Anthony Argenziano
asa2176
Prof. Sunil Gulati
W4911: Sports Economics

Pre-Arbitration Contract Extensions in Major League Baseball: A Retrospective Analysis

Abstract

Major League Baseball's pseudo-monopsonic rules governing player compensation have historically favored teams, depressing players' salaries below what their on-field production would demand and requiring players to resort to certain measures in order to secure what they consider improved annual wages. This study analyzes the properties and predicts the incidence of *pre-arbitration contract extensions*, a specific subcategory of historically team-friendly contract agreements made between Major League Baseball teams and players. While existing literature has aimed to pinpoint the determinants of player salary in baseball, this study attempts to isolate the factors that determine the value of pre-arbitration extensions specifically. Furthermore, it aims to build a predictive model of which types of players will receive these extensions. Collecting data from each pre-arbitration contract extension from 2008-2019 (n=104), heteroskedasticity-robust linear regression was performed to determine the predictive effect of certain player performance metrics and characteristics on the adjusted annual value (AdjAnnVal) of each pre-arb extension. Major League OPS proved to have the most significant positive effect for hitters' AdjAnnVal (additional $26,600 per OPS point added) while pitchers' *minor league* ERA and strikeout-to-walk ratio proved more predictive of what teams paid per season. Additionally, a case-control study was performed on the hitters in the dataset (n=64) and a randomly sampled group of 64 hitters from the same era, in which logistic regression and Random Forest classification techniques were used to predict the probability of a player receiving a pre-arbitration extension. Results suggested a positive correlation between a higher OPS in the previous MLB season and extension probability, as well as an increased probability of an extension if a player was born internationally. The study concludes that teams' decision-making in offering extensions to hitters is most specifically a product of a player's recent MLB performance, rather than a prior track-record of success from the minor leagues or other harbingers of value. Because these extensions have historically capped player's future potential wages, and given a recent spike in their incidence, the MLB Player's Association should be concerned that extensions are going to offensive players that performed significantly better than the control group, thus making them ineligible to receive the most lucrative free-agent contracts.

I. The Need to Investigate Pre-Arbitration Contract Extensions in Major League Baseball

This study aims to better understand the factors (age, performance, birthplace) that determine players' average annual salary in contract extensions made during their pre-arbitration seasons. In addition, it aims to incorporate these factors into a model that predicts whether or not a player will receive a pre-arbitration extension.

Occasionally, Major League Baseball free agents strike gold – as an example, in only a four-day span at the 2019 MLB Winter Meetings, three players signed contracts totaling $814 million over 23 seasons (average annual value of $35.4 million). However, for every free agent mega-deal that is signed, there are budding stars elsewhere who sacrifice earning potential for security, agreeing to extended-length contracts at a reduced salary before reaching free agency. At the start of 2019, and in his second season in the MLB, Curaçao-born All-Star Ozhaino Albies signed a 7-year/$35 million deal with Atlanta. While not an insignificant amount of compensation, many in the baseball industry clamored that Albies could have made three times as much if he had negotiated his contract year-to-year until free agency. Inspired by agreements such as Albies', this study attempts to answer the question of how teams and players come to valuations in pre-arbitration contract extensions, despite a player's relative lack of major-league performance data. Furthermore, it aims to profile the type of player that is most likely to accept such an extension, in an effort to educate player representatives about the types of top-tier players susceptible to potential lost earnings in these deals. The results of these analyses have the potential to educate not only players about what teams consider in determining contract value, but also rival organizations, as predicting contract extensions of other teams' players may provide an advanced outlook of which players are likely to eventually be available on the open market.

This report begins with an overview and explanation of important baseball concepts, such as service time, arbitration, and player compensation. This discussion will also set the stage for the existing power dynamic between players and teams. Next, the focus of the study -- the concept of the "pre-arbitration contract extension" -- will be introduced, as well as its common properties, practical purposes for players and teams, and increased prominence in recent baseball news. Finally, relevant literature in the field of baseball economics will be introduced, along with an explanation of where this analysis fits within the canon.

