Forecasting Batting Averages in MLB

by

Sarah Reid Bailey

B.Sc. University of the Pacific, 2015

Project Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in the

Department of Statistics and Actuarial Science

Faculty of Science

© Sarah Reid Bailey 2017

SIMON FRASER UNIVERSITY

Fall 2017

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Approval

Name:	**Sarah Reid Bailey**
Degree:	**Master of Science (Statistics)**
Title:	**Forecasting Batting Averages in MLB**
Examining Committee:	**Chair:** Jinko Graham Professor
	Tim Swartz Senior Supervisor Professor Simon Fraser University
	Jason Loeppky Supervisor Associate Professor UBC Okanagan
	Peter Chow-White Internal Examiner Associate Professor
Date Defended:	**November 14, 2017**

Abstract

We consider new baseball data from Statcast which includes launch angle, launch velocity, and hit distance for batted balls in Major League Baseball during the 2015, and 2016 seasons. Using logistic regression, we train two models on 2015 data to get the probability that a player will get a hit on each of their 2015 at-bats. For each player we sum these predictions and divide by their total at bats to predict their 2016 batting average. We then use linear regression, which expresses 2016 actual batting averages as a linear combination of 2016 Statcast predictions and 2016 PECOTA predictions. When using this procedure to obtain 2017 predictions, we find that the combined prediction performs better than PECOTA. This information may be used to make better predictions of batting averages for future seasons.

Keywords: Batting Average, MLB, Logistic Regression, Big Data, Forecasting

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sports analytics is becoming an ever increasing component of sports. Different teams and sports use data in a variety of functions. The main use of analytics and the focus of this project is on predicting player performance. However analytics is also prominent in the sports business sector, digital media, and for injury prevention (Davenport, 2014). By combining statistical analyses and sport, researchers are able to provide deeper insight into player and team abilities for both coaches and fans. From a team's perspective, analyzing data can give coaches and scouts a different view of the strengths and weaknesses of players and recruits. Not only can analytics be used on a player level, but also on a team level to highlight insufficient areas in both their own team and their competitors. This can give teams a competitive advantage by filling missing spots in their roster and exploiting other team's weaknesses. Both players and managers have the opportunity to use data to negotiate contracts and salaries based on player performance and progression. Managers can also use it when hiring or firing coaches by looking at the history of their coaching performance from an analytical mindset (Brousell, 2014).

Fans have welcomed sports analytics to enhance their game experience. Fans with a statistical inclination have scraped their own data to explore and analyze. Entire websites have been created for researchers to share their findings. One such website is baseball-prospectus.com, which consists of a variety of analytics in all aspects of baseball including player performance, transaction records, fantasy teams, and salary analysis. Fans have also used data to increase their chances of success in sports betting. Whether they use their own methods or other findings, the hope is to gain an advantage in the gambling world. We discuss a multitude of applications in sports analytics, as well as how analytics have impacted technology and player tracking. Specifically, we focus on an application to baseball and batting averages, which is the concentration of this project.

1.1 Analytics in Sport

From a business perspective, analytics can be used to optimize sales performance of both tickets and merchandise as well as optimize fan experience through promotions and social media (Davenport, 2014). Davenport discusses that many teams are apprehensive to invest a lot into analytics from a business perspective, as the main focus of a team is on sport performance and winning. The most popular source of business analytics is in ticket pricing. One common approach to ticket sales is through variable pricing, which allows teams to charge different prices for different tickets and games. However, the variable price does not change throughout the season. This was initially primarily used in Major League Baseball (MLB), but starting in 2015 half of the NFL teams had also implemented variable pricing (Smith, 2015).

An alternative to variable pricing is dynamic pricing. This allows teams to change their variable price throughout the season depending on who they are playing and how their team is currently doing. Another form of business analytics involves fan engagement through their online websites. Most teams have a tracking system to see where fans are visiting on their website. They also utilize social media platforms. However, few sports are actually using these metrics to improve and customize engagement on their site (Davenport, 2014).

Health analytics and injury prevention are also becoming prevalent with sports teams. AC Milan, a professional soccer team, uses analytics from their lab MilanLab, to study both mental and biomechanical components of players. Using this data, they create alerts when a player is out of an accepted range. This team has a consistently low injury rate since fully implementing their system in 2003 (Davenport, 2014). Despite the clear benefits of injury prevention with player analytics, many teams are apprehensive to use it. Players worry they will be punished or brought out of a game when their data shows that they are on the verge of injury. Players also fear that their data will show potential injuries during a contract year, which regardless of performance, could affect the evaluation of their worth. As well, there is some discussion as to privacy issues and who will be able to access the data and to what extent it should be used (Karkazis and Fishman, 2017). Nonetheless, there can clearly be a benefit to using data for injury prevention, but it is still in the early stages of being fully developed and implemented.

The most popular use of sports data includes player analysis and evaluation. This kind of data can come in many forms including video analysis, player tracking, box scores, and other summary statistics. With analyses of these datasets, teams can find optimal line-ups for games, use findings when drafting and recruiting, and work with players to optimize their game and discuss salary options based on analytical findings (Davenport, 2014). The opportunity for applications is endless, but using analytics for player evaluation is still a work in progress. One major backlash that sports science has faced is that decision makers sometimes have little statistical experience. From their perspective, making decisions based

on analytics is risky and goes against the sports culture of intuition and gut (Alamar and Mehrotra, 2012). Despite the backlash, teams that can find a balance between traditional methods and analytics can be very successful.

Teams across all disciplines have found a place for data. The NBA is one example where advanced analytics have begun to play a prominent role in decision making. Although not the best example of sport performance, the Philadelphia 76ers are one such team that have recently embraced analytics. In May 2013, they hired General Manager and President of Basketball Operations Sam Hinkie, who has an analytics first mindset. Hinkie began rebuilding the team from scratch using advanced statistics to build the roster (Torre, 2015). Hinkie has since been let go, but his decision making will influence the future of the Philadelphia 76ers. Some players have also begun to take an interest in their own personal data. The New England Patriots star quarterback Tom Brady, is one who takes an active approach in using analytics to enhance his game. At one point, he looked at what went wrong when he threw incomplete and intercepted passes by analyzing game day videos (Davenport, 2014).

