Analyzing and Modeling MLB Draft Biases using Wins Above Replacement

STATS 141XP Final Project

Peter DePaul, Anish Ravilla, Robin Lee
Kevin Kim, Alan Wong, Honye Zheng

2024-03-07

Abstract

[Abstrast here]

Contents

1 Introduction

In the world of professional baseball—specifically Major League Baseball (MLB)—the First Year Player Draft is a significant moment where teams meticulously evaluate and select amateur baseball players from high schools, colleges, and other amateur baseball clubs. However, even though the selection process should technically be based on objective and impartial reasoning, we are confident that various multifaceted biases subtly influence the decisions made by scouts and team executives (Caporale and Collier 2013).

1.1 Background

The baseball draft is a process in which the MLB's 30 professional baseball franchises gather together to draft amateur players, which mainly consist of players from the high school and college circuits (Staudohar, Lowenthal, and Lima 2006). Teams take turns picking players for their roster. Keep in mind that some players are drafted more than once. In this case, players are usually drafted out of high school, but they decide not to sign and instead play college baseball; they re-enter the draft after completing college.

Teams are awarded draft picks based on a **draft lottery**, with the teams that did not make the playoffs the previous year being entered into the lottery, and are given the opportunity to go first in the draft. The draft currently undergoes 20 rounds of selections, where each team gets to pick a player of interest for their roster. This draft, occurring in June, is also known as the **First-Team Player Draft**, where players who enter are typically high school or community college/four-year college graduates, and have never played for any professional baseball team (Garmon 2012).

1.2 Variable Overview

By implementing statistical data analysis, we seek to investigate how a scout's perception of talent and potential is influenced by a prospective player's attributes—such as their age, position, dominant hand, and educational background (Conforti, Crotin, and Oseguera 2022). Furthermore, in order to quantify the success of drafted players, we utilize their **Wins Above Replacement (WAR)**. As such, we hope to identify potential biases that influence when a player is drafted, and accurately determine whether or not these decisions are meritorious based on said player's future performance (Crotin et al. 2023).

We accessed our data from Bill Petti's BaseballR package:

Label	Description	Unit of Measure
fg_playerID	Player ID	Numeric
Name	Player name	Character
fWAR	Wins Above Replacement	Numeric
pick_round	Which round a player is picked	Numeric
pick_number	Which number a player is picked	Numeric
year	Year a player is picked	Numeric
person_birth_state_province	State or province the player is born	Character
person_height	Height of player	Numeric
person_weight	Weight of player	Numeric
person_primary_position_abbreviation	Player's primary position	Character
person_bat_side_code	Player's batting side (L/R)	Binary
person_pitch_hand_code	Player's pitching hand (L/R)	Binary
mlb_played_first	Year of first MLB game	Numeric
mlb_played_last	Year of last MLB game	Numeric
high_school	Player went to high school	Binary
home_state	Player's home state	Character

Table 1: Variable Overview

2 Exploratory Data Analysis

2.1 Relationship between drafting High School or College Players

In order to analyze the relationship between draft round and high school status, we created a barplot showing the total number of draftees by round with colors denoting high school status when drafted. By analyzing this barplot [see Figure 1], we can see that as you get further into the first 10 rounds that siginficantly less high school players are drafted.