Following the expositional portion of this report, the questions of interest will be restated, followed by a series of hypotheses. Then, there will be a discussion of the data that was used, as well as any distributions of interest, assumptions made, and limitations present in the sample. Once the dataset is explained, two models will be introduced to test each of the two given hypotheses. This will be followed by an explanation of the methods (multiple linear regression followed by logistic regression/random forest classification). Methods will be followed by results and visualizations of these results. Finally, conclusions will be made, to be followed by a discussion of implications of the study, and further areas of research to be explored.

II. Key Baseball Concepts Explained

The current study focuses on predicting the average annual value of pre-arbitration contract extensions in Major League Baseball. In order to understand what a pre-arbitration extension is, one must understand the laws that govern labor mobility and compensation in baseball. First, a "major leaguer" is defined as a player that is currently, or has ever been, on an MLB team's 25-man major league roster. Once a player become a major leaguer, he begins to accrue "service time," a measure attached to a player that counts how long (in years) a player has been on a major league roster. For example, a service time of 1.00 belongs to a player that has been a major leaguer for exactly one season.

Service time is an important marker for each major leaguer, as his service time determines his labor mobility (i.e. the ability to choose his employer) as well as his bargaining power with his team. It is important to note that, because contract renewals occur during the MLB offseason, a player's service time at the *end of each season* is what determines his eligibility for each of the three categories to be discussed. Disregarding rule exceptions, this metric separates players into three-distinct categories of 'labor power': (1) Players with service times ranging from 0 to 2.99 are considered *arbitration ineligible*—according to the rules, these players have not accrued enough service time to have the right to negotiate contracts with their club without the team's reciprocated interest in doing so. Instead, the club has the power to renew these players' contracts anywhere at, or above, the league-mandated minimum (which was $555,000 in 2019). (2) Players with service times ranging from 3.00 to 5.99 are considered *arbitration eligible*. While they, too, cannot sign with any other team unless their controlling team releases them, they are able to jockey for a better wage in one of two ways—either they can negotiate a one-year contract with their controlling team, *or* they can select a wage that they believe they deserve, and face their controlling team in an arbitration hearing, where a panel selects between the player's proposed salary and the team's proposed salary. (3) Players with greater than 6.00 years of service time are *free agent-eligible*—for the first time, they are free to negotiate a contract with any of the 30 MLB clubs.

Succinctly, Major League Baseball's labor compensation model is one that generally rewards veterans of the system with lucrative contracts nearer to the value of their production, provided that they bide their time earlier in their career, and accept wages generally way below what their production has earned them. Given the descriptions of the above three labor stages, it is apparent that more service time is associated with more individual labor power, and thus relatively higher wages as a player moves from stage to stage. Previous literature has been able to quantify how labor power shifts from stage to stage. In a study of all major leaguers' annual performance statistics and salaries from 2000-2011, Brad Humphreys and Hyunwoong Pyun calculate "Monopsony Exploitation Ratios," formulated by the operation $[\frac{deserved\ wage - awarded\ wage}{deserved\ wage}]$, in which "deserved wage" is based on players' estimated marginal revenue product. They calculate an average estimated exploitation ratio of 0.89 for arbitration ineligible players, 0.75 for arbitration eligible players, and 0.21 for free-agent eligible players, accentuating the remarkably large gap in fairness between pre-arbitration wages (players paid 11% of total deserved compensation) and free agent wages (when they earn 79% of total deserved compensation). With a labor system that, on average, underpays all its laborers, but especially its newest laborers, young players have been conditioned to either preserve their worth for late-arbitration and free agency, or be creative in other ways to secure reasonable career earnings.

III. The Pre-Arbitration Contract Extension, Discussion of Literature

While players can hope that they remain productive to free agency, recent developments in the Collective Bargaining Agreement (CBA) between the Commissioner's Office and the MLBPA have actually made free agency tougher on those who are eligible. Notably, for players who are offered and reject a "qualifying offer" from their team before entering free agency, whichever team eventually signs the player is levied a tax in the form of a lost draft pick. Because of teams' growing emphasis on player development and accumulation of minor-league talent, teams have begun to shy away from, or discount, free agents that would strip them of picks if they were to sign them. Therefore, even free agency, to some extent, has lost its luster as the guaranteed reward for six years of tolerance of wage suppression.

