CAN YOU STACK THE DEFENSE IN YOUR FAVOR? EXAMINING THE EFFECT OF THE INFIELD SHIFT ON BATTING AVERAGE AND SLUGGING PERCENTAGE IN MAJOR LEAGUE BASEBALL

A THESIS

Presented to

The Faculty of the Department of Economics and Business

The Colorado College

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Arts

By

Justine Miller

May 2021

CAN YOU STACK THE DEFENSE IN YOUR FAVOR? EXAMINING THE EFFECT OF THE INFIELD SHIFT ON BATTING AVERAGE AND SLUGGING PERCENTAGE IN MAJOR LEAGUE BASEBALL

Justine Miller

May 2021

Economics

Abstract

The infield shift is a defensive strategy used in baseball to decrease opponents' batting success by moving fielders to positions where the batter is most likely to hit the ball. This strategy has existed since the 1920s, but has increased in use in the last decade, aided by the new Statcast technology installed in the Major League Baseball stadiums in 2015. Although the simplest way for the batter to counteract the shift is to hit to where there are fewer fielders, the majority of batters attempt to hit over the fielders. This suggests that even if the shift successfully decreases batting average, it may consequently increase slugging percentage, as more players are changing their behavior to hit to the outfield. No peer-reviewed journal articles were found investigating the effect of the shift on batting performance, indicating a need for research in this area. Ordinary least squares regression was used to determine the effect of the shift on batting average in one model and the effect on slugging percentage in another. The results demonstrated that a one standard deviation increase in the percent of plate appearances facing a shift leads to a decrease of approximately 0.009 or 25% of a standard deviation in batting average, but an increase of approximately 0.008 or 10% of a standard deviation in slugging percentage. Therefore, the effect of the infield shift on batting average is greater than the effect on slugging percentage, suggesting that teams should continue to use the shift to decrease their opponents' success.

KEYWORDS: (Infield Shift, Batting Average, Slugging Percentage, Ordinary Least Squares Regression)
JEL CODES: (L83, Z21)

ON MY HONOR, I HAVE NEITHER GIVEN NOR RECEIVED
UNAUTHORIZED AID ON THIS THESIS

Signature

ACKNOWLEDGMENTS

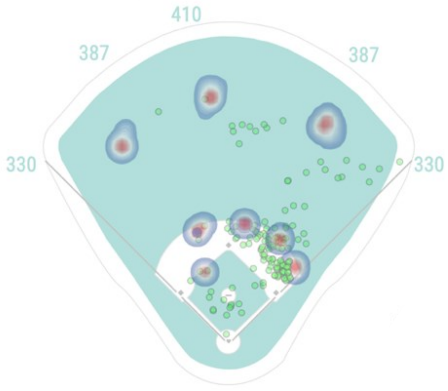
I would like to thank my thesis advisor, John Mann, who encouraged me to pursue my passion for this research topic and has guided me throughout every step of this process, including entertaining me with long conversations about baseball. I would also like to extend a thank you to Kevin Rask, Ph.D. who has fielded all of my questions, ranging from simple to complicated and has been an incredibly generous unofficial additional advisor for this project. Finally, this work and all my studies would not be possible without the encouragement and support of my family for which I am forever grateful.

TABLE OF CONTENTS

Introduction

It's a beautiful, sunny day in Minneapolis, Minnesota. A slight breeze wafts through Target Field as right fielder, Max Kepler, steps up to the plate at the start of the inning. The Houston Astros infielders take a quick look at their wrist strapped play-cards and shortstop, Carlos Correa, moves across the infield to stand slightly to the right side behind second base, leaving only third baseman, Alex Bregman, on the left side of the infield. What is going on? Why would a team leave so much space to hit the ball on the left side of the field? What some may not know is that Max Kepler almost always hits to the right side of the infield when there are no runners on base (Figure 1). If Kepler hits a ground ball to the right side of the infield, which he is statistically very likely to do in this situation, he is almost guaranteed to hit into an out and the Astros will have successfully defended against Kepler's hitting power with what is referred to as *the infield shift*.

Figure 1: Max Kepler's 2019 Spray Chart Against the Astros with No Runners on Base



Source: MLB Player Positioning vs Batter, 2020

The infield shift is a defensive strategy in baseball used to decrease base hits by moving infielders to positions where the current batter is most likely to hit the ball. The shift was first used in the 1920s, but did not make headlines until 1946 when it was infamously used by Lou Boudreau of the Cleveland Indians to diminish the hitting

success of Hall of Famer Ted Williams (Levine & Bierig, 2017). Some complain that this feature of defensive strategy has taken away from the athleticism of baseball and the MLB commissioner, Rob Manfred, has even considered banning the shift (Sullivan, 2015). The infield shift as it is used today, was inspired by the recent era of home runs in an attempt to stop the success of players pulling the ball (Sullivan, 2015). Use of the infield shift has increased in the past two decades (Sheehan, 2015) and has been aided by the installation of Statcast technology in all 30 Major League Ball Parks in 2015 (Levine & Bierig, 2017). This new technology gives teams the opportunity to examine statistics from every pitch of every at-bat in every single game, allowing them to know exactly where to position themselves to have the highest probability of making an out. However, like any defensive strategy, the infield shift may not be foolproof.

The seemingly simplest way to beat the infield shift is to hit a ground ball to the side of the field that is less defended; however, most players attempt to hit over the shift instead (Levine & Bierig, 2017). As stated by MLB player Josh Donaldson, who averaged 37 homers and 100 RBIs from 2015 to 2017, “In the big leagues these things they call groundballs are outs” (Levine & Bierig, 2017, p. 12). If batters are changing their behavior when the shift is employed by hitting over the infield, it is possible that while the shift may be taking away groundball singles, it may also be creating more extra base hits to the outfield. Thus, while the purpose of the shift is to take away hits, the spillover effects on other aspects of batting performance are critical in evaluating its impact. This study aims to provide insight on the efficiency of the infield shift on both batting average and slugging percentage (SLG) by asking the question: is the infield shift

successful at decreasing hits and if so, does it consequently create more opportunities for extra base hits?

Literature Review

Thorough research uncovered some opinion pieces and online published research discussing the effect of the infield shift on batting performance. Although there are no peer-reviewed academic journals directly addressing the focus of this study, some academic journal articles discuss the variability in batting performance due to other variables.

One study comparing batting average prediction strategies used physical attributes of a batters' swing, including launch angle, exit velocity, and distance of a hit as independent variables (Bailey et al., 2020). These variables were chosen based on a hypothesis that they were the significant physical variables impacting players' hitting success and were found, using a logistic regression model, to be statistically significant in predicting the probability of getting a hit (Bailey et al., 2020). However, the mean absolute error which measured the discrepancy between predicted batting average and actual batting average for this prediction method was determined to be 0.0208 batting average points (Bailey et al., 2020). Thus, Bailey et al. suggest that this batting average prediction could be improved by including other variables in the model such as age, injuries, and player speed (2020).

Another study focusing on the variable of injuries determined a negative relationship between concussions and batting performance (Wasserman et al., 2015). In this study, 66 instances of leave due to concussions and 68 instances of bereavement or paternity leave were compared within a retrospective cohort study design in order to

determine the effect of concussions on batting performance when players return after a leave of absence. The study determined that within two weeks after their return, players' batting averages, on-base percentages (OBP), slugging percentages (SLG), and on base plus slugging percentages (OPS) were significantly lower from leaves of absence due to concussion than leaves due to bereavement or paternity (Wasserman et al., 2015). This significant difference was present even when controls were implemented for pre-leave batting performance, player position, and number of days missed during leave (Wasserman et al., 2015). The study concluded that further research was necessary to determine exactly how concussions affect batting performance as a means to create better return-to-play protocols within the MLB (Wasserman et al., 2015). Although this study does not relate batting performance to defensive formation, it does suggest that there are other variables that may affect batting performance other than the physical attributes of each hit such as launch angle, exit velocity, or distance of a hit.

Batting performance and the variables that affect performance are important because they impact wins. Winning games is important for a team because it increases fan attendance and therefore increases profit and salaries alike. Both batting average and SLG have been demonstrated to have a significant effect on the number of runs teams score and their winning percentages. Changes in defensive positioning are built to disrupt batting performance and success as a means to increase team wins.

Supporting research analyzed the winning success of Major League Baseball teams in 2014 found that both earned run average (ERA) and on base plus slugging percentage (OPS) were statistically significant variables in predicting number of wins for an MLB team during the 2014 season (Peach et al., 2016). The ERA variable was

determined to have a negative correlation to the number of wins a team earned throughout the season while the OPS variable was determined to have a positive correlation (Peach et al., 2016). Although the OPS statistic is not exactly the same as the SLG statistic, they are directly correlated as OPS includes the SLG statistic and therefore the effect of the OPS variable is likely similar to a SLG variable. Subsequently, OPS also includes on-base percentage (OBP), which is equal to the number of hits and the number of walks a player has throughout the season divided by their total number of at bats. This means that the OBP statistic is similar to the batting average statistic (number of hits divided by total number of at bats) and may suggest that that a batting average variable may also have a similar effect to the OPS variable. Peach et al. confirm these comparisons by stating both OBP and SLG have been shown to significantly increase team performance (2016).

Similar to the findings by Peach et al., a study examining salaries, performance and owners' goals in the MLB during the 1999 season determined that home run hitting ability, batting average, ability to hit runs in, and the ability to draw walks were all player skills that were statistically significant in increasing the number of wins for a team (Yilmaz & Chatterjee, 2003). The study also examined the relationship between batting performance and fan attendance, which was used as a proxy for financial success, as well as the relationship between batting performance and player salaries (Yilmaz & Chatterjee, 2003). The best model determined in the study for predicting Log salary for players with salaries equal to or more than $1million included home runs, walks, and batting average as independent variables, which were able to explain 32.2% of the variability in Log salary (Yilmaz & Chatterjee, 2003). The best model determined for

predicting number of wins included mean runs batted in and maximum number of walks as independent variables, which were able to explain 57.1% of the variability in number of wins (Yilmaz & Chatterjee, 2003). Finally, the best model for fan attendance included maximum batting average, maximum home runs and maximum walks which explained 55.2% of the variability in fan attendance (Yilmaz & Chatterjee, 2003). Although this study is more dated and possibly comes to different beta estimate conclusions than what may be found with current data, it successfully demonstrates the relationship between batting performance and team success as well as financial success for both the team and individual players.

Another study investigating team revenues and MLB salaries found using a hierarchical linear model that a player's individual characteristics are significant predictors of their salary (Brown & Jepsen, 2009). OBP and SLG were found to be particularly important at predicting player salaries as both variables had positive beta estimates and were statistically significant at better than the 99% level (Brown & Jepsen, 2009). The study also determined that teams do not pay differently for individual player statistics as the Moneyball theory suggested (Brown & Jepsen, 2009), however, this may be because franchises have adjusted their spending in reaction to Oakland's success at exploiting this differential payment in the early 2000s. Brown and Jepsen also found that fielding average had a positive beta estimate in predicting player salary which other studies have not examined (2009). The study also concluded that teams with higher total revenues succeed more often than teams with lower payrolls due to their ability to purchase more players with desirable characteristics (Brown & Jepsen, 2009). This

conclusion demonstrates that teams should be motivated to obtain higher revenues in order to further increase their winning success.