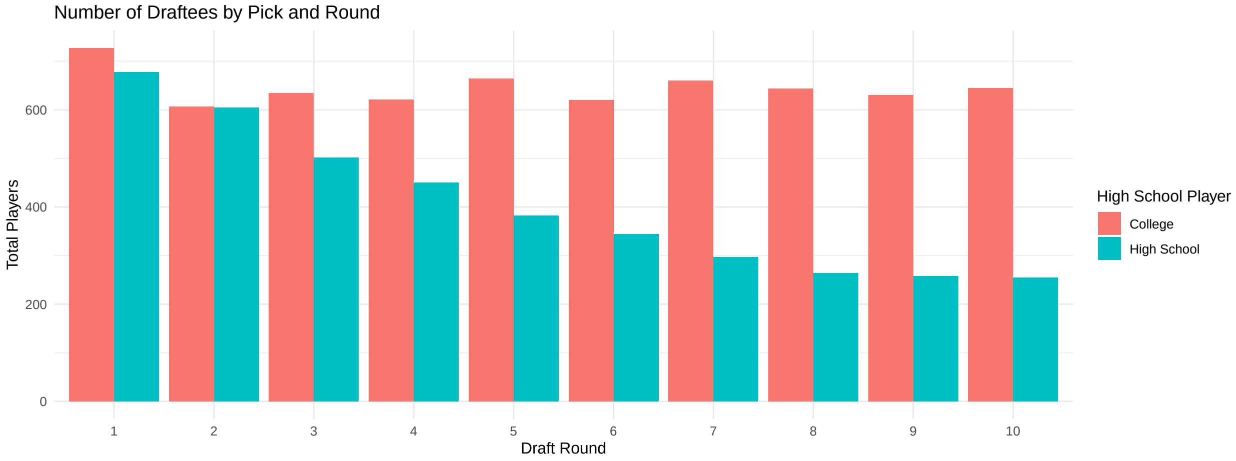

Figure 1: Frequency of High School Players by Round

2.2 Histogram of Time to Play First MLB Game

In order to analyze the number of years before a draftee played their first MLB game, we calculated the number of years between a player's first MLB game and the year they were drafted.

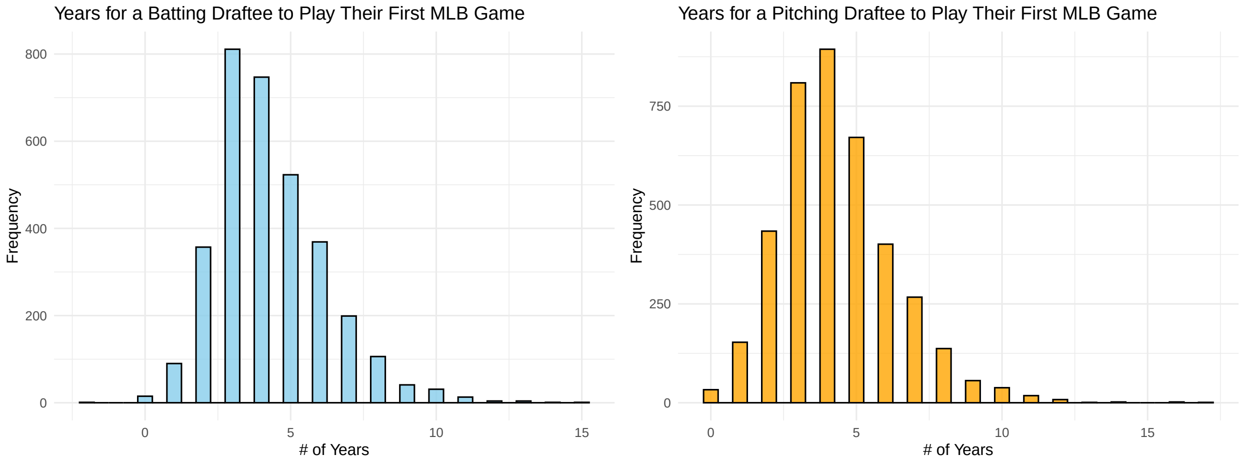

Figure 2: Histograms of Years before First MLB Game

By analyzing these two histograms [see Figure 2], the number of years for a batting draftee to play their first MLB game is similar to the number of years for a pitching draftee. Both datasets are slightly right-skewed, with most people playing their first game between 3-5 years after their draft.

2.3 Relationship between Weight and Height

We also investigated the relationship between weight and height for both batting and pitching draftees. By creating a scatterplot of both datasets, we not only mapped each individual person's attributes, but also plotted the mean weight/height and performed a simple linear regression to determine the overall trend.