Having perceived player's recent pessimism surrounding free agency, teams have opted to offer contract extensions to their free agency-ineligible players. Rather than setting a salary year-to-year as per convention, the team and player can agree to a contract over multiple seasons. Such an agreement benefits the player in that it allows him to secure a portion of his expected future earnings now, but hurts him in that requires him to sacrifice potentially higher wages in the future, since he will forfeit a few free agent seasons. This structure is also beneficial for teams, since those clubs that can effectively locate talent at an early stage (perhaps even before the player himself) are incentivized to "lock up" a player before he can either (a) become a free agent and sign with a competitor, or (b) become aggravatingly expensive in the later stages of arbitration, and again in free agency.

At this point, we have reached the concept of the **pre-arbitration contract extension**. We will define this type of contract as a multi-year pact that occurs between player and team, under the condition that it is signed before a player reaches arbitration eligibility. Therefore, of the 439 contract extensions that have occurred since 2008, this study will only analyze the 104 (23.6%) that fall in this category—a relatively small subset (MLBTradeRumors.com). While few baseball scholars have written about this type of extension in particular, those that have were inspirational in the creation of this analysis. In his piece for the Society for Baseball Research's Research Journal, Jim Turvey describes a player's act of signing a pre-arbitration extension as "guaranteeing himself a great deal of money, but at the same time… putting a ceiling on his earnings." Fangraphs writer Dave Cameron believes equally in the team-friendliness of such deals, but argued in 2017 that they were starting to become less frequent than they were in the early 2010s. Furthermore, he sensed a decline in the quality of players that received these extensions, arguing that the league's top young players were becoming "so good so fast that the leverage teams have over [them] expired before negotiations could even really begin." Thus, if it were up to Cameron in 2017, investigating these contracts might be a secondary exercise, given his observations of their declining prevalence among the league's best young players.

However, since Cameron's article in April 2017, much has changed. While only 16 pre-arbitration extension contracts were signed between 2015 and 2017, twelve were signed from January through April of 2019 alone. Therefore, Cameron's declaration of these contracts' "demise" was in fact somewhat premature. Furthermore, against Cameron's hypothesis, these contracts have recently been signed not just by average major leaguers, but even by the league's biggest young stars; not only did the aforementioned Albies sign a long term deal, but so too did his teammate, and 2018 Rookie of the Year, Ronald Acuña (8 years/$100M, and another heavy discount, according to pundits). Both the increase in the frequency of these extensions, and the fact that they are being accepted by some of the league's top talent, provided enough motivation to investigate this issue further.

While the concept of a pre-arbitration extension on the surface seems fairly mundane, there are a few reasons why investigating these contracts may yield further insights about the current power balance between the players and their controlling teams, as well as about how players are valued in the market in terms of statistical production. Because pre-arbitration extensions occur during the lowest-earning period of player's career, he is often willing to accept less money than he would as a free agent or an arbitration-eligible player, as long as it is slightly better than what he currently earns (which, as mentioned, is close to the league minimum).

Given these facts, there are much higher stakes surrounding pre-arbitration extensions than one might expect. Not only are these extensions occurring more frequently than ever, but they appear to be usurping much earning potential from the league's top young stars. This study attempts to analyze such deals that have been consummated, in order to better understand the conditions that promote them and which types of players are most prone to signing them.

IV. Hypotheses for Analysis I and Analysis II

In Analysis I, we attempt to answer the question, “**what components of a player’s past performance and personal profile determine the average annual value (AAV) of his pre-arbitration contract extension?”** One hypothesis is that the average annual value of a player’s pre-arbitration extension, unlike free agent contracts, will be positively correlated with minor league performance, rather than major league performance. This is because teams will be wary of the relatively small sample size of major league performance (each player has only a few MLB seasons at most). Another hypothesis is that a player in whom the team has already invested heavily (either a top-3 round pick, or if internationally born, a $100,000+ signing bonus), will likely have a higher average annual value on his extension, as teams may be susceptible to the “sunk cost fallacy,” a basic economic principle.

