Constructing Optimal MLB Teams with Linear Programming

Nathan Schor

August 28, 2022

Abstract

Major League Baseball teams face trade-offs in assembling talented rosters without overspending. Teams need to make difficult choices about which players to roster, how many players to keep at each position, and how much to pay those players. This paper provides insights into managing these trade-offs thorough setting up a constrained optimization problem that seeks to maximize total team performance, subject to player salary and position constraints.

Keywords— MLB, Constrained Optimization, Roster Construction

Contents

1 Introduction

The purpose of this research is to investigate the relationship between salary and team performance in Major League Baseball (MLB). We do this by looking at actual outcomes for MLB teams in 2021 and compare those results to optimal rosters constructed via linear programming. Our goal is to analyze the gap between current team performance and their potential optimal performance.

We begin with 2021 salary data for each MLB team from Brown (2021). This data is supplemented with each player's 2021 salary, team, position, and JEFFBAGWELL (our performance metric, abbreviated JB). JB stands for **J**oint **E**stimate **F**eaturing **F**anGraphs and **B**-R **A**ggregated to **G**enerate **W**ar, **E**qually **L**eveling **L**ists and is provided by Pane (2021). WAR is a metric that quantifies each player's value in terms of how many wins they provide MLB (nd). Next, we seek to maximize JB for each of the 30 teams subject to salary and player position constraints, and visualize the results. Lastly, we discuss the implications of the gap between teams' current roster and their optimal roster.

2 Literature Review

The relationship between spending and performance is a topic of discourse among both casual fans and diehard fans, such as members of the SABR[1] baseball community. The book *Moneyball* by Lewis (2003) launched the baseball analytics revolution. One example of using linear programming in baseball is McIntyre et al. (2016). Another example of linear programming in baseball is Adler et al. (1999). An instance of linear programming in other sports is Aramouni (2021).

[1]SABR is the Society for American Baseball Research, an organization dedicated to discovering objective baseball knowledge `https://sabr.org/`

3 Methodology/Data

3.1 Data Cleaning

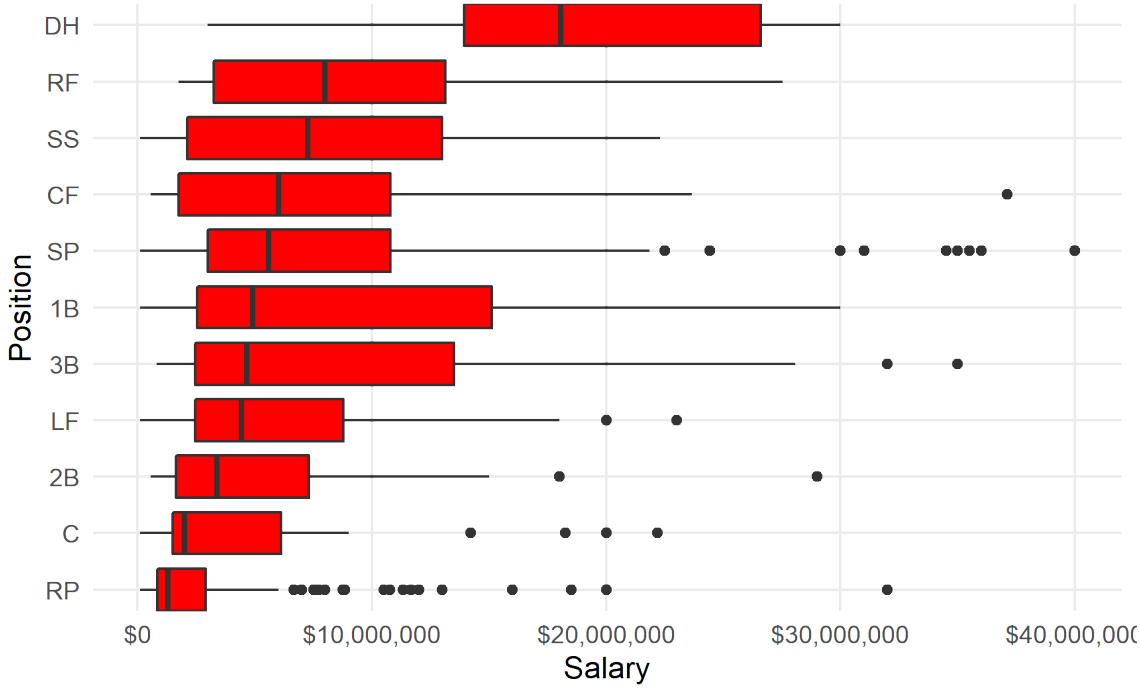
To clean the player dataset, we first assign positions to each player. A player's position is recorded as the position they played most frequently during the season (if they played two positions an equal amount of times, their position was randomly assigned to one of the two positions). Players with a $0 salary or a missing salary are removed. A pitcher is classified as a Starting Pitcher (SP) if $\geq$ 50% of their appearances were as a starter, and otherwise are classified as a Relief Pitcher (RP). Furthermore, First Basemen (1B), Second Basemen (2B), Third Basemen (3B), and Shortstops (SS) are classified as Infielders (IF). Left Fielders (LF), Center Fielders (CF), and Right Fielders (RF) are classified as Outfielders (OF). Catchers (C) and Designated Hitters (DH) are left as their own individual categories.