Player analysis is also heavily used in fantasy sports and gambling. Many fans belong to fantasy leagues in which they select various players to a team that will go head to head each week with other members of their league. The team that gets the most points wins the week and the process is repeated with other members of the league throughout the season. This type of gambling is seen across most sports including hockey, football (European), baseball, basketball, and American football. There are many other methods to sports gambling, with the three main ones being moneyline, spread, and over/under wagering. Betting on the moneyline simply means the team you bet on must win. Betting with a spread requires the team you bet on must win by more than the point spread that was assigned. Finally, betting over/under allows the gambler to bet whether the combined number of points in a game will be greater than or less than the predicted value where the prediction comes from the betting platform you are using. Swish Analytics is an example of a website that gives fans (for a price) both sport predictions and betting tools to improve their Return on Investment (ROI) (Swish, 2017). Swish analytics claims their algorithms have been successful in making profit over all three main styles of betting (Connolly, 2014). Statisticians who also have a passion for sports and gambling often use their own methods when placing wagers. Whatever source fans use to gamble and play in fantasy leagues, there is a clear benefit to having access to the data and analyses.

With the influx in information, there has also been an increase in technology used to record data and increase its accuracy. Prior to the technology boom, data would consist of simple summary statistics mainly from box scores. However, when many of these box scores became digitized and made publicly available, the interest in the data also increased. Many professional sports teams use a player monitoring system consisting of video cameras and tracking devices that track both players and balls or other equipment used. SportVU is the provider of game data for the NBA and was installed in all NBA arenas in 2013. This

setup consists of six cameras in the catwalk area of the arena and has the ability to track both the movements of players and the ball 25 times per second ("STATS", n.d.). Similarly, Sportvision is the technology used in all MLB Parks. It's initial installation into all stadiums began in 2007, where its cameras captured all pitch trajectories. In 2015, Statcast was implemented, which not only tracked pitches, but also what players in the field were doing and the result of the ball after it had been hit (Kagan and Nathan, 2017). More on Statcast and the type of data it provides will be explained in Chapter 2. One major provider of sports data and technology is STATS.com. This website provides data for almost all professional sports including the NFL, NBA, MLB, NCAA, and PGA. They are responsible for the technologies described previously, but also provide information to increase fan engagement, including data feeds, digital products, and fantasy sport predictions ("STATS", n.d.).

1.2 Analytics in Baseball

The rise of technology and data has been seen across many sports, however, Major League Baseball has been at the forefront of analytics since its introduction into sport. Although baseball analytics have been a part of the game since its beginnings, sabermetrics was only first officially defined by Bill James in 1980 as "the search for objective knowledge about baseball." The name stemmed from the "Society for American Baseball Research", or SABR, which is an organization dedicated to sharing baseball ideas and research. Members come from a variety of backgrounds including broadcasters, former players, and baseball officials (Birnbaum, 2013). From that definition 37 years ago to now, sabermetrics has seen rapid growth in both the academic and the sports world. After the annual publication of "The Bill James Baseball Abstract" began in 1982, the general public became aware and interested in the benefits of analyzing sports. In 1989, baseball data became easier to access when Retrosheet was created by David Smith, which had the goal of making every box score in MLB computerized ("Sabermetrics", n.d.). This allowed easier, more reliable data for scientists to analyze. One of the most famous users of sabermetrics was Billy Beane, who was hired as the general manager of the Oakland Athletics MLB team in 1989 and became known for his eccentric methods in recruiting, that utilized sabermetric findings. In 2003, Michael Lewis published "Moneyball: The Art of Winning an Unfair Game", which gained notoriety worldwide. In 2011 the movie "Moneyball" was released that was based on the book by Lewis. This movie broadened the audience of baseball analytics. In particular, casual fans, gamblers, and non-numerically inclined people began to see the benefits of baseball analytics ("Sabermetrics", n.d.). Sabermetrics has large variation in its complexity, from basic data counting strategies, to the invention of unique metrics including on-base plus slugging (OPS), which is the sum of a players on-base percentage and slugging percentage ("OPS", n.d.). There is also more data than many MLB teams know what to do with including the trajectories of every pitch and outcomes in MLB. This data has been

used for multiple purposes including predicting game outcomes, on base percentages, and batting averages.

Albert (2016) created a method for predicting a player's batting average. In his paper he demonstrated that instead of using a standard binomial setting where the outcome of a plate appearance is a hit or an out, the result can be represented as a multinomial distribution with four outcomes: strikeout (SO), home-run (HR), hit-in-play (HIP), and out-in-play (OIP). Albert emphasized a luck component in batting averages and showed that around half of the variability in player's batting averages can be contributed to luck. By breaking up a hit into different outcomes, he addressed the issue of luck and limited it's impact when predicting batting averages. A Bayesian random effects model was used to estimate each of the posterior probabilities for the different batting outcomes. Combining these probabilities, such that

$$p_H = (1 - p_{SO}) \cdot (p_{HR} + (1 - p_{HR}) \cdot p_{HIP}) \tag{1.1}$$

gives the probability of a hit. The estimate for the probability of a hit, $\hat{p}_H$ is found by replacing the exact probabilities in (1.1) with their estimates from the Bayesian random effects model (Albert, 2016). In this project, we wish to consider a luck aspect as well. A batted ball can be classified as one of four outcomes. These include ground ball (GB), line drive (LD), fly ball (FB), and pop up (PU). We represent these outcomes in our model with various explanatory variables. In particular, the launch angle of a batted ball can be used to classify these types of outcomes. A ground ball is measured as less than 10 degrees, a line drive is 10-25 degrees, a fly ball 25-50 degrees, and a pop up is greater than 50 degrees ("MLB", n.d.b). Typically a fly ball will produce more runs than a ground ball. Therefore if a batter gets an abnormally large number of hits from a ground ball, his hits would be considered "lucky". By including launch angle in our model, the coefficient should be fitted such that a batted ball with a small launch angle will have a lower probability of getting a hit. Using this and the other variables in the model, we hope to minimize the effect luck will have on predicting batting averages. We explain launch angle and the other variables in more detail in Section 2.3.