This finding is supported by the results of a study which used a data envelopment analysis technique to measure franchise payroll efficiency in both the NFL and the MLB (Einolf, 2004). From analyzing data in the NFL from 1981 to 2000 and the MLB from 1985 to 2001, Einolf determined that due to the lack of a salary cap and less revenue sharing, the MLB has less payroll efficiency than the NFL (2004). The study found that big spending and inefficient MLB teams often come from large media market, while small spending and efficient MLB teams come from small markets (Einolf, 2004). This was suggested to be the case because large market MLB teams receive greater revenue from their decisions and tend to overspend for on field performance (Einolf, 2004). In his study, Einolf claimed that the MLB economic structure creates a significant advantage for large market teams and therefore encourages inefficiency (2004). The study concludes with the statement that in the MLB as opposed to the NFL, winning is more important than efficiency, otherwise teams would not spend as much as they do trying to ensure success (Einolf, 2004).

However, in contrast to the conclusion that winning is optimal for teams, one study performed at the University of Indianapolis found that too much success may actually have a negative consequence for baseball teams (Zimmer, 2018). This study demonstrated that as the number of previous World Series Championships increased for Major League Baseball teams, fan attendance decreased (Zimmer, 2018). Zimmer hypothesized that this correlation may occur due to increases in fan apathy from having a very successful team (2018). In addition to this finding, however, the study determined

that within a given season, increases in a team's winning percentage increased fan attendance (Zimmer, 2018) which agrees with the findings within the other studies reviewed.

An additional study, performed at the University of Alberta, determined using an ordinary least squares regression model that fan attendance is not only impacted by individual team success but is influenced by competitive balance within the league as well (Soebbing, 2008). The study used an actual to idealized standard deviation ratio (AISDR) in reference to team win percentage to measure competitive balance and found that the variable had a negative beta estimate and was significant at the 99% level (Soebbing, 2008). This finding supports the uncertainty of outcome hypothesis which assumes fans gain more utility from watching games with unpredictable outcomes and therefore more fans will attend games in which the teams playing are more evenly matched (Soebbing, 2008). The games behind from a playoff appearance variable was also found to be negative and significant at the 99% level, demonstrating that individual team performance directly impacts fan attendance (Soebbing, 2008) which supports the findings from the previous studies analyzed.

As seen from the analysis of previous research, winning and scoring runs is vital for a team and therefore the infield shift may be a very important factor in baseball if it influences batting performance and consequently winning for a team. However, when looking at the impact of different batting performance statistics on winning, there has been a continuous debate of which, if any, is the most significant. For example, *Moneyball* by Michael Lewis claimed that player skills in the MLB were valued very inefficiently in terms of salaries and this is what allowed Billy Beane of the Oakland

Athletics to have a successful season with a very minimal budget (Lewis, 2003). This theory also claimed that the OBP statistic was actually more significant to winning games than SLG (Lewis, 2003). In a later study that evaluated this theory, the OBP and SLG were compared in terms of their effect on winning for a team and concluded that a one-point change in a team's OBP makes a more significant contribution to team winning percentage than a one-point change in SLG (Hakes & Sauer, 2006). This study also supported the Moneyball theory claim that OBP was an undervalued skill financially in the MLB during the 2000-2004 period, however, state that the market seems to have corrected this inefficiency after the findings of Lewis were published (Hakes & Sauer, 2006).

Although it was confirmed by Deli that the Moneyball theory was correct in claiming that certain characteristics of players were undervalued financially during the early 2000s, within his study assessing relative inputs in a production function, it is argued that OBP may not be more significant than SLG as the Moneyball theory suggested (2013). Deli's study demonstrated that OBP and SLG come from different distributions and do not have the exact same unit of measure (2013). Deli stated that when comparing variables within regression, the relative variability of each variable must be considered (2013). The study determined that there was much more variability in the SLG variable and therefore increasing OBP by 1% was much more difficult than increasing SLG by 1%, concluding that OBP is not necessarily more significant in predicting the number of runs scored (Deli, 2013).

Unlike Deli's findings however, Lee found within his study examining the Korean baseball league that while SLG and OBP are both significant in increasing number of

runs scored at the 95% level, OBP was approximately 2-3 times more important than SLG (2011). The study used a panel data analysis of a stochastic production frontier model and the results demonstrated that a ten-percentage point rise in OBP increased the number of runs scored by 41.7% while the same percentage rise in SLG only increased it by 18.6% (Lee, 2011). Lee's study also evaluated the effectiveness of small ball in the Korean baseball league, an offensive strategy that is used to get runners on base and move them into scoring position through methods such as stealing bases, bunting, using pinch hitters, hit-and-run-plays, and other related plays which often include sacrificing an out in order to advance runners (2011). The results of the small ball variables, which included stolen base attempts, sacrifice hits, and number of players used in a game, were mixed with stealing attempts found to be beneficial to scoring runs while sacrifice hits and number of players used were detrimental (Lee, 2011). However, based on the magnitude of all the variables, the overall effect was negative on runs scored, which demonstrates that letting batters hit may be more efficient than using small ball techniques (Lee, 2011).

A more recent study using a Markov decision process model determined in contrast to Lee that sacrifice bunts were more beneficial than previously thought (Hirotsu & Bickel, 2019). Hirotsu and Bickel claim that this study examined situations that had not been studied before which could explain why their conclusion was different than other studies (2019). Additionally the study investigated the effect of sacrifice bunts on the probability of winning a game as opposed to number of runs scored which followed the reasoning that the objective of a game is to win and not just score runs and therefore using the probability of winning as the dependent variable is a better measure to

determine the success of the sacrifice bunt (Hirotsu & Bickel, 2019). The study concluded that sacrifice bunts are found to have a positive impact on the probability of winning a game in specific situations such as when a team has a large lead or during an early inning of a game (Hirotsu & Bickel, 2019). Cumulatively, these studies show that there is not a definite variable that is has been shown to be the most important in affecting wins.

The studies reviewed demonstrate that there are many variables that affect batting performance including factors such as physical attributes of a swing and player injuries. It is possible that the infield shift could also be a variable that significantly impacts batting. Batting performance is important because it positively impacts runs scored and winning percentage for a team which in turn positively impacts fan attendance and revenue. Although it is clear that batting performance affects winning, there is not one definite variable that has been demonstrated to be the most important in predicting wins.

Theory

This study will use an ordinary least squares regression model to determine the effect of the infield shift on the variability in both batting average and SLG. As opposed to focusing on the physical aspects of a swing in looking at the variability in batting performance, this study focuses on the physical attributes of players as well as the situational aspects of plate appearances. These variables include: total number of plate appearances, player age, total number of pitches faced, types of pitches faced, average speed of pitches faced, percent of pitches in the strike zone, player sprint speed, player bat handedness, players' most common position in the batting order, division, league, and

percentage of plate appearances when an infield shift is in play. This leads us to the equation:

$$\begin{aligned}\gamma = {} & \beta_1 PlateAppearances + \beta_2 PlayerAge \\ & + \beta_3 BreakingBalls + \beta_4 FastBalls \\ & + \beta_5 TotalPitches + \beta_6 AveragePitchSpeed \\ & + \beta_7 InZonePercent + \beta_8 SprintSpeed \\ & + \beta_9 BattingFirst + \beta_{10} BattingSecond \\ & + \beta_{11} BattingThird + \beta_{12} BattingFourth \\ & + \beta_{13} BattingFifth + \beta_{14} BattingSixth \\ & + \beta_{15} BattingSeventh + \beta_{16} BattingEighth \\ & + \beta_{17} American + \beta_{18} East + \beta_{19} Central \\ & + \beta_{20} BatSide + \beta_{21} PercentageShifts\end{aligned}$$

Equation 1

One model implements batting average as the dependent variable in this equation and the other uses SLG. The number of off-speed pitches, batting ninth, and west division variables are all omitted from the model due to redundancy. Although not every variable used in the model has evidence supporting a direct effect on batting average or SLG, each has evidence supporting its effect on winning percentage or runs scored, which existing research suggests is in some ways correlated with batting performance.

Total number of plate appearances. It is expected that as players face more plate appearances, they will perform better as they get more experience, though at a certain point it is possible that too many plate appearances may cause physical fatigue for a batter. A study performed by Demiralp et al. found that within a fixed effects regression model, the number of games played has a negative impact on a player's OPS statistic

(2012). It was hypothesized that this result is likely due to the physical fatigue a player faces when playing higher numbers of games (Demiralp et al., 2012). Additionally, the number of games played squared variable also had a negative impact on OPS, demonstrating that this negative effect increases with more games played (Demiralp et al., 2012). Unlike this result, however, the number of games played was found to positively impact batting average (Demiralp et al., 2012). This result may be due to the experience that players gain throughout the season as they play more games. While the number of games played is not exactly proportional to the number of plate appearances a player faces, as players play more games they will have higher number of plate appearances and therefore it can be assumed that the effect of the number of plate appearances will likely be very similar to number of games played on batting performance.

Player age. As players age, it would be expected that they are able to gain skill and knowledge from their experience in the MLB, however, after a period of time it would also be expected that their physical abilities decrease. A study examining this experience-productivity relationship in the MLB found that age does significantly impact batting average (Krohn, 1983). This study used a linear regression model with batting average as the dependent variable and age and age squared as the independent variables (Krohn, 1983). The results demonstrated that gaining experience helps increase players' batting averages, but at a certain point, age causes players' physical abilities, and therefore their batting averages, to decline (Krohn, 1983). The study determined that the peak of a player's batting average is 28 years old with a standard error of about 2 years (Krohn, 1983). Demiralp et al. found similar results using a fixed effects regression

model examining the effect of age on OPS and batting average as well as other variables such as stealing bases (2012). Their regression results demonstrated that age has a positive impact on OPS, batting average, and stealing bases (Demiralp et al., 2012). However, the results also found that the age squared variable had a negative impact on OPS, batting average, and stealing bases, illustrating that the productivity of a player as they age increases at a decreasing rate (Demiralp et al., 2012). These results support the hypothesis that as players age they gain experience and skill, however, their physical abilities eventually decline. From their results, Demiralp et al. claimed that batting performance peaks at the age of 30, which is similar to Krohn's results (1983), while base stealing peaks at the age of 27 (2012). Hakes and Turner also demonstrated a relationship between batting productivity and age, with all players increasing in productivity as they age, but at a decreasing rate (2011). In addition, Hakes and Turner used quintile analysis to determine that the most skilled players peak in performance about two years later than lower skilled players (2011). This literature demonstrates that within this regression, age will likely have a significant effect on both batting average and SLG.