3.2 Salary and JEFFBAGWELL by Position

Baseball teams need to make important decisions about which players to have on their team, and how much to pay them. An interesting aspect is that each position is not equally valuable to creating a winning team, and, consequently, players at different positions have varying salaries (controlling for skill level).

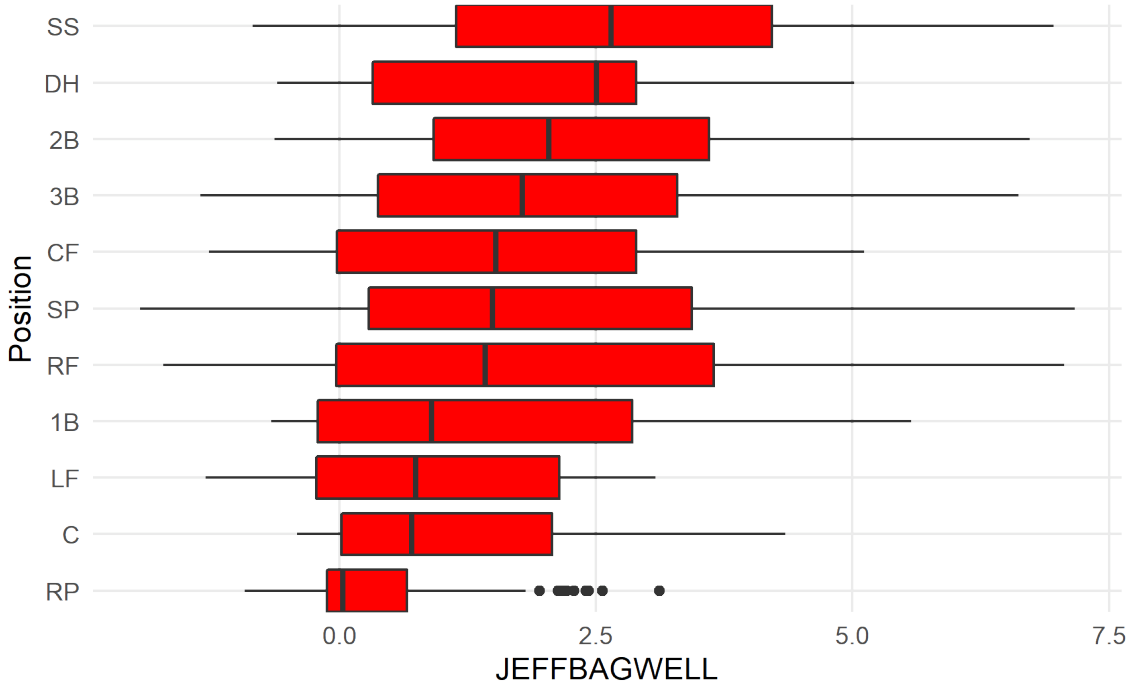
In Figure 1, we see how salary varies by position. The median salary for a Starting Pitcher is around $5,000,000 while the median salary for a Relief Pitcher is closer to $1,000,000. The salary distributions also vary drastically within infield and outfield positions. Starting Pitchers and Relief Pitchers also have the largest number of outliers.