1.3 Organization of the Project

This paper will look at PECOTA, a popular algorithm for predicting batting averages, in combination with Statcast data to see if PECOTA can be improved. In Chapter 2, we will describe the data in both PECOTA and Statcast and the different variables that were used in building our model. In Chapter 3, we explain the method used to create our own Statcast model and combine this with PECOTA to create a super model for predicting batting averages. We also discuss the predictions and compare them to various methods used for predicting batting averages. In Chapter 4 we conclude by summarizing our findings

and noting difficulties that were encountered as well as improvements that can be made. The purpose of this project is to try to use new information available via Statcast to see if the data can be used to improve PECOTA's predictions. However, with the limited data availability from Statcast, this may not be feasible at this point. Regardless, we hope to provide some insight into various metrics in Statcast and how these can be used when analyzing data and making predictions.

Chapter 2

Data

2.1 Batting Averages

The primary goal of this project is to predict batting averages. Batting averages (BA) are one method used in baseball to evaluate a players hitting performance, but can also be used to value a pitcher's skill by analyzing the batting average of batters faced. MLB.com defines batting average (BA) as $BA = \frac{H}{AB}$, where H is a players number of hits and AB is the number of the player's at-bats. An at-bat is officially defined as when a batter reaches base by a hit, error, or a fielders choice. An at-bat also includes instances when a batter is out, but not on a sacrifice hit (sacrifice fly or bunt). A hit is defined as getting to base on a fair hit, but not through a fielders choice or an error ("MLB", n.d.a). On average, MLB players have a .260 batting average with elite players having batting averages exceeding .300. Most sporting sites that keep track of MLB data record and post batting averages. Common websites include baseball-reference.com, espn.com/mlb, and MLB.com. However, virtually all websites that provide baseball statistics include batting average, or at the minimum, the number of at-bats and hits per player. Providing records on batting averages has been used for over 100 years. For example, currently the best single season batting average record belongs to Hugh Duffy, who in 1894 had a batting average of .4397 ("baseball reference", n.d.).

2.2 Prediction Systems

There are many systems in place to predict player performances including batting averages, where the top algorithms range from simple to complex. Henry Druschel from Beyond the Boxscore describes the main systems in place; Marcel, PECOTA, Steamer, and ZiPS (Druschel, 2016). Marcel the Monkey Forecasting System or simply Marcel, is considered the most simple of the four. It was developed in 2004 by Tom Tango and is often used as an initial model for other algorithms to build on. It's algorithm is available to anyone and involves using data from the last three years of a player, but gives a heavier weight to

the most recent year. It then shrinks a player's prediction to the current league average. PECOTA is the second major prediction system and the basis for this project. It will be described in Section 2.2. Steamer is a prediction algorithm that was developed in 2008 as a high school project by teacher Jared Cross and his students Dash Davidson and Peter Rosenbloom. It uses a weighted average of past performances, but also adjusts the predictions so they are closer to the league average. Steamer determines these weights and the degree of alteration based on regression analysis from previous players. The final commonly used prediction method is ZiPS or the sZymborski Projection System, that was created by Dan Szymborski and uses Voros McCracken's discovery of Defense Independent Pitching Statistics (DIPS). DIPS is based on the notion that pitchers have very little control of an opponent's batting result on balls in play (BABIP). ZiPS uses a weighted regression analysis on four years of data for experienced players and three years for newer players or players reaching the end of their careers. It then pools players together (loosely) based on similar characteristics (Druschel, 2016). Each method has its various strengths and weaknesses and performs with different efficiencies depending on the desired statistic you wish to predict.

2.3 PECOTA

PECOTA, or formally known as Player Emprical Comparison and Optimization Test Algorithm is a proprietary prediction system created by Nate Silver in $2002 - 2003$ and was bought by Baseball Prospectus in 2003 (PECOTA, 2017). Silver continued to manage the predictions until 2009. In 2008, Silver created FiveThirtyEight.com, a blog that covers politics, sports, science and health, economics, and culture from an analytical perspective. In 2008, Silver's predictions for the U.S. presidential election gained him and his website notoriety when he correctly predicted the winner for 49 of the 50 States. His prediction also accurately forecasted the winner of all 50 U.S. Senate members ("NateSilver", n.d.). PECOTA utilizes prior seasons to predict player performance using three elements. These include Major-league equivalencies, baseline forecasts and a career-path adjustment ("BaseballProspectus", n.d.). PECOTA contains predictions for seven baseball statistics, each of which has an associated confidence interval (Schwarz, 2005). PECOTA is similar to ZiPS in that it regresses on prior performances to get a baseline prediction. It then uses that prediction to cluster players into pools based on different characteristics. PECOTA is more advanced with the clustering as it considers age, position, body type, and other features. The exact comparison structure is not known due to the proprietary nature of the predictions (Druschel, 2016). This project considers the PECOTA predictions of batting average for 2016 and 2017. We only consider players who are strictly batters since pitchers only bat in the National League and often have less at-bats, as well as more fluctuation in their batting averages.

2.4 Statcast

Statcast is a relatively new data source that was implemented in 2015 across all MLB parks. Prior to Statcast, MLB only allowed the general public to access data through PITCHf/x, which measured a multitude of parameters including pitch speed, trajectory, spin, and release point. It was created by Sportvision and was implemented in every MLB stadium before the 2008 season (Fast, 2010). In 2015, Statcast was introduced, which has the capability to track pitches, balls batted, and players in the field. Statcast uses two different types of technology to do this. The first is Trackmans Doppler radar, which follows the ball and the second is ChyronHego video monitoring, which tracks the players (Kagan and Nathan, 2017). Statcast measures over 30 variables related to defensive, offensive, and pitching statistics. The model used in this project includes three variables from Statcast, which we believe are important predictors of hits. These variables are launch angle, exit velocity, and hit distance. Launch angle is the vertical angle, measured in degrees, of a batted ball when it leaves a players bat. Exit velocity is the speed of the batted ball and is measured in miles per hour (mph). Hit distance is the distance in feet from the home plate that a ball either hits the ground, or reaches its final position ("MLB", n.d.b). In Chapter 3 we describe the model as well as the transformations and interactions of the variables that were used.

Chapter 3

Data Analysis

The analysis involved three years of data from both Statcast and PECOTA as well as building two models using this data to form our predictions of the batting averages for 2017. This chapter describes the process in detail, but Figure 3.1 is provided below as a guideline to keep track of the different years of data and their role in the prediction process.