Sprint speed. In this study, sprint speed refers to the feet per second a player can run in their fastest one-second window. While a player's running speed may not directly affect how often or how well they hit the ball, it may help to increase how many times they can get on base by outrunning the throw to first base, thus increasing their batting average. It may also increase the number of bases they can gain on a hit, thereby increasing their SLG. Bailey et al. support this idea in their study, which focused on the prediction of batting averages, by suggesting in their study conclusions that running speed is a variable that could improve batting average predictions (2020). Additionally,

Demmink demonstrated through a linear regression model that stolen base attempts positively impact number of wins at a significance level above 99% (2010). Similarly, Lee determined that stolen base attempts were also beneficial to scoring runs in the Korean baseball league (2011).

Bat handedness. In this study, bat handedness is represented by a dummy variable in which a value of zero represents a right-handed batter and a value of one represents a left-handed batter. Although players are talented regardless of their handedness and there are many very successful left-handed and right-handed batters, bat handedness may have an effect on batting averages and SLG due to the opposite hand advantage that occurs when a batter is facing a pitcher with opposing handedness. A study investigating the opposite hand advantage determined that there is an advantage for opposite handed batters for OPS, SLG, strikeouts, and walks (Chu et al., 2016). However, Chu et al. claimed that the skill cut-off point for left-handed batters is likely lower than right-handed batters because right-handed batters consistently perform better than left-handed batters in every statistic except walks when facing a same handed pitcher (2016). By using a fixed effects regression model, Chu et al. found that opposite hand advantage explains about 15% of the variability in OPS on average for left-handed hitters, but only 7% for right-handed (2016). This result leads to the conclusion that there should be more left-handed batters in the MLB because it would increase the frequency of opposite handed batting and therefore increase batter advantage and performance (Chu et al., 2016). Although this study is unable to explain why left-handed batters have a more significant opposite hand advantage, it does demonstrate that bat handedness does effect batting performance.

Batting order. In this study, the batting order dummy variable represents the position in the batting order at which a player has the most plate appearances in a single season. The batting order in the MLB is usually designed so that a team has the best chance to get runners on base by placing their best hitters early in the lineup. This means that players are often put into the batting order based upon their existing batting statistics. However, a specific position in the batting order may impact how well a batter will hit due to the specific situations they are faced with along with the performance of the players who bat after them. While many studies assume that a player's batting performance is independent of other players, Bradbury and Drinen demonstrated that the quality of the on-deck batter negatively impacts the preceding batter (2008). This effect was found to be very small with a change in one standard deviation from the mean on-deck batter OPS only changing the batter's batting average by about 0.0028 which is around 1% of the mean batting average, however, the effect is still significant (Bradbury & Drinen, 2008). Bradbury and Drinen claimed that this effect likely occurs because pitchers often change their behavior due to the player on-deck, meaning that the pitcher will try harder to get a player out or force them to hit weakly when a good hitter is going to hit next (2008). Another study looking at the effect of anxiety on batting performance in softball in critical situations, found that anxiety has an effect on batting performance especially in very critical versus non-critical situations (Krane et al., 1994). The study found that as the criticality of a situation rose, so did anxiety which has been shown to reduce strategic thinking and negatively impact performance (Krane et al., 1994). Although this study looks at softball as opposed to baseball, due to the similarity of the sports, it is likely that there is a similar effect of critical situation anxiety on batting

performance in baseball. If baseball players are in a certain position in the batting order that often faces highly critical situations, they may have increased anxiety and lower batting performance.

Pitching variables. Various pitching variables are included in this model due to the direct impact of pitching on batting performance. Bradbury and Drinen stated within their study that a player's batting performance is positively correlated with their own ability and negatively correlated with the pitcher's ability (2008). Rotating between different pitch types and pitch speeds has been shown to affect players' batting performance. Fortenbaugh et al. emphasized within their study the importance of shifting weight for a batter in ensuring correct timing and balance in their swing (2011). The study examined this weight shift against fastballs and changeups through maximum horizontal and vertical ground reaction forces of professional baseball players (Fortenbaugh et al., 2011). The results demonstrated that hitters shift their weight differently on different pitch types and pitch speeds and that changing pitch type and speed was able to disrupt the coordination of this weight shift for batters (Fortenbaugh et al., 2011). The in-zone percent variable is also included as it is more likely for a player to swing at a ball in the zone and therefore more of these balls are expected to be put in play. A study examining coordination of hitting movement found that trunk and arm movements differed as pitch location changed (Katsumata et al., 2017). The study also found that the time taken to hit the ball differed based on pitch location, with inside balls taking the most time to hit (Katsumata et al., 2017). These results demonstrate that timing adjustments are required to succeed at hitting the ball in different locations (Katsumata et

al., 2017) and therefore pitch location and in-zone percent may impact batting performance.

Division/League. In this study, division and league are represented by dummy variables. Although there usually are very successful hitters on all teams, division and league are included as variables in the model because they determine which teams are played the most and subsequently which pitchers and fielders are faced the most within a given season. It has been demonstrated that pitching ability impacts batting performance (Bradbury & Drinen, 2008) and therefore players who are in a division or league with better pitchers will likely have lower batting statistics. In addition, team ERA has been shown to significantly explain the variability in winning for a team, meaning the teams with the best (lowest) ERAs are also likely the most successful teams (Peach et al., 2016). Along similar lines, there may also be better fielding players in some divisions compared to others, which could impact the batting performance of opposing teams or certain team matchups that lead to higher or lower batting success.

Split models. Finally, the data from each season from 2016 to 2019 will be split into separate models. As the infield shift, in its current use, has been implemented consistently in the league for the past decade, it is possible that batters have begun to adapt to the defensive formation, which may mean each season will have unique effects from the percentage shifts variable. This is supported by the study performed by Peach et al., which demonstrated that predictions in one baseball season may be very different from another season and therefore the data from one specific season cannot necessarily be used to predict the next (2016).

Data and Results

Data was collected from baseballsavant.mlb.com and fangraphs.com, two reliable and up to date public data sources for the MLB. Controls were implemented within the dataset to make the model more precise. Data was only collected for players who had more than or equal to 100 plate appearances within a given season. This limitation was used to control for temporary players, such as pinch hitters or pinch runners, pitchers, who only hit in the National league, bench players or playoff callups, and players who were injured most of the season. Switch hitters were also omitted from the dataset as many of their statistics were shown cumulatively and were not split between their left-handed and right-handed performance. In total, data for 398 players were collected for the 2019 season, 390 for the 2018 season, 374 for the 2017 season, and 379 for the 2016 season. Descriptive statistics were collected for both models in each season to determine the means and standard errors of the variables, with dummy variables omitted (Tables 1-4).

Table 1. 2016 Season Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
batting_avg	379	.2537916	.0345302	.164	.348
SLG	379	.4140475	.075113	.197	.657
player_age	379	28.86807	3.93714	21	43
b_total_pa	379	394.3615	187.935	101	744
pitch_coun~d	379	173.8602	93.92341	23	452
pitch_coun~l	379	932.6491	447.0905	217	1914
pitch_coun~g	379	409.2823	205.6019	78	889
pitch_count	379	1534.536	737.8915	353	3014
sprint_speed	379	26.9942	1.460365	22.6	30.8
in_zone_pe~t	379	48.78839	3.00588	40.4	56.8
PitchMPH	379	88.75119	.5678954	86.9	90.4
Percentage~t	379	13.06966	19.30703	0	93.8

Table 2. 2017 Season Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
batting_avg	374	.2549091	.0353194	.144	.346
SLG	374	.4255455	.0789761	.203	.69
player_age	374	28.72995	3.851021	21	44
b_total_pa	374	398.7995	178.1855	102	725
pitch_coun~d	374	173.3182	91.03576	23	436
pitch_coun~l	374	937.2647	415.676	213	1930
pitch_coun~g	374	427.2888	206.2663	69	1002
pitch_count	374	1557.04	697.0639	382	2989
sprint_speed	374	27.08342	1.436201	21.9	30.5
in_zone_pe~t	374	49.00615	2.64583	42.1	57.6
PitchMPH	374	88.65642	.4993819	86.8	90.2
Percentage~t	374	11.21898	18.58313	0	93.8

Table 3. 2018 Season Descriptive Statistics

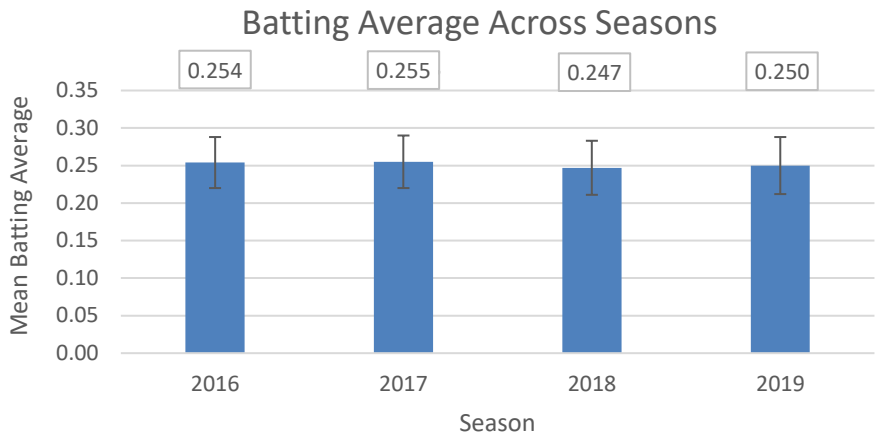
Variable	Obs	Mean	Std. Dev.	Min	Max
batting_avg	390	.2467769	.035955	.117	.346
SLG	390	.4055385	.077248	.165	.671
player_age	390	28.61026	3.717795	20	40
b_total_pa	390	381.0846	180.0348	101	740
pitch_coun~d	390	169.3538	91.58582	28	476
pitch_coun~l	390	887.1256	420.6109	209	1839
pitch_coun~g	390	418.7179	209.2016	84	945
pitch_count	390	1483.944	702.9834	361	2942
sprint_speed	390	27.07744	1.447595	22.2	30.2
in_zone_pe~t	390	49.03615	2.394435	41.5	56.2
PitchMPH	390	88.59795	.5048311	86.6	90.2
Perce~eShift	390	16.34795	20.59619	0	92.1

Table 4. 2019 Season Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
batting_avg	398	.2497965	.0375143	.124	.344
SLG	398	.4303065	.0862466	.144	.671
player_age	398	28.44724	3.684125	20	39
b_total_pa	398	381.1307	174.5384	101	747
pitch_coun~d	398	179.6709	95.32439	31	498
pitch_coun~l	398	870.505	401.4517	215	1907
pitch_coun~g	398	436.0503	209.1301	92	1029
pitch_count	398	1499.088	688.0538	368	3223
sprint_speed	398	27.01206	1.454466	22.2	30.8
in_zone_pe~t	398	47.75427	2.40895	39.8	56.7
PitchMPH	398	88.64799	.4604056	87.3	89.9
Percentage~s	398	24.78141	23.90565	0	95.9

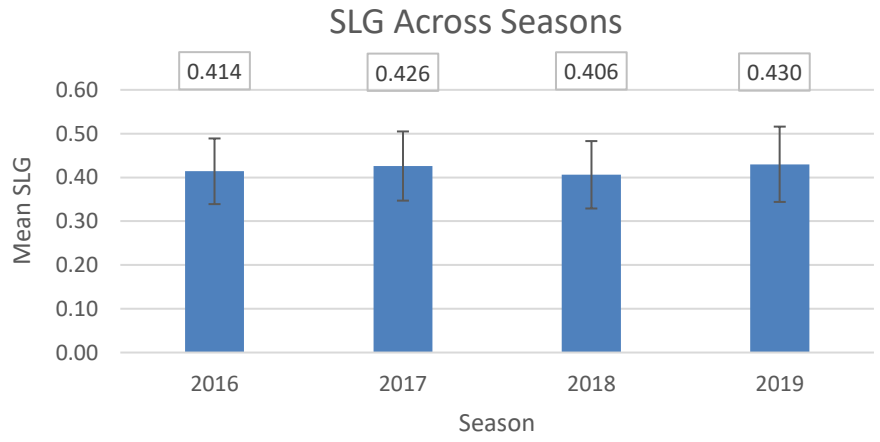
Bar graphs compared various statistics across the four seasons that were examined. These graphs were created to determine if the means for the two dependent variables, batting average and SLG, differed significantly from year to year as well as the means for the percent shift variable, as it is the variable at the basis of the research question.