Figure 1: Boxplot of Salary (for Players with Salary > $0) by Position



It is interesting to compare this with Figure 2. C and RP have the lowest median values in both. However, IF and OF positions are much closer together in the JB plot. This suggests a potential discrepancy between how much players are valued and how much they are paid–if players were perfectly paid according to their value, we would expect the position ordering on the y-axis of the two graphs to be identical.

Figure 2: Boxplot of JEFFBAGWELL (for Players with Salary > $0) by Position



3.3 Solving the Optimization Problem

We address the discrepancy between player performance and player salary by solving a constrained optimization problem. The decision variables (x_i) are which MLB players will be selected for each team T. x_i is a binary variable that equals 1 if player i is chosen, and 0 if they are not. The objective function is to maximize the JB $\forall T \in \{1, 2, \ldots, 29, 30\}$:

$$\sum_{i=1}^{N} x_i * JB_i \tag{1}$$

where N is the total number of eligible players in 2021 (547 players)

subject to the following constraints:

$$\sum_{i=1}^{N} x_i = 25 \tag{2}$$

$$\sum_{i=1}^{N} x_i = 5 \quad \forall x_i \in SP \tag{3}$$

$$\sum_{i=1}^{N} x_i = 7 \quad \forall x_i \in RP \tag{4}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in CF \tag{5}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in RF \tag{6}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in LF \tag{7}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in 2B \tag{8}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in 3B \tag{9}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in 1B \tag{10}$$

$$\sum_{i=1}^{N} x_i \geq 1 \quad \forall x_i \in SS \tag{11}$$

$$\sum_{i=1}^{N} x_i = 2 \quad \forall x_i \in C \tag{12}$$

$$\sum_{i=1}^{N} x_i = 1 \quad \forall x_i \in DH \tag{13}$$

$$\sum_{i=1}^{N} x_i = 5 \quad \forall x_i \in IF \tag{14}$$

$$\sum_{i=1}^{N} x_i = 5 \quad \forall x_i \in OF \tag{15}$$

$$\sum_{i=1}^{N} x_i * x_{salary} \leq T_{salary} \tag{16}$$

Equation (2) constrains each team to have exactly 25 players. Equations (3) -

(15) stipulate the number of players at each position and the total number of players allowed for grouped positions. (16) requires that each team spend no more on players than they did in the actual 2021 season.

4 Computational Experiment and Results

We solve the constrained optimization problem for all 30 teams to generate each team's optimal 25 player roster. In Table 1, we see there are 5 players who are chosen for each team, and 18 players who are chosen more than 20 times. In Table 2, we see that teams have at most 3 players who were on both their actual and optimal teams. The modal amount of players to have on both the optimal and actual team is 0.

In Figure 3, we sum the salary and JB for each of the actual 30 teams' rosters' in the graph on the left, and we sum the salary and JB for each of the 30 optimized teams' rosters' on the right. Each point represents a team.

Figure 3: Scatterplots of a Team's Total JEFFBAGWELL vs. Total Dollars Spent for the Actual Team (Left)and Optimal Team (Right). The Red and Blue lines are constructed using LOESS. Note the difference in magnitude of the 2 y-axes