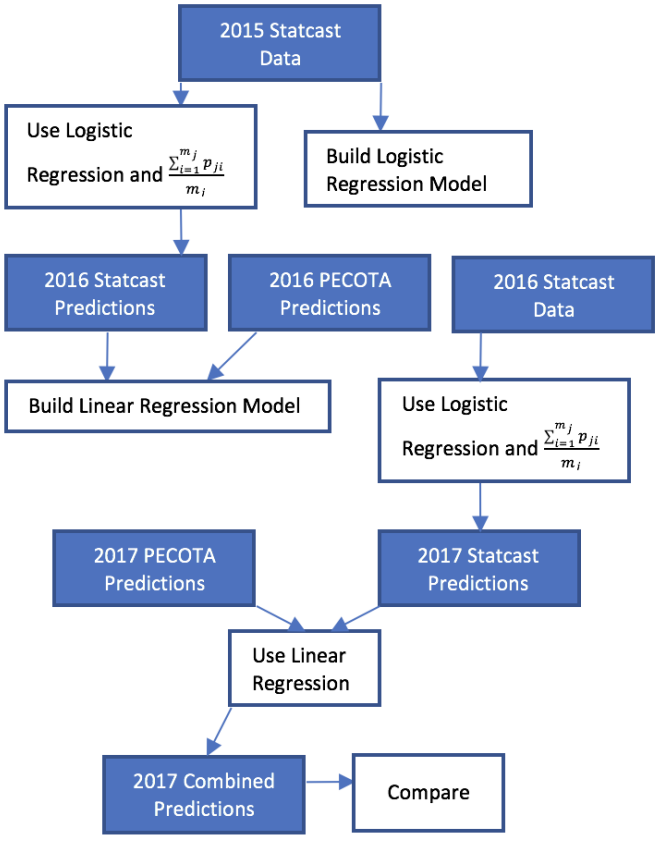

Figure 3.1: Analysis overview

3.1 Verification of Statcast Accuracy

Because it is a relatively new tracking system, we wanted to check the accuracy of Statcast data to ensure it contained all at-bats. As well, Statcast does not have a separate indicator for a hit, out, or walk. Instead it has an event column and a description column. The event column describes the outcome of each at-bat, such as field error, strikeout, double play, and more. The description column is a play by play sentence that indicates the players involved and the result of the play as well as details on the type of pitch, batted ball, and fielding play. Using a string matcher in R, we searched through the event and description to classify each at-bat as either a hit, out, or other. The other column consisted of sacrificial at-bats, walks, and hits by pitches. We grouped these together because they will not be considered in the analysis.

We randomly selected five players from 2015 and five from 2016 and found their total number of at-bats and total number of hits for the regular season of their respective year. This was then checked with www.baseball-reference.com to ensure accuracy. Eight of the players had the same number of at-bats and hits as Baseball-Reference, however there were two players that were off by one due to unique classifiers. These were Clint Robinson, who in 2016 had a double play that should've been classified as an out and Troy Tulowitzki who had an additional restraint for a fielders choice that was incorrectly labelled. The accuracy of the data is an important component of our prediction procedure. The check is shown in Table 3.1.

	Player	Aramis Ramirez	Omar Infante	Troy Tulowitzki	Bobby Wilson	Trevor Plouffe	Clint Robinson	Martin Prado	Alex Rodriguez	Andrew McCutchen	Pedro Alvarez
	Year	2015	2015	2015	2015	2015	2016	2016	2016	2016	2016
BaseballReference	Hits	117	97	136	25	140	46	183	45	153	84
	At-Bats	475	440	486	132	573	196	600	225	598	337
Statcast	Hits	117	97	136	25	140	46	183	45	153	84
	At-Bats	475	440	486+	132	573	197*	600	225	598	337
	* Off by one due to uniquely classified double play that should not be an out										
	+ additional restraint for removing fielders_choice_out included'										

Table 3.1: Checking Statcast accuracy

In this research we were interested in launch angle, exit velocity, and hit distance. We believe these variables are important indicators of whether a batted ball will result in a hit or an out. We believe launch angle will have an impact on the probability of a hit as balls with a large or small angle are more likely to result in an out. As well, exit velocity is important as faster batted balls give fielders less time to react and respond. We believe hit distance is important as a ball batted out of the park will always result in a hit (home

run). As well, we believe batted balls with a small hit distance will result in an out more often than not.

Due to it's new technology, some of Statcast's cameras do not capture every element of every batted ball. Therefore, there are some at-bats where our variables of interest were not measured. This could decrease the efficiency of our model, but leaves room for improvement in the future as the technology and accuracy improves. In particular, Statcast sometimes loses ball location for extreme cases of batted balls, including high angled pop-ups and very low, sharp angled balls or ground balls. This can be due to the system being down which Statcast classifies as an unbiased source of error, or because of the radar losing track of the ball which is considered a biased source of error. If the missed data is due to the system being down, they will input the average launch angle and speed based on the type of batted ball and the outcome. If the missing data is due to the cameras losing sight of the ball, then they have to adjust the average such that it is slightly more extreme in order to correspond to the typical batted balls that are missed (Tangotiger, 2017).

It is also possible to have missing data due to the nature of the at-bat. For example, strikeouts necessarily do not have an exit velocity covariate. This type of missingness requires special consideration in our estimation procedure. In this analysis, we consider two different models to address systematic missingness. The first logistic regression model removed batted balls that contained empty data for any of our variables and removed batted balls that were classified as either a ground ball, pop-up, or when the result of the ball was a strikeout. The second logistic regression model used ground balls and only considered exit velocity as a predictor of the probability of a hit. In our approach, we believe we handle systematic missingness in a sensible way. The other types of missingness (i.e. missing at random and due to technical problems) account for only 2.3% of our data (3280 cases from 140896 at-bats)

3.2 Using 2015 Statcast Data to Predict 2016 Batting Averages

Two logistic regression models were used to predict the probability of a hit. A logistic regression was chosen for several reasons including the simplicity and interpretability of the model while still maintaining a relatively high efficiency. Logistic regression is particularly useful in the case of predicting the probability of success and assumes each event, in our case each at-bat is independent of other at-bats. This assumption is a common approach taken by statisticians analyzing baseball data.