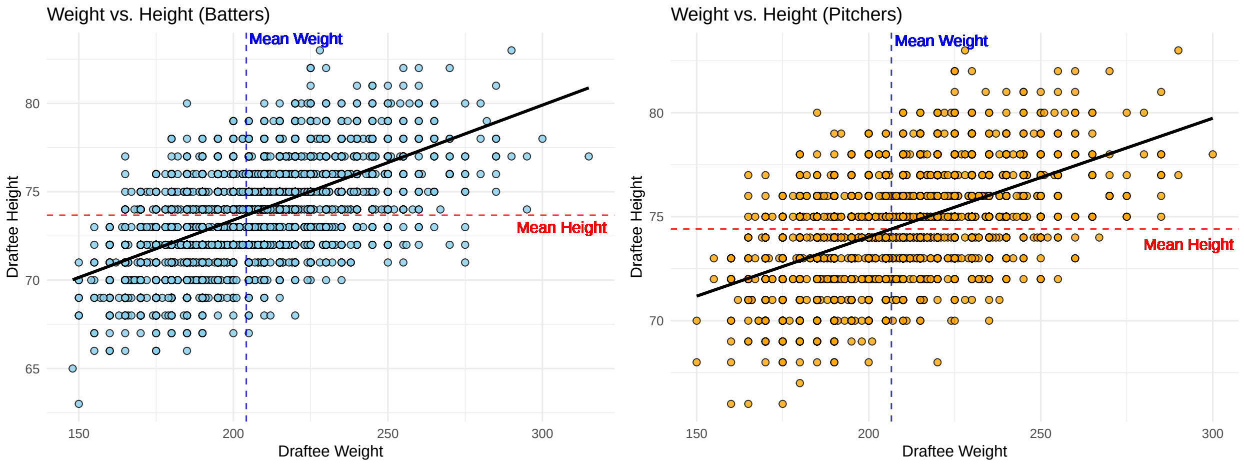

Figure 3: Scatterplots of Mean vs. Weight

Comparing these two plots [see Figure 3], the mean weight for batting draftees is slightly smaller than the mean weight for pitching draftees. Likewise, the mean height of batting draftees is also slightly smaller than the mean height of their pitching counterparts. We can expect the average batting draftee to be slightly shorter and lighter than the average pitching draftee.

2.4 Correlation Heatmap

The correlation between pick round, year, height, weight, and WAR for both player categories is:

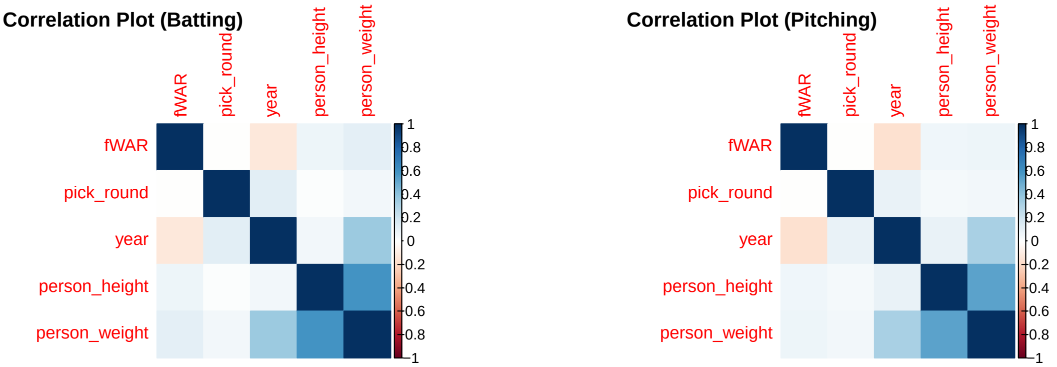

Figure 4: Correlation Plots for Pitching and Batting

As expected [see Figure 4], there is a strong correlation between a person's height and their weight. For pitching, there is a small positive correlation between a player's height and their WAR; for batting, there is a negative correlation between a player's height and their WAR. We see there is no correlation between a batting draftees' weight and WAR, but there is a slight positive correlation for a pitching draftee.

