**Distribution Agreement**



Signature:

Lauren S. Treiman 14 April 2020

Predicting Baseball Player Performance with OLS Regression and Out-of-Sample Forecasting

By

Lauren Treiman

Robert Roth, Ph.D.
Adviser

Department of Mathematics

Robert Roth, Ph.D.
Adviser

Bree Ettinger, Ph.D.
Committee Member

Seunghwa Rho , Ph.D.
Committee Member

2020

Predicting Baseball Player Performance with OLS Regression and Out-of-Sample Forecasting

By

Lauren Treiman

Robert Roth, Ph.D.Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Department of Mathematics

2020

Abstract

Predicting Baseball Player Performance with OLS Regression and Out-of-Sample Forecasting
By Lauren Treiman

Objective: To assist Major League Baseball (MLB) teams in contract negotiations by better predicting players' future performance.

Methods: Players from 1871 – 2018 with at least 7 years of MLB experience were analyzed to determine the most important factors affecting their performance. I used wins above replacement (WAR) as my dependent variable to measure players' value and Ordinary Least Squares (OLS) regression to predict players' future WAR. Initially, players from the 2010s were analyzed with out of sample forecasting by comparing players with similar WAR. Multiple regression models of comparable players were then developed from different decades with 1-6 years of past experience. Future performance for multiple seasons were then predicted for players competing in the early 2010s by using comparable players who played in the last 3 decades (1990s-2010s) with 6 years of past experience. To best reflect the contract negotiation process, only the sample's actual WAR from their first 6 years in the MLB was considered to predict the rest of their career. Thus, WAR predictions for their 7th, 8th, … years were used to predict performance towards the end of their career.

Results: The model developed was most accurate when only analyzing the 3 most recent decades of past players (players since the 1990s for batters in 2010s) in conjunction with the past 6 WAR values. The regression model constructed was within 2 WAR from the actual WAR and was able to accurately predict player's performance trends throughout their career. My model should help teams by providing additional information that will improve evaluation of a player's performance for the next four years after seven years in MLB.

# Predicting Baseball Player Performance with OLS Regression and Out-of-Sample Forecasting

By

Lauren Treiman

Robert Roth, Ph.D.
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors

Department of Mathematics

2020

Acknowledgments

I would like to thank Richard Anderson, assistant director of quantitative analysis for the Los Angeles Angels, for teaching me baseball analytics and guiding and mentoring me through this process. I would also like to thank the Los Angeles Angels Baseball Operations Department, including Kevin Ferris, for giving me the opportunity to do a summer internship in Baseball Operations with their organization. I would like to thank my brother, Scott Treiman, for forcing me to watch baseball games instead of cartoons and for making me play baseball with him and his friends. Without him, I would have never learned the rules of the game or eventually fallen in love with the sport.

# Contents

# List of Figures

# List of Tables

# 1. Introduction

Miguel Cabrera was an all-star for the Detroit, Tigers. When Cabrera was eligible for free agency, the Tigers re-signed him for 248 million dollars for 8 years. After signing the contract, however, Cabrera's performance declined and he was no longer maintaining his prior and expected level of performance (Trachtman 2019). The Detroit Tigers had spent millions of dollars that could have been better used on other players performing at a higher level and leading to more victories. The Tigers are not alone. Almost all Major League Baseball (MLB) teams have spent millions of dollars on players who do not perform at their expected level. A better method to predict future performance would allow allocation of financial resources to players more likely to meet or exceed expectations.

This analysis was performed to help teams better predict the future performance of players and assist in contract negotiations. I hope my findings improve contract negotiations by providing an objective, statistical model that is reproducible and more accurate than current approaches.

## 1.1 Contracts

To understand the model, it is necessary to have a background in baseball contract negotiations. When a player signs a contract with a baseball organization, that organization has contractual rights to him for 6 years (Sievert 2018). A contract is divided into three stages. For the first two years of his contract, a player is considered to be a "pre-arbitration player" and is usually paid close to the MLB minimum salary. Between $3-6$ years of service, a player is eligible for arbitration. However, an arbitration agreement can only be signed between the player and his current team.

These contracts can be substantially more than the MLB minimum, but are usually significantly less than free agent contracts. (Sievert 2018). An arbitration deal can also extend the initial contract allowing the player to continue to play for his current team. After 6 years of service or the end of his arbitration contract, players become free agents and are able to sign with any MLB team. During free agency, players can engage in ”competitive bidding” between organizations, often receiving much higher salaries (Ng 2017). It is therefore not surprising that younger players almost invariably make less money than older players with similar performance levels (Hakes and Turner 2011).

## 1.2 Alternative Projection Methods

Since predicting player performances is critical to creating appropriate contracts, there have been many methods developed trying to meet this goal. Many researches have investigated aging trends attempting to determine the age when players peak (referred to as peak age), and if the age of peak performance varies depending on the player's ability. Researchers agree that performance declines following initial improvement (Ng 2017). Ng found that batter's peaked at age 29 (Ng 2017). In contrast, Shultz et al. concluded that a player's peak age is 27 and that his initial age when starting his MLB career was not important in determining future performance (Schultz et al. 1994). Shultz et al. also found that better players improved faster, (Schultz et al. 1994) and Hakes and Turner concluded that better players peak later in their careers relative to players of lower ability (Hakes and Turner 2011).

Researchers have created projecting systems that are well-known and are currently used by teams and many sports fans. These projection systems include the following:

## Delta Method

The Delta method, created by Mitchel Lichtman and Tom Tango, is primarily used to determine aging patterns. It calculates the change in performance for players who have played in two consecutive seasons (M. Lichtman 2009), and then groups these differences based on players' ages for both seasons (for example, age $25-26$ is a group). This method then computes an average of these differences, weighting players with higher PA more heavily (M. Lichtman 2009).

## Marcel

Marcel is a simple projection system created by Tom Tango. To project a player's performance for the following year, Marcel uses his last three years of data, regresses to the mean, and adds an age factor (Drushcel 2016). The three years are weighted by multiplying each performance by $5, 4$ or $3$, weighting the recent years more heavily (Drushcel 2016). The regression to the mean (RTTM) factor attempts to remove any "luck" or random variation that may be seen in players with a low number of plate appearances (PA). In other words, the less a player plays, the less we know about his ability and the more we must regress his ability to the mean. (Alley 2010). The RTTM factor is calculated by multiplying the MLB's average by 1200 PA. (Drushcel 2016). The projection is then multiplied by an age factor. A projection for a player's ability (denoted as $M$) for year $t$ using the Marcel method is shown in Equation 1.1.

$$M = \frac{5 * PA_{t-1} * M_{t-1} + 4 * PA_{t-2} * M_{t-2} + 3 * PA_{t-3} * M_{t-3} + 1200 * M_{avg}}{5 * PA_{t-1} + 4 * PA_{t-2} + 3 * PA_{t-3} + 1200} \quad (1.1)$$

$M$ represents the batting metric chosen to measure a player's ability. These batting metrics include wOBA, OPS, OBP, SLUG, ISO, and AVG. I describe these metrics in detail in the Methods section. $M_{avg}$ denotes the league average.

## PECOTA

The Player Empirical Comparison and Optimization Test Algorithm (PECOTA), created by Nate Silver, first calculates a baseline performance level using past performances, weighting recent years heavier. It then compares a player's baseline performance level, age, position, body type (height and weight) to other players. PECOTA then constructs a career trajectory for this player by comparing career trajectories of other players who are closely related (Drushcel 2016).

## Steamer

Steamer is a projection system that is very similar to Marcel. The only difference is Steamer uses regression analysis to determine the weights of past performances and the number of PA for the RTTM factor instead of using arbitrary weights $5, 4$, and 3 and 1200 PA for RTTM factor (Drushcel 2016).

## ZiPS

The sZymborski Projection System (ZiPS), created by Dan Szyborski, uses the past 4 years of experience, weighting past years more heavily with weights $8, 5, 4$, and 3, and aging trends to develop future projections (Drushcel 2016). ZiPS compares a player's baseline performance to players with similar baseline performances at every point in their careers. ZiPS also creates a player's probable aging curve by using players who are the same age, play the same position, bat and throw using the same hand, and have a comparable height and weight compared to the averages for the given time period (Szymborski 2010). Height and weight are compared to the average height and weight for a specific time period since the average player's body type has changed dramatically through the years (Szymborski 2010).

# 2. Methods

## 2.1 Datasets

In this study, I sought to develop a model based upon a player's current value that will more accurately predict his future performance and improve subsequent contract negotiations. To measure a player's value, I used the metric wins above replacement (WAR) that was created by Baseball References. These data sets were readily accessible and have been validated (*Baseball-References.com* 2019).

To obtain data sets regarding information about players and their performance, I accessed data produced by Sean Lahman covering the time period from $1871 - 2019$. From Lahman's database, I compiled data on a player's age, length of career, position, and hand he uses to throw and bat. The database also provided metrics, including batting average, (AVG), on base percentage (OBP), slugging percentage (SLUG), weighted on base average (wOBA), isolated power (ISO), that help measure a player's performance (Lahman 2019). For this analysis, I will refer to AVG, OBP, SLUG, wOBA, and ISO as the performance metrics.

I downloaded both databases into R programming language and merged the datasets together based on a player's ID (which is consistent across all datasets from different reference sources) and year. To deal with players who switched teams mid-season and would have multiple metrics for a single season, I calculated a weighted average based on plate appearances (PA). Thus, all players only had one value for a specific metric in the dataset for a each season.

