VALUING MLB PLAYERS: OPTIMAL ALLOCATION OF SALARY

Jonah Lubin | Rice University

ABSTRACT

How should a salary cap be allocated; in terms of both assessing how much each player should get, as well as how much a position group should get? Using batting, fielding, base running, and pitching to determine the value of all MLB players.

BACKGROUND AND IMPORTANCE:

The allocation of salaries in Major League Baseball (MLB) is a very important decision-making process for teams. The MLB does not consist of a formal salary cap, as teams operate under some budgetary restrictions depending on the market size, ownership, and luxury tax thresholds. This difference that the MLB has that other sport leagues do not have influenced me to raise the question of how should these teams' financial resources be allocated among players and positional groups to optimize success?

Baseball itself is a complicated sport, given there are many ways that players can contribute to the team, such as batting, fielding, base running, and pitching. These four facets of baseball are not equivalent in its importance to the success of a team, as instinctively batting and pitching are the main two components. A framework that quantifies these components would allow teams to quantify these contributions to assess a player's worth as one value, which would then align with a salary figure. There are 26 players on each MLB roster, typically consisting of 13 batters and 13 pitchers. Deciding a value for every player is not the only aim of this research, as deciding an allocation for position groups is also a goal. In 2024, the top ten highest paid players consist of 5.5 pitchers and 4.5 batters (the .5's are due to Shohei Ohtani playing in both position groups). This top ten shows that teams slightly favor paying pitchers than batters, at least in terms of the ones that are viewed to be elite.

The topic of salary allocation is important in all sports, but there is a clear additional importance in the MLB, as teams with a lower payroll must be almost perfect with their salary allocation in order to compete with teams that have a high payroll. My goal is to use analytical methods to find the optimal position group payroll distribution, as well as a value for each MLB player.

DATA

The data that I used for batting, fielding, and pitching was play by play data, where each row is a plate appearance from the 2024 season. Columns consist of ids of the game, batter, pitcher, fielders, and base runners as well as the event that occurred on that play, and ball tracking data such as launch speed, launch angle, and spray angle, which are only available for balls that are contacted with. In addition to this data, I also used a linear weights data set, which has every event's linear weight.

There are several data sets I used to get player names and ids, as depending on the platform the data comes from, they give players different ids, so I used the Chadwick's Bureau's public register function in the baseballR package in R, as that had the ids from several platforms.

For the framing section, I used data that was on a pitch by pitch basis instead of a plate appearance basis. The columns I used from this data set were the balls x and z locations when it crosses the plate and the ball and strike count. For the stolen base section, I used the same data from the batting, fielding, and pitching sections, but only has rows for stolen base attempts. A data set with a base out run expectancy was used in conjunction with this. Finally, Lahman's Pitching and Batting data were used to get the expected plate appearances, games, and innings in order to scale players to a full season.

METHODOLOGY

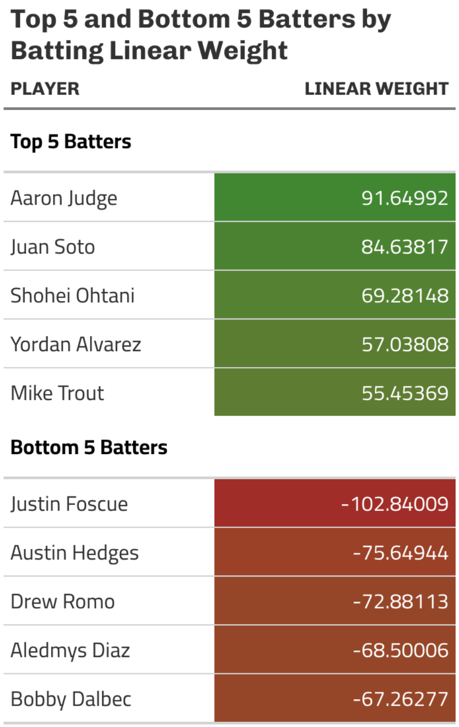
When I first started theorizing how I would accomplish my goal of giving every player a value, I wanted to put all players on the same metric scale, so I decided to put everyone on a run scale, so a combination of runs added and runs saved, depending on if it is an offensive or defensive statistic. Linear weights is a metric that allocates a value to every outcome of a plate appearance that is independent of the game context. For example, a double with bases loaded that scores 3 runs has the same linear weight as a double with no one on base that scores 0 runs. This metric, as opposed to a metric that is context dependent, has a greater accuracy in predicting future production statistics, so I decided to use this metric because it captures a player's value the most. Batting, fielding, base running, and pitching will all use linear weights, albeit in different ways.

