

Hitting around the shift: Evaluating batted-ball trends across Major League Baseball

Michael W. Model

Department of Statistical Science Honors Thesis

Duke University

Durham, North Carolina

2020

Abstract

The infield shift has negatively affected Major League hitters who formerly thrived on ground balls through the gaps in the infield. Nearly a quarter of plate appearances during the 2019 season featured infield shifts, up from 13.8 percent just three seasons prior. I analyzed both the evolution of shift implementation and whether batters hit differently with and without the shift using hierarchical Bayesian regression methods on both pitch-level and batter-tendency data from 2015 to 2019. Since most of the recent talk surrounding the infield shift has been related to a drastic increase in fly balls and players hitting over the shift, I looked specifically at adaptation on the ground. Not a single batter was found to have had a significant difference between their batted-ball distributions for either a given season or throughout the entire five-year period, suggesting the increase in shifting is unlikely to end in the near future.

Introduction

An infield shift occurs when a baseball team deviates from the standard symmetric two-and-two infielder alignment in favor of an asymmetric infielder alignment with three players on one side of second base and one on the other. The infield shift was first popularized by former Tampa Bay Rays manager Joe Maddon in the early 2010s. The concept behind the shift is as follows: by focusing solely on the area of the infield where batters have a tendency to hit the ball, batted balls are less likely to get through gaps in the infield (Lewis & Bailey, 2015).

For the first 100-plus years of Major League Baseball, infield positioning remained constant with two fielders on the left side and two fielders on the right side of second base. Players specialized their abilities to take advantage of the gaps between the spread-out infielders. Amidst the recent rise of analytics, however, teams realized that many hitters' batted-ball distributions did not span the entire infield. Therefore, defending teams could more strategically place their fielders based on batter tendencies and eliminate the gaps those players had previously exploited in the process.



Figure 1.1: The above figure displays how fielding positions change for a standard infield shift. The silhouettes show where the fielders are normally positioned and the arrows show the movement to shifted positioning. In this case, the third baseman (3B), second baseman (2B) and shortstop (SS) are all moved up and to the right and the shortstop is now positioned to the right of second base (the square behind the figure on the brown circle) (Kawahara, 2016).

My interest in this topic stems from the fact that players initially claimed that they could not control their process nor could they aim where they hit the ball. This is especially notable because having three defenders on one side of the infield opens up a glaring hole on the other side. Adapting to the shift is not as easy as aiming for the other side of the infield, but rather requires an entirely different hitting mindset which can take numerous offseasons developing (Crasnick, 2018).

Current Colorado Rockies second baseman Daniel Murphy said the following about the difficulty of aiming for the gap: “If any of us could control hits, we would get more of them. But you can’t. You can only control the process” (Crasnick, 2018). The reason adaptation to the shift is crucial is because the shift has derailed the career trajectories of certain hitters (Arthur, 2017). While teams are easily able to understand where a player tends to hit the ball and fill those gaps, batters cannot respond quickly at all. This has hurt hitters because their batted ball strategy is no longer as successful, decreasing their offensive production.

The effect of the shift is evident in the drastic change in [Batting Average on Balls in Play](#)—the percentage of batted balls in play, i.e. not strikeouts or home runs that are hits—(BABIP) since 2017. According to Petriello (2018), across Major League Baseball, batters see an .018 difference in BABIP between facing the shift vs. not facing the shift, or .281 vs. .299. This change essentially signifies the fact that ground balls are getting through the infield more frequently without a shift in place. Teams across the league have adopted shifts as the

percentage of plate appearances with shifts implemented has nearly doubled since 2016. The shifting rate is up to 25.6 percent from 13.8 percent according to Baseball Savant—Major League Baseball’s hub for analytics. Shifting behavior has also disproportionately impacted left-handed batters, who were shifted upon 41.9 percent of the time in 2019 compared to 14.3 percent for right-handed hitters.

Nearly five years after the shift’s leaguewide implementation, have new-age hitters altered their approach at the plate? While numerous papers have measured the effects of the shift, this analysis seeks to fill the research paucity of measuring whether players have begun to adapt and have been successful a hitting against the shift, particularly on ground balls. More specifically, this project hopes to achieve two goals. First, I hoped to model team shifting behavior from 2015 through 2019 in order to understand which factors teams take into consideration and how they have changed over time. Second, I wanted to uncover whether or not hitters have successfully altered their approach at the plate, and if so, understand whether or not they have gotten better at hitting the ball against the shift over time.

The main reason I focused on ground balls is because the question as to whether or not hitters have improved at beating the shift on the ground has been overshadowed by the “Fly Ball Revolution.” Rather than attempt to hit around the shift, many players have been trying to hit the ball over the shift. Sawchik (2017) explains this phenomenon as numerous players have discovered newfound success with this approach. The main reason this approach is effective is because fly balls have higher expected Weighted On-Base Average (wOBA). wOBA is defined by Slowinski (2010) as the number of bases per plate appearance, weighted by the relative value of each outcome (for example, a double is not worth the same as two singles). This phenomenon occurs because hitting the ball in the air can lead to doubles, triples and home runs, whereas a ground ball is almost always a single and rarely a double. The fly ball movement has been the primary way, that is currently known, that the shift has been combatted by Major League hitters.

When shifts were first being implemented, numerous hitters affected contemplated bunting (or simply tapping the ball away from the shift without swinging) to the open side which, if laid down correctly, has a high chance of being a free base (Shaw, 2015). However, by opting to bunt, hitters were eliminating their chances of getting a double, triple or home run—all of which are better outcomes. Thus, the strategy of bunting against the shift never gained widespread traction.

Literature Review

Jensen et al., 2009

The “Bayesball” paper, written by Jensen and colleagues (2009) pioneered the usage of Bayesian methods in baseball in order to establish a novel metric to evaluate Major League Baseball fielders’ defensive ability. Jensen and colleagues utilized probit regression in order to model the probability that a given batted ball was fielded successfully. The researchers hoped to understand the differences in fielding behavior for each position, for each batted-

ball type in a given year between 2002 and 2005. Excluding pitchers and catchers, Jensen and colleagues identified 18 possible batted-ball/position combinations. They got 18 because all seven fielders can record outs on fly balls and line drives, but only the four infielders can record outs on ground balls.

The researchers built a probit model for each of the 18 batted-ball/position combinations for each of the four years to get a total of 72 different models. The covariates in the fly ball and line drive models included the distance the player had to travel to the ball, the direction the player traveled and the velocity of the batted-ball. The ground ball model was comprised of the angle between the fielder and the batted-ball, the velocity of the ball and the direction the player had to travel. Jensen and colleagues also generated player-specific coefficients by utilizing random effects to share information across players at a particular position. The researchers estimated the posterior distribution of the unknown parameters by implementing Gibbs sampling.

Lewis & Bailey, 2015

Bailey and Lewis worked with the Tampa Bay Rays in order to improve upon their infield shifting model. According to the researchers, the Rays had formerly divided the infield up into a few small sections and had used linear regression models to predict the percentage of batted balls that would go into each region. The model was flawed, however, because the sub-regions of the infield were wider than the range of the infielders. In addition, the Rays established a binary threshold (e.g. 40 percent) at which point they would implement a shift. The researchers instead hoped for a more granular model since shifting at 41 percent, but standing pat at 39 percent, was likely suboptimal.

To do this, they first divided the field into nine sub-regions—each 10 degrees wide—and fit logistic regression and random forest models. However, the desired accuracy level was not met. Lewis and Bailey opted for Bayesian methodology, which allowed the researchers to predict the batted-ball distribution rather than the most likely categorical zone. Lewis and Bailey found a mixed effects model to be the best solution. Random effects on both the batter and the pitcher reduced a significant amount of variance in spray distributions, while information on the counts (e.g. two balls, two strikes) and base-out state (e.g. runner on first base, no outs) also improved the predictability of the model.

Hawke, 2017

Hawke attempted to quantify the effects of the shift through the lens of the fielders. The researcher took a similar spatial modelling approach to Jensen and colleagues (2009), however, the modelling was conducted through a frequentist framework. Hawke built a pair of datasets using data from the 2016 season. The first dataset included batted-ball information from the 42 most shifted upon hitters in Major League Baseball as a proxy for shifted plate appearances, while the other dataset included batted balls from the rest of the hitter pool. Hawke fit frequentist probit models for each outfield position to understand the probability that a fly ball or line drive would be successfully caught. Like Jensen and colleagues (2009), Hawke used the angle between the batted-ball and the fielder, the exit velocity of the batted-ball and the direction the fielder had to travel. Unfortunately, the

researcher did not have time to evaluate the effects of ground balls or calculate individual player effects. These are two gaps in the literature that my thesis aimed to fill.

Data

There were two different data types of interest for this analysis: data at the pitch level and data at the hitter level. The pitch-level data provides in-game context around the at-bat and information on the outcome of the at-bat. The batter-tendency data gives context to the hitter's batted-ball tendencies on average over the course of the entire season.

Pitch-Level Data

The pitch level data utilized in this analysis was collected from Major League Baseball's [Statcast database](#). The database was queried to collect every ground ball hit over the five-year period from 2015 to 2019. Due to the fact that public data downloads are limited to 40,000 rows, I extracted a total of 10 CSV files—one for each half-season. Within these files, only data from players with at least 30 ground balls during a given half-season was included. Using `dplyr` manipulations, I adjusted the formatting of the data and joined the rows of the CSV files in order to get a complete dataset. The variables used and created from the Statcast information can be seen in Table 1.