Figure 2. Batting Average Means Across Seasons with Standard Deviation Error Bars



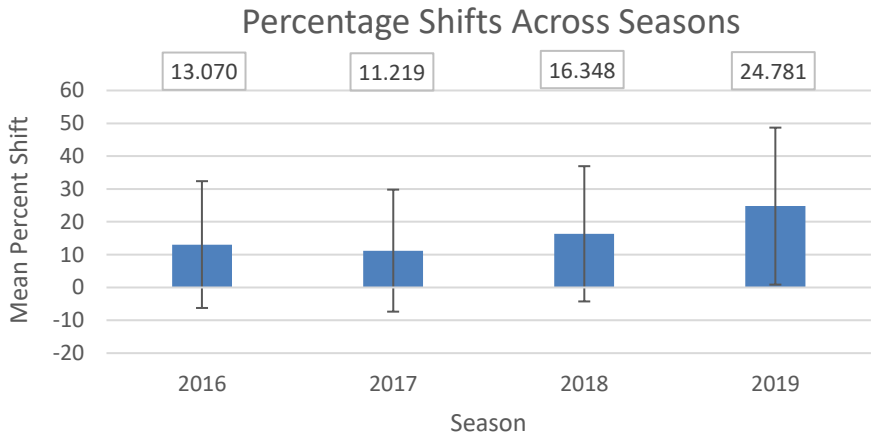
As demonstrated in the graph comparing batting average means across seasons, the mean player averages do not appear to vary significantly from year to year (Figure 2). The means are very similar and the standard deviation error bars overlap, likely reflecting that the means are not statistically significant from each other. In addition, there does not appear to be an increasing or decreasing pattern between the means from 2016 to 2019.

Figure 3. SLG Means Across Seasons with Standard Deviation Error Bars



Similar to batting average, as demonstrated in the graph comparing the mean SLG statistic across seasons, the mean player SLG does not appear to vary significantly from year to year (Figure 3). Like batting average, the means are similar and the standard deviation error bars also overlap. Similarly, there also does not appear to be an increasing or decreasing pattern across the seasons.

Figure 4. Percentage Shifts Means Across Seasons with Standard Deviation Error Bars



Unlike the graphs comparing batting average and SLG means across seasons, the means of the percentage shifts variable visually appear to vary across seasons (Figure 4). However, the standard deviation error bars are very large and all overlap, indicating that the means likely are not statistically significant from each other. Similar to batting average and SLG, there also does not appear to be an increasing or decreasing pattern of the means for the percentage shifts across seasons.

Ordinary least square regression (OLS) was performed for each model for the 2016 to 2019 seasons (Figures 5-12).

Figure 5. OLS 2016 Season Regression Results with SLG as the Dependent Variable

Source	SS	df	MS	Number of obs	=	379
				F(21, 357)	=	14.52
Model	.982543547	21	.046787788	Prob > F	=	0.0000
Residual	1.1501156	357	.003221612	R-squared	=	0.4607
				Adj R-squared	=	0.4290
Total	2.13265915	378	.005641955	Root MSE	=	.05676

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0006962	.0008989	-0.77	0.439	-.002464	.0010717
b_total_pa	-.0000859	.0001138	-0.75	0.451	-.0003098	.000138
pitch_count_fastball	-.0003526	.0001098	-3.21	0.001	-.0005685	-.0001367
pitch_count_breaking	-.000133	.0001176	-1.13	0.259	-.0003643	.0000983
pitch_count	.000307	.0000988	3.11	0.002	.0001128	.0005012
sprint_speed	.0009889	.0024948	0.40	0.692	-.0039174	.0058952
in_zone_percent	-.0024017	.0013563	-1.77	0.077	-.0050691	.0002656
PitchMPH	.00649	.0083421	0.78	0.437	-.0099159	.0228958
batting1	.0160187	.0139366	1.15	0.251	-.0113894	.0434268
batting2	.0251554	.0143741	1.75	0.081	-.0031132	.053424
batting3	.051366	.0163711	3.14	0.002	.01917	.0835619
batting4	.0282994	.0149659	1.89	0.059	-.0011331	.0577319
batting5	.0234143	.0139807	1.67	0.095	-.0040805	.0509091
batting6	.023176	.0133688	1.73	0.084	-.0031156	.0494676
batting7	-.0007342	.0129033	-0.06	0.955	-.0261103	.0246418
batting8	-.0241832	.012884	-1.88	0.061	-.0495213	.0011549
American	-.0112666	.0073865	-1.53	0.128	-.0257931	.0032598
East	-.0089528	.0072147	-1.24	0.215	-.0231415	.0052359
Central	.0006227	.0078363	0.08	0.937	-.0147884	.0160338
bat	-.0272224	.0092525	-2.94	0.003	-.0454186	-.0090261
PercentageShift	.0002746	.000224	1.23	0.221	-.000166	.0007151
_cons	-.1046895	.7443055	-0.14	0.888	-1.568464	1.359085

Figure 6. OLS 2016 Season Regression Results with Batting Average as the Dependent Variable

Source	SS	df	MS	Number of obs	=	379
				F(21, 357)	=	9.46
Model	.161138851	21	.007673279	Prob > F	=	0.0000
Residual	.289563682	357	.000811103	R-squared	=	0.3575
				Adj R-squared	=	0.3197
Total	.450702533	378	.001192335	Root MSE	=	.02848

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0003731	.000451	-0.83	0.409	-.0012601	.000514
b_total_pa	.000245	.0000571	4.29	0.000	.0001326	.0003573
pitch_count_fastball	-.0000448	.0000551	-0.81	0.416	-.0001531	.0000635
pitch_count_breaking	-.0000498	.000059	-0.84	0.399	-.0001659	.0000662
pitch_count	-5.83e-06	.0000496	-0.12	0.906	-.0001033	.0000916
sprint_speed	.0004886	.0012518	0.39	0.697	-.0019732	.0029504
in_zone_percent	.0000731	.0006805	0.11	0.915	-.0012653	.0014115
PitchMPH	.0048157	.0041858	1.15	0.251	-.0034162	.0130476
batting1	.0178084	.0069929	2.55	0.011	.0040559	.0315609
batting2	.0193346	.0072125	2.68	0.008	.0051503	.0335188
batting3	.0308039	.0082145	3.75	0.000	.0146491	.0469587
batting4	.013454	.0075094	1.79	0.074	-.0013143	.0282222
batting5	.0094543	.007015	1.35	0.179	-.0043417	.0232503
batting6	.0111599	.006708	1.66	0.097	-.0020323	.0243521
batting7	.0032614	.0064744	0.50	0.615	-.0094715	.0159942
batting8	-.0029414	.0064648	-0.45	0.649	-.0156552	.0097724
American	-.0041646	.0037063	-1.12	0.262	-.0114535	.0031243
East	-.0022702	.0036201	-0.63	0.531	-.0093896	.0048493
Central	-.0011587	.003932	-0.29	0.768	-.0088914	.0065741
bat	.0004404	.0046426	0.09	0.924	-.0086899	.0095706
PercentageShift	-.000356	.0001124	-3.17	0.002	-.000577	-.0001349
_cons	-.2081475	.3734672	-0.56	0.578	-.9426198	.5263249

Figure 7. OLS 2017 Season Regression Results with SLG as the Dependent Variable

Source	SS	df	MS	Number of obs	=	374
				F(21, 352)	=	12.98
Model	1.01542152	21	.048353406	Prob > F	=	0.0000
Residual	1.31106121	352	.003724606	R-squared	=	0.4365
				Adj R-squared	=	0.4028
Total	2.32648273	373	.006237219	Root MSE	=	.06103

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0022435	.0009762	-2.30	0.022	-.0041633	-.0003236
b_total_pa	-.0000178	.0001305	-0.14	0.892	-.0002744	.0002388
pitch_count_fastball	-.0000438	.0001071	-0.41	0.683	-.0002544	.0001668
pitch_count_breaking	.0000424	.0001124	0.38	0.706	-.0001787	.0002635
pitch_count	.0000485	.0000955	0.51	0.612	-.0001394	.0002363
sprint_speed	-.0022601	.0027227	-0.83	0.407	-.007615	.0030948
in_zone_percent	-.004827	.0014781	-3.27	0.001	-.0077341	-.0019199
PitchMPH	.0014197	.0094091	0.15	0.880	-.0170855	.0199249
batting1	.0342448	.014835	2.31	0.022	.0050684	.0634212
batting2	.0452315	.0142481	3.17	0.002	.0172093	.0732537
batting3	.0551502	.0160869	3.43	0.001	.0235117	.0867887
batting4	.0594202	.0162625	3.65	0.000	.0274364	.091404
batting5	.0317407	.0149608	2.12	0.035	.0023168	.0611646
batting6	.0235141	.0146738	1.60	0.110	-.0053453	.0523735
batting7	.0052442	.0136989	0.38	0.702	-.0216977	.0321861
batting8	-.0169015	.0137074	-1.23	0.218	-.0438603	.0100572
American	-.0222675	.0071064	-3.13	0.002	-.0362439	-.0082912
East	-.001561	.0078619	-0.20	0.843	-.0170232	.0139012
Central	-.0063361	.0079108	-0.80	0.424	-.0218944	.0092223
bat	-.0172518	.0098254	-1.76	0.080	-.0365757	.0020721
PercentageShift	.0006586	.0002351	2.80	0.005	.0001962	.001121
_cons	.6040805	.8433543	0.72	0.474	-1.054566	2.262727

Figure 8. OLS 2017 Season Regression Results with Batting Average as the Dependent Variable