I also wanted to make sure that every player would be placed on the same scale in terms of opportunities, so all batters would be scaled to the same number of plate appearances, fielders would be scaled to the same number of balls hit in play, base runners would be scaled to same number of games, and pitchers would be scaled to the same number of batters faced, depending on the type of pitcher they are.

BATTING

I first started with batting; there are many ways to incorporate linear weight in batting performance, so determining which method to use was very important. Using every player's total linear weight among the season, where all doubles would be given a certain value, all strikeouts would be given a certain value, etc., is better than context dependent metrics, but it is still not the best means to capture a player's value, as one batter's double could be another batter's fly out depending on the ball park, defense, weather, etc. Therefore, I decided to make use of the launch speed, launch angle, and spray angle metrics provided for all batted balls. Deciding which combination of those three metrics was based on the combination that produced the most accuracy in predicting future success, which was determined to be the combination that includes launch speed and launch angle, also referred as xLW2. I used a random forest model to predict the linear weight of each batter based on launch angle and launch speed and then allocated that linear weight to each batter. Since this only takes batted balls into account, I also got the linear weights of walks, strikeouts, and hit by pitches to give each batter a linear weight for balls not hit into play. These linear weights were then combined and divided by the number of plate appearances they had during the season.

As mentioned in the previous section, I want to scale all batters to the same number of plate appearances. After getting the linear weight per plate appearance, I decided to scale it to a season's length, so I needed to find out how many plate appearances a batter would have in a single season, which was determined as 675 for a player who would play every game in a season. Every player's linear weight per plate appearance was then multiplied by 675 to get the expected runs added to a team from batting in a season (Figure 1).



Top 5 and Bottom 5 Batters by Batting Linear Weight

PLAYER	LINEAR WEIGHT
Top 5 Batters	
Aaron Judge	91.64992
Juan Soto	84.63817
Shohei Ohtani	69.28148
Yordan Alvarez	57.03808
Mike Trout	55.45369
Bottom 5 Batters	
Justin Foscue	-102.84009
Austin Hedges	-75.64944
Drew Romo	-72.88113
Aledmys Diaz	-68.50006
Bobby Dalbec	-67.26277

Figure 1: The top 5 and bottom 5 batters based on expected linear weight in a season.

FIELDING:

The next aspect I wanted to include in this evaluation was fielding. Baseball is played with an offense and a defense, and for defense, although most of the responsibility is intuitively on the pitcher, there are 8 other fielders that control what happens on a plate appearance once the ball is contacted. If a ball is not put in play, the fielders have no responsibility on the play, so I only included plays in which a ball was hit into play. Depending on the launch angle, launch speed, and spray angle, the play will have different expected linear weights, as explained in the "Batting" section. In addition to it having different linear weights, it also changes the likelihood that a given player will get an out. A soft ground ball hit to third base has an extremely low likelihood that the right fielder will touch the ball and subsequently have responsibility for getting a runner out. Therefore, depending on the ball tracking metrics, each player is given a likelihood that they get an out on a play, and depending on if they do get the out or not are credited with runs saved or runs allowed, in which the runs saved and runs allowed would be calculated from linear weights. Similar to batting, all fielders were put on the same scale, but instead of plate appearances, it is balls in play. Every player was given a runs saved per ball in play, so then I needed to get how many balls in play a fielder can expect in a game, which equated to 24.4 a game, meaning 3953 balls in play in a full season, so I multiplied every player's runs saved per ball in play by 3953.