Table 1: Pitch-Level Variables

| Covariate | Description |
|--------------------|--|
| Season | Year the game was played |
| Batter Name | Name of the batter |
| Stand | Whether the batter hit right-handed or left-handed |
| Pitcher Throws | Whether the pitcher threw right-handed or left-handed |
| Balls | Number of balls prior to the given pitch |
| Strikes | Number of strikes prior to the given pitch |
| Runner on 1b | Whether there was a runner on first base |
| Runner on 2b | Whether there was a runner on second base |
| Runner on 3b | Whether there was runner on third base |
| Outs | Number of outs in the inning |
| Inning | Current inning of the game |
| Infield Alignment | Whether an infield shift was implemented |
| Shifting Team | Name of the fielding team |
| Score Differential | Absolute value of the difference in score between the two teams |
| Spray Angle | Angle formed between home plate and the location the ground ball was first fielded (<i>First base is 45 degrees, second base is 0 degrees and third base is -45 degrees</i>) |

While much of the above data came directly from Statcast, a few of the variables were modified for ease of interpretation:

- Each of the **Runner on** covariates initially was populated with a unique player identifier for the individual runner on base. Those variables were converted to binaries since knowing whether or not someone is on base is more relevant than understanding the effect the individual runners have.
- **Infield Alignment** initially had three values (shift, no shift, strategic shift). A strategic shift occurs when the players shift slightly, but there still remain two fielders on each side of second base. Due to the subjectivity of a “strategic shift” with regard to the size of the gap it creates, I consider “strategic shifts” to be the same as “no shifts” throughout my thesis.
- **Score Differential** was created in order to capture the effect of a game competitiveness on shifting and hitting behavior. That is the reason why I opted for the absolute value of the difference in runs, rather than sticking with the sheer difference.
- The **Spray Angle** for a given batted ball is calculated by conducting a transformation using the coordinates of where it is fielded. The methodology implemented was inspired first utilized by Jim Albert (2018). The formula behind the transformation can be found in the appendix.

Batter-Tendency Data

The pitch-level data was complemented by batter-tendency data from [FanGraphs](#). FanGraphs’ leaderboard contains cumulative statistics and advanced hitting metrics, which were collected for each individual player-season between 2015 and 2019. Data was only extracted from FanGraphs for qualified hitters—hitters who averaged at least 3.1 plate appearances per team game during the season. The variables used and created from the FanGraphs information can be seen in Table 2.

Table 2: Batter-Tendency Variables

| Covariate | Description |
|----------------------|--|
| Season | Year the game was played |
| Batter Name | Name of the batter |
| Home Runs | Number of home runs hit during the entire season |
| Strikeout Percentage | Percentage of plate appearances ending in a strikeout |
| wOBA | Weighted on-base average (wOBA)—an advanced metric that weights plate appearance outcomes by their expected change in run expectancy for an inning relative to an out. |
| Fly-ball Percentage | Percentage of batted balls that are hit in the air |

| | |
|---------------------|---|
| Pull Percentage | Percentage of batted balls hit to the “pull” side (<i>left side for right-handed hitters, right side for left-handed hitters</i>) |
| Hard-hit Percentage | Percentage of batted balls hit hard (<i>as defined by Baseball Info Solutions</i>) |
| Age Category | Age group of the player (<i>younger than 25, 25-29, 30-34, 35 or older</i>) |

Once again, a few modifications were made to the initial FanGraphs’ information with the hope of improving predictability and interpretability of the models later on:

- The values of **Fly-ball Percentage**, **Pull Percentage** and **Hard-hit Percentage** used in this analysis are not actually raw values, but rather are scaled to league average by FanGraphs. This allows for standardization across the measures. In each case, a value of 100 represents the league average, and a value of X is (X - 100) percent above league average if positive or below if negative.
- **Fly-ball Percentage**, **Pull Percentage**, **Hard-hit Percentage** and **wOBA** were all mean-centered such that the intercept represents a league-average hitter. League-average **wOBA** is 0.320.
- In order to capture the effect of age, I binned player-seasons into four brackets. **Age Category** <25 generally represents newcomers to the league, 25-29 are players in their prime and rising, 30-35 are players in their prime and declining, and 35+ are those on their way out of the league.

Once I had cleaned and collected the data from both sources, I joined the dataframes on both **Batter Name** and **Season**. This provided the desired outcome of each row representing a ground ball with both situational data on the game and season-long data on the hitter. In total, the final dataset had 133,899 ground balls hit over the course of five seasons from 303 unique players. More information on the specifics of the data can be found in the appendix.

Modelling

There were two main classes of modelling conducted for this analysis: Bayesian logistic regression and Bayesian linear regression. I used logistic regression in order to model the probabilities that shifts were implemented. Linear regression was utilized to model batters’ spray angle distributions. Regression methods were preferred to more complex modelling approaches due to their interpretability. For this project, my focus was primarily on understanding the factors behind shifting behavior and hitting patterns rather than predictability.

Bayesian methods are advantageous due to their ability to provide posterior distributions for both the coefficients and of the predictions. This allows for a stronger quantification of uncertainty throughout the modelling process. To conduct these analyses, I utilized the **brms**–Bayesian Regression Models using Stan–package. For each model, four chains with 2000 observations each were used. 1000 samples in each chain were removed as burn-in,

resulting in a total of 4000 draws per model. I also used the **brms** default noninformative prior for the covariates. Models were then addressed for convergence by evaluating the **R-hat** values of every parameter. R-hat addresses convergence by comparing the between- and within-chain estimates for the model parameters and values above 1.05 are concerning. If the model converged, I then conducted numerous checks for each model type, such as residual analysis and area under the curve calculations to assess the validity of the models. More information on model specifics can be found below.

Shift Probability Models

Model Overview

In order to understand the factors teams take into consideration when deploying shifts and also learn how they've changed over time, I ran a separate Bayesian logistic regression model for each season. Simply adding a parameter for **season** into the model and running a single model would have provided information solely on how the shifting rate changed, not how the individual factors changed. I could have added season and an interaction with season for each covariate, but I opted against doing so for the purpose of clarity. With the goal of understanding how the factors changed over time at the forefront, the same model structure was used for each season of data.

Due to ease of interpretation, I converted the numeric variables **balls**, **strikes**, **outs** and **inning** to factors for this analysis. In addition to many fixed effects describing the game situation and cumulative batter-tendencies, random effects were implemented on both the hitter and the shifting team. Some hitters and teams deal with shifts constantly, while others have sparse interactions with shifts. These random effects provide a clearer picture of where hitters and teams lie on the spectrum and how their positions changed over time. The full shift probability model structure can be seen below:

$$\begin{aligned} \text{logit}(\text{Pr}(\text{infield.alignment} = \text{shift})) = & \alpha_{j[i]}^{\text{hitter}} + \alpha_{k[i]}^{\text{shifting.team}} + \beta_{\text{stand}} \text{stand}_i + \\ & \beta_{p.\text{throws}} p.\text{throws}_i + \beta_{\text{balls}1} \text{balls}1_i + \beta_{\text{balls}2} \text{balls}2_i + \beta_{\text{balls}3} \text{balls}3_i + \beta_{\text{strikes}1} \text{strikes}1_i + \\ & \beta_{\text{strikes}2} \text{strikes}2_i + \beta_{\text{on.}3b} \text{on.}3b_i + \beta_{\text{on.}2b} \text{on.}2b_i + \beta_{\text{on.}1b} \text{on.}1b_i + \beta_{\text{outs}1} \text{outs}1_i + \\ & \beta_{\text{outs}2} \text{outs}2_i + \beta_{\text{inning}2} \text{inning}2_i, \dots, + \beta_{\text{inning}18} \text{inning}18_i + \beta_{\text{score.diff}} \text{score.diff}_i + \\ & \beta_{\text{home.runs}} \text{home.runs}_i + \beta_{\text{strikeouts}} \text{strikeouts}_i + \beta_{\text{wOBA}} \text{wOBA}_i + \beta_{\text{fly.balls}} \text{fly.balls}_i + \\ & \beta_{\text{pull}} \text{pull}_i + \beta_{\text{hard.hit}} \text{hard.hit}_i + \beta_{\text{age.cat} < 25} \text{age.cat} < 25_i + \beta_{\text{age.cat} 25-29} \text{age.cat} 25 - \\ & 29_i + \beta_{\text{age.cat} 30-34} \text{age.cat} 30 - 34_i + \beta_{\text{age.cat} 35+} \text{age.cat} 35+_i \end{aligned}$$

$$\alpha_j^{\text{hitter}} \sim N(\mu_a^{\text{hitter}}, \sigma_{\text{hitter}}^2)$$

$$\alpha_k^{\text{shifting.team}} \sim N(\mu_a^{\text{shifting.team}}, \sigma_{\text{shifting.team}}^2)$$

for $i = 1, \dots, N$ $j = 1, \dots, J$ $k = 1, \dots, 30$ where i represents a unique ground ball for the given year, j represents a unique hitter with ground balls in the given year and k represents a unique Major League team.

A subset of the overall dataset filtered for the given year was used for each of the models. The 2019 and 2018 models converged on the first run, while the 2017, 2016 and 2015 had

R-Hat violations. After taking a further look into the data, I realized the problem occurred due to the fact that data for the late extra innings (innings 13 through 18) was limited since games rarely go that long. Thus, for the 2017, 2016 and 2015 models, I filtered out innings 13 through 18 and the models converged on the second attempt.