2.5 Home State Frequency Heatmap

In order to investigate which state produces the most draftees, we construct a frequency heatmap:

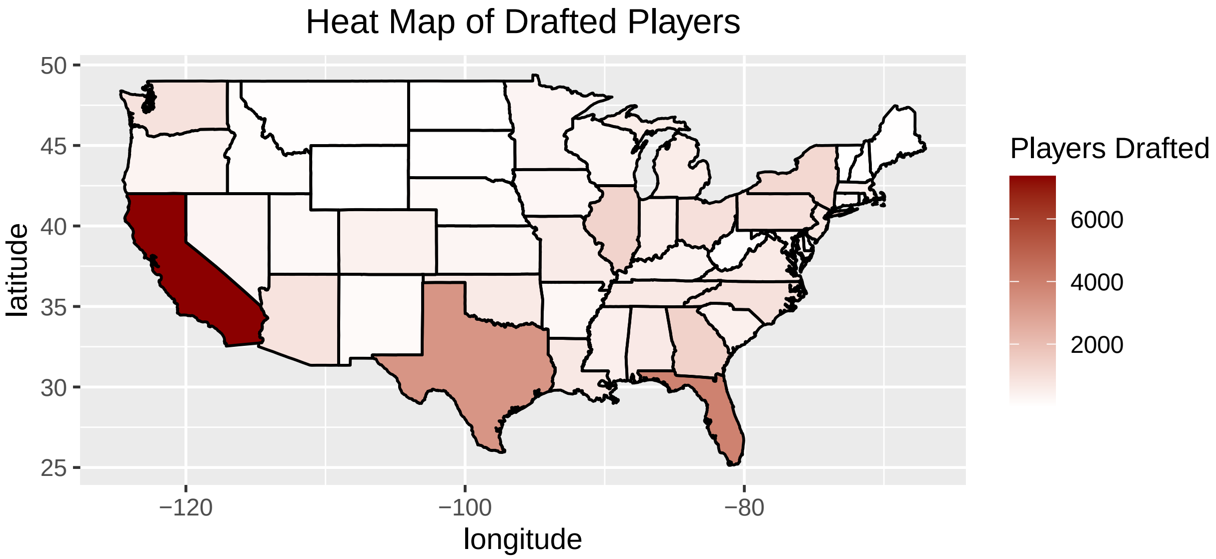

Figure 5: Frequency Heatmap of Players per State

From our state map [see Figure 5], we can see that most players drafted are originally from California, with Texas and Florida as the other two primary home states. We can attribute this trend to weather, as these states tend to have higher temperatures and little-to-no snow compared to East, Midwest, and Northwest.

2.6 Relationship between Time in MLB and WAR

Exploring how the length of player's career affects their WAR, we construct the following scatterplot:

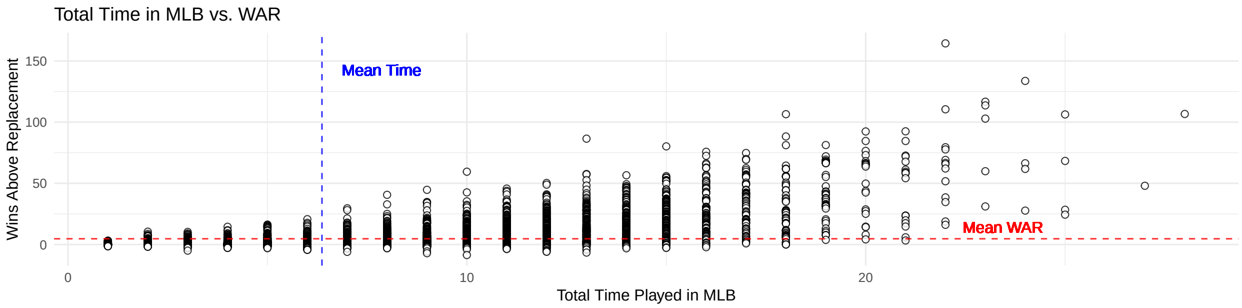

Figure 6: Relationship between Time Played in MLB vs. WAR

By comparing each player's total time played in MLB and WAR [see Figure 6], we can see that the length of a player's career does not always equate to a high WAR. However, there does seem to be an upward trend, which makes sense as we expect a player to win more as they play longer.