Catchers have an additional form of fielding, where they have framing, which is the act of moving the catching glove in a way that tries to convince the umpire to call a pitch a strike, regardless of if it is or not. The methodology behind this was to get the linear weight per ball strike count, as framing a ball on a 3-2 count has a lot more ramifications than on a 0-0 count. Depending on the x and z locations of the ball as it passes the plate, each pitch was given a likelihood of it being a ball or strike, and the expected value of a pitch was then given as the likelihood of a strike multiplied by the run expectancy if it was called a strike plus the likelihood of a ball multiplied by the run expectancy if it was called a ball. Every catcher was given a run saved per framing opportunity, which was then multiplied by 5863, which was the 75th percentile of all catchers' framing opportunities in 2024, as catchers do not play every game like most other positions can. Due to the context dependent nature and the fact that this process can vary depending on many external factors, I ran a mixed effects model to get the signal and noise variances of framing depending on the catcher. This was then incorporated to regress them to the mean, making an adjusted runs saved for all catchers for a season.

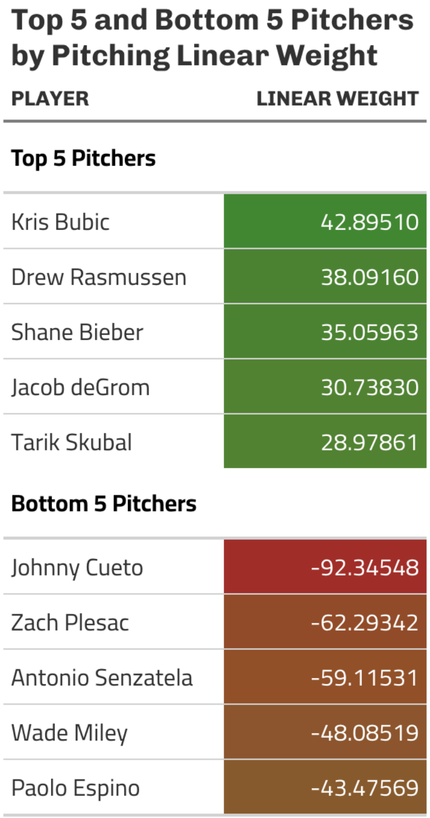
BASE RUNNING:

The last component of batter production is base running, specifically stolen bases. Base out run expectancies were very important in this section because a stolen base has different run values depending on if there are runners on base and the number of outs. When a batter steals a base where the run expectancy changes by .4 runs, they are attributed with .4 runs added, but if they get caught stealing and the run expectancy changes by .7 runs, then they are charged with -.7 runs added. I then scaled these to runs added per game from stolen base attempts and multiplied that by 162 games.

PITCHING:

For pitching, I decided to mirror the process that I did with batting, where I used xLW2 to estimate the number of runs given up on a batted ball and attributed that to the pitcher instead of the batter. Because that is only for balls in play, I included the linear weight for strikeouts, walks, and hit by pitches for balls not in play. These were then put on a per plate appearance scale to get linear weight per plate appearance.

The difference between pitching and batting is that relief pitchers and starting pitchers aren't going to have the same number of plate appearances in a season. I calculated that relief pitchers typically play 32 games per season and starting pitchers play 25 games per season. I also calculated the number of batters faced per game for each player, as great pitchers might go 5 innings facing only 20 batters while some bad pitchers might go 5 innings facing 30 batters, so that cannot be fairly scaled. A pitcher's expected linear weight per season is their linear weight per plate appearance multiplied by the batters they face per game, then multiplied by 32 games for relievers or 25 games for starters.



Top 5 and Bottom 5 Pitchers by Pitching Linear Weight

PLAYER	LINEAR WEIGHT
Top 5 Pitchers	
Kris Bubic	42.89510
Drew Rasmussen	38.09160
Shane Bieber	35.05963
Jacob deGrom	30.73830
Tarik Skubal	28.97861
Bottom 5 Pitchers	
Johnny Cueto	-92.34548
Zach Plesac	-62.29342
Antonio Senzatela	-59.11531
Wade Miley	-48.08519
Paolo Espino	-43.47569

Figure 2: The top 5 and bottom 5 pitchers based on expected linear weight in a season.