In Analysis II, we attempt to retrospectively answer the question, “**what factors influence the probability that a player will have received a pre-arbitration extension?”** If we believe that players are, in general, unwilling to bet on themselves in free agency, then we can hypothesize that the probability that a player receives a pre-arbitration extension should be positively correlated with how he performed in the previous season. Another hypothesis stems from an intuition drawn from the Ozhaino Albies contract (7 years, $35M) introduced earlier. Critics of the deal alleged that Atlanta exploited Albies’ immediate need for cash flow, which was especially necessary for a player whose family resides in Curaçao, an island with a GDP per capita which would rank 67th in the world (3x smaller than that of the U.S.). Therefore, it is hypothesized that international players, who often come from areas of the world with less earning potential and financial prosperity, are more likely to take money up front than risk never earning a large payday, thus increasing the likelihood they sign a pre-arbitration extension. Finally, it is hypothesized that former top picks and high signing bonus players, conversely, are less likely to have signed a pre-arbitration extension, as they have already achieved relative financial security, and can bear to take the risk in maximizing their pay versus curbing their earning potential.

V. The Sources and the Nature of the Data

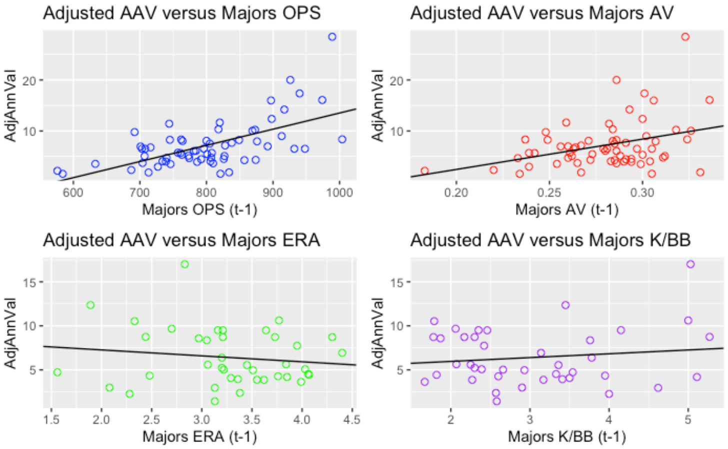
Data was compiled from various public digital sources. First, using MLBTradeRumors.com's "Extension Tracker," information was collected about all 104 pre-arbitration extensions between 2008-2019, serving as the skeleton of the dataset. To ensure that the data only included arbitration-ineligible players, the dataset was filtered by service time -- only deals signed by players with fewer than 3.00 years of service time (or fewer than 2.00 years of service time if they were "*Super Two"*) were considered. Each row of the raw dataset represents a pre-arbitration extension, and columns list factors such as the contract signing date, player's name, his current team, contract length, total value of the contract, and the player's current service time. An additional column, "AAV" (average annual value), was created by dividing the total value of the contract by its length in years. To account for year-over-year inflation in baseball contract values, a column called "Adj. Ann. Val." was created by multiplying AAV by an inflation-neutralizing factor (based on average annual salary in the MLB from year to year), with base year 2019. For example, the mean annual baseball salary was $4.36M in 2019 and $3.84M in 2015, so contracts signed in 2015 were corrected by a factor of 4.36/3.84, or increased by 13.5%. This method proved an efficient way to avoid bias, where otherwise greater contract value would have been heavily correlated with year. This method also avoids the use of 11 binary variables for year, which would have resulted in overfitting of the data.