3 Modeling and Analysis

3.1 Who is Successful?

3.1.1 How often do Draftees make it to the MLB?

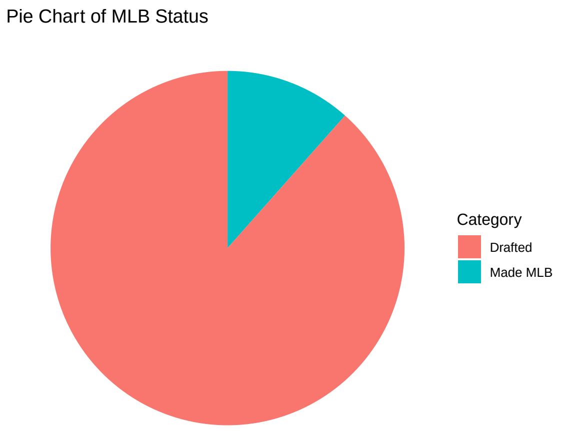

Figure 7: Pie Chart of Draftees MLB Status

3.1.2 What is the Best Draft Class in MLB History?

We can see from the barplot that the 1965, 1985, and 2002 draft classes are highly successful with each garnering over 900 total fWAR across the draft class. 1965 included the likes of Johnny Bench, Nolan Ryan, and Tom Seaver. Meanwhile 1985 includes Barry Bonds, John Smoltz, and Randy Johnson. 2002 includes Zack Greinke, Prince Fielder, and Cole Hamels. The thing these drafts share in common is that they have an abundance of talent, both in terms of Hall of Famers and depth.

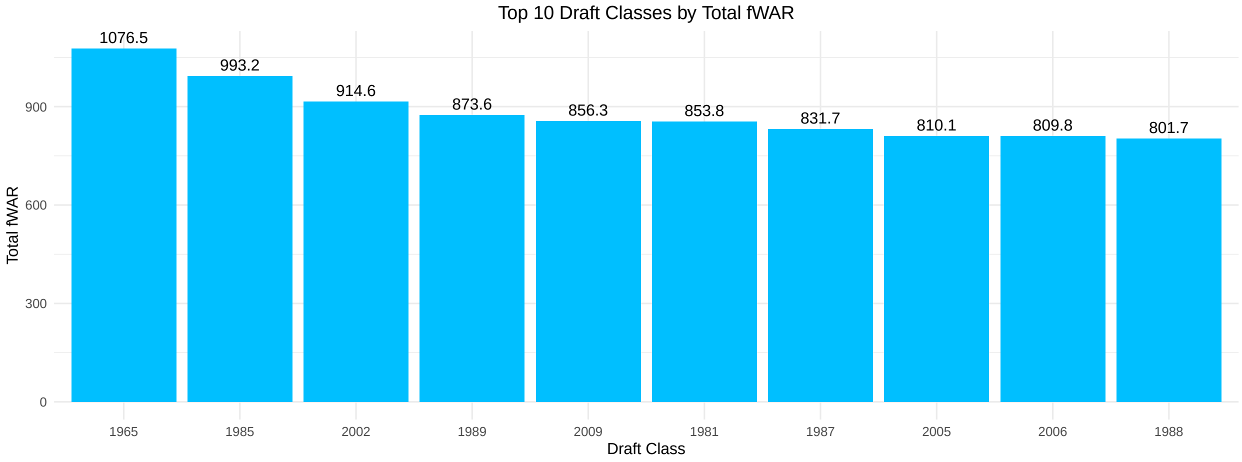

Figure 8: Best Draft Classes by fWAR

3.1.3 What Positions have been Successful?

We are able t osee that the positions which are most successful have often been outfielders (CF or LF), First Basemen (1B), or Third Basemen (3B). This can be explained by First Basemen possessing game changing amount of power hitting, and the high level fielding roles of outfield and third base.