## 2.2 Variables

Baseball abbreviations used in this paper are listed in the Appendix.

### WAR

Wins above replacement (WAR) is a statistic summarizing a player's offensive and defensive value to his team. WAR is defined as the number of additional games a team wins because they have that player in the game compared to a replacement player: a minor league player or free agent that is readily available when a starter is unable to play. By definition, all replacement players have a WAR of 0. Since WAR represents the number of wins, if Player A has a 5.4 WAR, then his team won 5.4 games he was in the game (Slowinski 2010e). It is possible for a player to have a negative WAR value, indicating that he is less valuable than a replacement player (Slowinski 2010e).

WAR is the only baseball metric that uses both offensive and defensive measurements and is calculated differently for position players and pitchers. WAR is calculated by using six different components for position players. These components are batting runs, base running runs, runs added or lost due to grounding into a double play, fielding runs, positional adjustable runs, and replacement level runs (*WAR Explained* n.d.). Batting runs, base running runs, and runs added or lost due to grounding into a double play measure a player's offensive performance by factoring in how well he bats, runs the bases, steals bases, and hits into double plays. Fielding runs measure a player's defensive performance by determining how many runs he prevents. Positional adjustments are the adjustments made based on the difficulty of the position. Replacement level runs is the expected value of runs a replacement player will have compared to a starter. This is used as an adjustment factor when calculating WAR (*WAR Explained* n.d.).

WAR has been standardized to adjust for context, league, and park effects (Slowinski 2010e) so it can compare players across teams, leagues and years. WAR should be interpreted as a range rather than an exact value. A position player's value to his team based on his WAR statistic is shown in Table 2.1 (Slowinski 2010e).

| | |
|---|---|
| Scrub | 0-1 |
| Role Player | 1-2 |
| Solid Starter | 2-3 |
| Good Player | 3-4 |
| All-Star | 4-5 |
| Superstar | 5-6 |
| MVP | 6+ |

Table 2.1: Wins Above Replacement (WAR)

## Age and Experience

Age is defined as a player's age at the start of the season while experience is the number of seasons a player has participated. Age and experience are closely correlated but may vary depending on when a player starts his major league career, returns to the minor leagues for an extended period of time, or takes a leave of absence from MLB. I only included the variable experience in my model since baseball contracts and free agency status are based on the number of years of service (Hakes and Turner 2011).

I constructed a new dataset creating multiple lag variables that display past performances of the metrics WAR, OBP, wOBA, SLUG and ISO for each player. These variables were defined as metric1, metric2,... where the number after each metric corresponds to the number of past years before the given season. Figure 2.1 is an example of creating lag WAR variables for Mike Trout's first five seasons.

| Player | Experience | Age | Year | Bats | Throws | PA | WAR | WAR1 | WAR2 | WAR3 | WAR4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Trout | 1 | 20 | 2011 | R | R | 135 | 0.5 | NA | NA | NA | NA |
| Trout | 2 | 21 | 2012 | R | R | 639 | 10.5 | 0.5 | NA | NA | NA |
| Trout | 3 | 22 | 2013 | R | R | 716 | 8.9 | 10.5 | 0.5 | NA | NA |
| Trout | 4 | 23 | 2014 | R | R | 705 | 7.7 | 8.9 | 10.5 | 0.5 | NA |
| Trout | 5 | 24 | 2015 | R | R | 682 | 9.6 | 7.7 | 8.9 | 10.5 | 0.5 |

Figure 2.1: Ex. Mike Trout (first five years)

The variable WAR refers to Trout's WAR value for the given season. WAR1-WAR4 represent the WAR values for the last 1-4 years prior to the given season. For example, in his second year, he would only a metric for WAR1 representing his performance in his prior season but there is no WAR value for WAR 2- WAR 4 since he had not played that long. By creating these variables, I was able to include past performances in my model.

I expected that both age and experience would have a diminishing returns effect. In other words, players will get better at the start of their career but will eventually decline. This implies that there is a quadratic relationship between age or experience and WAR (Hakes and Turner 2011)). Therefore, I used the variables experience and $\text{experience}^2$ in my model. I tested the finding that better players tend to peak later in their careers (Ng 2017). I also evaluated these conclusions by grouping players based on their skill level (I explain how I grouped players later), created a linear regression model, estimated each group's WAR value for each year of experience, and graphed the results. I determined the year (number of years of experience) their WAR value was highest and ran t-tests between groups to determine if there was a statistically significant difference between when the performance of players in each group started to decrease.

## Position

Pitchers and position players' WAR were calculated differently (Slowinski 2010e) since pitchers and position players were used differently to help their team. In this analysis, I only predicted position players future performances. WAR adjusts for the difficulty of a position so I did not include a position variable in my model.

## Handedness

Bats is a categorical variable that represents the hand a player predominantly uses when hitting. Players are left-handed, right-handed, or both. Players who predominantly use both their left and right hand are called switch-hitters. Throws is a dichotomous variable that shows if a person throws with his left or right hand. (There were no players in my sample who throw using both hands). I tested whether left-handed players, right-handed players, and switch-hitters perform differently throughout their careers. Several players bat and throw using opposite hands, so I treated these two variables separately. I ran two separate linear regression models where I regressed WAR on experience and $experience^2$ and interacted both terms with either throws or bats. I then estimated each group's WAR value for each year of experience and graphed the results.

## Batting Average (AVG)

Batting average (AVG) measures how frequently a batter hits the ball and makes it on base safely. Even though AVG is the most recognizable baseball statistic, it has two major flaws and should thus be used with caution when analyzing players (Weinberg 2015). First, batting average does not take into account other possible outcomes, such as a walk or hit-by-pitch (HBP), that result in a batter making it on base. In general, baseball teams care more that a player makes it on base rather than the method he uses since safe bases lead to scoring runs (Weinberg 2015). Second, AVG does not take into account hit-type as hitting home runs are more valuable than singles (MLB n.d.(a)). Statisticians have created new metrics that account for these flaws so I did not use AVG for this analysis.

## On Base Percentage (OBP)

On-Base-Percentage (OBP) represents how frequently a player gets on base safely (Tango, M. G. Lichtman, and Dolphin 2006). It is more accurate than AVG since it takes into account hits, walks and HBP, all possible ways a batter gets on base. OBP can be calculated using Equation 2.1 (Slowinski 2010b).

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF} \tag{2.1}$$

Similar to AVG, OBP values range from 0 to 1. A rough estimate of a player's batting performance based on his OBP is shown in Table 2.2 (Slowinski 2010b).

| | |
|---|---|
| Excellent | 0.390 |
| Great | 0.370 |
| Above Average | 0.340 |
| Average | 0.320 |
| Below Average | 0.310 |
| Poor | 0.300 |
| Awful | 0.290 |

Table 2.2: On Base Percentage (OBP)

OBP weights all hits and walks equally (Slowinski 2010c), so it is necessary to use another statistic, slugging percentage (SLUG), to accurately measure a player's batting performance.

## Slugging Percentage (SLUG)

Slugging Percentage (SLUG) measures how many bases a batter gains for himself by weighting hit-types differently (Tango, M. G. Lichtman, and Dolphin 2006). SLUG only accounts for hits and can be calculated using Equation 2.2 (MLB n.d.(b)).

$$SLUG = \frac{1 * (1B) * 2(2B) * 3(3B) * 4(HR)}{AB} \tag{2.2}$$

Even though doubles are not twice as valuable as singles (MLB n.d.(b)), SLUG is a useful metric since it helps better determine a value of a hit. However, it is not the best statistic to measure a batter's power as I will discuss later in this analysis.

## Weighted On Base Percentage (wOBA)

Together, OBP and SLUG are two important metrics for understanding batting performance but are limited when analyzed separately since OBP is a summary of how frequently a batter gets on base while SLUG is a summary of how many extra bases he averages per at bat. These two metrics should be used together to accurately measure a player's batting performance (Slowinski 2010c). Statisticians have created the metric On-Base Plus Slugging (OPS) that is computed by adding OBP and SLUG, but this metric has been criticized since OBP and SLUG are weighted equally when OBP is 1.8 times more important (Slowinski 2010c). To address these limitations, I used the metric weighted On-Base Average (wOBA), developed by Tom Tango, that weights OBP more than SLUG (Tango, M. G. Lichtman, and Dolphin 2006). The formula to calculate wOBA is displayed in Equation 2.3 (Tango, M. G. Lichtman, and Dolphin 2006).

$$wOBA = \frac{WH}{PA} \tag{2.3}$$

where

$$WH = .72(NIBB) + .75(HBP) + .9(1B) + .92 * RBOE + \\ 1.24(2B) + 1.56(3B) + 1.95(HR)$$

wOBA uses the same scale as OBP, so it is easy to compare the two statistics. A rough estimate for a player's performance based on his wOBA is displayed in Table 2.3 (Slowinski 2010f).

| | |
|---|---|
| Excellent | 0.400 |
| Great | 0.370 |
| Above Average | 0.340 |
| Average | 0.320 |
| Below Average | 0.310 |
| Poor | 0.300 |
| Awful | 0.290 |

Table 2.3: weighted On Base Average (wOBA)

### Isolated Power (ISO)

SLUG is a useful tool for measuring power, but is not the best statistic since it changes when AVG changes and does not only measure a batter's power (Humphrey 2014). To more accurately calculate a player's power, I only measured a player's extra-base hits per at bat. Subtracting AVG from SLUG to calculate the extra-base hits is referred to as Isolated Power (ISO) (Humphrey 2014). I used ISO since I think players considered power hitters tend to decline at a faster rate than other players. ISO ranges from 0 to 1 and remains constant when AVG changes. A rough estimate of a player's performance based on his ISO is shown in Table 2.4 (Slowinski 2010a).

| | |
|---|---|
| Excellent | 0.250 |
| Great | 0.200 |
| Above Average | 0.170 |
| Average | 0.140 |
| Below Average | 0.120 |
| Poor | 0.100 |
| Awful | 0.080 |

Table 2.4: Isolated Power (ISO)

## 2.3 Predictive Metric Analysis

I used the Pearson Correlation Coefficient to calculate the correlations between WAR and the performance metrics. I ran a linear regression model using all performance metrics and other variables such as handedness and experience to determine those variables that affected a player's WAR. I used the coeftest() function to use robust standard errors to assure there was no heteroskedasticity. I ran F-tests and t-tests to determine if there was multicollinearity. Multicollinearity occurs when two or more explanatory variables are individually insignificant but jointly significant to the dependent variable. This occurs when the variables are highly correlated. If there is multicollinearity between the explanatory variables, then the OLS regression model is less precise since the variance is very high. Thus, the model would not be accurate. To ensure the model accuracy, I did not run a model between variables that were highly correlated.