We began with three years of Statcast data, 2015, 2016, and 2017. The first step was to use 2015 data to predict 2016 batting averages. Initially we had considered all players. However, we found that some players had very few at-bats, which would affect the final

results (Table 3.2). Therefore we chose to use batters with 200 or more at-bats to build the model. This resulted in using 84.1% of all the at-bats (140896 cases of the 167606 total at-bats). The logistic regression where all variables were recorded used 49845 at-bats or 35.4% of the data for 2015. For the ground ball logistic regression, 50717 at-bats were used or 36.0% of the 2015 data. Including the at-bats for pop-ups (7341 at-bats or 5.21% of data) and strikeouts (29713 at-bats or 21.09% of data) resulted in the use of 97.7% of 2015 at-bats. This resulted in the prediction of 334 players' batting averages for 2015 and 333 for 2016.

Player	ABs	Hits
Ronald Belisario	1	0
Branden Pinder	1	1
Chin-hui Tsao	1	1
Austin Romine	2	0
Donn Roach	1	1
Brandon Kintzler	1	0
Tyler Wilson	2	0
Matt Andriese	1	0
Donovan Hand	1	0
Sam LeCure	1	0

Table 3.2: Some examples of players during the 2015 season who had very few at-bats

3.2.1 Predicting the probability of a hit with complete 2015 Statcast data

Due to the "missingness" of data as described in Section 3.1, we limited model training for the first regression to be on complete cases of the data. We initially considered 10 variables. These included hit distance, launch angle, exit velocity, launch angle^2, launch angle^3, hit distance^2, hit distance^3, hit distance $\times$ launch angle, hit distance $\times$ exit velocity, and launch angle $\times$ exit velocity. After analyzing the correlation between these variables, we determined that launch angle $\times$ exit velocity, hit distance $\times$ exit velocity, launch angle^2, hit distance^2, and hit distance^3 should be removed from consideration due to their high correlation with other variables.

We chose the above interaction terms due to our physical understanding of the original variables having a non-linear effect. For example, launch angle can vary from -90 degrees to 90 degrees, where a negative angle represents a ground ball and a large angle a fly ball. We expect the probability of a hit to increase as the angle increases until it reaches a certain point. This is because the probability of a hit for both a ground ball and fly ball, which are at opposite ends of the range, will be lower than those angles in the middle of

the range. Therefore, we would expect some curvature in the relationship of a probability of a hit and launch angle.

After removing the correlated variables, we conducted stepwise variable selection using a likelihood ratio test at each step to determine the included variables in the model. Stepwise variable selection is an appropriate choice as it allows variables to enter and exit the model at any point. We began with the simplest model of just including an intercept term and moved forward towards the model with all terms included. At the end of the process four variables and the intercept were selected. The R output for the model is shown in Table 3.3.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.37E+00	9.19E-02	-14.869	< 2.00E-16 ***
Launch Angle^3	-2.64E-05	1.51E-06	-17.524	< 2.00E-16 ***
Exit Velocity	2.15E-02	1.32E-03	16.346	< 2.00E-16 ***
Hit Distance * Launch Angle	-1.52E-04	1.35E-05	-11.238	< 2.00E-16 ***
Hit Distance	3.12E-03	4.70E-04	6.645	3.04E-11 ***

Table 3.3: Logistic regression output

We can see from the table that all of the variables have small p-values. This is due to the large size of the dataset. To verify the effect sizes, we went through each variable and their corresponding coefficients to determine whether the coefficients made sense in terms of increasing the probability of a hit or not. For example, we assume a batted ball with a fast exit velocity will have a higher probability of a hit than one with a slow exit velocity. Inputting one slow exit velocity (20mph) and one fast exit velocity (120mph) for x_1 in (3.1), where $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the estimated coefficient for exit velocity, we get the probability of a hit p.

$$p = \frac{e^{\hat{\beta}_0 + x_1\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + x_1\hat{\beta}_1}} \tag{3.1}$$

For the slow batted ball, we get the probability of a hit to be 0.28 and for the fast batted ball 0.77. These calculations seem reasonable according to our baseball intuition. Note that the simple calculation does not account for launch angle nor hit distance. Similar calculations were carried out for the other variables in the model.

Furthermore, we can use the ANOVA function to conduct a likelihood ratio test. In particular, we focus on the deviance of the variables. The deviance column in Table 3.4 shows how much the residual deviance will decrease when the term is added to the model (in decreasing order of magnitude).

	DF	Deviance	Residual DF	Residual Deviance
NULL	49844	68799		
Launch Angle^3	1	13502	49843	55297
Exit Velocity	1	502	49842	54795
Hit Distance * Launch Angle	1	174.2	49841	54621
Hit Distance	1	43.5	49840	54577

Table 3.4: Anova output

With the exception of hit distance^3, most of the variables have a relatively high impact on the deviance of the model. Because Hit distance^3 was selected in the stepwise process used to build the model, we chose to keep it in the final regression despite its smaller deviance. After verifying the logical nature of the coefficients and variables, we end up with the final fitted model shown in (3.2).

$$\log\left(\frac{p}{1-p}\right) = -1.37 - 2.645^{-5}x_2^3 + 2.15^{-2}x_1 - 1.52^{-4}x_2x_3 + 3.12^{-3}x_3 \tag{3.2}$$

where p is the probability of a hit, x_1 is exit velocity, x_2 is the launch angle and x_3 is the hit distance. Using this model and the 2015 and 2016 Statcast data, we predict the probability of a hit for each at-bat.

3.2.2 Predicting the probability of a hit for ground balls

Because of missing hit distances in the Statcast data for ground balls, a separate logistic regression was used. We considered exit velocity as the only predictor for the probability of a hit for ground balls. We excluded launch angle because we believed the minimum variation in their values for ground balls would not greatly affect the probability of a hit. As well, we believe that velocity is the most important predictor in ground balls as faster balls are harder for fielders to deal with and could have more unpredictable bounces compared to slower ground balls. Fitting the binary logistic regression model, we have

$$\log\left(\frac{p}{1-p}\right) = -4.515 + 0.0392x_1. \tag{3.3}$$

We again confirm the logic of these values with an application to baseball. With respect to ground balls, we would expect a slow exit velocity at-bat to have a very little probability of hit success. Using an exit velocity of 20mph and 120 mph, we have $P(hit) = 0.023$ and $P(hit) = 0.547$ respectively. These values are logical as we would expect the probability of a hit for a ground ball at any exit velocity to be lower than that of other batted ball types.