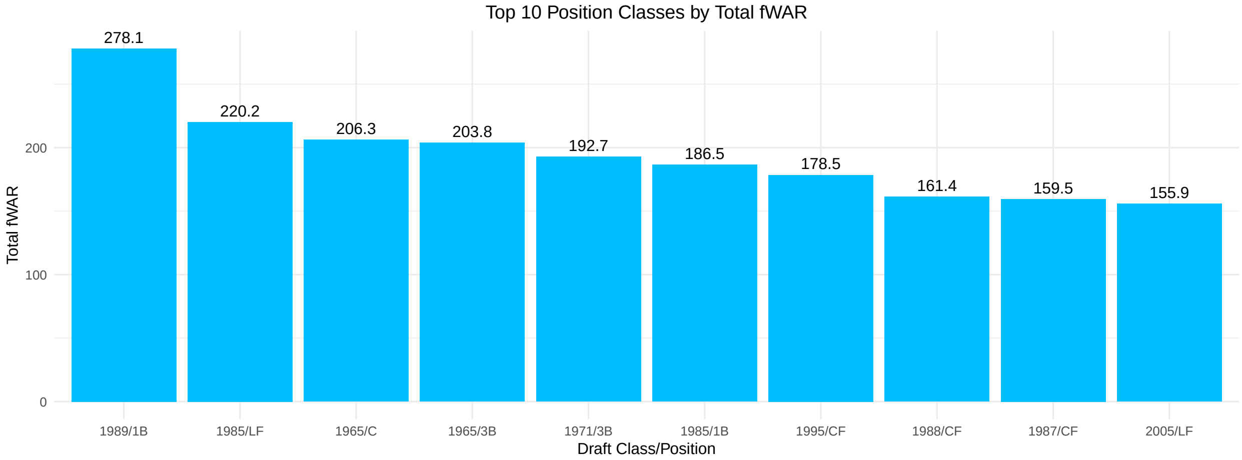

Figure 9: Best Draft Classes by Year and Position

3.2 Possible Model

By defining each player's WAR as the definition of success,

4 Results

5 Conclusion

Bibliography

Caporale, Tony, and Trevor C. Collier. 2013. “Scouts Versus Stats: The Impact of*Moneyball*on the Major League Baseball Draft.” *Applied Economics* 45 (15): 1983–90. https://doi.org/10.1080/00036846.2011.641933.

Conforti, Christian M., Ryan L. Crotin, and Jordan Oseguera. 2022. “Major League Draft WARs: An Analysis of Wins Above Replacement in Player Selection.” *Journal of Sports Analytics* 8 (1): 77–84. https://doi.org/10.3233/jsa-200586.

Crotin, Ryan L., Christian M. Conforti, David J. Szymanski, and Jordan Oseguera. 2023. “Anthropometric Evaluation of First Round Draft Selections in Major League Baseball.” *Journal of Strength & Conditioning Research* 37 (8): 1609–15. https://doi.org/10.1519/jsc.0000000000004442.

Garmon, Christopher. 2012. “Major League Baseball’s First Year Player Draft.” *Journal of Sports Economics* 14 (5): 451–78. https://doi.org/10.1177/1527002511430229.

Staudohar, Paul D, Franklin Lowenthal, and Anthony K Lima. 2006. “The Evolution of Baseball’s Amateur Draft.” *NINE: A Journal of Baseball History and Culture* 15 (1): 27–44. https://doi.org/10.1353/nin.2006.0056.