## 2.4 Player Classification

A stabilization point refers to the number of PA required to know that the player's metric is accurate. The purpose of a stabilization number is to prevent people from interpreting measurements that are highly susceptible to random chance (Slowinski 2010d). Since WAR is both an offensive and defensive statistic, there is no point in which it is known that the WAR value is stabilized. For this analysis, I chose 50 PA as the stabilization point to ensure that there was still a big sample size and players who rarely played were removed.

I then filtered the data to only include players who played for at least seven seasons in the MLB. I chose to use seven seasons, an idea used in "Analyzing Major League Baseball Player's Performance Based on Age and Experience" (Ng 2017), as a minimum requirement since this required players to have been eligible for free agency and have signed a second contract following their starting contract. This allowed for using these players' performances to predict other players' future performance currently in the process of signing a second contract.

I wanted to compare players to other players with similar performance levels. I used the 3rd highest WAR value, a concept developed by Beirig, Hollenbeck, and Stoud, to compare players based on their 3rd highest season (Hakes and Turner 2011). This was done since using the highest WAR might not accurately measure a player's value since some players have one or two phenomenal seasons but then do not perform nearly as well for the rest of their career. In addition, I did not use their career average due to possible outliers that might skew the average. I also filtered the data to only include seasons where a player had at least 50 plate appearances when determining the 3rd highest WAR value. This removed players not playing enough resulting in a WAR value that did not accurately reflect the player's performance. I then grouped these players based on the percentile of their 3rd value WAR. I divided the percentiles into $0-25\%$, $25-50\%$, $50-75\%$, $75-100\%$.

I filtered the data to only include the top 75% of players since the majority of these players signed the biggest contracts. This model will work for the other three groups but I only reported the results for the top group of players.

I assigned each player to a decade based on the decade he played in the middle of his career. For example, if he played from $1996-2010$, 2003 marks the middle of his career and he would be assigned to the 2000 decade. I displayed the number of players assigned to each decade. I then assigned "decade numbers" based on how recently they played. You can interpret this decade number as "Player A's middle season occurred "decade number" decades ago. This "decade number" was used later in my model to address the issue of how many past decades I should use to compare players.

## 2.5 The Model

I used Ordinary Least Squares (OLS) regression out-of-sample forecasting to predict the most recent players (players assigned to decade 2010). I first divided these players into two groups. I predicted WAR values for the first group using the second group's WAR values. I then created multiple models that used a different number of past years (ranging from $1-6$), dummy variables such as throws and bats since left-handed players might perform perform differently than right-handed players, and other metrics using the same number of past years that affect WAR. I added an additional decade of players to determine how many players (or years back) could be used since there is a trade-off of increasing sample size versus comparing players from two different time periods.

I calculated the sum of square residuals (SSR) by subtracting the first group's actual WAR value from their predicted value, squaring the difference, and then adding all the residuals. This calculation can be represented in Equation 2.4.

$$SSR = \sum_{i=1}^{n}(WAR_i - \widehat{WAR_i})^2 \tag{2.4}$$

In Equation 2.4, $n$ represents the number of players in the sample, $WAR_i$ and $\widehat{WAR_i}$ are the actual and predicted values of WAR for player$_i$.

I then calculated the standard error of the residuals (referred to as error) to determine the difference between the predicted WAR values and actual WAR values. Equation 2.5 shows how I calculated the error.

$$error = \sqrt{\frac{SSR}{n}} \tag{2.5}$$

I graphed the error in respect to the number of past years used, number of decades used, and the number of additional metrics used. The best model was the model with the smallest error.

# 3. Results

## 3.1 Groups

This study analyzed 4,100 position players over 148 years to help predict a player's future performance. I divided the players into four groups based on their $3^{rd}$ highest WAR. This provided four groups of approximately 1,050 each (Table 3.1). As can be seen in Table 3.1, there was the most variation in group 4 and the the least variation in group 2.

| **Percentile** | **Group** | **Num. of Players** | **WAR Range** |
|---|---|---|---|
| 0-25 | 1 | 1049 | 0.001 - 2.84 |
| 25-50 | 2 | 1046 | 2.84 - 3.68 |
| 50-75 | 3 | 1047 | 3.68 - 5.88 |
| 75-100 | 4 | 1046 | 5.88 - 12.24 |

Table 3.1: Division of Groups

Tables 3.2, 3.3, 3.4, and 3.5 display the summary statistics for the variables *war*, *experience*, and *age*, for each group.

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| war | 0.01 | 0.16 | 0.49 | 1.25 | 2.74 | 8.33 |
| experience | 1.00 | 3.00 | 5.00 | 5.82 | 8.00 | 27.00 |
| age | 16.00 | 25.00 | 28.00 | 28.48 | 31.00 | 50.00 |

Table 3.2: Summary Statistics for Group 1

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| war | 0.01 | 0.37 | 2.78 | 2.13 | 3.18 | 9.69 |
| experience | 1.00 | 3.00 | 6.00 | 6.60 | 9.00 | 25.00 |
| age | 16.00 | 26.00 | 29.00 | 28.91 | 32.00 | 57.00 |

Table 3.3: Summary Statistics for Group 2

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| war | 0.01 | 1.41 | 3.32 | 3.18 | 4.36 | 11.00 |
| experience | 1.00 | 3.00 | 6.00 | 6.91 | 10.00 | 31.00 |
| age | 17.00 | 26.00 | 26.00 | 29.00 | 32.00 | 50.00 |

Table 3.4: Summary Statistics for Group 3

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| war | 0.01 | 3.24 | 5.44 | 5.18 | 7.14 | 12.41 |
| experience | 1.00 | 4.00 | 8.00 | 8.37 | 12.00 | 29.00 |
| age | 17.00 | 26.00 | 29.00 | 29.60 | 33.00 | 55.00 |

Table 3.5: Summary Statistics for Group 4

As shown in Figure 3.1, there are some players who's highest WAR was significantly above the average for their group. These players had one or two excellent seasons compared to the rest of their career. These players were classified into the group based on the $3^{rd}$ highest WAR but were removed from the model since these outliers (WAR $\geq WAR_{Q_3} + (1.5 * WAR_{Q_3} - WAR_{Q_1})$, where $WAR_{Q_1}$ and $WAR_{Q_3}$ are WAR values at the $25^{th}$ and $75^{th}$ percentile) might skew the data.

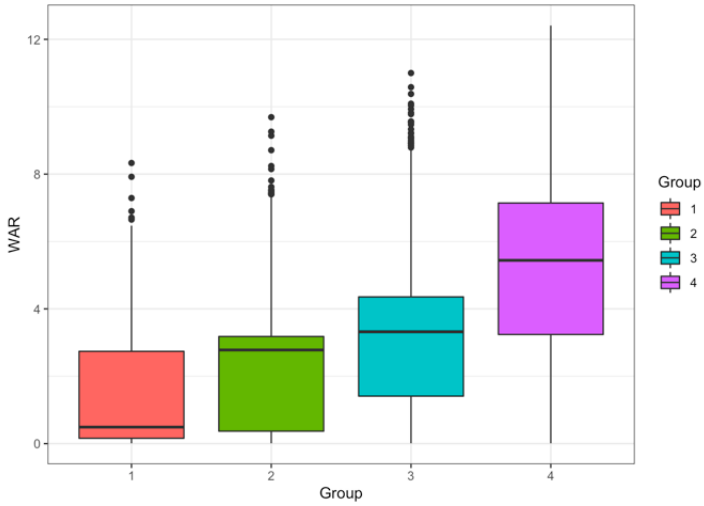



Figure 3.1: Distribution of $3^{rd}$ Highest War for Each Group

Table 3.6 shows the number of players removed from each group. There were no players removed from Group 4 since Group 4 has a very large range of performance.

| **Group** | **Num. Removed Players** |
|---|---|
| 1 | 49 |
| 2 | 134 |
| 3 | 319 |

Table 3.6: Players Removed

Figure 3.2 shows the change in the average WAR for each group of players (groups $1-4$) with increasing length of years in major league baseball. While some players had careers that extended beyond 12 years, the model was restricted to the first 12 years of a player's career since the data is stronger during this time period. This provides more data (as shown in Table 3.7), increasing the power of the study, and lowering the variance of results. I evaluated the model using all years of a player's career, but the few players with very long careers skewed the data and the quadratic basis for the model lead to uninterpretable results.