3.2.3 Predicting batting averages

After using (3.3) to find the probability of a hit for each ground ball in 2015 and 2016, we combined these probabilities with the ones computed by (3.2) as well as the zero

probabilities for pop-ups and strikeouts. Let m_j be the number of at-bats for the jth player during the season of interest. We have $m_j > 200$ and $1 \leq j \leq 334$ for 2015 and $1 \leq j \leq 333$ for 2016. Therefore, to predict the batting averages for the subsequent year for player j, we take the sum of the predicted probability of a hit for the jth player, $\hat{p}_{ji}$ during the ith at-bat, and divide this sum by the total number of at-bats for that player, m_j, to give his Statcast prediction.

$$\widehat{y_{sj}} = \frac{\sum_{i=1}^{m_j} \hat{p}_{ji}}{m_j} \tag{3.4}$$

We believe this method is logical as we assume the majority of players will perform similarly in the following year as they did during the current year. Furthermore, we hope (3.2) and (3.3) focus on the quality of a player's at-bats and not the luck or bad luck component that we wish to remove from the prior season when making our prediction.

3.3 Combining 2016 PECOTA and 2016 Statcast to predict 2016 batting averages

Using the Statcast prediction for a player's 2016 batting average shown in (3.3) and PECOTA's 2016 batting average predictions, we wish to combine them in a linear regression that will be used to predict 2017 batting averages. Although advanced in nature, PECOTA doesn't use the at-bat qualities from the Statcast data that we use for our model. We hope this additional information given by a player's at-bats will improve the PECOTA predictions.

Initially we had wanted to assign weights to the PECOTA and Statcast predictions that utilized the variance of each prediction method to get a combined predicted batting average for 2016 and 2017. However upon further consideration, we wanted our combined prediction to reflect both the variance and bias of the predictions. Therefore we decided to use a linear combination of the predictions. Let $\widehat{y_{pj}}$ be the 2016 PECOTA predicted batting average for the jth player and y_{aj} be the 2016 actual batting average for the jth player. Then we regress

$$y_{aj} = \beta_0 + \beta_1 \widehat{y_{pj}} + \beta_2 \widehat{y_{sj}}$$

with unknown coefficients β_0, β_1, and β_2. Fitting this model, we have

$$\widehat{y_{aj}} = 0.0257 + 0.6482\widehat{y_{pj}} + 0.2577\widehat{y_{sj}}. \tag{3.5}$$

The non-zero intercept term reflects bias in the estimations. As well, we notice that the PECOTA predictions are more heavily weighted than the Statcast predictions. This is not concerning as we expect the more advanced algorithm to contribute most of the final prediction. We show the relationship of the two predictions in Figure 3.2. Clearly there

exists some linear relationship between the two predictions. Therefore we are confident in our decision to use regression to combine the predictions.

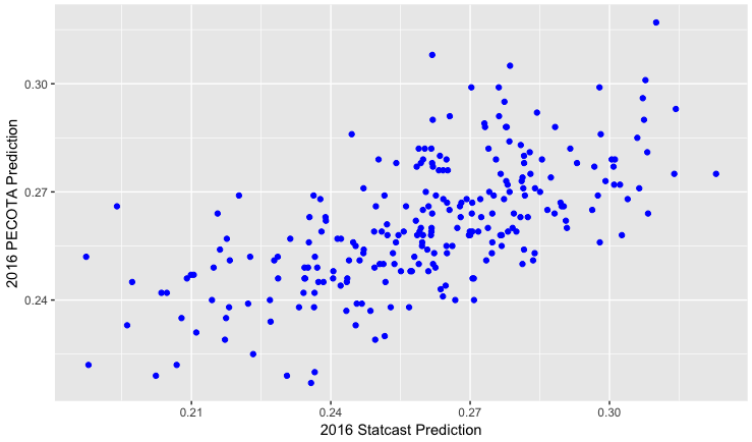

Figure 3.2: Statcast and PECOTA predictions

3.4 2017 Predictions and Comparisons

Using the 2016 Statcast data and (3.2), (3.3), along with (3.4), we get the 2017 Statcast predictions for 246 players. The decrease from the original 2017 Statcast predictions (333 players) to the final 2017 predictions comes from players leaving the league in 2017, or having less than 200 at-bats for the 2017 season. We then combine the Statcast predictions with the 2017 PECOTA predictions using (3.5) to get the final predictions for 2017.

To compare the predictions, we used mean absolute error (3.6), where k represents one of the three different prediction methods, Statcast, PECOTA, and their linear combination (3.5), and y_j is the actual batting average for the jth player in 2017.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \mid \widehat{\mathrm{y_{kj}}} - \mathrm{y_j} \mid \tag{3.6}$$

We began by comparing the 2016 PECOTA and Statcast predictions. For PECOTA, we got a mean absolute error of 0.0209 and for Statcast 0.0236. We expected PECOTA to perform better than the original Statcast predictions. However, the difference between the two errors is relatively small and impressive when comparing a proprietary method like PECOTA to openly available data. Statcast also does not take into account important variables such as age. It is also evident that the prediction of batting average is difficult. A MAE of 25 points is a large difference in the perceived quality of players.

The main focus of this project was to see if an improvement could be made with the 2017 predictions. Again, using mean absolute error, we compare 2017 PECOTA predictions with the 2017 combined predictions. The mean absolute error for PECOTA was 0.0208 and for the combined model 0.0207. In terms of mean absolute error, there is a slight

improvement in the combined model. Looking at the absolute errors of the combined model for every prediction, it appears a lot of the larger mean errors occurred where PECOTA also had large errors. However, in many cases, the combined version was better than PECOTA. In cases where PECOTA outperformed the combined version, larger differences of prediction performance ($i.e. > 0.005$) occurred for only 37 players. Looking at individual mean errors, it was found that the combined predictions outperformed the PECOTA predictions 56% of the time, implying an improved prediction performance for the combined model on an individual level. Looking at individuals is appealing in baseball as many players have "off" years that, due to the limited Statcast data, our model does not consider.

Using either error, the combined version does outperform the PECOTA prediction which suggests that the additional information provided from Statcast is valuable in predicting batting averages. As well, we believe as the data from Statcast will become more accurate, and with additional years, the predictions will continue to improve.