POSITION GROUP PAYROLL ALLOCATION

Now that I calculated all the components of player production, I can combine them all into one metric because they were all scaled to runs added per season. For batters, I took their batting linear weight minus the runs allowed fielding minus the adjusted runs allowed framing for catchers plus the runs added from stolen bases, and for pitchers it is simply their pitching linear weights.

In order to calculate how much each position group should be paid as a whole, I wanted to see the distribution of each position group's linear weight (Figure 3). This graph shows a wider distribution for the batters, meaning they have greater variability in their production, leading to a stronger emphasis on paying them more than pitchers, as the difference between a good and bad batter seems to be way more apparent than a good and bad pitcher. The positional group payroll allocation formula I used was the average absolute value of the position group's linear weight times divided by the total absolute value of both position groups' linear weight, which resulted in 73% of a payroll going to batters and 27% going to the pitchers, reflecting the pattern shown in Figure 3.

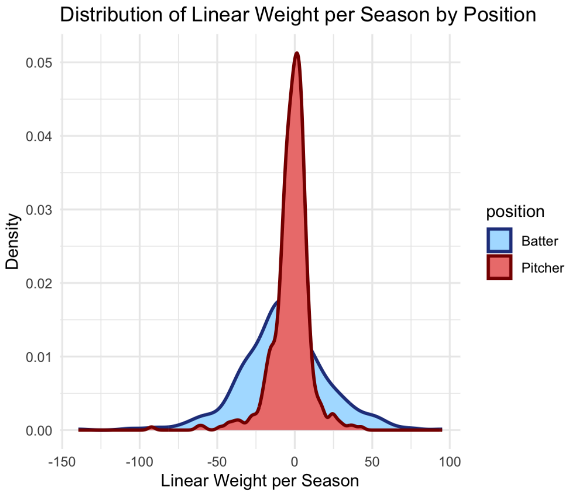


Figure 3: The distributions of linear weight by position group.

INDIVIDUAL PLAYER PAYROLL ALLOCATION

My first thought in calculating the value of every player was that the players needed to be given values based on how they compared with each other. For example, if most players had around a linear weight of 50, then a linear weight of 60 is not that impressive, but if most players had around a linear weight of -10, then a linear weight of 60 is elite. The issue with this current data set of all players is that some players have extremely negative values, such as Johnny Cueto in Figure 3, which then throw off the scaling portion because, in theory, the worst player in the data set will still have to be paid the minimum salary, but I decided that any player outside the top 390 players at that position group, which is 13 players per team times 30 teams, should be assumed to be out of the league due to poor performance. The distributions of the linear weights for each position group of only the top 390 players still showcase a larger variability in batters than in pitchers (Figure 4).

I then wanted to shift every player's linear weight to be on a scale that does not include negative numbers, so it would be easier to assign a dollar value; therefore, I shifted their linear weights to be their original linear weight minus the minimum linear weight of the position group, which were negative in both instances.

Due to the fact that not all teams have the same payroll, I used the total MLB payroll as the baseline for salary allocation. Since we determined that batters should be paid 73% of a team's payroll, and pitchers should be paid 27%, I multiplied those percentages by the MLB payroll of roughly 5 billion dollars to get each position group's total money allocated to its top 390 players.

Each player's salary would then be calculated as their shifted linear weight divided by the total linear weight for that position group multiplied by the total calculated payroll for that position group. The top end batters are paid rather substantially more than the top end pitchers (Figures 5 & 6).