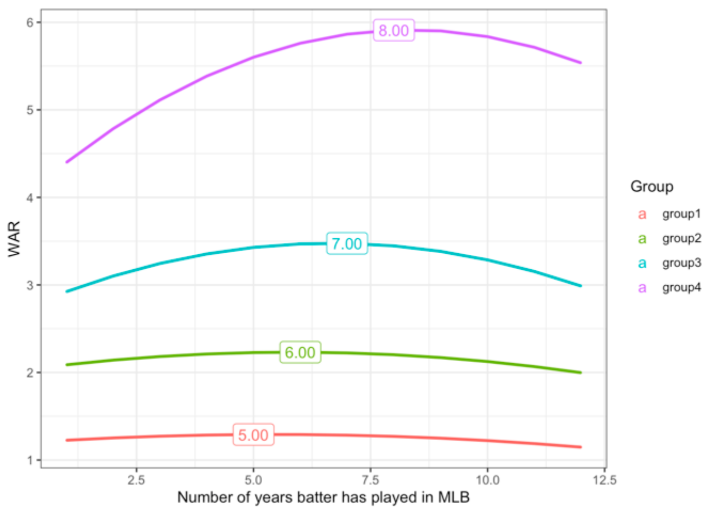



Figure 3.2: WAR vs Years of Experience

| | Groups | | | |
|---|---|---|---|---|
| Duration | **1** | **2** | **3** | **4** |
| 7 | 935 | 942 | 926 | 1006 |
| 8 | 805 | 864 | 886 | 974 |
| 9 | 604 | 745 | 798 | 949 |
| 10 | 482 | 613 | 690 | 899 |
| 11 | 323 | 471 | 594 | 830 |
| 12 | 231 | 388 | 492 | 782 |

Table 3.7: Players For Each Time Period

Between each group, I compared the year the highest WAR value was obtained. I ran two tailed t-Tests to determine if each group's best year was statistically different. These results are shown in Table 3.8 as well as the average best year (rounded to the nearest year) and 95% confidence intervals (CI).

| | | | | | | CI | |
|---|---|---|---|---|---|---|---|
| **Group** | **Best Yr. Avg** | **Group** | **Best Yr. Avg** | **t-Stat** | **p-value** | **LB** | **UB** |
| 1 | 5.3 | 2 | 5.6 | -2.34 | 0.02 | -0.56 | -0.05 |
| 1 | 5.3 | 3 | 5.8 | -4.30 | 0.00 | -0.80 | -0.30 |
| 1 | 5.3 | 4 | 6.7 | -10.97 | 0.00 | -1.71 | -1.19 |
| 2 | 5.6 | 3 | 5.8 | -1.91 | 0.06 | -0.50 | 0.01 |
| 2 | 5.6 | 4 | 6.7 | -8.62 | 0.00 | -1.41 | -0.89 |
| 3 | 5.8 | 4 | 6.7 | -6.88 | 0.00 | -1.16 | -0.64 |

Table 3.8: Comparison of Best Performance Year Between Groups

Figure 3.2 and Table 3.8 show that the peak year (defined as the highest WAR) was different for each group.

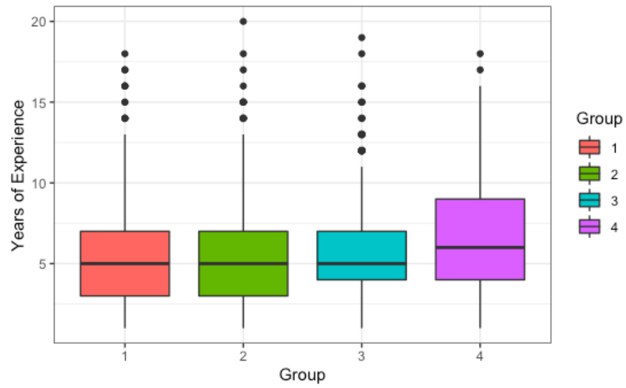



Figure 3.3: Summary Statistics of Best Year for each group

Figure 3.3 indicates that better players peak later in their careers. The median (the horizontal black line in the box plot) and $75^{th}$ percentile (the top of the boxplot) are greater for group 4 than all other groups. In addition, there are fewer outliers in group 4 compared to other groups. Group 4 was normally distributed, in contrast to the other groups that had a right-skewed distribution, reflecting that more players in group 4 peaked later in their careers than players in the other groups.

The performance of players in group 1 and group 2 was relatively constant throughout their careers, as seen in Figure 3.2. Therefore, there was no consistency in the year players in these groups peaked. This explains why group 1 and group 2 had so many outliers and thus had an IQR range similar to group 3.

Since a primary goal of this study is to assist players and management in contract negotiations and especially those involving highly valued players, the remainder of this analysis will focus only on group 4 since these players have the highest WAR and are likely to sign the largest contracts. However, this analysis should be applicable to the other groups as well.

## 3.2 Handedness

Table 3.9 provides the demographics for the number of players that bat and throw with each hand (B - both, R - right, L - left).

| Bats | Throws | Number | Percent |
|---|---|---|---|
| B | L | 8 | 0.78 |
| B | R | 94 | 9.11 |
| L | L | 154 | 14.92 |
| L | R | 207 | 20.06 |
| R | L | 4 | 0.39 |
| R | R | 565 | 54.75 |

Table 3.9: Handedness

Figure 3.4 shows the average WAR throughout a player's career based on how he hits.

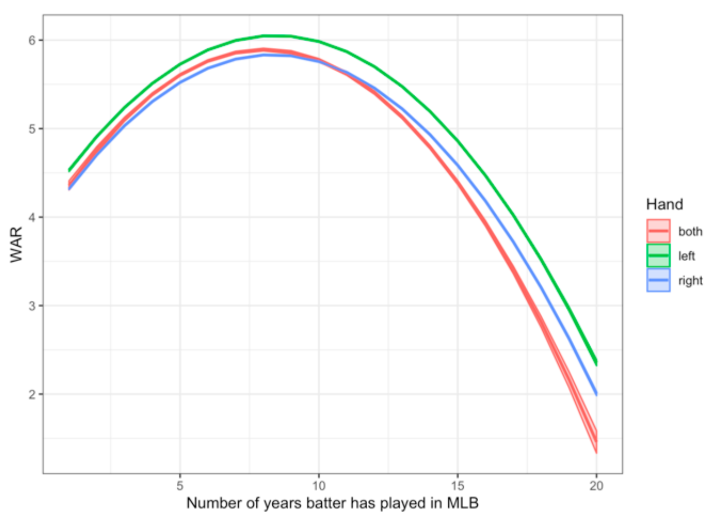



Figure 3.4: WAR based on Handedness (Bats)

Figure 3.5 shows the average WAR throughout a player's career based on how he throws.

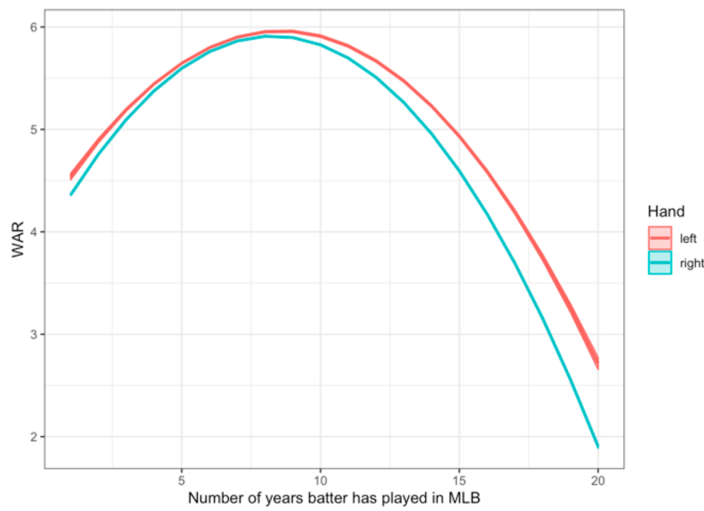



Figure 3.5: WAR based on Handedness (Throws)

Both Figure 3.4 and Figure 3.5 illustrate that the trend of WAR throughout a player's career does not differ based on which hand the player uses to bat or to throw.