Source	SS	df	MS	Number of obs	=	374
				F(21, 352)	=	11.49
Model	.189192632	21	.009009173	Prob > F	=	0.0000
Residual	.276110277	352	.000784404	R-squared	=	0.4066
				Adj R-squared	=	0.3712
Total	.465302909	373	.001247461	Root MSE	=	.02801

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0000963	.000448	-0.22	0.830	-.0009774	.0007847
b_total_pa	.0003554	.0000599	5.94	0.000	.0002377	.0004732
pitch_count_fastball	-.0000209	.0000491	-0.43	0.670	-.0001176	.0000757
pitch_count_breaking	-.0000223	.0000516	-0.43	0.666	-.0001238	.0000792
pitch_count	-.0000564	.0000438	-1.29	0.199	-.0001425	.0000298
sprint_speed	-.0001125	.0012495	-0.09	0.928	-.0025699	.002345
in_zone_percent	-.0000239	.0006783	-0.04	0.972	-.001358	.0013102
PitchMPH	.0065801	.004318	1.52	0.128	-.0019122	.0150724
batting1	.0191491	.006808	2.81	0.005	.0057597	.0325385
batting2	.0280336	.0065386	4.29	0.000	.0151738	.0408933
batting3	.0287987	.0073825	3.90	0.000	.0142794	.043318
batting4	.0229657	.007463	3.08	0.002	.008288	.0376435
batting5	.013331	.0068657	1.94	0.053	-.000172	.026834
batting6	.0064183	.006734	0.95	0.341	-.0068257	.0196622
batting7	.0013669	.0062866	0.22	0.828	-.0109971	.0137309
batting8	-.0016813	.0062905	-0.27	0.789	-.014053	.0106904
American	-.0054802	.0032612	-1.68	0.094	-.0118941	.0009338
East	.0039672	.0036079	1.10	0.272	-.0031286	.011063
Central	.0020399	.0036304	0.56	0.575	-.0051	.0091799
bat	.0068141	.004509	1.51	0.132	-.0020539	.0156821
PercentageShift	-.0004629	.0001079	-4.29	0.000	-.0006751	-.0002507
_cons	-.3556102	.3870257	-0.92	0.359	-1.116784	.4055635

Figure 9. OLS 2018 Season Regression Results with SLG as the Dependent Variable

Source	SS	df	MS	Number of obs	=	390
				F(21, 368)	=	14.43
Model	1.0480989	21	.049909471	Prob > F	=	0.0000
Residual	1.27316003	368	.003459674	R-squared	=	0.4515
				Adj R-squared	=	0.4202
Total	2.32125892	389	.005967247	Root MSE	=	.05882

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0028296	.0009615	-2.94	0.003	-.0047204	-.0009388
b_total_pa	-.0001231	.0001281	-0.96	0.337	-.0003751	.0001288
pitch_count_fastball	-.000229	.0001169	-1.96	0.051	-.000459	8.98e-07
pitch_count_breaking	-.0001275	.0001137	-1.12	0.263	-.0003512	.0000961
pitch_count	.0002356	.0001011	2.33	0.020	.0000368	.0004345
sprint_speed	-.0015396	.0025075	-0.61	0.540	-.0064704	.0033913
in_zone_percent	-.0046784	.0015349	-3.05	0.002	-.0076967	-.00166
PitchMPH	.0128158	.0082934	1.55	0.123	-.0034925	.0291241
batting1	.0537722	.0143036	3.76	0.000	.0256452	.0818992
batting2	.0812748	.0147252	5.52	0.000	.0523187	.1102308
batting3	.0779076	.0156367	4.98	0.000	.0471591	.108656
batting4	.0834609	.0154366	5.41	0.000	.0531059	.1138159
batting5	.06106	.0142995	4.27	0.000	.032941	.0891791
batting6	.0416382	.0142697	2.92	0.004	.0135778	.0696987
batting7	.0407479	.0135386	3.01	0.003	.0141251	.0673706
batting8	.0089238	.01285	0.69	0.488	-.0163448	.0341924
American	.0051029	.006446	0.79	0.429	-.0075728	.0177785
East	-.0072794	.0079381	-0.92	0.360	-.0228891	.0083304
Central	.0000765	.0075257	0.01	0.992	-.0147223	.0148752
bat	-.0195841	.0101973	-1.92	0.056	-.0396364	.0004682
PercentageShift	.00008	.0002012	0.40	0.691	-.0003157	.0004756
_cons	-.4671083	.7372522	-0.63	0.527	-1.916864	.9826476

Figure 10. OLS 2018 Season Regression Results with Batting Average as the Dependent Variable

Source	SS	df	MS	Number of obs	=	390
				F(21, 368)	=	13.26
Model	.216617688	21	.010315128	Prob > F	=	0.0000
Residual	.286267904	368	.000777902	R-squared	=	0.4307
				Adj R-squared	=	0.3983
Total	.502885592	389	.001292765	Root MSE	=	.02789

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0005917	.0004559	-1.30	0.195	-.0014882	.0003049
b_total_pa	.0002483	.0000608	4.09	0.000	.0001288	.0003678
pitch_count_fastball	-.0000621	.0000554	-1.12	0.263	-.0001711	.0000469
pitch_count_breaking	-.0000785	.0000539	-1.46	0.146	-.0001845	.0000276
pitch_count	.0000125	.0000479	0.26	0.794	-.0000817	.0001068
sprint_speed	-.0020428	.001189	-1.72	0.087	-.0043809	.0002953
in_zone_percent	.0006109	.0007278	0.84	0.402	-.0008204	.0020421
PitchMPH	.0059757	.0039326	1.52	0.129	-.0017574	.0137089
batting1	.0294781	.0067825	4.35	0.000	.0161407	.0428154
batting2	.04344	.0069824	6.22	0.000	.0297096	.0571705
batting3	.0366256	.0074146	4.94	0.000	.0220453	.051206
batting4	.0394039	.0073198	5.38	0.000	.0250101	.0537977
batting5	.0297378	.0067806	4.39	0.000	.0164043	.0430714
batting6	.023186	.0067665	3.43	0.001	.0098802	.0364917
batting7	.0171064	.0064198	2.66	0.008	.0044823	.0297304
batting8	.0035837	.0060932	0.59	0.557	-.0083982	.0155656
American	.0038667	.0030566	1.27	0.207	-.0021439	.0098772
East	.0027921	.0037641	0.74	0.459	-.0046097	.010194
Central	.0053298	.0035685	1.49	0.136	-.0016875	.0123471
bat	.0034734	.0048354	0.72	0.473	-.006035	.0129819
PercentageShift	-.0005781	.0000954	-6.06	0.000	-.0007657	-.0003905
_cons	-.2861724	.3495916	-0.82	0.414	-.9736201	.4012754

Figure 11. OLS 2019 Season Regression Results with SLG as the Dependent Variable

Source	SS	df	MS	Number of obs	=	398
				F(21, 376)	=	12.78
Model	1.22984105	21	.05856386	Prob > F	=	0.0000
Residual	1.72323155	376	.004583063	R-squared	=	0.4165
				Adj R-squared	=	0.3839
Total	2.9530726	397	.00743847	Root MSE	=	.0677

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0009581	.001071	-0.89	0.372	-.003064	.0011478
b_total_pa	.0000342	.0001386	0.25	0.805	-.0002382	.0003067
pitch_count_fastball	-.0003158	.0001269	-2.49	0.013	-.0005654	-.0000663
pitch_count_breaking	-.0001246	.0001127	-1.11	0.270	-.0003462	.0000971
pitch_count	.0002489	.0001036	2.40	0.017	.0000452	.0004526
sprint_speed	.0044856	.0028185	1.59	0.112	-.0010565	.0100276
in_zone_percent	-.004565	.001771	-2.58	0.010	-.0080472	-.0010827
PitchMPH	.0263815	.0110474	2.39	0.017	.0046589	.048104
batting1	.0448949	.0163677	2.74	0.006	.0127112	.0770787
batting2	.0414377	.0169772	2.44	0.015	.0080556	.0748198
batting3	.0629118	.017727	3.55	0.000	.0280554	.0977682
batting4	.0462451	.0171764	2.69	0.007	.0124712	.0800189
batting5	.0495536	.0160398	3.09	0.002	.0180147	.0810925
batting6	.0349377	.0162093	2.16	0.032	.0030655	.0668099
batting7	.0001919	.0150619	0.01	0.990	-.0294242	.029808
batting8	-.00279	.0151902	-0.18	0.854	-.0326584	.0270784
American	.0084217	.0076232	1.10	0.270	-.0065677	.0234111
East	.0020779	.0085422	0.24	0.808	-.0147185	.0188743
Central	-.0081831	.0087655	-0.93	0.351	-.0254187	.0090524
bat	-.0317665	.0112963	-2.81	0.005	-.0539783	-.0095548
PercentageShifts	.0005481	.0002041	2.69	0.008	.0001468	.0009495
_cons	-1.874724	.9945019	-1.89	0.060	-3.830207	.080758

Figure 12. OLS 2019 Season Regression Results with Batting Average as the Dependent Variable

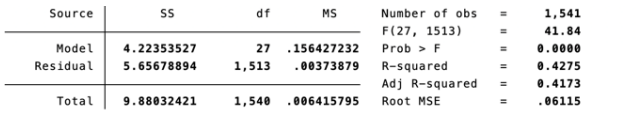
Source	SS	df	MS	Number of obs	=	398
				F(21, 376)	=	11.79
Model	.221869774	21	.010565227	Prob > F	=	0.0000
Residual	.336838741	376	.000895848	R-squared	=	0.3971
				Adj R-squared	=	0.3634
Total	.558708515	397	.001407326	Root MSE	=	.02993

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0004394	.0004735	-0.93	0.354	-.0013704	.0004917
b_total_pa	.0002665	.0000613	4.35	0.000	.000146	.0003869
pitch_count_fastball	-.0000871	.0000561	-1.55	0.121	-.0001974	.0000232
pitch_count_breaking	-.0000566	.0000498	-1.14	0.257	-.0001546	.0000414
pitch_count	.0000174	.0000458	0.38	0.704	-.0000726	.0001075
sprint_speed	.0010037	.0012461	0.81	0.421	-.0014465	.003454
in_zone_percent	-.0003426	.000783	-0.44	0.662	-.0018821	.001197
PitchMPH	.0102848	.0048843	2.11	0.036	.0006809	.0198887
batting1	.020849	.0072365	2.88	0.004	.00662	.0350781
batting2	.0186981	.0075059	2.49	0.013	.0039392	.033457
batting3	.0261599	.0078374	3.34	0.001	.0107492	.0415706
batting4	.0164891	.007594	2.17	0.031	.0015571	.0314212
batting5	.0194127	.0070915	2.74	0.006	.0054688	.0333566
batting6	.01087	.0071664	1.52	0.130	-.0032212	.0249613
batting7	-.0012794	.0066592	-0.19	0.848	-.0143732	.0118145
batting8	-.0057371	.0067159	-0.85	0.394	-.0189425	.0074683
American	.0021264	.0033703	0.63	0.528	-.0045007	.0087535
East	.0016262	.0037767	0.43	0.667	-.0057998	.0090522
Central	-.0034243	.0038754	-0.88	0.377	-.0110445	.0041958
bat	.0045253	.0049943	0.91	0.365	-.005295	.0143455
PercentageShifts	-.000462	.0000902	-5.12	0.000	-.0006394	-.0002846
_cons	-.6892487	.4396882	-1.57	0.118	-1.553805	.1753073

A RESET test (Ramsey, 1969) was performed for each model to test for omitted variable bias and specification errors within the models (Figures A1-A8). All the p-values from each model in four seasons were above 0.10 meaning that we can fail to reject the null hypothesis that there is no omitted variable bias within the models. A White test (White, 1980) was performed on each model to test for heteroskedasticity within the regression models (Figures A9-A16). Heteroskedasticity violates the classical assumption of ordinary least squares regression that the variances of the error term are constant. While it does not cause bias in the estimates, it may underestimate the standard errors and possibly make the model seem better than it actually is. All the p-values from each model in the four seasons were above 0.10 meaning that we can fail to reject the null hypothesis that there is no heteroskedasticity in the models.