Fit Diagnostics

Prior to interpreting the model coefficients, two diagnostic tests were conducted to assess the model fit. First, I conducted posterior predictive checks using the `pp_check` function. The idea behind a posterior predictive check is to make sure the data generated by the model looks similar to the observed data. If there is a mismatch in the fits, it is likely that there is a modelling error or data mismatch.

I also conducted external validation for each season, by testing the models on pitches that resulted in fly balls during the second half of the respective seasons. The test datasets ranged from 6400-7300 rows, about 20 percent of the size of the training data for each season. An area under the curve (AUC) analysis was run once shifting probabilities were predicted in order to assess whether or not the models were an improvement on random chance. The resulting plots for each model can be viewed below:

2019

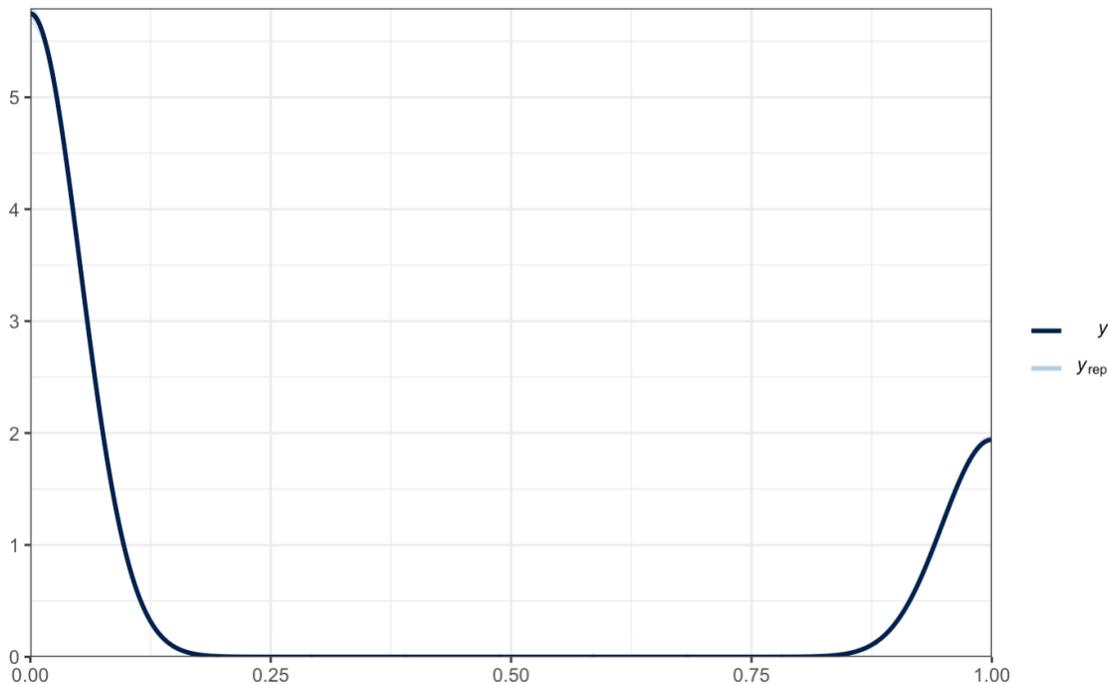


Figure 4.1: The above plot shows a posterior predictive check for the 2019 model. The x-axis represents the probability of a shift and the y-axis represents the density. The dark-blue line represents the density of the observed data and the thin light-blue line represents the density of the data sampled from the posterior predictive distribution. The light-blue line is difficult to see due to its close overlap with the dark-blue one.

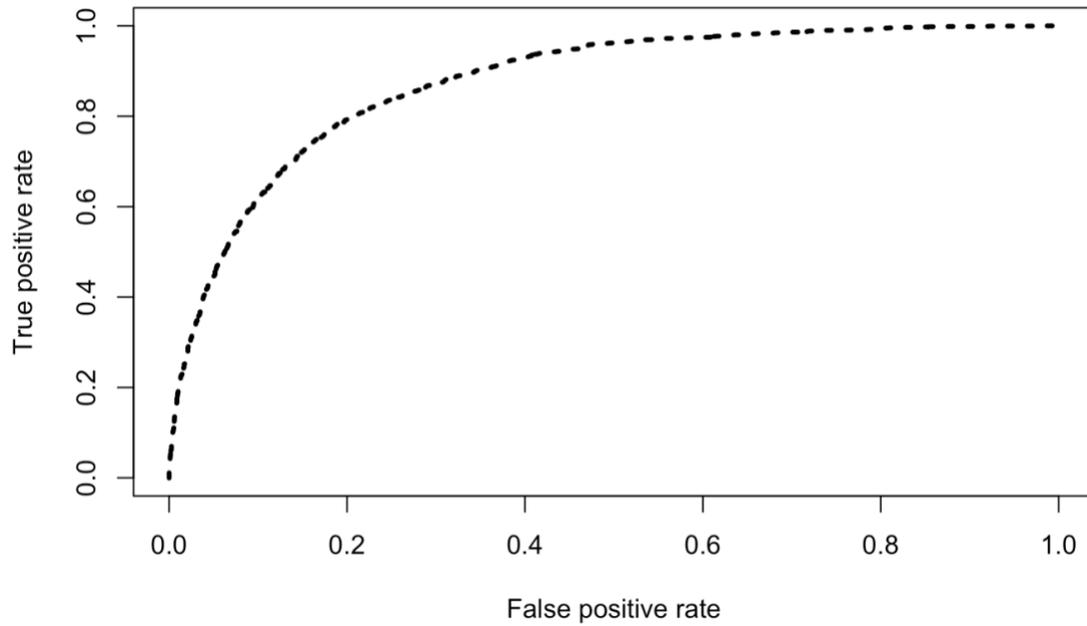


Figure 4.2: The above plot shows a ROC curve for the 2019 data. The x-axis represents the false positive rate and the y-axis represents the true positive rate.

2018

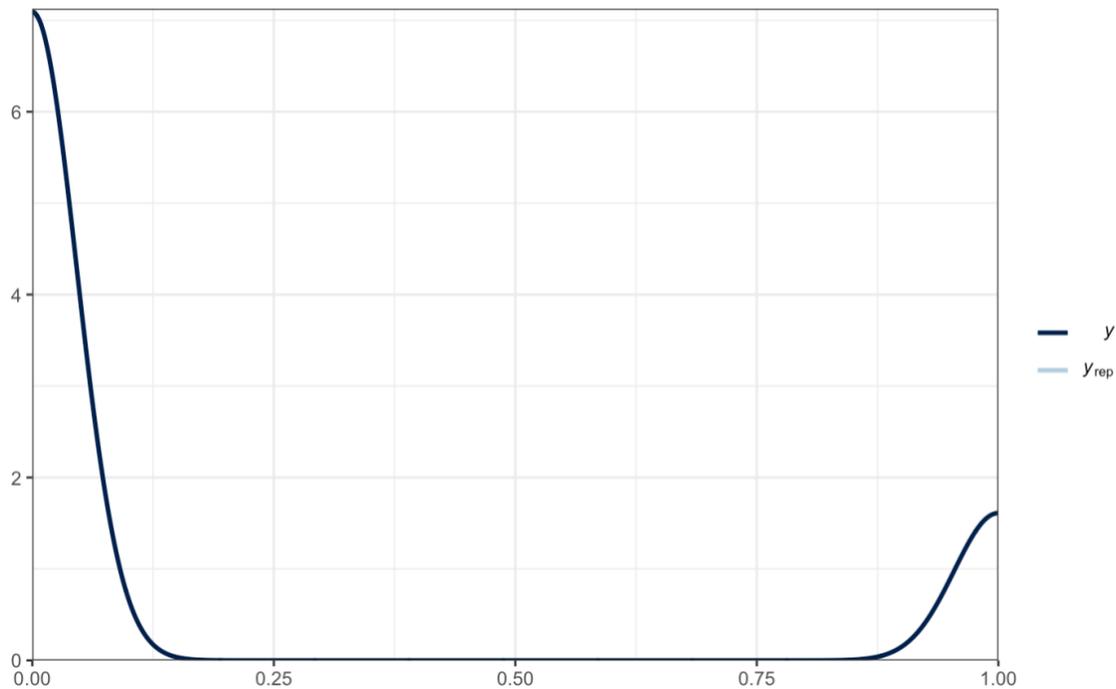


Figure 4.3: The above plot shows a posterior predictive check for the 2018 model. The x-axis represents the probability of a shift and the y-axis represents the density. The dark-blue line represents the density of the observed data and the thin light-blue line represents the density of the data sampled from the posterior predictive distribution. The light-blue line is difficult to see due to its close overlap with the dark-blue one.