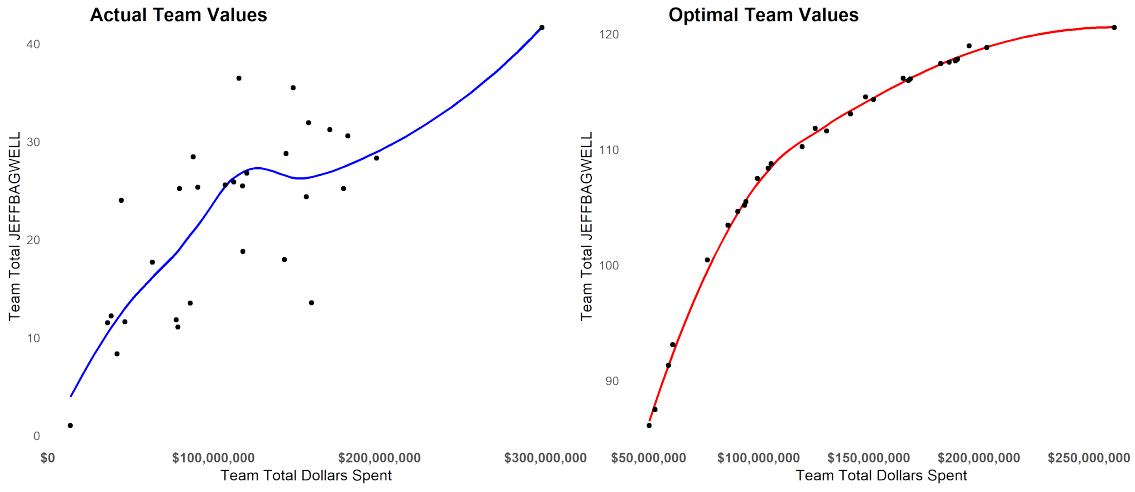


Table 1: Number of teams (max of 30) selecting a given player for their optimal roster

Player	Teams Picked
Brandon Woodruff	30
Fernando Tatis Jr.	30
Mike Zunino	30
Shohei Ohtani	30
Walker Buehler	30
Andrew Chafin	29
Carlos Rodon	29
Juan Soto	29
Matt Olson	28
Jesse Winker	27
Jose Ramirez	26
Aaron Loup	25
Carlos Correa	25
Jacob Stallings	25
Aaron Judge	24
Josh Hader	23
Kendall Graveman	23
Enrique Hernandez	22
Chad Green	19
Robbie Ray	17
Liam Hendriks	16
Ryan Tepera	16
Brandon Lowe	15
Marcus Semien	15
Zack Wheeler	14
Blake Treinen	12
Freddy Peralta	12
Adam Cimber	11
Julio Urias	11
Brad Boxberger	10
Harrison Bader	10
Joey Gallo	9
Bryce Harper	8
Luke Jackson	6
Teoscar Hernandez	6
Tyler Mahle	5
C.J. Cron	4
Caleb Thielbar	4
JT Chargois	4
Luis Robert	4
Mark Melancon	4
Ryan Pressly	4
Salvador Perez	4
Byron Buxton	3
Jeimer Candelario	3
Kyle Schwarber	2
Michael A. Taylor	2
Michael Fulmer	2
Starling Marte	2
Tony Kemp	2
Adam Engel	1
AJ Pollock	1
Buster Posey	1
Craig Kimbrel	1
Frankie Montas	1
Joey Wendle	1
Jorge Polanco	1
Luis Cessa	1
Max Scherzer	1

Table 2: For each team, the number of players selected for their optimal team who are also on their actual team

Team	Number of Players
LAD	3
MIL	3
CHW	2
HOU	2
NYY	2
PHI	2
TBR	2
BOS	1
CHC	1
LAA	1
NYM	1
OAK	1
PIT	1
SDP	1
SEA	1
TEX	1
TOR	1
WSN	1
ATL	0
SFG	0
KCR	0
STL	0
MIN	0
CLE	0
CIN	0
MIA	0
BAL	0
COL	0
ARI	0
DET	0

5 Discussion and Conclusions

5.1 Actual vs. Optimal Plots

The shape of the two graphs in Figure 3 is interesting. The points on the left are relatively scattered, but with a general upward trend–JB and dollars spent are positively correlated. With the smoother, we see that there is a roughly linear relationship up until the salary hits $100,000,000. JB then *decreases* up until $150,000,000, where it then continues to increase in a concave upwards fashion.

Nearly all the points on the right fall along a single curve that is almost entirely non-decreasing, and roughly linear up until around $200,000,000. Some takeaways from examining these two curves:

- The marginal value of a player is non-constant
- Almost all teams that are at the bottom of the spending distribution would benefit from spending to acquire higher JB players, but additional spending is not worthwhile for teams at the top of the distribution
- Teams can use their current status (left curve) and their optimal status (right curve) to move to a more desirable point on the JB-Salary curve

5.2 Actual vs. Optimal Table

It is surprising how few players are on the same actual and optimal team in Table 2. This table gives a rough sense of bargain star power for each team. Assuming that players are paid what they are worth (which often does not happen until free agency), the high costs of having an elite player on the team might not offset the increase in JB. Teams are able to have the most desirable combination of high JB and low cost by developing or acquiring players *before* they have attained star level, since they then pay a premium once the player is an established star.