## 3.3 Predictive Metrics

I then compared the predictive metrics for group 4 players to determine which metrics most strongly correlated with each other. There was a strong correlation (correlation $\geq 0.85$) between OBP and wOBA, SLUG and wOBA, and ISO and SLUG. These correlations can be seen in Table 3.10.

| | **WAR** | **SLUG** | **OBP** | **ISO** | **wOBA** |
|---|---|---|---|---|---|
| **WAR** | 1.00 | | | | |
| **SLUG** | 0.61 | 1.00 | | | |
| **OBP** | 0.57 | 0.71 | 1.00 | | |
| **ISO** | 0.47 | 0.88 | 0.41 | 1.00 | |
| **wOBA** | 0.64 | 0.90 | 0.93 | 0.66 | 1.00 |

Table 3.10: Correlation of Predictive Metrics

I ran a linear regression model to determine which variables predicted WAR. These results are displayed in Model 1 in Table 3.11. I omitted players who bat with both hands and players who bat with their left hand to avoid perfect collinearity. These variables are used as a comparison to the other variables in the model. For the predictive metrics that were not statistically significant ($p > 0.05$), I ran a linear regression model, displayed in Model 2, after removing the predictive metrics that were statistically significant ($p \leq 0.05$). This enabled me to determine if there was multicollinearity. Model 1 found that the variable ISO was not statistically significant, but after controlling for significant predictive variables (OBP, SLUG, and wOBA) the second linear regression found ISO was statistically significant in affecting WAR.

| | Dependent variable: | |
| --- | --- | --- |
| | war | |
| | Model 1 | Model 2 |
| experience | 0.249*** | 0.343*** |
| | (0.012) | (0.013) |
| $\text{experience}^2$ | −0.016*** | −0.021*** |
| | (0.001) | (0.001) |
| year | −0.007*** | −0.014*** |
| | (0.001) | (0.001) |
| OBP | 16.716*** | |
| | (1.624) | |
| SLUG | 12.629*** | |
| | (0.892) | |
| ISO | 0.962 | 19.667*** |
| | (0.668) | (0.292) |
| wOBA | −6.937*** | |
| | (2.485) | |
| throws (right) | 0.167*** | 0.105* |
| | (0.054) | (0.061) |
| bats (left) | −0.213*** | −0.277*** |
| | (0.064) | (0.072) |
| bats (right) | −0.106* | −0.445*** |
| | (0.059) | (0.065) |
| Constant | 9.552*** | 29.419*** |
| | (1.244) | (1.048) |
| Observations | 14,896 | 14,896 |
| $R^2$ | 0.464 | 0.318 |
| Adjusted $R^2$ | 0.463 | 0.318 |
| Residual Std. Error | 2.054 (df = 14885) | 2.316 (df = 14888) |
| F Statistic | 1,287.274*** (df = 10; 14885) | 993.795*** (df = 7; 14888) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |
| | | df := degrees of freedom |

Table 3.11: Linear Regression Results

Table 3.12 displays the number of players assigned to each decade. The decade beginning in 2010 is arbitrarily defined to be decade 1 since it's the most recent, the decade beginning in 2000 as decade 2 and each proceeding decade sequentially assigned a higher number.

| Year | 1870 | 1880 | 1890 | 1900 | 1910 | 1920 | 1930 | 1940 |
|---|---|---|---|---|---|---|---|---|
| Players | 4 | 38 | 38 | 56 | 54 | 55 | 56 | 53 |
| Year | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 | |
| Players | 63 | 96 | 94 | 95 | 112 | 126 | 106 | |

Table 3.12: Number of Players Assigned to Each Decade

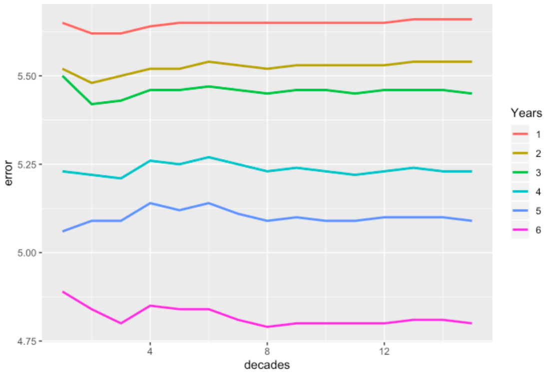
## 3.4 Single Year Predictions

I randomly divided the 106 players from decade 1 (2010) into two equal groups (53 players in each). My goal was to predict the second group's WAR values using the coefficients developed by modeling the first group. This method is referred to as out of sample forecasting. The model was constructed using past WAR performance and dummy variables representing the hand a batter uses to bat or throw. Equation 3.1 shows the model using 1 year of past experience.

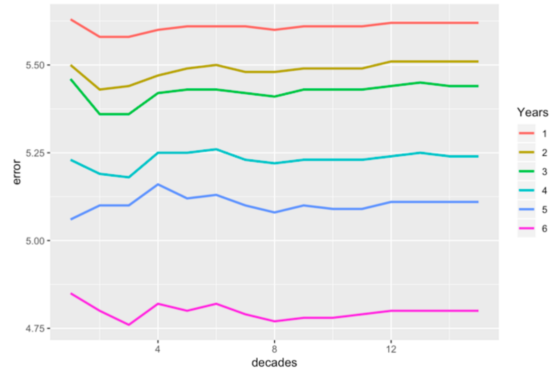
$$WAR_{t,i} = \beta_0 + \beta_1 exp_i + \beta_2 exp_i^2 + \beta_3 TL_i + \beta_4 BL_i + \beta_5 BR_i + \beta_6 WAR_{t-1} \tag{3.1}$$

To use $2, 3, 4, 5$ and 6 past years, I added the variables $\beta_7 WAR_{t-2}$, $\beta_8 WAR_{t-3}$, $\beta_9 WAR_{t-4}$, $\beta_{10} WAR_{t-5}$, and $\beta_{11} WAR_{t-6}$. In Equation 3.1, $t$ represents the year of experience I wanted to predict, $exp$ is abbreviated for experience, and $TL, BL$ and $BR$ are dummy variables for Throws Left-handed, Bats Left-handed, and Bats Right-handed. I dropped Bats Both and Throws Right-handed from the model to avoid perfect collinearity.

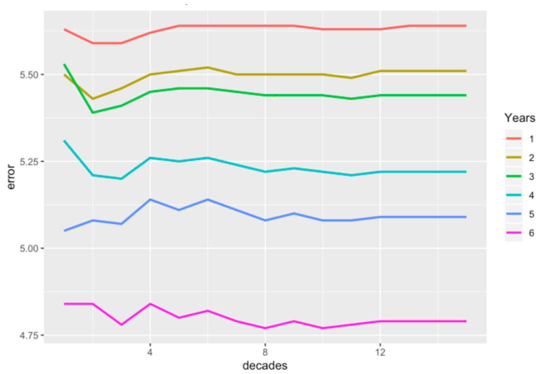
To determine the accuracy of my predictions, I calculated the second group's Sum of Squares Residuals (Referred to as Error). I wanted to determine how many past years of data from the first group would best predict the second group's WAR value. I repeated this process and created multiple models using different numbers of past years. To increase the power, I increased the sample size by including the previous decade (2000 decade was added to group 1) and then repeating the procedure previously described. I repeated this step until players from all 15 decades were used to predict the WAR for players in group 1. Using this approach, I was able to determine how many past years and how many past decades I should use by choosing the model that had the smallest error. I then added other metrics to improve the model by hopefully reducing error and thus increasing accuracy. I iterated this process 50 times to increase precision (repeat-ability). Figure 3.6 displays one of the iteration's results. Similar figures could be constructed for each of the other 49 iterations. For this iteration, using 6 years of past performance produced the most accurate predictions. I selected the best model (defined as the model with the smallest error) using one or more different predictive metrics. These results are shown in Table 3.13. The error value did not change when adding additional predictive metrics (ISO, OBP, SLUG, and wOBA) to the model constructed before, indicating that using these past metrics did not increase the accuracy of predicting a player's WAR. I analyzed whether incorporating one or more of these additional metrics increased the accuracy of the model but found the error did not change with every iteration.