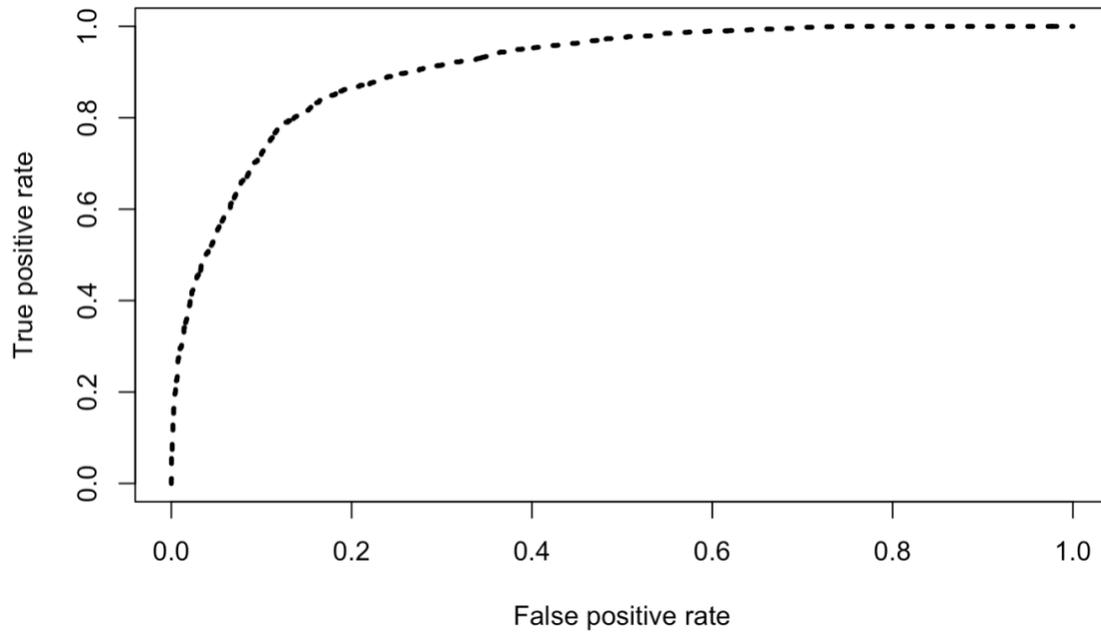


Figure 4.4: The above plot shows a ROC curve for the 2018 data. The x-axis represents the false positive rate and the y-axis represents the true positive rate.

2017

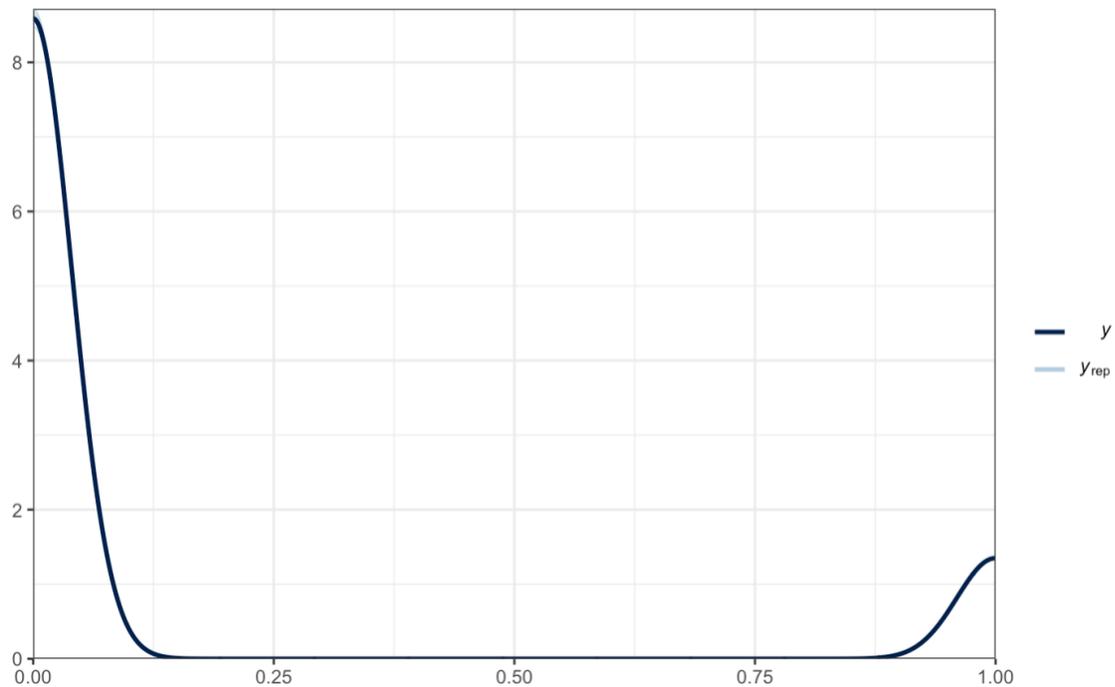


Figure 4.5: The above plot shows a posterior predictive check for the 2017 model. The x-axis represents the probability of a shift and the y-axis represents the density. The dark-blue line represents the density of the observed data and the thin light-blue line represents the density of the data sampled from the posterior predictive distribution. The light-blue line is difficult to see due to its close overlap with the dark-blue one.

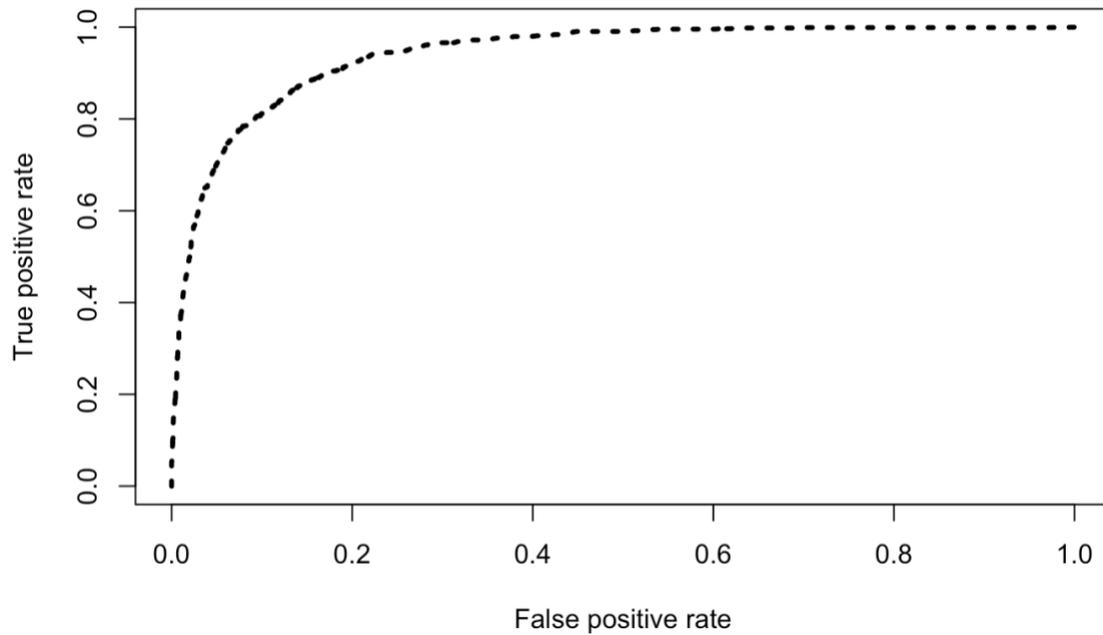


Figure 4.6: The above plot shows a ROC curve for the 2017 data. The x-axis represents the false positive rate and the y-axis represents the true positive rate.

2016

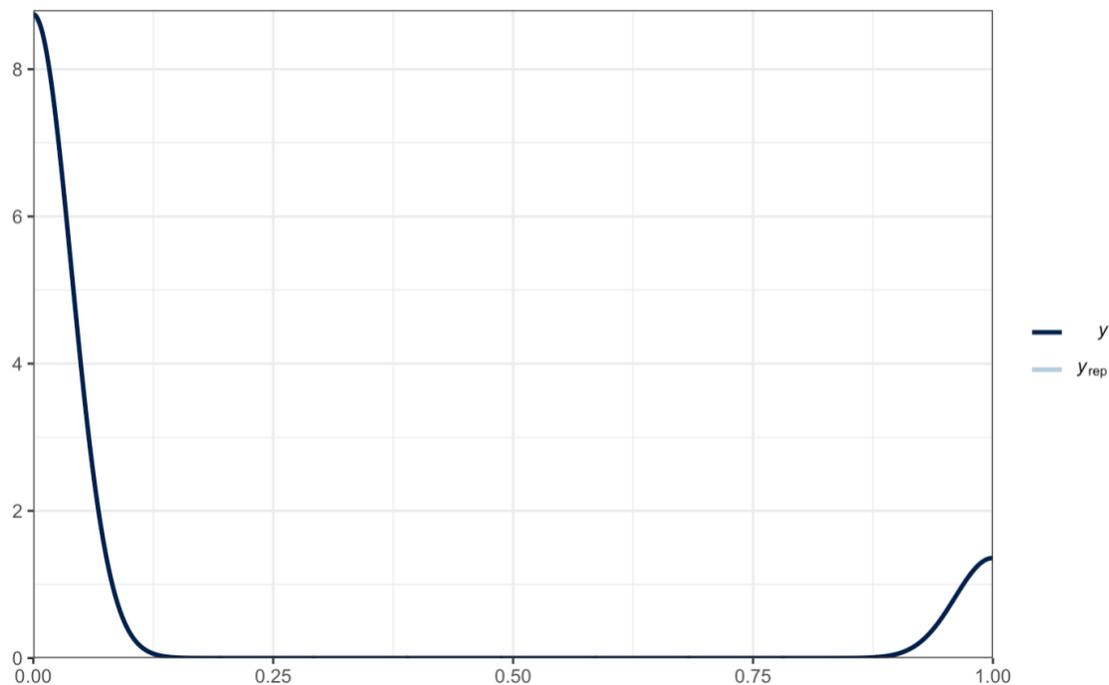


Figure 4.7: The above plot shows a posterior predictive check for the 2016 model. The x-axis represents the probability of a shift and the y-axis represents the density. The dark-blue line represents the density of the observed data and the thin light-blue line represents the density of the data sampled from the posterior predictive distribution. The light-blue line is difficult to see due to its close overlap with the dark-blue one.