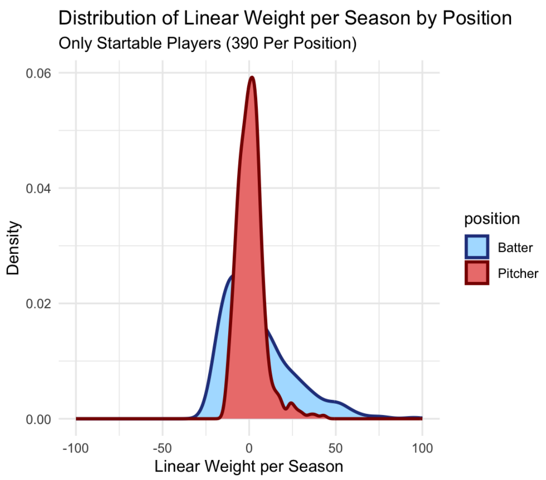
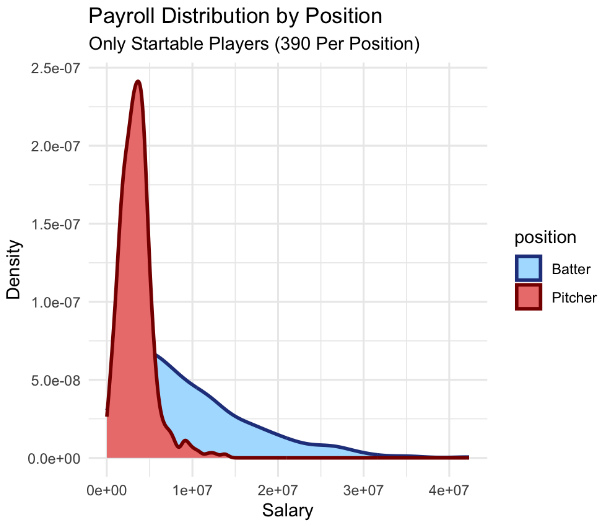
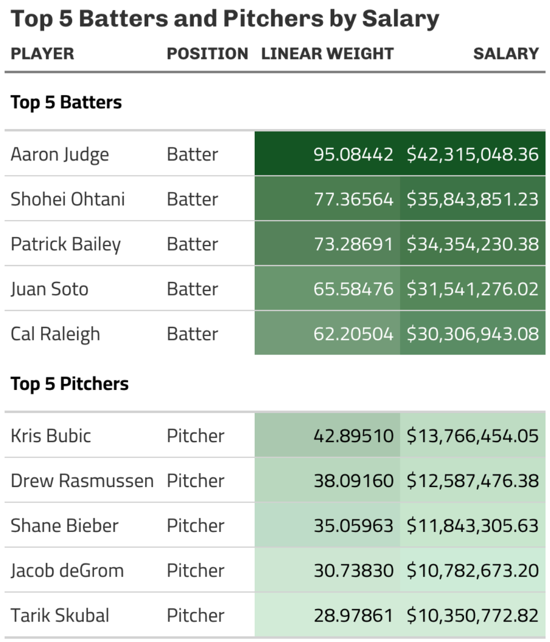


Figure 4: Distribution of linear weight by position group for top 390 players per position group





Top 5 Batters and Pitchers by Salary

PLAYER	POSITION	LINEAR WEIGHT	SALARY
Top 5 Batters			
Aaron Judge	Batter	95.08442	$42,315,048.36
Shohei Ohtani	Batter	77.36564	$35,843,851.23
Patrick Bailey	Batter	73.28691	$34,354,230.38
Juan Soto	Batter	65.58476	$31,541,276.02
Cal Raleigh	Batter	62.20504	$30,306,943.08
Top 5 Pitchers			
Kris Bubic	Pitcher	42.89510	$13,766,454.05
Drew Rasmussen	Pitcher	38.09160	$12,587,476.38
Shane Bieber	Pitcher	35.05963	$11,843,305.63
Jacob deGrom	Pitcher	30.73830	$10,782,673.20
Tarik Skubal	Pitcher	28.97861	$10,350,772.82

Figure 5: Distribution of payroll distribution by position group

Figure 6: Top 5 batters and pitchers by their salary.

CONCLUSION

Ultimately, this research shed light on the idea that batters should be prioritized in payroll allocation given their much higher variability of linear weight. The optimal allocation of teams' payrolls have yet to be fully taken advantage of in the MLB, meaning that there is still room for improvement in this field.