6 Limitations

There are a number of limitations for this work. The metric JB is an imperfect measure of player performance. Teams know *ex-ante* how much they will pay a player

for the season, but they do not know what the player's performance will be prior to the season. We also use data only for the 2021 season and thus only have 30 data points for teams. It is possible that the relationship between JB and salary varies by season, and that is left as an area for future work. Furthermore, we have assumed that all players are currently available on the free agent market. This assumption is clearly unrealistic since most of the players who are selected for the optimal teams are currently under contract with particular MLB teams.

References

Adler, I., Erera, A. L., Hochbaum, D. S., and Olinick, E. V. (1999). Baseball, optimization and the world wide web.

Aramouni, N. (2021). Playing moneyball: Creating an efficient nba team with linear programming. https://towardsdatascience.com/playing-moneyball-creating-an-efficient-nba-team-with-linear-programming-ef14f6383861.

Brown, M. (2021). 2021 mlb final player payrolls show $168 million drop from last full season; here's every team's number. https://www.forbes.com/sites/maurybrown/2021/12/22/2021-mlb-final-player-payrolls-show-168m-drop-from-last-full-season-heres-every-team/?sh=674b3e663999.

Lewis, M. (2003). *Moneyball*. W.W. Norton & Company.

McIntyre, J., Sharma, M., and Khitalishvili, K. (2016). Linear optimization and baseball teams. https://rpubs.com/Koba/linear-opt-baseball.

MLB (n.d.). Wins above replacement (war). https://www.mlb.com/glossary/advanced-stats/wins-above-replacement.

Pane, N. (2021). Mlb historical war data. https://github.com/NeilPaine538/MLB-WAR-data-historical.

7 Appendix

Figure 4: Histogram of Salary for Players with Salary > $0

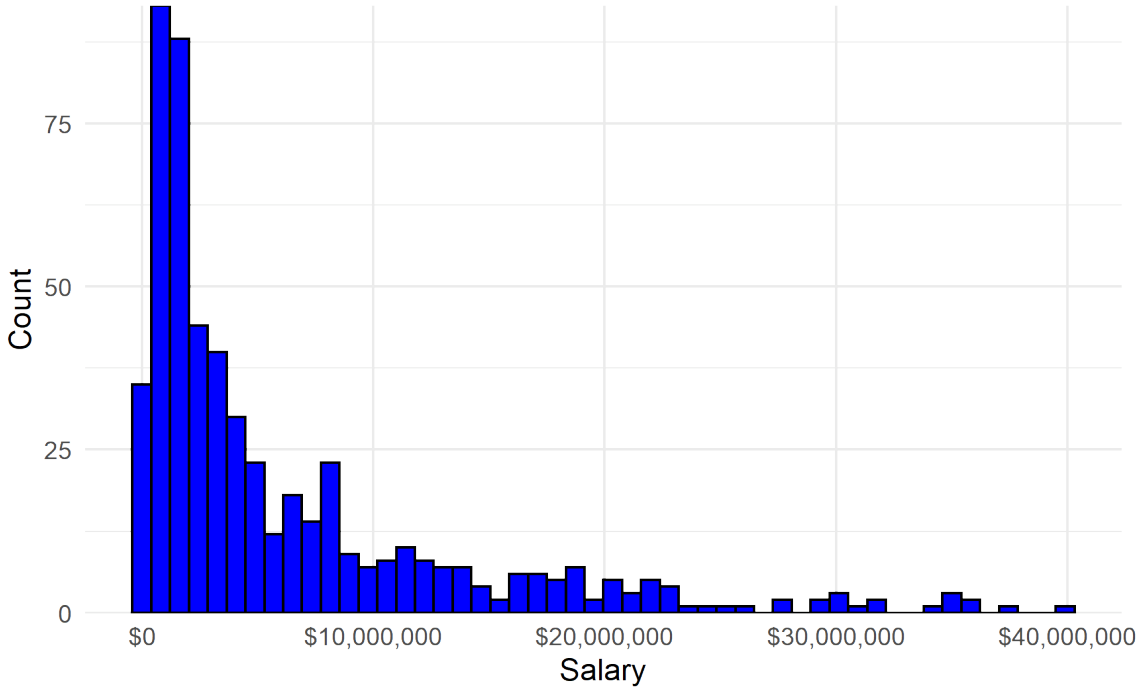


Figure 5: Histogram of JEFFBAGWELL for Players with Salary > $0