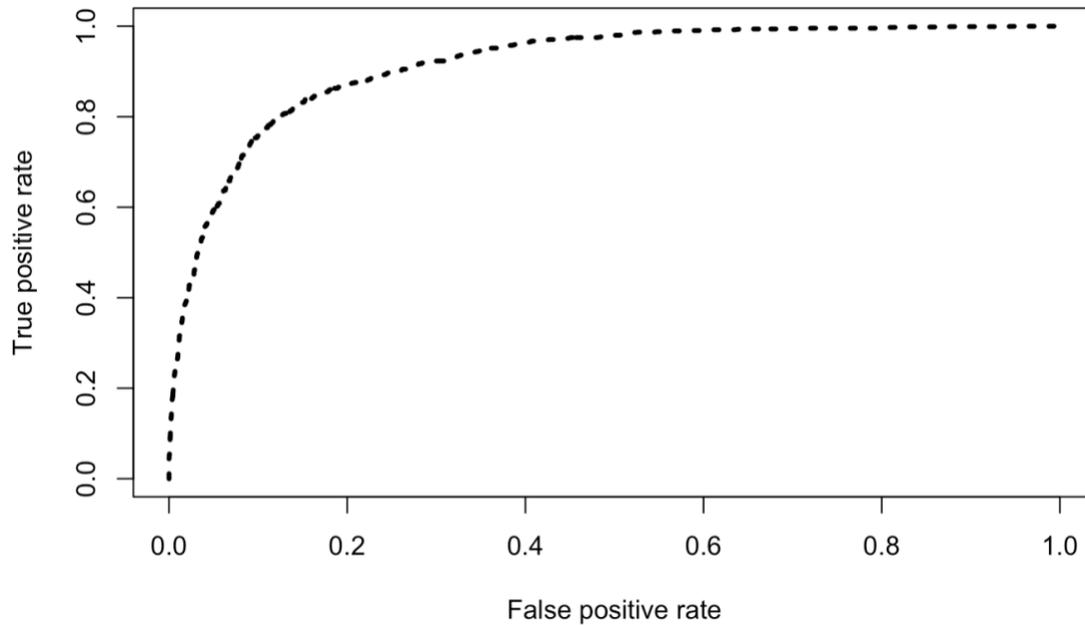


Figure 4.8: The above plot shows a ROC curve for the 2016 data. The x-axis represents the false positive rate and the y-axis represents the true positive rate.

2015

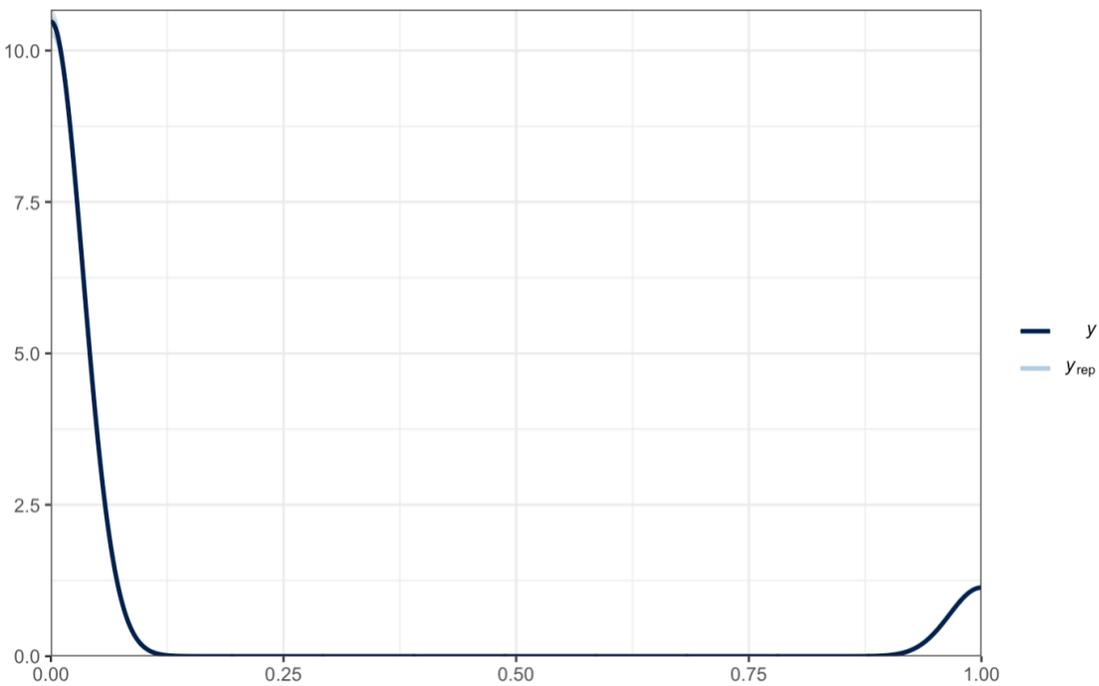


Figure 4.9: The above plot shows a posterior predictive check for the 2015 model. The x-axis represents the probability of a shift and the y-axis represents the density. The dark-blue line represents the density of the observed data and the thin light-blue line represents the density of the data sampled from the posterior predictive distribution. The light-blue line is difficult to see due to its close overlap with the dark-blue one.

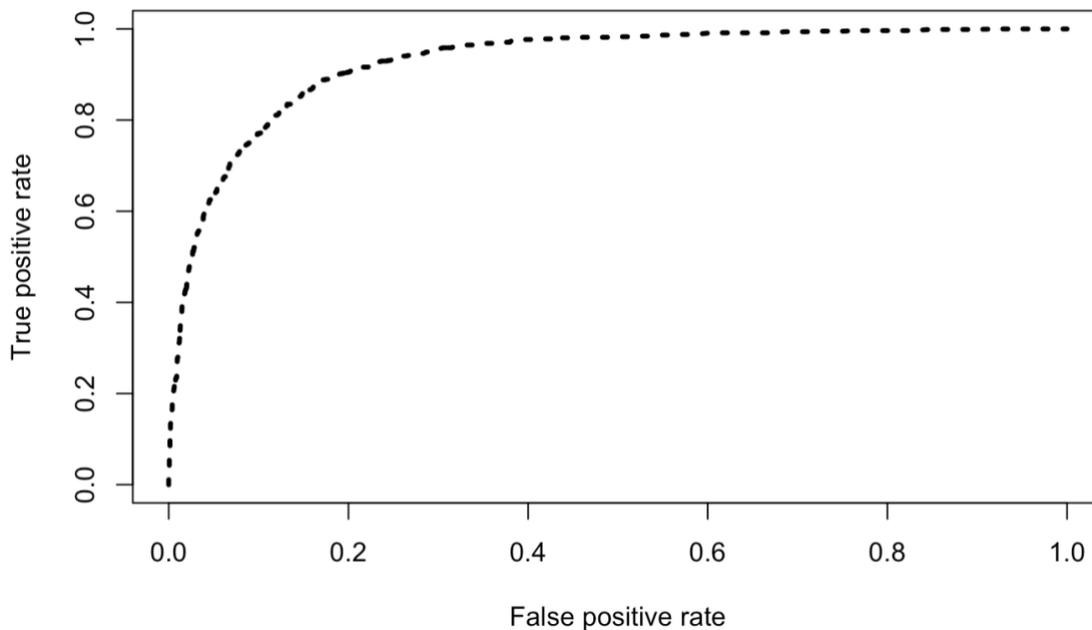


Figure 4.10: The above plot shows a ROC curve for the 2015 data. The x-axis represents the false positive rate and the y-axis represents the true positive rate.

For each of the models, both plots looked optimal. The light and dark-blue lines on the posterior predictive checks have near perfect overlap and the ROC curve extends well into the upper-left corners.

The goal of an AUC analysis is to make sure that the model performs better than random chance by measuring the area under the ROC curve. Table 3 displays the accuracy rate of the model, the percentage of at-bats in the test dataset that do not have shifts and the AUC. Notice that the model accuracy decreases slightly as time increases due to the fact that shifts have become much more common.

Table 3: Shift Probability Model Diagnostics

| Season Model | Accuracy | Percentage of ABs without shifts | AUC |
|--------------|----------|----------------------------------|------|
| 2019 | 81% | 69% | 0.88 |
| 2018 | 86% | 77% | 0.91 |
| 2017 | 91% | 84% | 0.94 |
| 2016 | 89% | 84% | 0.92 |
| 2015 | 91% | 87% | 0.93 |

Spray Angle Model

Model Overview

In order to understand how Major League hitters have adapted to the infield shift, I implemented both fixed and random effects within a Bayesian linear regression framework.

In this case, the response variable is the spray angle of a given batted-ball. The fixed effects in the spray angle model serve two roles. The batter-tendency data (i.e. cumulative season statistics) is instrumental in understanding how different types of hitters' distributions differ, while the pitch-level (i.e. situational) data provides insight on how batters act differently depending on the game situation. Unlike the shift probability models, only one model was run for spray angle and season was a covariate.

Due to the fact that shifts effect right-handed and left-handed hitters differently, interactions were added between `stand` and the following: `on 1b`, `on 2b`, `on 3b` and `infield alignment`. Without these interactions, left-handed batters going against the shift would be indistinguishable from right-handed batters hitting directly into the shift. Offensive strategies also vary by whether or not there are runners on base, and therefore, the fielding alignments in these situations could also have an effect.