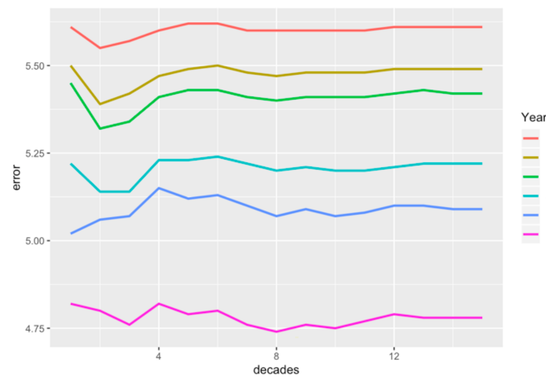

Predictive Metrics: None



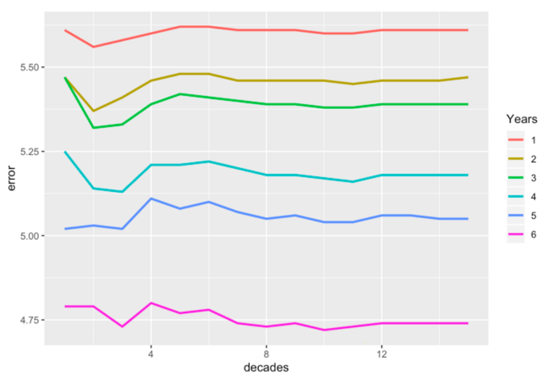

Predictive Metrics: ISO



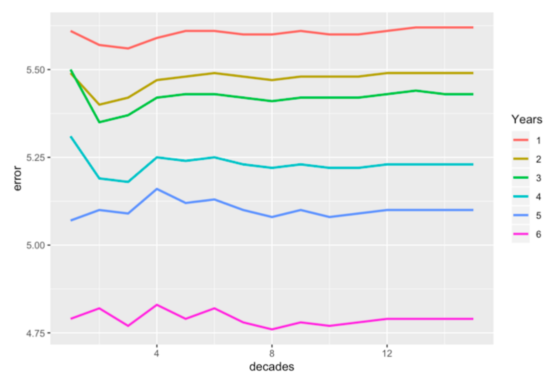

Predictive Metrics: OBP



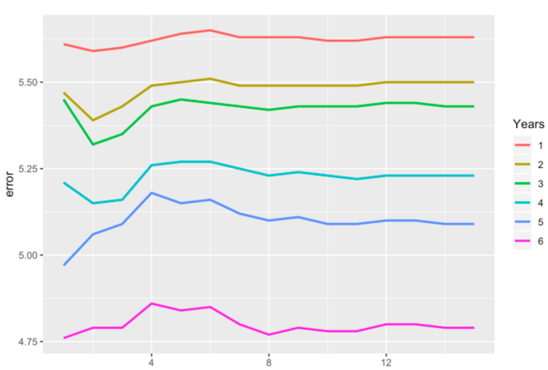

Predictive Metrics: SLUG



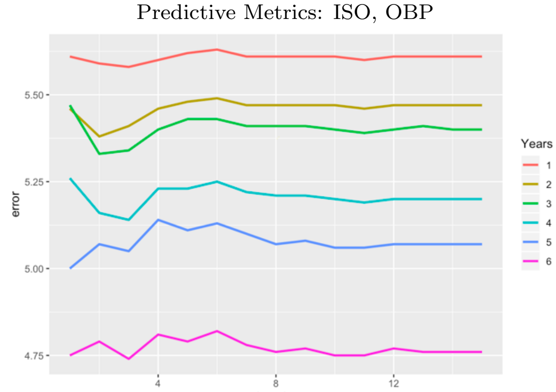

Predictive Metrics: wOBA



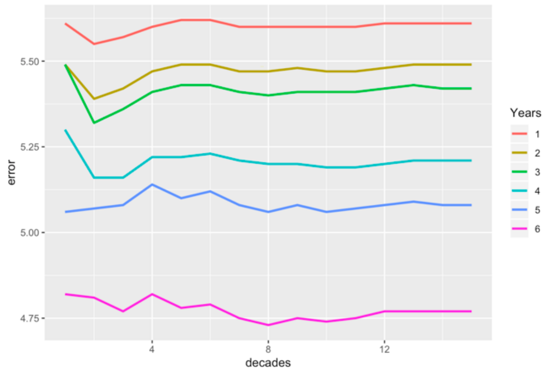

Predictive Metrics: ISO, OBP



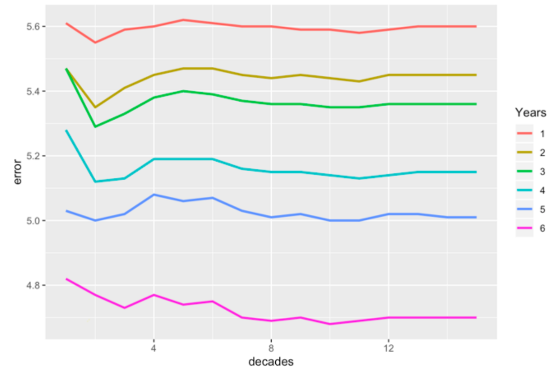

Predictive Metrics: ISO, SLUG



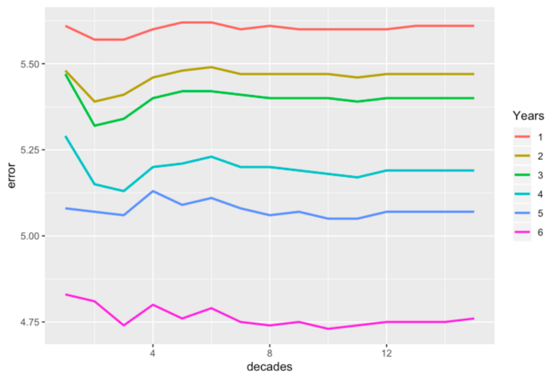

Predictive Metrics: ISO, wOBA



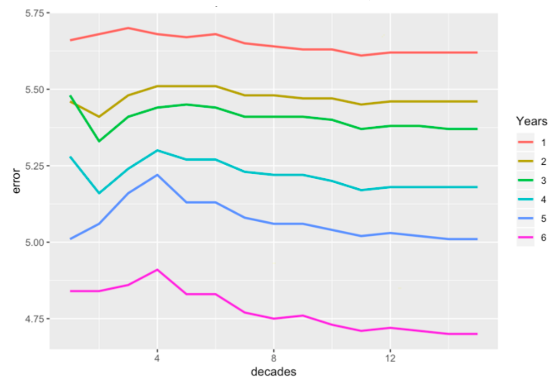

Predictive Metrics: OBP, SLUG



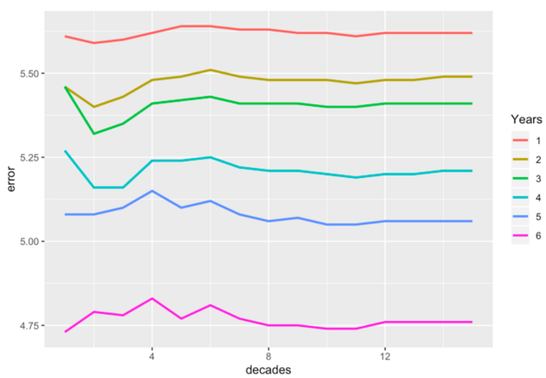

Predictive Metrics: OBP, wOBA




Predictive Metrics: SLUG, wOBA




Predictive Metrics: ISO, OBP, wOBA




Predictive Metrics: ISO, SLUG, wOBA

Figure 3.6: Error vs. Number of Decades and Years Used

| **Predictive Metric** | **Decades** | **Past Years** | **Error** |
|---|---|---|---|
| None | 3 | 6 | 2.19 |
| ISO | 3 | 6 | 2.18 |
| OBP | 8 | 6 | 2.18 |
| SLUG | 3 | 6 | 2.18 |
| wOBA | 8 | 6 | 2.17 |
| ISO, OBP | 8 | 6 | 2.18 |
| ISO, SLUG | 1 | 6 | 2.18 |
| ISO, wOBA | 1 | 6 | 2.18 |
| OBP, SLUG | 3 | 6 | 2.18 |
| OBP, wOBA | 8 | 6 | 2.16 |
| SLUG, wOBA | 3 | 6 | 2.18 |
| ISO, OBP, wOBA | 11 | 6 | 2.17 |
| ISO, SLUG, wOBA | 1 | 6 | 2.18 |

Table 3.13: Results for one iteration

As mentioned previously, I selected the model with the lowest error for each iteration and displayed the results in Table 3.14

| Predictive Metric | Decades | Past Years | Error |
|---|---|---|---|
| wOBA | 1 | 2 | 2.24 |
| ISO, wOBA | 1 | 6 | 2.15 |
| ISO, SLUG, wOBA | 2 | 6 | 2.12 |
| ISO, OBP, SLUG | 2 | 6 | 2.21 |
| OBP, wOBA | 3 | 6 | 2.18 |
| OBP | 1 | 6 | 2.14 |
| wOBA | 2 | 3 | 2.20 |
| ISO, OBP, SLUG | 1 | 6 | 2.30 |
| SLUG, wOBA | 3 | 6 | 2.01 |
| OBP, wOBA | 7 | 6 | 2.26 |
| ISO, OBP, SLUG | 2 | 6 | 2.13 |
| None | 2 | 6 | 2.03 |
| SLUG | 1 | 5 | 2.19 |
| ISO, OBP, SLUG | 2 | 6 | 2.09 |
| OBP | 1 | 6 | 2.19 |
| SLUG | 2 | 3 | 2.24 |
| None | 1 | 1 | 2.26 |
| wOBA | 3 | 6 | 2.14 |
| ISO, OBP, SLUG | 3 | 6 | 2.18 |
| ISO, OBP, SLUG | 1 | 6 | 1.97 |
| ISO, OBP, wOBA | 2 | 4 | 2.10 |
| OBP, wOBA | 1 | 6 | 2.28 |
| wOBA | 2 | 5 | 2.06 |
| ISO, SLUG, wOBA | 3 | 6 | 2.04 |
| OBP | 1 | 4 | 2.22 |
| None | 2 | 6 | 2.25 |
| wOBA | 2 | 3 | 2.32 |
| ISO | 1 | 2 | 2.33 |
| OBP, wOBA | 11 | 6 | 2.33 |
| ISO, SLUG, wOBA | 2 | 6 | 2.14 |
| None | 1 | 4 | 2.23 |
| None | 2 | 6 | 2.26 |
| ISO, SLUG, wOBA | 5 | 6 | 1.98 |
| OBP | 7 | 6 | 2.13 |
| OBP, wOBA | 10 | 6 | 2.15 |
| OBP, wOBA | 8 | 6 | 2.22 |
| OBP | 1 | 6 | 2.18 |
| OBP | 1 | 6 | 2.03 |
| None | 1 | 6 | 2.32 |
| SLUG | 1 | 4 | 2.22 |
| ISO, OBP, SLUG | 3 | 6 | 2.26 |
| OBP | 1 | 4 | 2.34 |
| ISO, SLUG | 1 | 4 | 2.12 |
| ISO, OBP, SLUG | 1 | 6 | 2.10 |
| OBP, wOBA | 5 | 6 | 2.18 |
| ISO, OBP, wOBA | 3 | 6 | 2.22 |
| wOBA | 3 | 6 | 2.16 |
| OBP | 3 | 6 | 2.11 |
| ISO, SLUG, wOBA | 3 | 6 | 2.07 |
| ISO | 3 | 6 | 2.19 |
| ISO, OBP, SLUG | 2 | 6 | 2.38 |