The results from the bar graphs which demonstrated that there did not appear significant differences in the means for batting average, SLG, or percentage shifts across seasons led to a question of whether the data for each season should be analyzed separately or together. Two tests were performed to determine whether the data should be kept in separate models divided by year or if they could be incorporated together. The first test determined if the percentage shifts variable differed year to year by creating interaction terms. Ordinary least squares regression was run on the pooled data including year dummy variables and interaction variables between each year dummy and the percentage shifts variable. The results found that the three interaction terms (the 2016 season was omitted due to redundancy) all had high p values in both the SLG and batting average models, demonstrating that the percentage shifts variable did not vary significantly from year to year (Figures 13 and 14).

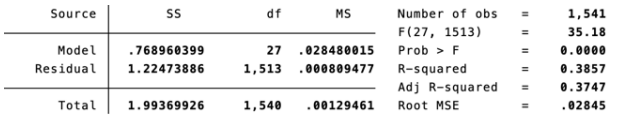
Figure 13. OLS Pooled Model Regression Results Including Interaction Terms with SLG as the Dependent Variable



Source	SS	df	MS	Number of obs	=	1,541
				F(27, 1513)	=	41.84
Model	4.22353527	27	.156427232	Prob > F	=	0.0000
Residual	5.65678894	1,513	.00373879	R-squared	=	0.4275
				Adj R-squared	=	0.4173
Total	9.88032421	1,540	.006415795	Root MSE	=	.06115

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0018172	.0004821	-3.77	0.000	-.0027628	-.0008715
b_total_pa	-.0000469	.0000629	-0.75	0.456	-.0001702	.0000765
pitch_count_fastball	-.0002331	.0000544	-4.29	0.000	-.0003398	-.0001265
pitch_count_breaking	-.0000967	.0000539	-1.79	0.073	-.0002024	8.99e-06
pitch_count	.0002117	.0000474	4.47	0.000	.0001188	.0003046
sprint_speed	.0007134	.0013014	0.55	0.584	-.0018393	.0032662
in_zone_percent	-.0040449	.0007291	-5.55	0.000	-.0054751	-.0026147
PitchMPH	.0120216	.0043398	2.77	0.006	.0035089	.0205343
batting1	.0364598	.0073758	4.94	0.000	.021992	.0509276
batting2	.0470162	.0074749	6.29	0.000	.032354	.0616785
batting3	.0610145	.0081302	7.50	0.000	.0450668	.0769622
batting4	.0547635	.007895	6.94	0.000	.0392772	.0702497
batting5	.0426718	.0073421	5.81	0.000	.0282701	.0570735
batting6	.0310982	.0072614	4.28	0.000	.0168547	.0453418
batting7	.0108678	.0068461	1.59	0.113	-.0025611	.0242967
batting8	-.0090879	.0067619	-1.34	0.179	-.0223515	.0041758
American	-.0031235	.0034183	-0.91	0.361	-.0098286	.0035817
East	-.0039444	.0038625	-1.02	0.307	-.0115209	.0036321
Central	-.0040078	.0039249	-1.02	0.307	-.0117066	.0036909
bat	-.0242798	.0049064	-4.95	0.000	-.0339038	-.0146558
PercentageShifts	.0003014	.0001876	1.61	0.108	-.0000666	.0006693
y2	.0088471	.0053373	1.66	0.098	-.0016221	.0193163
y3	-.0061058	.0055863	-1.09	0.275	-.0170636	.004852
y4	.0003469	.006032	0.06	0.954	-.0114851	.0121789
interactiony2	.0002798	.0002384	1.17	0.241	-.0001878	.0007474
interactiony3	-.0000251	.0002258	-0.11	0.911	-.000468	.0004178
interactiony4	.0001645	.0002145	0.77	0.443	-.0002562	.0005853
_cons	-.4918717	.387955	-1.27	0.205	-1.252858	.2691149

Figure 14. OLS Pooled Model Regression Results Including Interaction Terms with Batting Average as the Dependent Variable



Source	SS	df	MS	Number of obs	=	1,541
				F(27, 1513)	=	35.18
Model	.768960399	27	.028480015	Prob > F	=	0.0000
Residual	1.22473886	1,513	.000809477	R-squared	=	0.3857
				Adj R-squared	=	0.3747
Total	1.99369926	1,540	.00129461	Root MSE	=	.02845

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0004265	.0002243	-1.90	0.057	-.0008665	.0000136
b_total_pa	.0002763	.0000293	9.44	0.000	.0002189	.0003337
pitch_count_fastball	-.0000492	.0000253	-1.94	0.052	-.0000988	4.24e-07
pitch_count_breaking	-.0000509	.0000251	-2.03	0.042	-.0001001	-1.74e-06
pitch_count	-.0000104	.000022	-0.47	0.636	-.0000537	.0000328
sprint_speed	-.0000105	.0006055	-0.02	0.986	-.0011983	.0011773
in_zone_percent	.0000409	.0003393	0.12	0.904	-.0006246	.0007064
PitchMPH	.0070286	.0020193	3.48	0.001	.0030676	.0109896
batting1	.0214952	.003432	6.26	0.000	.0147633	.0282271
batting2	.0273081	.0034781	7.85	0.000	.0204857	.0341305
batting3	.0301869	.003783	7.98	0.000	.0227663	.0376074
batting4	.0233566	.0036736	6.36	0.000	.0161508	.0305625
batting5	.0179478	.0034163	5.25	0.000	.0112466	.024649
batting6	.0131341	.0033788	3.89	0.000	.0065065	.0197617
batting7	.0051407	.0031855	1.61	0.107	-.0011079	.0113892
batting8	-.0022549	.0031463	-0.72	0.474	-.0084265	.0039167
American	-.0007188	.0015906	-0.45	0.651	-.0038387	.0024011
East	.0015895	.0017972	0.88	0.377	-.0019359	.0051148
Central	.0006335	.0018263	0.35	0.729	-.0029488	.0042158
bat	.0036657	.002283	1.61	0.109	-.0008124	.0081438
PercentageShifts	-.0004124	.0000873	-4.73	0.000	-.0005837	-.0002412
y2	.0022172	.0024834	0.89	0.372	-.0026542	.0070885
y3	-.0020137	.0025993	-0.77	0.439	-.0071124	.003085
y4	.0033813	.0028067	1.20	0.229	-.0021242	.0088867
interactiony2	-.000101	.0001109	-0.91	0.363	-.0003186	.0001166
interactiony3	-.000071	.0001051	-0.68	0.499	-.0002771	.000135
interactiony4	-9.07e-06	.0000998	-0.09	0.928	-.0002048	.0001867
_cons	-.3966168	.180517	-2.20	0.028	-.7507069	-.0425268

The second test used was a Chow test which was performed to determine if the separate season models were statistically significant or confirm that all the data could be pooled together into an inclusive model (Chow, 1960).

Figure 15. OLS Pooled Model Regression Results with SLG as the Dependent Variable

Source	SS	df	MS	Number of obs	=	1,541
				F(21, 1519)	=	52.26
Model	4.14441705	21	.197353193	Prob > F	=	0.0000
Residual	5.73590715	1,519	.003776107	R-squared	=	0.4195
				Adj R-squared	=	0.4114
Total	9.88032421	1,540	.006415795	Root MSE	=	.06145

SLG	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0018041	.0004838	-3.73	0.000	-.0027531	-.000855
b_total_pa	-.0000584	.0000631	-0.93	0.354	-.0001822	.0000653
pitch_count_fastball	-.0002541	.0000541	-4.70	0.000	-.0003602	-.000148
pitch_count_breaking	-.0001132	.0000534	-2.12	0.034	-.0002179	-8.50e-06
pitch_count	.0002324	.0000472	4.93	0.000	.0001399	.0003249
sprint_speed	.0006804	.0013057	0.52	0.602	-.0018808	.0032416
in_zone_percent	-.0040321	.0007142	-5.65	0.000	-.005433	-.0026311
PitchMPH	.0128063	.0042392	3.02	0.003	.004491	.0211215
batting1	.0356363	.0073982	4.82	0.000	.0211246	.050148
batting2	.0467008	.0075032	6.22	0.000	.031983	.0614185
batting3	.0601657	.0081582	7.37	0.000	.0441632	.0761681
batting4	.0532649	.00791	6.73	0.000	.0377493	.0687806
batting5	.0423867	.0073592	5.76	0.000	.0279513	.056822
batting6	.0308015	.0072899	4.23	0.000	.016502	.0451009
batting7	.0105563	.0068781	1.53	0.125	-.0029352	.0240479
batting8	-.0094432	.0067913	-1.39	0.165	-.0227644	.0038781
American	-.0029968	.0034296	-0.87	0.382	-.0097242	.0037305
East	-.0039192	.0038779	-1.01	0.312	-.0115259	.0036875
Central	-.0041514	.0039402	-1.05	0.292	-.0118802	.0035773
bat	-.0244363	.0049065	-4.98	0.000	-.0340606	-.0148121
PercentageShifts	.0003664	.0001025	3.57	0.000	.0001654	.0005675
_cons	-.5606754	.3786879	-1.48	0.139	-1.303482	.1821311

Figure 16. OLS Pooled Model Regression Results with Batting Average as the Dependent Variable