The spray angle model is also more complex than the shifting probability models due to the inclusion of a nested random effect. Nesting ground balls within seasons, within infield alignments and within players allowed me to answer the following crucial questions:

- Within a given season, do individual players have different batted-ball distributions depending on the infield alignment?
- Over the five-year period, do individual players have different batted-ball distributions depending on the infield alignment?
- Have individual players gotten better at hitting against the shift over time?

By nesting the ground balls in this manner, the model output a coefficient for each hitter overall, a coefficient for each batter with the shift and without the shift, and a coefficient for each batter with the shift and without the shift for each individual season that they are present in the dataset. The full model can be seen below:

$$\begin{aligned}
 angle_i = & \alpha_{j[i]}^{hitter} + \alpha_{k[j][i]}^{hitter*infield.alignment} + \alpha_{l[k][j][i]}^{hitter*infield.alignment*season} + \beta_{stand}stand_i + \\
 & \beta_{p.throws}p.throws_i + \beta_{balls1}balls1_i + \beta_{balls2}balls2_i + \beta_{balls3}balls3_i + \beta_{strikes1}strikes1_i + \\
 & \beta_{strikes2}strikes2_i + \beta_{on.3b}on.3b_i + \beta_{on.2b}on.2b_i + \beta_{on.1b}on.1b_i + \beta_{on.3b*stand}on.3b * \\
 & stand_i + \beta_{on.2b*stand}on.2b * stand_i + \beta_{on.1b*stand}on.1b * stand_i + \beta_{outs1}outs1_i \\
 & + \beta_{outs2}outs2_i + \beta_{inning}inning_i + \beta_{score.diff}score.diff_i + \beta_{home.runs}home.runs_i + \\
 & \beta_{strikeouts}strikeouts_i + \beta_{wOBA}wOBA_i + \beta_{fly.balls}fly.balls_i + \beta_{pull}pull_i + \\
 & \beta_{hard.hit}hard.hit_i + \beta_{age.cat<25}age.cat < 25_i + \beta_{age.cat25-29}age.cat25 - 29_i + \\
 & \beta_{age.cat30-34}age.cat30 - 34_i + \beta_{age.cat35+}age.cat35+_i + \beta_{season2016}season2016_i + \\
 & \beta_{season2017}season2017_i + \beta_{season2018}season2018_i + \beta_{season2019}season2019_i + \\
 & \beta_{infield.alignment}infield.alignment_i + \beta_{infield.alignment*stand}infield.alignment * stand_i
 \end{aligned}$$

$$\alpha_j^{hitter} \sim N(\mu_a^{hitter}, \sigma_{hitter}^2)$$

$$\alpha_{k[j]}^{hitter*infield.alignment} \sim N(\mu_{a_j}^{hitter*infield.alignment}, \sigma_{hitter*infield.alignment}^2)$$

$$\alpha_{l[k][j]}^{\text{hitter*infield.alignment*season}} \sim N(\mu_{\alpha_{k[j]}}^{\text{hitter*infield.alignment*season}}, \sigma_{\text{hitter*infield.alignment*season}}^2)$$

for $i = 1, \dots, N$ $j = 1, \dots, J$ $k = \text{Shift, No Shift}$ $l = 2015, \dots, 2019$ where i represents a given ground ball, j represents a unique player, k represents a whether a shift was implemented during the given at-bat and l represents the season the at-bat took place.

Fit Diagnostics

The spray angle model converged after a single run as none of the parameters had R-hat values equal to or greater than 1.05. In order to make sure that the model met the regression assumptions, I constructed residual plots which can be seen in Figures 4.11 and 4.12. The model appears to pass the diagnostics as the residuals appear to be random, noncorrelated and the errors appear to have an approximate normal distribution.

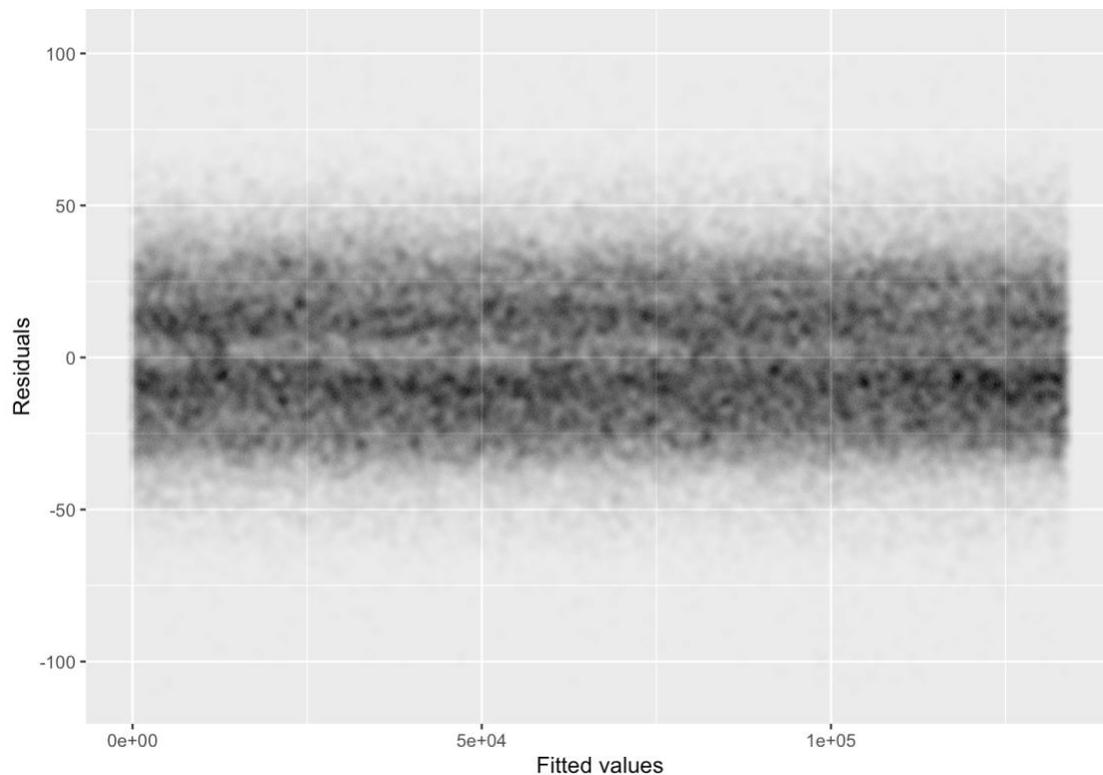


Figure 4.11: The above plot shows a residual-fitted values plot. The x-axis represents the fitted values and the y-axis represents the residuals.

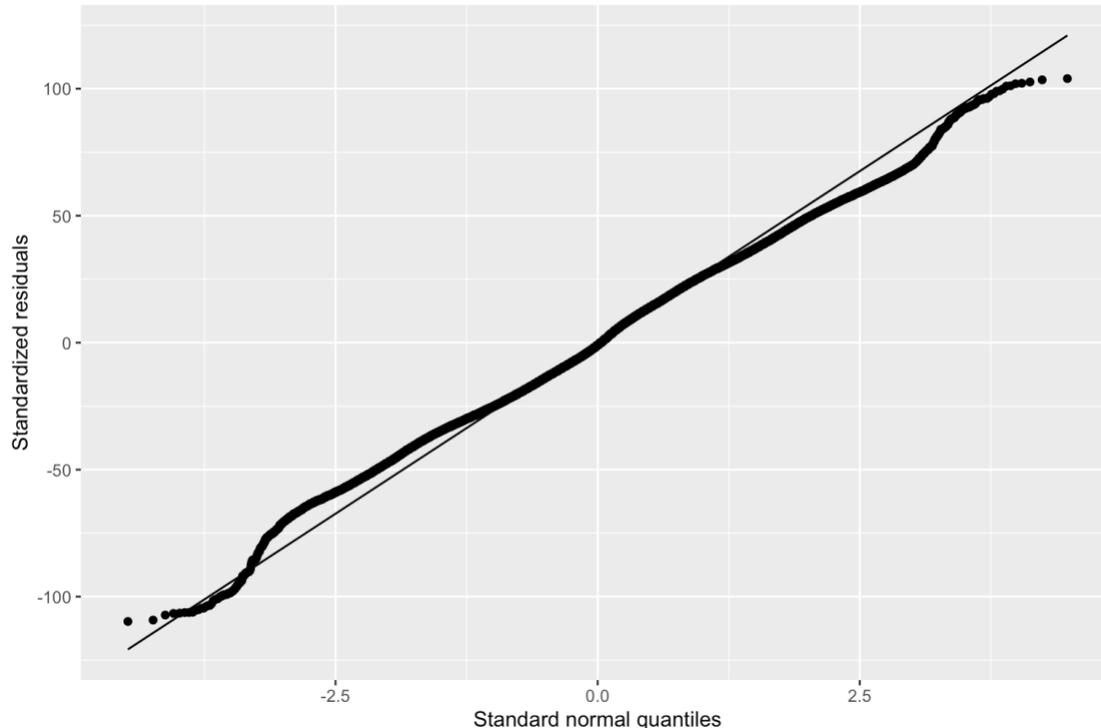


Figure 4.12: The above plot shows a Quantile-Quantile plot. The x-axis represents the standard normal quantiles and the y-axis represents the standardized residuals.