Table 3.14: Lowest Error for each iteration

## 3.5 Multi-Year Predictions

To check the accuracy of my model, I predicted players whose $7^{th}$ season occurred between $2010 - 2015$. I used their actual WAR values from their first 6 seasons and the WAR values I predicted from their $7^{th}, 8^{th}, ...$ season. My predictions for 3 players are displayed in Table 3.15.

| **Player** | **Year** | **PA** | **Year of Exp.** | **Actual WAR** | **Predicted WAR** |
|---|---|---|---|---|---|
| Melky Cabrera | 2011 | 706 | 7 | 7.59 | 3.44 |
| Melky Cabrera | 2012 | 501 | 8 | 8.02 | 3.95 |
| Melky Cabrera | 2013 | 372 | 9 | 2.80 | 4.68 |
| Melky Cabrera | 2014 | 621 | 10 | 6.51 | 6.00 |
| Melky Cabrera | 2015 | 683 | 11 | 4.07 | 5.46 |
| Melky Cabrera | 2016 | 646 | 12 | 5.71 | 4.46 |
| Melky Cabrera | 2017 | 428 | 13 | 2.19 | 2.63 |
| Robinson Cano | 2011 | 681 | 7 | 9.02 | 7.87 |
| Robinson Cano | 2012 | 697 | 8 | 11.54 | 9.50 |
| Robinson Cano | 2013 | 681 | 9 | 10.93 | 8.10 |
| Robinson Cano | 2014 | 665 | 10 | 9.55 | 7.03 |
| Robinson Cano | 2015 | 674 | 11 | 6.55 | 5.56 |
| Robinson Cano | 2016 | 715 | 12 | 10.52 | 5.87 |
| Robinson Cano | 2017 | 648 | 13 | 6.72 | 5.93 |
| Robinson Cano | 2018 | 348 | 14 | 6.44 | 6.88 |
| Justin Upton | 2013 | 643 | 7 | 5.77 | 6.15 |
| Justin Upton | 2014 | 640 | 8 | 6.34 | 6.79 |
| Justin Upton | 2015 | 619 | 9 | 7.46 | 6.33 |
| Justin Upton | 2016 | 626 | 10 | 4.54 | 6.70 |
| Justin Upton | 2017 | 520 | 11 | 5.83 | 5.78 |

Table 3.15: Predictions for 3 Players

I calculated the difference between the actual and predictive WAR values for years $7 - 15$ (depending on the number of years they played). I then computed the average of these differences (referred to as error) and the number of times I overestimated or underestimated the WAR values for each year. These results are shown in Table 3.16.

| Year of Exp. | Avg. Error | Num. Players | Overestimate | Underestimate |
|---|---|---|---|---|
| 7 | 2.35 | 58 | 28 | 30 |
| 8 | 2.81 | 54 | 28 | 25 |
| 9 | 2.60 | 52 | 26 | 26 |
| 10 | 2.88 | 44 | 24 | 20 |
| 11 | 2.57 | 33 | 24 | 9 |
| 12 | 2.44 | 24 | 17 | 7 |
| 13 | 2.42 | 15 | 9 | 6 |
| 14 | 2.49 | 7 | 5 | 2 |
| 15 | 1.75 | 2 | 1 | 1 |

Table 3.16: Cross Validation Check: Predicting Players Performance

# 4. Discussion

## 4.1 Background

Baseball organizations frequently expend significant time and financial resources signing players to lucrative contracts who then do not perform close to expectations. The ability to better predict player performance and more accurately identify athletes likely to meet or exceed expectations would provide teams with substantial cost savings and dramatically reduce their investment of time and coaching on athletes not able to contribute significantly to the team.

To help organizations predict player performance and more accurately develop appropriate contracts, I developed a model based on past performance, player attributes, and predictive metrics. The model was developed using multiple iterations, and is most accurate for players with at least six years of major league experience. I compared various models after evaluating three approaches previously developed.

## 4.2 Approach

Comparing similar players enables me to refine the model and increase the accuracy of the predictions (Koop 2002). I therefore divided the players into four approximately even groups based on past performance to analyze and determine if top players performed differently than others. I compared the performances of each group of players to the other groups.

## 4.3 Rate of Improvement

I found players performance varied over time in initial rate and overall duration of improvement, as well as length of time they performed consistently until deterioration. Players in the top group (group 4) improved at a faster rate and peaked later in their careers, which can be seen in Figure 3.2.

The finding that stronger players improved faster and peaked later in their careers, (results are displayed in and Figure 3.2 and Table 3.8) support those of Schultz et al. who also identified the importance of the rapid rate of improvement (Schultz et al. 1994), as well as those of Hakes and Turner's that better players peak 2 years later than players with lower performance (Hakes and Turner 2011).

These results can be helpful for teams when developing arbitration arrangements. Rate of improvement early on is important since the first few years can help predict when a player peaks and predict his performance throughout his career. Since better players peak later in their careers, they are more likely to continue to improve for a longer duration and therefore maintain a higher level of performance for a longer period of time.

In contrast, players with very high initial WAR values or strong rookie seasons who did not show improvement over the next few years, even if they were playing well, were not as likely to maintain that level of performance. Duration of early improvement, therefore, was not as important as rate. These results indicate players with a slower rate of improvement early in their careers will more likely decline sooner and will not be as valuable as they were in the beginning of their careers compared to players who rapidly improved.

## 4.4 Handedness

Table 3.9 shows the majority of players in the MLB (65% of players) throw right-handed. Of the players who bat right-handed, 99% throw right-handed. This is not the case for players who bat left-handed since only 43% of these players throw left-handed. In fact, there are more left-handed batters who throw right-handed (15% more) than left-handed batters who throw left-handed. Since there is no direct relationship between batting and throwing with the same hand for left-handed players, I analyzed the variables *bats* and *throws* separately. Figure 3.4 and figure 3.5 depict a player's performance based on which hand he uses. In both figures, the trends were similar, indicating that the hand a player used did not affect his performance over time. This supports Bierig, Hollenbeck, and Stroud's findings that handedness did not affect a players' performance (Bierig, Hollenbeck, and Stroud n.d.). In contrast, both the variables *bats left* and *throws* were statistically significant in predicting WAR ($p < 0.05$) in Model 1 shown in Figure 3.11. The variable *bats right* was not statistically significant ($p > 0.05$) in Model 1 but was statistically significant in Model 2 ($p < 0.05$). I am unsure why there would be a statistically significant relationship in only one model. However, I decided to use both *bats* and *throws* in my model for predicting players' performances since adding these variables can only make my model more accurate.

## 4.5 Predictive Metrics

I calculated the correlations between WAR and all predictive metrics and presented the results in Table 3.10. wOBA was highly correlated with SLUG (0.90) and OBP (0.93). I expected these results since wOBA is an adjusted calculation of OPS, which is a linear combination of OBP and SLUG. To test for multicollinearity, I regressed

WAR on all predictive metrics as well as other variabes including experience and handedness (shown in Figure 3.11). In Model 1, ISO was not statistically significant in predicting WAR ($p > 0.05$). However, after removing OBP, SLUG, and wOBA from the linear regression model, ISO became statistically significant with WAR ($p < 0.05$). This indicated that ISO was not jointly significant but individually significant with WAR. I expected that there would be multicollinearity between ISO, SLUG, and OBP since OBP was an adjusted calculation of AVG and there was a perfect linear relationship between AVG, ISO, and SLUG. Even though wOBA, OBP, and SLUG were all jointly significant with WAR, I suspected there was multicollinearity since these variables were highly correlated with each other and have a similar relationship as ISO, SLUG, and OBP. Since I either found or suspected multicollinearity, I did not test a model where ISO, SLUG, and OBP were included or OBP, SLUG, and WOBA were included.

From Figure 3.6, it is apparent that the subgraphs were nearly identical, indicating that adding predictive metrics ISO, OBP, SLUG, and wOBA to the model did not increase its accuracy. These results can also be seen in Table 3.13 as the lowest error using different predictive metrics ranged by only 0.03. Figure 3.6 and Table 3.13 represent a single iteration, but I found similar results for all iterations. In addition, the lowest error from each iteration used different predictive metrics (shown in Table 3.14), indicating that the predictive metrics used in the iteration with the lowest error was random. For these reasons, I did not include any predictive metrics in my model. The determination of WAR already accounted for these predictive metrics, so adding these models as separate variables did not improve the model.

From Figure 3.6, it can be seen that the error decreased as I added additional past years of performance. From Table 3.14, the majority of iterations used 6 past years to compute the most accurate predictions. This indicated that using additional past years led to more accurate predictions. This supports Healy's research that teams should be using multiple past years, instead of just one season, when predicting player performances and offering salaries (Healy 2008). For players undergoing arbitration contracts or with less than 6 years of service, teams should be using all past performances when predicting future years. Teams tend to weight a player's last year's performance twice as much as a player's performance from two and three years ago combined (Healy 2008). This is a major mistake since adding past years significantly improves the ability to predict a player's performance.

Table 3.14 shows that using only $1-3$ past decades produces the most efficient model. There were 19 iterations that only used one decade, 14 iterations that used the last two decades, and 11 iterations that used the last three decades. Thus, only players who played from the $1990's$, at the latest, should be used when predicting a player who played in the $2010's$. To increase the power of the model, using either 2 or 3 (if wanting to predict players performance at the beginning of the decade) past decades is optimal. Since I predicted players in the early $2010's$, I used the past 3 decades.