Unlike the PECOTA predictions, we are able to provide estimates of the standard error for each player. Consider the Statcast prediction for a player's batting average as given by (3.4). Because of the assumption that each at-bat is independent, we are able to estimate the variance of the Statcast predicted batting averages,

$$\mathrm{Var}(\widehat{y_{sj}}) = \frac{\sum_{i=1}^{m_j}(\mathrm{SE}(\hat{p}_{ji}))^2}{m_j^2} \tag{3.7}$$

where $\mathrm{SE}(\hat{p}_{ji})$ is the standard error of the probability of a hit for the ith at-bat from the logistic regression output. To estimate the variance of the combined estimator in (3.5), we have

$$\hat{\mathrm{Var}}(\widehat{y_{aj}}) = \hat{\beta}_1^2\mathrm{Var}(\widehat{y_{pj}}) + \hat{\beta}_2^2\mathrm{Var}(\widehat{y_{sj}}) + 2\hat{\beta}_1\hat{\beta}_2\mathrm{Cov}(\widehat{y_{pj}}, \widehat{y_{sj}}) \tag{3.8}$$

where $\mathrm{Var}(\widehat{y_{pj}})$ can be estimated by squaring the standard error of the PECOTA predictions. Note we are treating $\hat{\beta}_1$ and $\hat{\beta}_2$ as constants when they are in fact random. The $\mathrm{Cov}(\widehat{y_{pj}}, \widehat{y_{sj}})$ can be estimated by calculating the sample covariance of the PECOTA and Statcast predictions. This covariance is positive as can be seen in Figure 3.2. Therefore we have

$$\begin{aligned}\mathrm{Var}(\widehat{y_{aj}}) &= 0.6482^2(0.000295) + 0.2577^2\mathrm{Var}(\widehat{y_{sj}}) + 2(0.6482)(.0.2577)(0.000304) \\ &= 0.066\mathrm{Var}(\widehat{y_{sj}}) + 0.000226.\end{aligned} \tag{3.9}$$

Due to the small nature of the variances, we report players standard errors by squaring (3.9).

3.5 Players with Large Absolute Errors

We consider the players who had the highest errors for the combined predictions. We look at 13 players who's combined prediction had an absolute error of 0.05 or more. Table 3.5 lists the combined absolute difference per player along with their corresponding batting average and combined, Statcast, and PECOTA prediction along with the combined standard deviation and the statcast standard error. Note that the standard errors for the combined estimators are smaller and preferred to the standard errors of both PECOTA and Statcast.

Players	Team	Batting Average	Combined Prediction	Combined Standard Error	PECOTA Prediction	Statcast Prediction	Statcast Standard Error	Absolute Combined Error
Tyler Saladino	Chicago White Sox	0.178	0.260	0.0164	0.257	0.264	0.0251	0.082
Avisail Garcia	Chicago White Sox	0.330	0.262	0.0159	0.263	0.256	0.0208	0.068
Rougned Odor	Texas Rangers	0.204	0.265	0.0157	0.271	0.246	0.0175	0.061
Miguel Cabrera	Detroit Tigers	0.249	0.307	0.0157	0.310	0.311	0.0183	0.058
Trevor Plouffe	Minnesota Twins	0.198	0.255	0.0163	0.250	0.262	0.0242	0.057
Bryce Harper	Washington Nationals	0.319	0.263	0.0158	0.270	0.243	0.0190	0.056
John Jaso	Pittsburgh Pirates	0.211	0.267	0.0161	0.268	0.261	0.0228	0.056
Jose Bautista	Toronto Blue Jays	0.203	0.258	0.0159	0.255	0.258	0.0200	0.055
Hyun Soo Kim	Baltimore Orioles	0.230	0.283	0.0164	0.278	0.300	0.0259	0.053
Marwin Gonzalez	Houston Astros	0.303	0.251	0.0158	0.252	0.242	0.0196	0.052
Josh Reddick	Houston Astros	0.314	0.264	0.0161	0.254	0.285	0.0227	0.050
Ryan Zimmerman	Washington Nationals	0.303	0.253	0.0159	0.247	0.262	0.0206	0.050

Table 3.5: Players with large absolute errors based on 2017 predictions

The largest error occurred with Tyler Saladino who only had a 0.178 batting average in 2017. Saladino seemed to be plagued with injuries this season, which caused far lower than normal performance. After an average start to his season in April (batting average of 0.221) [1], Saladino experienced a performance decrease in May, where he batted 0.167. He would not play in June due to nerve issues in his back and in July he only played nine games where he again averaged 0.167. His performance looked to be increasing in August with a 0.205 batting average, but in September he only had three hits of his 40 at-bats (NBCSports, 2017)

Other poor performances included Trevor Plouffe, who had his first under 0.200 batting average since his first year in MLB in 2010, where he batted 0.146. Jose Bautista has seen a steady decline in batting averages since he peaked in 2011 with a 0.302 batting average. His performance in 2017 (0.203) marks the worst since 2005 (0.203 in only 11 games). Miguel Cabrera batted 0.249 in 2017, his first average under 0.300 since 2008 (0.292). Cabrera made headlines at the end of August for his involvement in a fight during the August 24th game against the New York Yankees. This fight resulted in a six game suspension for him.

[1] All player statistics in this section are from baseballreference.com

On the opposite end of the spectrum, Avisail Garcia batted 0.330, the third highest in the MLB for the 2017 regular season and a large improvement from his 0.245 batting average in 2016. Bryce Harper rebounded from a slow 2016 season (0.243) after he had a breakout season in 2015 (0.330). After signing with the Houston Astros this year, Josh Reddick had his best batting average (0.313) since he began playing in MLB in 2009.

For the 2017 regular season, Jose Altuve led MLB in batting average with 0.346. PECOTA had predicted Altuve to bat 0.307, while our Statcast model had predicted 0.313. The combined prediction was 0.305. In this case, the Statcast model did the best job of predicting his season. Interestingly enough, according to the Statcast model, Jose Altuve's predicted 2018 batting average is only 0.269. This implies he may have been lucky with a lot of his hits for the 2017 season.

Because PECOTA doesn't release their 2018 projections until February 2018, we consider the 2018 projections for the Statcast model. According to this model, the top three predicted batting averages for 2018 are Joe Mauer (0.319), DJ LeMahieu (0.316), and Melky Cabrera (0.315). In the 2017 regular season, Mauer batted 0.305, LeMahieu 0.310, and Cabrera 0.285. According to what we hope our model captures, these projections imply Mauer and LeMahieu had good hits in 2017 and, assuming similar performances in 2018, will continue to bat at a high level. As well, Cabrera is projected to be quite a bit higher than his 2017 performance. This suggests he had good hitting statistics, but perhaps some level of unluckiness with his number of outs in 2017.