Source	SS	df	MS	Number of obs	=	1,541
				F(21, 1519)	=	44.52
Model	.759586707	21	.036170796	Prob > F	=	0.0000
Residual	1.23411256	1,519	.000812451	R-squared	=	0.3810
				Adj R-squared	=	0.3724
Total	1.99369926	1,540	.00129461	Root MSE	=	.0285

batting_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
player_age	-.0004228	.0002244	-1.88	0.060	-.000863	.0000175
b_total_pa	.0002745	.0000293	9.38	0.000	.0002171	.0003319
pitch_count_fastball	-.0000599	.0000251	-2.39	0.017	-.0001091	-.0000107
pitch_count_breaking	-.0000542	.0000248	-2.19	0.029	-.0001027	-5.59e-06
pitch_count	-2.85e-06	.0000219	-0.13	0.896	-.0000458	.0000401
sprint_speed	8.36e-06	.0006057	0.01	0.989	-.0011796	.0011964
in_zone_percent	-.0000911	.0003313	-0.28	0.783	-.000741	.0005587
PitchMPH	.0080375	.0019663	4.09	0.000	.0041805	.0118945
batting1	.0213442	.0034316	6.22	0.000	.014613	.0280754
batting2	.0269422	.0034804	7.74	0.000	.0201154	.033769
batting3	.0296201	.0037841	7.83	0.000	.0221974	.0370428
batting4	.022647	.003669	6.17	0.000	.0154501	.029844
batting5	.0174161	.0034136	5.10	0.000	.0107203	.0241119
batting6	.0127478	.0033814	3.77	0.000	.006115	.0193806
batting7	.0049915	.0031904	1.56	0.118	-.0012665	.0112496
batting8	-.0024058	.0031501	-0.76	0.445	-.0085848	.0037733
American	-.0006369	.0015908	-0.40	0.689	-.0037574	.0024836
East	.0014021	.0017988	0.78	0.436	-.0021262	.0049305
Central	.0003761	.0018276	0.21	0.837	-.0032088	.0039611
bat	.0031299	.0022759	1.38	0.169	-.0013343	.0075941
PercentageShifts	-.0004418	.0000475	-9.29	0.000	-.0005351	-.0003486
_cons	-.4786476	.1756539	-2.72	0.007	-.8231975	-.1340978

The residual sum of squares was collected from each separate model (Figures 5-12) and from the pooled models (Figures 15 and 16) and used within the Chow test equation for four separate models:

$$F = \frac{\frac{RSS_p - (RSS_1 + RSS_2 + RSS_3 + RSS_4)}{3(k+1)}}{\frac{(RSS_1 + RSS_2 + RSS_3 + RSS_4)}{(N_1 + N_2 + N_3 + N_4) - \big(4(k+1)\big)}} \qquad \text{Equation 2}$$

Within this equation, RSS_p denotes the residual sum of squares from the pooled seasons model while RSS_1, RSS_2, RSS_3, and RSS_4 denote the residual sum of squares from each separate season model. N_1, N_2, N_3, and N_4 denote the number of observations from the four separate season models. And finally, k denotes the number of independent variables within the model. This test was performed for both the batting average models and SLG models. The F-statistic result for the batting average models was approximately 0.840 and the result for the SLG models was approximately 1.123. The numerator degrees of freedom for this model is equal to $3(k + 1)$ which in this case equals 66 and the denominator degrees of freedom is equal to $((N_1 + N_2 + N_3 + N_4) - (4(k + 1))$ which in this case equals 1453. The F-statistic values found are both below the F-distribution critical value at the 5% level for the given degrees of freedom and therefore the null hypothesis, which assumes that the models are not statistically significant from each other, fails to be rejected. These results lead to two finalized models for analysis, one for the combined seasons batting average and one for the combined seasons SLG. Descriptive statistics were collected for these pooled data models with dummy variables omitted (Table 5).

Table 5. Pooled Model Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
batting_avg	1,541	.2512557	.0359807	.117	.348
SLG	1,541	.4188838	.0800987	.144	.69
player_age	1,541	28.66061	3.79623	20	44
b_total_pa	1,541	388.6613	180.1728	101	747
pitch_coun~d	1,541	174.0889	92.99333	23	498
pitch_coun~l	1,541	906.1979	421.8728	209	1930
pitch_coun~g	1,541	422.9539	207.6293	69	1029
pitch_count	1,541	1518.038	706.4216	353	3223
sprint_speed	1,541	27.04153	1.448887	21.9	30.8
in_zone_pe~t	1,541	48.63686	2.671699	39.8	57.6
PitchMPH	1,541	88.66275	.5115052	86.6	90.4
Percentage~s	1,541	16.47502	21.37785	0	95.9

Discussion

The r-squared value for the SLG model was approximately 0.420, meaning that the variables in the model were able to explain 42% of the variability in SLG. The r-squared value for the batting average model was approximately 0.382, meaning that the variables in the model were able to explain 38.2% of the variability in batting average. The independent variables that were found to be statistically significant to the 90% level in the batting average model were player age, number of plate appearances, number of fastballs, number of breaking balls, pitch speed, batting positions first through sixth, and percent of plate appearances facing a shift. The independent variables that were found to be statistically significant to the 90% level in the SLG model were player age, number of fastballs, number of breaking balls, total number of pitches, percent of pitches in the strike zone, pitch speed, batting positions first through sixth, bat handedness, and percent of plate appearances facing a shift. The variables that were not statistically significant in either model were sprint speed, the seventh and eighth batting positions, league, and division. Sprint speed likely does not have a significant effect on either batting average or

slugging percentage because most players in the MLB have similar sprint speeds, which is demonstrated by the descriptive statistics showing the standard deviation for sprint speed to be only approximately 1.449 feet per second. Since the bases in an MLB field are 90 feet apart, a 1.449 foot difference per second would likely not create a major difference in batting success for most players. Because the batting order variables are dummy variables with the ninth batting position omitted, the results of the regression demonstrate that, unlike other batting positions, the seventh and eighth position in the batting order are not statistically significant in their effect on batting success in comparison the ninth position. This result is unsurprising as batting orders are often created so that the best performing players hit earlier in the order and the less performing players, who would likely have similar batting statistics, hit at the end. The division and league dummy variables not being statistically significant is also not very surprising, as mentioned previously, because it would be expected that all divisions and leagues have a mix of successful and less successful batters. However, this may also mean that the ERAs in individual leagues or divisions are less impactful on batting success than expected.

The direction of the beta estimates demonstrates the relationship between the independent variables and the dependent variables. Notably, the beta estimate directions for the percentage shifts variable are opposite for the batting average model and SLG model, with a negative estimate in the batting average model and a positive estimate in the SLG model. This result answers the core research question, demonstrating that the implementation of the shift is successful at decreasing batting average, but consequently also increases SLG.

Beyond examining the directional effect of the independent variables, the magnitudes of the estimates were evaluated to determine the amount of impact each variable has on batting average and SLG. Since the variables are recorded in many different types of units, the standard deviations for the continuous variables were investigated to determine their effect on the dependent variables. Because standard deviations do not have distinct units, they can be compared across variables that use different units of measure. The standard deviations for the variables were multiplied by their corresponding beta estimates and these values were then analyzed (Table 6).

Table 6. Continuous Independent Variable Standard Deviations, Beta Estimates, and Effect Magnitudes

Variable	Standard Deviation	Batting Average Estimate	Batting Average Effect	SLG Estimate	SLG Effect
Player Age	3.796	-0.0004228	-0.001604949	-0.0018041	-0.006848364
Plate Appearances	180.173	0.0002745	0.049457489	-0.0000584	-0.010522103
Fastballs	421.873	-0.0000599	-0.025270193	-0.0002541	-0.107197929
Breaking Balls	207.63	-0.0000542	-0.011253546	-0.0001132	-0.023503716
Total Pitches	706.422	-0.00000285	-0.002013303	0.0002324	0.164172473
Sprint Speed	27.042	0.00000836	0.000226071	0.0006804	0.018399377
In Zone Percent	2.672	-0.0000911	-0.000243419	-0.0040321	-0.010773771
Pitch Speed	0.512	0.0080375	0.0041152	0.0128063	0.006556826
Percentage Shifts	21.378	-0.0004418	-0.0094448	0.0003664	0.007832899

The results of this comparison demonstrate that a one standard deviation increase in the percentage shifts variable leads to a decrease of about 0.009 or 25% of a standard deviation in batting average and an increase of about 0.008 or 10% of a standard deviation in SLG. This means that a one standard deviation increase in the percent of a player's plate appearances facing a shift has a greater effect on batting average than on SLG. This trade off favoring the impact on batting average may explain why the shift is still valued and implemented today even though it likely leads to an increase in SLG. Although the infield shift is a significantly impactful variable on both batting average and SLG, there are other variables for which a one standard deviation increase has a greater

impact, specifically the number plate appearances and the total number of pitches. The number of plate appearances resulted in the greatest effect on batting average with an increase of one standard deviation resulting in an increase of about 0.049 or 136% of a standard deviation in batting average. This result supports the hypothesis that as a player has more plate appearances, they gain experience and skill throughout the season and therefore would increase their batting average. The total number of pitches resulted in the greatest effect on SLG with an increase of one standard deviation of pitches resulting in an increase of about 0.164 or 205% of a standard deviation in SLG. This result may mean that players who face more pitches are more selective on what pitches they choose to hit and therefore are more likely to get better hits on the pitches they choose to hit. When examining the effects of standard deviation changes in the independent variables, it is apparent that the pitch type and total pitch count have a greater effect on SLG than they do on batting average. This result makes sense as the pitch type will likely matter more in terms of getting an extra base hit than getting a hit in general.

Conclusion

From the regression model results it can be concluded that the infield shift has a significant impact on both batting average and SLG, with a negative effect on batting average and a positive effect on SLG. The resulting negative effect on batting average was found to be greater than the positive effect on SLG, meaning that the total effect on batting performance is likely negative. These results suggest that teams should continue to use the infield shift to decrease their opponents' batting success. Although there are other variables that have a greater impact on batting average and SLG, indicating that the

shift cannot completely overcome the ability of a batter, it is a successful way for defenses to decrease their opponents' chances. This conclusion is important because batting average and SLG can impact both the number of runs a team scores and the number of games that they win. Because the total effect of the infield shift on batting success appears to be negative, implementing the shift would also presumably decrease the number of runs an opposing team scores and decrease the chances of the opposing team to win the game. A future study determining the exact effect of batting average and SLG on number of wins would be required to confirm this assumption. As demonstrated by the previous studies analyzed, winning is important for a franchise because it increases fan attendance and team revenues, which can lead to more success in the future.

Winning has also been shown to positively impact housing values, income, and general well-being in the area in which the team is located. Team success typically leads to increased team spending, including building new stadiums. A study using hedonic model analyses compared the pricing of single-family homes in the immediate area around FedEx Field with comparable homes further from the field to determine the effect of a sports stadium on housing value (Tu, 2005). The study found that, in this case, the construction of a new stadium improved housing values in the area in close proximity to the field (Tu, 2005). Beyond housing values, a cross-section time-series analysis in another study determined the effect of the construction of NFL and MLB stadiums on income (Santo, 2016). The results of the study demonstrated that in some cases stadiums have positive effects on income in the local area, but that the context of the national economic conditions matter (Santo, 2016). In addition to impacting economic prosperity, successful teams could also have an impact on fan wellbeing. When examining the

effects of emotional shocks of college football teams' wins and losses on the wellbeing of local population, Janhuba found that unexpected wins positively affect the life satisfaction of local citizens (2019). It was also demonstrated that this effect increases with stadium size relative to the population, which suggests that the number of fans sharing the same experience increases the effect of the experience (Janhuba, 2019). These findings illustrate that the knock-on effect of using the infield shift may be more impactful than just altering batting success and have an influence on not just players' incomes and lives, but also on the lives of those around them.