To predict player's whose $7^{th}$ season occurred between $2010-2015$, I used 6 past years and 3 past decades (since these players played at the beginning of the decade). The model I used to predict players is shown in Equation 4.1.

$$WAR_{t,i} = \beta_0+\beta_1 exp_i+\beta_2 exp_i^2+\beta_3 TL_i+\beta_4 BL_i+\beta_5 BR_i+\beta_6 WAR_{t-1}+...+\beta_{11} WAR_{t-6} \tag{4.1}$$

In Equation 4.1, $t$ represents the year of experience I wanted to predict, $exp$ is abbreviated for experience, and $TL, BL$ and $BR$ are dummy variables for Throws Left-handed, Bats Left-handed, and Bats Right-handed.

From Table 3.16, my predictions are roughly $\pm 2$ off from the actual WAR. The model is unexpectedly consistent over time as I would expect the model would be most accurate when I used the actual WAR values to predict the $7^{th}$ year WAR values compared to using the predictive WAR values to predict the $13^{th}$ year WAR values. This indicates that my model is able to predict a player's trend over time and is thus useful for baseball predictions for predicting a player's WAR throughout his career. However, since the estimate is on average $\pm 2$ off, this model should not be used to get an exact prediction.

## 4.6 Implications

Teams waste financial resources when they overestimate WAR values and believe players are more valuable than their actual worth. In contrast, other players may be underpaid if their WAR value and contribution to the organization is higher than predicted. Baseball teams would rather underpay their players and hope they perform better than expected, but this is difficult given the competition to sign quality players. My model evenly overestimates and underestimates WAR values for years $7-10$ but tends to overestimate WAR values for years after year 10. I expected the model to overestimate WAR values for players at the end of their careers since players are more likely to get injured due to age and the cumulative risk inherent in playing. This indicates that the model should be used with caution since it is not robust to factors such as injuries and aging. Further, job security may affect performance as players might not make the same effort once they are signed into a long-term contract (Dorfman and Kuehl 1989/1995/2002).

## 4.7 Limitations

This model requires players to have played at least 6 consecutive years. Thus, this model cannot predict players who had to miss a season due to an injury, illness, or were demoted to the Minor Leagues. This model cannot be used to predict players undergoing arbitration contracts since arbitration deals occur between $3-6$ years of experience in the MLB (Sievert 2018). This model can be modified by only using $n$ past years of experience (where $n$ is the number of years the player under consideration had played), but these predictions are not as accurate as using 6 years of experience.

The model is on average $\pm 2$ WAR from the average WAR and cannot accurately predict drastic performance changes. This model should not be used to predict exact WAR performances for players but used as an estimate on how a player will improve and decline throughout his career. Since the model is unable to capture the "random noise" (injury, mental game, etc...) and the predicted WAR is approximately $\pm 2$ from the actual WAR, it is unreliable for predicting WAR for a specific season but can be used to get the overall trend of a player's performance.

## 4.8 Future Approaches

I would be interested in determining a stabilization point for WAR. For this analysis, I chose 50 PA to ensure a large sample. However, I am not sure if a player's WAR value has stabilized after 50 PA. To determine this stabilization point, I would graph a player's WAR value after each PA. I expect that at some point the slope of the graph levels, indicating that the WAR value has stabilized and now accurately measures a player's value. I would repeat this process for multiple players and calculate the average PA when the slope for each player begins to level. It is important to remember that there is no finite threshold when the WAR value is susceptible to random noise, so the average I calculate is still an arbitrary number but is still more accurate than using 50 PA.

Ideally, I would want to determine if I could lower the cutoff value of group 4 to increase the number of players in group 4 and thus increase the power of my model. However, a drawback to this approach is that it would decrease the power of models for groups $1 - 3$ so predictions for these groups would not be as accurate. I would be interested in changing the cut-off value of group 4 to 5.00 since players with WAR values $\geq 5$ are considered to be superstars (Slowinski 2010e). I would run similar tests to those in this analysis and determine if this new group performed differently than other groups. If this group performed similarly, then it could be concluded that players with WAR values between $5 - 5.88$ perform differently over time than players with WAR values 5.88+ and thus should not be included in the top group.

Since I've incorporated the best aspects of other methods, my model should improve prediction of player performance. I would be interested in comparing my results to other projection systems such as steamer and ZiPS to determine whether my model better predicts performance. Confirmation that my model provides more accurate predictions would be a significant benefit to teams in developing contracts.

# 5. Conclusion

My model was constructed to help predict a player's performance trend for the rest of his career. It is consistent in predicting players' WAR as the average predicted WAR is approximately $\pm 2$ from the actual WAR for any specific year. Better players improve faster relative to others, so it is important to take this into consideration when constructing an arbitration agreement. When attempting to forecast future performance, it is best to use at least 6 years of information when signing free agents and all past years when constructing arbitration packages. When comparing players, it is helpful to exclude players who played more than 3 decades before the player under consideration since the game is constantly evolving. My model should help teams by providing additional information that will improve evaluation of a player's performance after seven years in MLB for the next four years. While I have not compared this study with other approaches, I hope this will provide teams and players with objective information that can be used to improve contract negotiations.

# References

Alley, Triples (2010). *Marcel and Forecasting Systems*. URL: https://triplesalley.wordpress.com.

*Baseball-References.com* (2019). data retrieved from Baseball-References, https://www.baseball-reference.com/data/.

Bierig, Brian, Jonathan Hollenbeck, and Alexander Stroud (n.d.). *Understanding Career Progression in Baseball Through Machine Learning*. URL: http://arXiv:1712.05754v1arXiv:1712.05754v1.

Dorfman, H.A. and Karl Kuehl (1989/1995/2002). *The Mental Game of Baseball: A Guide to Peak Performance*. Lanham, MD: Diamond Communications.

Drushcel, Henry (2016). *A guide to the projection systems*. URL: https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun.

Hakes, Jahn and Chad Turner (2011). "Pay, productivity and aging in Major League Baseball". In: *Prod Anal* 35, pp. 61–74. DOI: http://dx.org/10.1007/s11123-009-0152-8.

Healy, Andrew (2008). "Do Firms Have Short Memories? Evidence From Major League Baseball". In: *Journal of Sports Economics* 9 (4), pp. 407–424.

Humphrey, Ben (2014). *SLG vs. ISO: Allen Craig's Power Outage*. URL: https://www.vivaelbirdos.com/2014/2/20/5428860/slg-vs-iso-allen-craigs-power-outage.

Koop, Gary (2002). "Comparing the Performance of Baseball Players: A Multiple Output Approach". In: *Journal of the American Statistical Association* 97.459, pp. 710–720.

Lahman, Sean (2019). *SeanLahman.com*. data retrieved from Lahman's Baseball Database, http://www.seanlahman.com/baseball-archive/statistics/.

Lichtman, Mitchel (2009). *How do baseball players age? (Part 1)*. URL: https://tht.fangraphs.com/how-do-baseball-players-age-part-1/.

MLB (n.d.[a]). *Batting Average (AVG)*. URL: http://m.mlb.com/glossary/standard-stats/batting-average.

— (n.d.[b]). *Slugging Percentage (SLUG)*. URL: http://m.mlb.com/glossary/standard-stats/slugging-percentage.

Ng, Kevin (2017). "Analyzing Major League Baseball Players Performance based on Age and Experience". In: *Journal of Sports Economics Management* 7.2, pp. 78–100. DOI: https://doi.org/10.1177/1527002507310440.

Schultz, Richard et al. (1994). "The Relationship Between Age and Major League Baseball Performance: Implications for Development". In: *American Psychological Association* 9, pp. 274–286.

Sievert, Justin (2018). *Breaking down the MLB salary arbitration process*. URL: https://www.sportingnews.com/us/mlb/news/mlb-salary-arbitration-process-breakdown-spring-training-2016/4jkawqkczi8i17cb4rhqjxseh.

Slowinski, Steve (2010a). *ISO*. URL: https://library.fangraphs.com/offense/ops/.

— (2010b). *OBP*. URL: https://library.fangraphs.com/offense/obp/.

— (2010c). *OPS and OPS+*. URL: https://library.fangraphs.com/offense/ops/.

— (2010d). *Sample Size*. URL: https://library.fangraphs.com/principles/sample-size/.

— (2010e). *What is WAR?* URL: https://library.fangraphs.com/misc/war/.

— (2010f). *wOBA*. URL: https://library.fangraphs.com/offense/woba/.

Szymborski, Dan (2010). *The 2020 ZiPS Projections*. URL: https://blogs.fangraphs.com/the-2020-zips-projections/.

Tango, Tom M., Mitchel G. Lichtman, and Andrew E. Dolphin (2006). *The Book*. Newark, DE: TMA Press.

Trachtman, Seth (2019). *The worst contracts in MLB history*. URL: https://www.yardbarker.com/mlb/articles/the_worst_contracts_in_mlb_history/s1__28002203#slide_1.

*WAR Explained* (n.d.). URL: https://www.baseball-reference.com/about/war_explained.shtml.

Weinberg, Neil (2015). *Stats to Avoid: Batting Average*. URL: https://library.fangraphs.com/stats-to-avoid-batting-average/.

# Appendix

| **Abbreviations** | **Definitions** |
|---|---|
| 1B | single |
| 2B | double |
| 3B | triple |
| AB | at-bat |
| BB | walk |
| H | hit |
| HBP | hit-by-pitch |
| HR | home run |
| NIBB | non-intentional walk |
| PA | plate appearances |
| RBOE | reached base on error |
| SF | sacrifice fly |