The players in Table 3.5 all experienced irregular seasons compared to their previous year. This caused our prediction to be closer to their 2016 batting average than what really occurred. We hope in the future when more Statcast data is available, that we can use more prior years to help predict a player's batting average. This will help reduce errors in players whose first prior year was not consistent with their usual batting statistics.

Chapter 4

Concluding Remarks

In this project, we wanted to explore new tracking data of batted balls by predicting the batting average of MLB players. The new data allowed us to remove a luck component from our predictions that naturally occurs in baseball. Using Statcast data and the logistic regressions from (3.2) and (3.3) along with (3.4), we formed Statcast predictions for 2016 and 2017. We then fit a linear model using 2016 Statcast and PECOTA predictions and obtained the combined 2017 predictions using 2017 Statcast and PECOTA predictions and (3.5).

In Chapter 1, we discussed multiple applications of sports analytics as well as their contribution to the growth in technology. We discussed the use of sports analytics for both teams and fans. By providing a better prediction of batting averages than one of the current best systems, we show the importance of considering the new Statcast data in analyses. As well, we show that there is still room for improvement in the proprietary systems. This information can be used to increase gambling success by fans or with coaches as an assessment tool in their current roster.

In Chapter 2, we discussed the two different data sources we used as well as other prediction systems that are available. In Chapter 3, we described the methodology behind forming the combined PECOTA and Statcast predictions for 2017 batting averages.

Although the results were promising, we believe they can be improved as additional years of Statcast data are available and the accuracy of the data improves. In the future we could refit the logistic regression model with newer years of data that are more complete. As well, we could include multiple years of comparison between our combined prediction and PECOTA.

Regardless, we believe this project can be used as valuable source of information in describing and using the Statcast data. As well, we have showed that adding additional features to a complex set of predictions in a simple manner can be very effective.

Bibliography

Alamar, B. and Mehrotra, V. (2012), "Beyond Moneyball: The Future of sports analytics. Analytics Magazine," Retrieved from `http://analytics-magazine.org/beyond-moneyball-the-future-of-sports-analytics/` August 5, 2017.

Albert, J. (2016), "Improved component predictions of batting and pitching measures," *Journal of Quantitative Analysis in Sports*, 12, 73–85.

"baseball reference" (n.d.), "Single-Season Leaders and Records for Batting Average," Retrieved from `https://www.baseball-reference.com/leaders/batting_avg_season.shtml` August 28, 2017.

"BaseballProspectus" (n.d.), "Baseball Prospectus | Gallery," Retrieved from `http://www.baseballprospectus.com/glossary/index.php?search=PECOTA` August 5, 2017.

Birnbaum, P. (2013), "A guide to sabermetric research," *Society for American Baseball Research (SABR)*.

Brousell, L. (2014), "8 Ways Big Data and Analytics Will Change Sports," Retrieved from `http://www.cio.com/article/2377954/data-management/data-management-8-ways-big-data-and-analytics-will-change-sports.html` July 20, 2017.

Connolly, J. M. (2014), "Analytics Brings Science to Sports Betting," Retrieved from `https://www.allanalytics.com/author.asp?section_id=3624&doc_id=275025` July 20, 2017.

Davenport, T. H. (2014), "Analytics in sports: The new science of winning," *International Institute for Analytics*, 2, 1–28.

Druschel, H. (2016), "A Guide to the Projection Systems," Retrieved from `https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun` August 21, 2017.

Fast, M. (2010), "What the heck is PITCHf/x," *The Hardball Times Annual*, 2010, 153–158.

Kagan, D. and Nathan, A. M. (2017), "Statcast and the Baseball Trajectory Calculator," *The Physics Teacher*, 55.

Karkazis, K. and Fishman, J. R. (2017), "Tracking US professional athletes: The ethics of biometric technologies," *The American Journal of Bioethics*, 17, 45–60.

"MLB" (n.d.a), "Glossary | Standard Stats," Retrieved from `http://m.mlb.com/glossary/standard-stats` August 18, 2017.

— (n.d.b), "Glossary | Statcast," Retrieved from `http://m.mlb.com/glossary/statcast` August 17, 2017.

"NateSilver" (n.d.), retrieved from `https://en.wikipedia.org/wiki/Nate_Silver` August 28, 2017.

NBCSports (2017), "Nerve issue sends Tyler Saladino to DL, White Sox bring up outfielder Adam Engel," .

"OPS" (n.d.), "OPS and OPS | FanGraphs Sabermetrics Library," Retrieved from `http://www.fangraphs.com/library/offense/ops/` July 15, 2017.

PECOTA (2017), retrieved from `https://en.wikipedia.org/wiki/PECOTA` July 15, 2017.

"Sabermetrics" (n.d.), "Sabermetrics," Retrieved from `https://en.wikipedia.org/wiki/Sabermetrics#Early_history`July 20, 2017.

Schwarz, A. (2005), "Keeping Score; Predicting Futures in Baseball, and the Downside of Damon," Retrieved from `http://query.nytimes.com/gst/fullpage.html?res=9C0CEEDA133EF930A25752C1A9639C8B63` August 17, 2017.

Smith, M. D. (2015), "Half of NFL Team snow using Variable Ticket Pricing," Retrieved from `http://profootballtalk.nbcsports.com/2015/07/11/half-of-nfl-teams-now-using-variable-ticket-pricing/` August 5, 2017.

"STATS" (n.d.), "STATS LLC," Retrieved from `https://www.stats.com` July 20, 2017.

Swish (2017), "Sports Predictions, Betting Tools, and Analysis | Swish Analytics," Retrieved from `https://swishanalytics.com/FAQ` August 5, 2017.

Tangotiger (2017), "Statcast Lab: No Nulls in Batted Balls Launch Parameters," Retrieved from `http://tangotiger.com/index.php/site/article/statcast-lab-no-nulls-in-batted-balls-launch-parameters` August 27, 2017.

Torre, P. S. (2015), "The 76ers' plan to win (yes, really)," Retrieved from `http://www.espn.com/nba/story/_/id/12318808/the-philadelphia-76ers-radical-guide-winning` August 17, 2017.