Random Effect Coefficient Densities

The final step in my analysis before interpreting the random effect coefficients was to calculate the density of coefficient differences for each individual draw of the 4000 total. Using the **tidybayes** package, I calculated three types of differences from which 95 percent credible intervals were created to assess significance:

- For each player-season, I calculated the difference in the coefficients with the shift and without the shift using the $\alpha_{l[k][j]}^{\text{hitter*infield.alignment*season}}$ intercepts.
- For each player, I calculated the difference in the coefficients with the shift and without the shift using the $\alpha_{k[j]}^{\text{hitter*infield.alignment}}$ intercepts.
- In order to see whether players improved at hitting against the shift, I calculated the difference between the player-season differences for each season combination (e.g. Player A's 2015 difference - Player A's 2018 difference).

Results

After checking the validity of the models through diagnostic plots in the previous section, this section provides further insight into the significance of the models and key behavioral trends involving the infield shift. Throughout this section, parameter significance is

calculated by evaluating whether or not zero is within the 95 percent credible intervals. Since Bayesian regression modelling provides the posterior distributions of the coefficients, the credible interpretation is that there is a 95 percent chance that the true parameter lies within the given interval.

Factors That Affect Shifting

The infield shift probability models provide clear insight into the factors considered by Major League Baseball teams when deciding whether or not to shift on a batter. Over the five-year period of shift behavior analyzed throughout this project, numerous covariates were significant in the same direction across every model. Interpretations and intuition as to why the trends might occur can be seen below. (*Note: The credible intervals are from the 2019 model*):

- **Right-handed hitters are much less likely to be shifted upon.** There is a 95 percent chance that the odds a right-handed hitter faces a shift are between 89 and 93 percent less than the odds a left-handed hitter faces one. About 42 percent of left-handed plate appearances had shifts in 2019 compared to about 14 percent for right-handed hitters. Considering the fact that about 30 percent of MLB hitters are left-handed, this result seems reasonable.
- **Teams shift more with right-handed pitchers on the mound.** There is a 95 percent chance that the odds a batter faces a shift increase between 30 and 57 percent when the opposing pitcher throws right-handed. This also makes intuitive sense because pitches from a right-handed batter will break in towards a left-handed hitter. Therefore, the batters would be more likely to pull the ball into the shift.
- **Shifts are less common when there are runners on base.** Regardless of the base being occupied, the probability of a shift is lower with runners on. There is a 95 percent chance that the odds a batter faced a shift with a runner on first base were between 32 and 42 percent less than if no one was on base. In addition, in the 2016, 2017, 2018 and 2019 models, the probability of a shift with a runner on second was significantly less than the situation with a runner on first. The reason for this is likely because shifts reduce the probability of force outs at bases by taking fielders out of position. For example, turning a double play is much more difficult in a shifted setting for a left-handed batter. If a left-handed batter is at the plate, a shift with a runner on second is even less likely because the fielding team would be leaving third base open for an easy stolen base.
- **Teams shift more often when there are two outs.** When teams do not need to get force outs, and a throw to first base will suffice, they are more likely to opt for a shift. There is a 95 percent chance that with two outs the odds a team shifts are between 12 and 34 percent greater than if there were no outs. It is worth noting that the coefficient for opportunities with one out was not significant, plausibly due to some of the factors mentioned about base-running.
- **Shifting is less common in the later innings.** There is a 95 percent chance that the odds a team shifts in the 9th inning are between 13 and 39 percent less than the odds a team does in the first inning. In 2019, the 6th through 9th innings were all significant. That is up from the 7th through 9th in 2018 and just the 8th and 9th in 2017 and 2016. One potential explanation

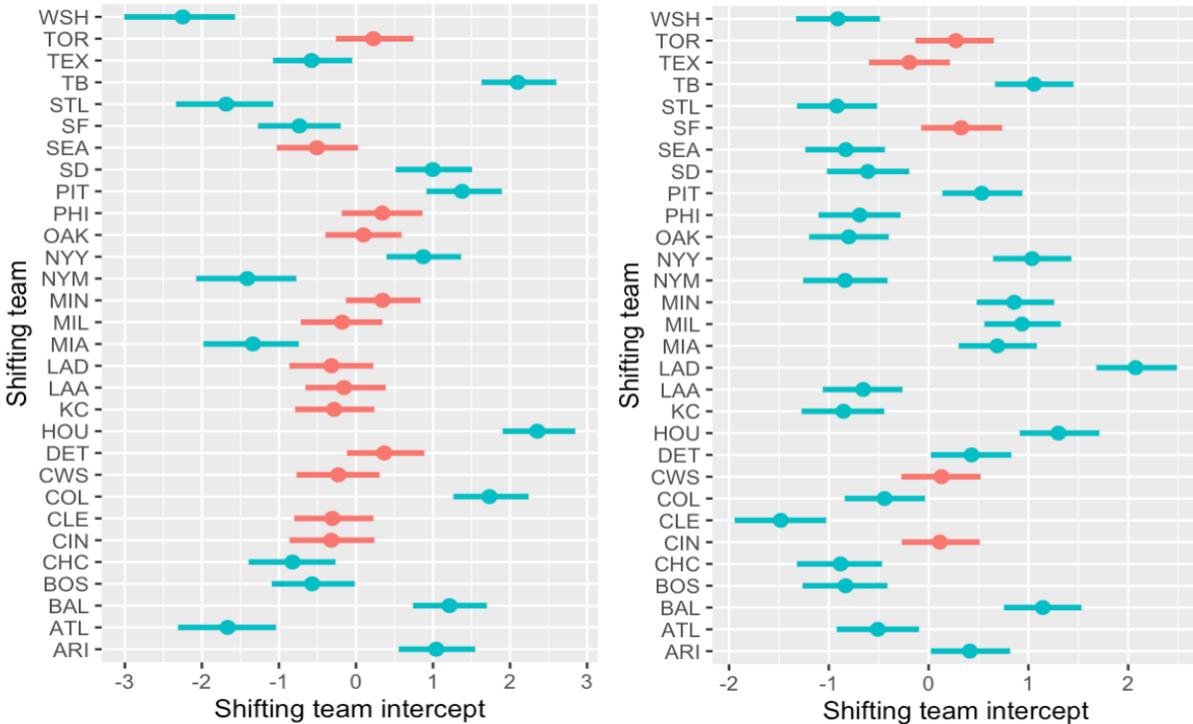
for this is that because pitchers at the end of games are more so high-strikeout pitchers than starters and middle-relievers, it's possible that teams could feel less of a need to shift. This also could occur because there are more runners on base in these high leverage situations. It is also possible this trend has to do with loss aversion. If a manager opts to shift and the team loses the game late because the ball gets through a big gap in the infield, there will be a lot more scrutiny from the fans than if a ball gets through an infield with a standard alignment.

- **Batters who pull the ball more get shifted on more.** The whole reasoning behind the infield shift is the fact that some players pull the ball more than others. Therefore, it would be concerning if this factor were not significant. According to the 2019 model, there is a 95 percent chance that the odds a batter whose pull percentage is 10 percent above league average will be shifted on are between 49 and 101 percent greater than a league average hitter.

In addition to the covariates that were significant across each of the past five seasons, a notable factor was significant across four of the five models. Score differential was significantly positive from 2016 to 2019. This would imply that teams shift more as the game gets less close. This is likely because in close games teams are often concerned with preventing runners from advancing bases or turning double plays. In blowouts, the priority is to record outs and shifting is a more optimal strategy.

Team Shifting Behavior

In addition to the significant fixed effects for the five shifting probability models, the random intercepts for the teams provide further insight. From the team perspective, the coefficients provide an understanding of the franchises that have fully bought into shifting and those that are still somewhat skeptical. Figures 5.1 and 5.2 are forest plots displaying the 95 percent credible intervals for the individual team intercepts from the 2015 and 2019 seasons, respectively. Intervals colored blue are significant whereas the red values are not significant.



Figures 5.1 and 5.2: Forest plots displaying the random intercepts for MLB teams during the 2015 season (left) and 2019 season (right).

Due to the fact that there was a steady increase in the number of significant intercepts throughout the period, only the 2019 and 2015 forest plots are shown. The increase in the number of significant values from 2015 to 2019 suggests that teams overall have picked their shifting strategy. On the 2015 plot, for example, most of the teams are hovered around zero with the Tampa Bay Rays and Houston Astros the two major adopters as the only teams with means greater than two. On the 2019 plot, it appears teams have found their path, with a large amount on either the right or the left, but few in the middle. Another interesting takeaway from the 2019 plot is the seeming divide amongst the analytics community. Analytic front-runners such as the Rays, Astros, New York Yankees and Los Angeles Dodgers appear to be all in on shifting as four of the top five implementers of shifts. Yet, the the Oakland “Moneyball” Athletics, are placed in the bottom half and the Cleveland Indians—who’re using analytics to revolutionize pitching—find themselves with the lowest coefficient of the 30 teams.

Spray Angle Model Interpretations

While teams can actually control when they apply shifts, the main takeaway from this thesis is that players still cannot fully control where they hit the ball. Neither the fixed effects nor random effects including or containing the infield alignment were statistically significant predictors for spray angle.

As described in the Methods section, the nested nature of the random effect allowed for coefficients specific to each player and player-season with and without the shift. This random effect structure made it possible to answer the following three questions:

- Did any players hit significantly differently with and without a shift?
- Did any players hit significantly differently with and without a shift within a given season?
- Have players gotten better at adapting to the shift over time?

After calculating the posterior densities, there is no evidence to suggest that even a single player had a significant difference in their batted-ball distribution with and without a shift either in a given season or over the five year period. This suggests that within a season players are not able to adapt to the shift on the fly and over the course of the season, controlling where they aim the ball is not possible.

