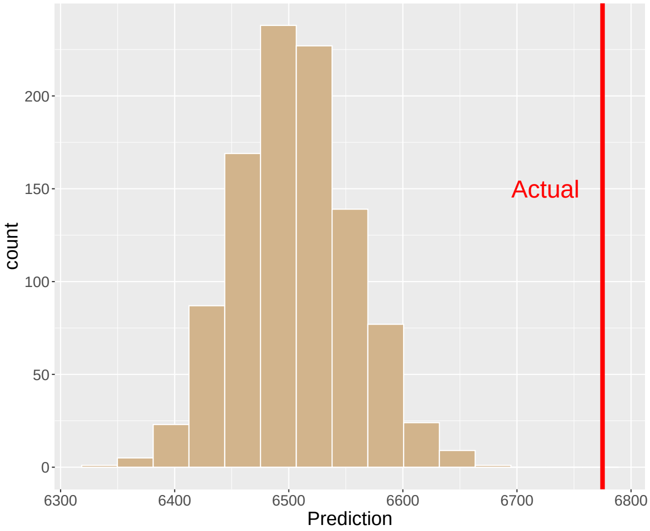

Figure 1.9: Prediction distribution from GAM for home run count in the 2019 season.

ter has focused on the launch conditions of batted balls, specifically the launch angle and launch speed measurements, and the relationship of these launch conditions with home runs. This chapter has demonstrated that the launch conditions of players have changed in recent seasons. Players are generally hitting balls harder and hitting at higher launch angles that would contribute to more home runs. But the composition of the ball appears to play an important role. For example, a contributing factor to the great increase in home run hitting in 2019 compared to 2018 appears to be additional carrying effect of the 2019 baseball.

There is an active effort to learn more about the changes in manufactured baseballs between seasons. For example, Wills (2018) has taken apart baseballs from different seasons and showed how the characteristics of baseballs have changed. Rogers and Ciaccia (2019) describe efforts of Lloyd Smith to measure baseballs with more precision to better understand which characteristics of the baseball would lead to more home runs.

Major League Baseball has been concerned with the increase in home run hitting, thinking that this increase will lead to a fundamental change to baseball and may be less attractive to the fans of the game. In the 2019 season Frank (2019) describes some rule change experiments by the MLB in the Atlantic League (an independent professional league) to see if any changes might lead to a decrease in home run hitting. Currently, baseball plate appearances are dominated by the so-called “three true outcomes” of home runs, strikeouts, and walks that only involve two players–the pitcher and the batter. Many people believe that baseball will be more popular in the future if there are more balls placed in-play that involve all the defensive players in the field. It remains to be seen if the game will eventually move away from the three true outcomes.