While the infield shift study results show significant impact of the shift on batting average and slugging, further controls could be implemented to make the regression models more precise. One possible control could be to examine the effects of the shift in specific base runner situations. This could help improve the precision of the beta estimate for the variables in the models as different situations with runners on base cause changes in batting behavior and, subsequently, batting performance. Another possible control could be to limit the data to only include regular season data. Strategies in post-season baseball can often be different than those used in the regular season, which also may influence the beta estimates within the models. Additionally, using more seasons of data would be able to increase the sample size of the study in order to make the overall model more accurate. However, because the recording mechanics of Statcast have only been used in the MLB since 2015, this would not be possible until a future time.

While this study is able to demonstrate the aggregate effect of the infield shift on batters' performance, it is not able to demonstrate the efficiency of the shift on individual batters. It may be the case that the implementation of the shift is much more successful

against some players than others, which could potentially be the focus of future research. In addition, this study has focused on the changes in batting behavior with the implementation of the infield shift and another future study could be performed to determine how possible pitcher behavior changes the shift's impact. Pitchers may attempt to throw the ball in specific locations so that hitters are more likely to hit into the shifted fielders. Some pitchers may be more successful at this than others and, even if the shift is successful against a specific batter, it may be detrimental for a given pitcher facing that batter. Along similar lines, specific fielders may perform better with the implementation of the shift than others and, therefore, using the shift may be more advisable with certain fielders than others. Future studies on these topics could potentially provide a deeper understanding for the real-world application of the infield shift. Another interesting finding that arose during the regression modeling was the beta estimate directions for the various pitch types. The regression results demonstrated that both the fastball pitches variable and the breaking ball pitches variable had negative estimates within the models, indicating that in comparison to off-speed pitches, both fastballs and breaking balls lead to lower batting performance. This surprising finding could be very important for pitchers' performance and should be investigated further in a future study to determine why this occurs. Although there are many possibilities for future research, this study has been able to show some effects of the infield shift and sets the groundwork for further research on the topic.

Additional Figures

Figure A1. 2016 Season RESET Test on SLG Model

```
Ramsey RESET test using powers of the fitted values of SLG
       Ho:  model has no omitted variables
                 F(3, 354) =      0.57
                  Prob > F =      0.6336
```

Figure A2. 2016 Season RESET Test on Batting Average Model

```
Ramsey RESET test using powers of the fitted values of batting_avg
       Ho:  model has no omitted variables
                 F(3, 354) =      0.40
                  Prob > F =      0.7512
```

Figure A3. 2017 Season RESET Test on SLG Model

```
Ramsey RESET test using powers of the fitted values of SLG
       Ho:  model has no omitted variables
                 F(3, 349) =      1.24
                  Prob > F =      0.2958
```

Figure A4. 2017 Season RESET Test on Batting Average Model

```
Ramsey RESET test using powers of the fitted values of batting_avg
       Ho:  model has no omitted variables
                 F(3, 349) =      1.02
                  Prob > F =      0.3840
```

Figure A5. 2018 Season RESET Test on SLG Model

```
Ramsey RESET test using powers of the fitted values of SLG
       Ho:  model has no omitted variables
                 F(3, 365) =      0.38
                  Prob > F =      0.7677
```

Figure A6. 2018 Season RESET Test on Batting Average Model

```
Ramsey RESET test using powers of the fitted values of batting_avg
       Ho:  model has no omitted variables
                 F(3, 365) =      0.53
                  Prob > F =      0.6642
```

Figure A7. 2019 Season RESET Test on SLG Model

```
Ramsey RESET test using powers of the fitted values of SLG
       Ho:  model has no omitted variables
                 F(3, 373) =      0.08
                  Prob > F =      0.9696
```

Figure A8. 2019 Season RESET Test on Batting Average Model

```
Ramsey RESET test using powers of the fitted values of batting_avg
       Ho:  model has no omitted variables
                 F(3, 373) =      0.67
                  Prob > F =      0.5684
```

Figure A9. 2016 Season White Test on SLG Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)     =    227.06
         Prob > chi2   =    0.2133
```

Figure A10. 2016 Season White Test on Batting Average Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)     =    223.23
         Prob > chi2   =    0.2687
```

Figure A11. 2017 Season White Test on SLG Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)     =    206.34
         Prob > chi2   =    0.5777
```

Figure A12. 2017 Season White Test on Batting Average Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)     =    203.16
         Prob > chi2   =    0.6383
```

Figure A13. 2018 Season White Test on SLG Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)     =    192.25
         Prob > chi2   =    0.8183
```

Figure A14. 2018 Season White Test on Batting Average Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)    =    204.00
         Prob > chi2  =    0.6226
```

Figure A15. 2019 Season White Test on SLG Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)    =    226.53
         Prob > chi2  =    0.2205
```

Figure A16. 2019 Season White Test on Batting Average Model

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(211)    =    232.39
         Prob > chi2  =    0.1491
```

References

Bailey, S. R., Loeppky, J., & Swartz, T. B. (2020). The prediction of batting averages in Major League Baseball. *Stats, 3*(2), 84-93. doi:10.3390/stats3020008

Bradbury, J. C., & Drinen, D. J. (2008). Pigou at the plate: Externalities in Major League Baseball. *Journal of Sports Economics, 9*(2), 211-224. doi:10.1177/1527002507300178

Brown, K. H., & Jepsen, L. K. (2009). The impact of team revenues on MLB salaries. *Journal of Sports Economics, 10*(2), 192-203. doi:10.1177/1527002508329858

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica, 28*(3), 591-605. doi:10.2307/1910133

Chu, C. Y. C., Chang, T. & Chu, J. (2016). Opposite hand advantage and the overrepresentation of left-handed players in Major League Baseball. *Academia Economic Papers, 44*(2), 171-205.

Deli, D. (2013). Assessing the relative importance of inputs to a production function: Getting on base versus hitting for power. *Journal of Sports Economics, 14*(2), 203-217.

Demiralp, B., Colburn, C., & Koch, J. (2012). The effects of age, experience and managers upon baseball performance. *Journal of Economics and Finance, 36*(2), 481-498. doi:10.1007/s12197-010-9141-z

Demmink, H. (2010). Value of stealing bases in Major League Baseball: “Stealing” runs and wins. *Public Choice, 142*(3), 497-505. https://doi-org.coloradocollege.idm.oclc.org/10.1007/s11127-009-9546-4

Einolf, K. W. (2004). Is winning everything: A data envelopment analysis of Major League Baseball and the National Football League. *Journal of Sports Economics, 5*(2), 127-151. doi:10.1177/1527002503254047

Fortenbaugh, D., Fleisig, G., Onar-Thomas, A., & Asfour, S. (2011). The effect of pitch type on ground reaction forces in the baseball swing. *Sports Biomechanics, 10*(4), 270-279. doi:10.1080/14763141.2011.629205

Hakes, J. K., & Sauer R. (2006). An economic evaluation of the Moneyball hypothesis. *Journal of Economic Perspectives, 20*(3), 173-186.

Hakes, J. K., & Turner, C. (2011). Pay, productivity and aging in major league baseball. *Journal of Productivity Analysis, 35*(1), 61-74.

Hirotsu, N., & Bickel, J. E. (2019). Using a Markov decision process to model the value of the sacrifice bunt. *Journal of Quantitative Analysis in Sports, 15*(4), 327-344. doi:10.1515/jqas-2017-0092

Janhuba, R. (2019). Do victories and losses matter? Effects of football on life satisfaction. *Journal of Economic Psychology, 75*, 102102. doi:10.1016/j.joep.2018.09.002

Katsumata, H. Himi, K., Ino, T., Ogawa, K., & Matsumoto, T. (2017). Coordination of hitting movement revealed in baseball tee-batting. *Journal of Sports Sciences, 35*(24), 2468-2480.

Krane, V., Joyce, D., & Rafeld, J. (1994). Competitive anxiety, situation criticality and softball performance. *The Sport Psychologist, 8*(1), 58-72. doi:10.1123/tsp.8.1.58

Krohn, G. A. (1983). Measuring the experience-productivity relationship: The case of Major League Baseball. *Journal of Business & Economic Statistics, 1*(4), 273-279. doi:10.1080/07350015.1983.10509351

Lee, Y. H. (2011). Is the small-ball strategy effective in winning games? A stochastic frontier production approach. *Journal of Productivity Analysis, 35*(1), 51-59.

Levine, B., & Bierig, J. (2017). Are defensive shifts changing hitters' swing level? *Baseball Digest*, *76*(6), 12-18.

Lewis, Michael. 2003. *Moneyball: The Art of Winning an Unfair Game.* Norton: New York.

MLB Player Positioning vs Batter. (2020). Retrieved October 12, 2020, from https://baseballsavant.mlb.com/visuals/batter-positioning?playerId=596146

Peach, J. T., Fullerton, S. L., & Fullerton, T. M. (2016). An empirical analysis of the 2014 major league baseball season. *Applied Economics Letters, 23*(2), 138-141. doi:10.1080/13504851.2015.1058898

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B, Methodological, 31*(2), 350-371. doi:10.1111/j.2517-6161.1969.tb00796.x

Santo, C. (2016). The economic impact of sports stadiums: Recasting the analysis in context. *Journal of Urban Affairs, 27*(2), 177-192. doi:10.1111/j.0735-2166.2005.00231.x

Sheehan, J. (2015). Reality Check. *Sports Illustrated. 122*(22), 42-45.

Soebbing, B. P. (2008). Competitive balance and attendance in Major League Baseball: An empirical test of the uncertainty of outcome hypothesis. *International Journal of Sport Finance, 3*(2), 119-126.

Sullivan, T.R. (2015). The infield shift. *Baseball Digest, 74(*6), 20-23.

Tu, C. C. (2005). How does a new sports stadium affect housing values? The case of FedEx field. *Land Economics*, *81*(3), 379-395. doi:10.3368/le.81.3.379

Wasserman, E. B., Abar, B., Shah, M. N., Wasserman, D., & Bazarian, J. J. (2015). Concussions are associated with decreased batting performance among Major League Baseball players. *The American Journal of Sports Medicine, 43*(5), 1127-1133.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817-838. doi:10.2307/1912934

Yilmaz, M.R., & Chatterjee, S. (2003). Salaries, performance, and owners' goals in Major League Baseball: A View Through Data. *Journal of Managerial Issues, 15*(2), 243-255.

Zimmer, T. E. (2018). The negative influence of prior World Series victory on attendance: Winning and increased fan apathy. *Journal of Global Sport Management, 3*(4), 369-388. doi:10.1080/24704067.2018.1442237