Through both the residual analysis conducted in the Methods section and the sensible significant fixed effects, I am confident that the lack of significance from the random effects is accurate. The following are some of the interpretations from the spray angle model:

- There is a 95 percent chance that ground balls hit by a right-handed hitter would have an angle between 27.30 and 28.66 degrees less than a left-handed hitter. This makes sense because the intercept is between 16 and 28 degrees—between first and second base—so right-handers then would hit the ball between -12 and 0 degrees, which is between the shortstop in a standard alignment and second base.
- There is a 95 percent chance that ground balls hit against a right-handed pitcher will have a spray angle between 0.78 and 1.41 degrees more than if the pitcher was left-handed. This is plausible because the pitches will break towards the first-base side and therefore are more likely to be hit there.
- There is a 95 percent chance that ground balls hit in two-strike counts will have a spray angle between 0.04 and 0.76 degrees more than if there were no strikes. This makes sense because in two-strike counts batters are more focused on putting the ball in play. They are also usually less selective and thus are more likely to be late on a pitch—a right-handed hitter who is late will hit the ball to the right side.
- There is a 95 percent chance that ground balls hit with a runner on third base will have a spray angle between 0.82 and 2.29 degrees more than if there were no runners on base. There is also a 95 percent chance that ground balls hit with a runner on second base will have a spray angle between 0.12 and 1.22 degrees more than if there were no runners on base. These are plausible because the runners will be able to advance if the ball is not hit near either the base they are occupying or the base they wish to go to.

Discussion

While Major League teams have remained consistent in their shifting patterns since 2015, professional hitters have shown no sign of improvement on the ground. One question this thesis unfortunately did not answer is whether or not affected batters have focused their

training on beating the shift on the ground. Numerous pull hitters across the league have revitalized their careers with the emphasis of beating the shift in the air (Sawchik, 2017). If a player is able to hit the ball far enough consistently enough, no defensive scheme can stop them. A potential next step would be to incorporate Statcast's launch angle measurements and further understand adaptation to the infield shift in the z-dimension.

I hope to rerun this analysis when newly collected spray angle data becomes available. Although Statcast records batted-ball launch angles, the spray angle data is collected from stringers. Stringers are essentially people sitting at baseball games who put dots on a screen to mark where a batted-ball is hit. The transformation mentioned in Chapter 3 addresses this issue by converting the screen-tapped coordinates to spray angles. It is possible that a more automatic collection method would lead to stronger results.

Another future analysis point would be to better understand the implementation of strategic alignments. Strategic alignments occur when teams partially shift, but do not have three players on one side of the infield. For the sake of this analysis, I only looked at full infield shifts, but as fielding patterns become more dynamic with increasing data, understanding strategic alignments could be crucial to remain ahead of the curve.

If I had more time, I would have liked to conducting more rigorous model selection with the hope of increasing the predictive viability of both modelling tasks. The shift probability models took about seven hours to run and the spray angle model took more than 48 hours. I would consider include implementing informative priors for the coefficients and testing other methods to account for potential nonlinearities in the data. I would also allow for the variance of a given player's batted-ball distribution to vary since understanding the shape of the predictive distribution is crucial for teams trying to implement shifts. Lastly, for the shift probability models, I would like to include interactions for each team in order to understand if different teams value different factors when deciding whether or not to shift on a given hitter.

Appendix

Spray Angle Transformation

In order to convert the spray angle values from x-y coordinates as provided by Statcast to an angle, I used the following transformation (Albert, 2018):

$$angle = \tan^{-1}((x - 125.42)/(198.27 - y)) * 180/\pi$$

Shift Probability Model Output

Table 4 displays the covariates implemented in the five shift-probability models and compares the significant factors across the models. Values of Yes correspond to parameters with credible intervals that did not include zero.

Table 4: Shift Probability Model Fixed-Effects

| Covariate | 2019 | 2018 | 2017 | 2016 | 2015 |
|-----------|------|------|------|------|------|
| Intercept | Yes | Yes | Yes | Yes | Yes |
| standR | Yes | Yes | Yes | Yes | Yes |
| pthrowsR | Yes | Yes | Yes | Yes | Yes |
| balls1 | No | No | No | No | No |
| balls2 | No | No | No | No | No |
| balls3 | No | No | No | No | No |
| strikes1 | No | No | No | No | No |
| strikes2 | Yes | No | Yes | No | No |
| on.3b | Yes | Yes | Yes | Yes | Yes |
| on.2b | Yes | Yes | Yes | Yes | Yes |
| on.1b | Yes | Yes | Yes | Yes | Yes |
| outs1 | No | No | No | No | Yes |
| outs2 | Yes | Yes | Yes | Yes | Yes |
| inning2 | No | No | No | No | No |
| inning3 | No | No | No | No | No |
| inning4 | No | No | No | No | No |
| inning5 | No | No | No | No | No |
| inning6 | Yes | No | No | No | No |
| inning7 | Yes | Yes | No | No | No |
| inning8 | Yes | Yes | Yes | Yes | No |
| inning9 | Yes | Yes | Yes | Yes | Yes |
| inning10 | No | No | No | No | No |

| | | | | | |
|------------|-----|-----|-----|-----|-----|
| inning11 | No | No | No | No | No |
| inning12 | No | No | No | No | No |
| score.diff | Yes | Yes | Yes | Yes | No |
| home.runs | No | No | Yes | Yes | No |
| strikeouts | Yes | No | Yes | No | No |
| wOBA | No | No | No | No | No |
| fly.balls | Yes | No | No | Yes | No |
| pull | Yes | Yes | Yes | Yes | Yes |
| hard.hit | No | Yes | Yes | Yes | Yes |
| age 25-29 | No | No | No | No | Yes |
| age 30-34 | No | Yes | No | No | No |
| age 35+ | No | Yes | Yes | No | Yes |

Spray Angle Model Random Effects

In order to calculate the credible intervals for the differences between the coefficients with the shift and without the shift, I extracted each draw using the `spread_draws` function and created a tidy dataframe so that each row represented either a player or a player season. I then subtracted the coefficients with the shift and then grouped by either the player or the player and the season. The summary statistics for those groupings were the 2.5 percent and 97.5 percent posterior quantiles. Example rows of the dataframes for the main random effect questions can be seen below:

Within a given season, do individual players have different batted-ball distributions depending on the infield alignment?

| Batter Name | Season | 2.5 percent | 97.5 percent |
|--------------|--------|-------------|--------------|
| A.J. Pollock | 2015 | -4.7739800 | 4.012689 |

Over the five-year period, do individual players have different batted-ball distributions depending on the infield alignment?

| Batter Name | 2.5 percent | 97.5 percent |
|--------------|-------------|--------------|
| A.J. Pollock | -1.312993 | 1.294864 |

References

- Albert, J. (2018, January 15). Chance of Hit as Function of Launch Angle, Exit Velocity and Spray Angle. Retrieved February 25, 2019, from Exploring Baseball Data with R website: <https://baseballwithr.wordpress.com/2018/01/15/chance-of-hit-as-function-of-launch-angle-exit-velocity-and-spray-angle/>
- Arthur, R. (2017, June 6). Ryan Howard's Career Is Dead. The Shift Killed It. Retrieved April 19, 2019, from FiveThirtyEight website: <https://fivethirtyeight.com/features/ryan-howards-career-is-dead-the-shift-killed-it/>
- Crasnick, J. (2018, July 10). MLB hitters explain why they can't just beat the shift. Retrieved March 30, 2019, from ESPN.com website: http://www.espn.com/mlb/story/_/id/24049347
- Hawke, C. J. (2017). Quantifying the Effect of The Shift in Major League Baseball. *Senior Projects Spring 2017*. 191. http://digitalcommons.bard.edu/senproj_s2017/191
- Jensen, S. T., Shirley, K. E., & Wyner, A. J. (2009). Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, 3(2), 491–520. <https://doi.org/10.1214/08-AOAS228>
- Kawahara, M. (2016, June 25). Retrieved March 30, 2019, from Sacramento Bee website: https://www.sacbee.com/sports/mlb/34igqh/picture86023232/alternates/FREE_1140/Bballshift2
- Lewis, M., & Bailey, R. (2015). Batted ball spray charts: a system to determine infield shifting. *2015 Systems and Information Engineering Design Symposium*, 206–211. <https://doi.org/10.1109/SIEDS.2015.7116975>
- Petriello, M. (2018, May 16). 9 things you need to know about the shift. Retrieved February 24, 2019, from MLB.com website: <https://www.mlb.com/news/9-things-you-need-to-know-about-the-shift/c-276706888>
- Sawchik, T. (2017, February 6). Can More MLB Hitters Get Off the Ground? Retrieved February 24, 2019, from FanGraphs Baseball website: <https://blogs.fangraphs.com/can-more-mlb-hitters-get-off-the-ground/>
- Shaw, S. (2015, August 5). Beating Defensive Shifts with Game Theory | Banished to the Pen. Retrieved March 2, 2019, from Banished to the Pen website: <http://www.banishedtothepen.com/beating-defensive-shifts-with-game-theory/>
- Slowinski, S. (2010, February 15). wOBA | Sabermetrics Library. Retrieved April 28, 2019, from <https://library.fangraphs.com/offense/woba/>
- Slowinski, S. (2010, February 16). BABIP | Sabermetrics Library. Retrieved April 28, 2019, from <https://library.fangraphs.com/pitching/babip/>

Sullivan, J. (2014, March 21). More About Bunting and Beating the Shift | FanGraphs Baseball. Retrieved March 30, 2019, from FanGraphs.com website:
<https://blogs.fangraphs.com/more-about-bunting-and-beating-the-shift/>