With the main dependent variables set, independent variable columns were then manually added to each row. One of these columns is a binary variable which takes the value 1 if a player is a hitter, and 0 if a pitcher. Sorting by this variable yielded 64 hitters and 40 pitchers. Because hitter and pitcher performances are evaluated using different sets of statistics, they were kept separate in this analysis, and different metrics were compiled for each. From baseballreference.com, two main performance statistics were collected from each hitter's last full season before receiving a contract extension: *On-base-plus-slugging percentage* (OPS), a proxy for a hitter's power and plate discipline, and *Batting Average* (AV), a proxy for a player's contact and "pure hitting" ability. Two statistics were also compiled from each pitcher's last full season prior to extension: *Earned run average* (ERA), a proxy for the pitcher's ability to prevent runs (lower is better), and *Strikeout-to-walk rate* (K/BB), the ratio of the total strikeouts a pitcher accumulates (correlated with higher pitch velocity, break, and “raw ability”) to the total walks he allows (a measure of a pitcher’s ability to locate pitches accurately and consistently). After compiling their major-league statistics, TheBaseballCube.com was accessed to scrape the same two metrics for each player, this time from their career *minor league* totals (prior to the extension). Finally, two additional binary variables were collected -- “Top3 Round”, which takes the value 1 if a player was selected in the top 3 rounds of the Rule 4 amateur draft and 0 if not, and “International,” which takes the value 1 if a player was born in a country other than the United States. Note that because only domestic players enter the Rule 4 Amateur draft, internationally-born players also received a value of 1 for “Top3 Round” if their initial signing bonus exceeded $100,000 (which is on the higher end of the international bonus spectrum, and allows a measure of prospect status for both drafted and internationally signed players). Below lies a snapshot of the columns and properties of each variable in the data.

Column	Name	Contract Length	Total Val. ($)	Adj Ann Val ($)	Int’l	Hitter or Pitcher	MLB Stat. 1	MLB Stat. 2	MiLB Stat. 1	MiLB Stat. 2	Top3 Round	Age
Data Type	Char.	Numeric	Numeric	Numeric	Binary (1 if yes)	Binary (1 if hitter)	H: OPS P: ERA (numeric)	H: AV P: K/BB (numeric)	H:OPS P: ERA (numeric)	H:AV P:K/B B(num	Binary (1 if yes)	Numeric

VI. Model Selection and Methods, Analysis I

In order to justify a linear model, scatter plots were built displaying AdjAnnVal against various predictors, to determine if a linear relationship existed.



While MLB hitting statistics (top row) showed a stronger linear relationship with AdjAnnVal than did pitching statistics (bottom row), a look at residual plots (not pictured) show slight heteroskedasticity, as the values of the residuals grow as do both Majors OPS and Majors AV. To correct for this and ensure that results were a product of variable relationships and not heteroskedasticity, the rlm() function from the MASS package in R was used to conduct heteroskedasticity-robust multiple linear regression.

For hitters (n=64), AdjAnnVal was regressed against the following variables of interest: MajorsOPS, MajorsAV, MinorsOPS, MinorsAV, International, and Top3Round. Meanwhile, for pitchers (n=40), AdjAnnVal was regressed against the following variables of interest: MajorsERA, Majors.K.BB, Minors.ERA, Minors.K.BB, International, and Top3Round.

In both of these regressions, Age and ServiceTime were included as control variables. Plotting AdjAnnVal against ServiceTime (not pictured) shows a positive trendline, suggesting that players' leverage to negotiate higher adjusted annual values in their extensions increases as they get closer to becoming arbitration-eligible. Furthermore, there is a correlation between a player's service time and his age, because a player's service time (assuming he is a major leaguer for the entire year) increases at the same speed as his age does. While the coefficients on these two are not of interest, they do contribute to variation in AdjAnnVal and must be included to remove omitted variable bias. The linear models for this part are shown below:

$$\textit{Hitters}: \widehat{\boldsymbol{AdjAnnVal}} = B_0 + B_1\boldsymbol{MajorsOPS} + B_2\boldsymbol{MajorsAV} + B_3\boldsymbol{MinorsOPS} + B_4\boldsymbol{MinorsAV} + B_5\boldsymbol{International}_{\boldsymbol{Binary}} + B_6\boldsymbol{Top3Round}_{\boldsymbol{Binary}} + B_7\boldsymbol{Age} + B_8\boldsymbol{ServiceTime} + \textit{Error}$$

$$\textit{Pitchers}: \widehat{\boldsymbol{AdjAnnVal}} = B_0 + B_1\boldsymbol{MajorsERA} + B_2\boldsymbol{Majors.K.BB} + B_3\boldsymbol{MinorsERA} + B_4\boldsymbol{Minors.K.BB} + B_5\boldsymbol{International}_{\boldsymbol{Binary}} + B_6\boldsymbol{Top3Round}_{\boldsymbol{Binary}} + B_7\boldsymbol{Age} + B_8\boldsymbol{ServiceTime} + \textit{Error}$$

VII. Results, Analysis I

Hitters

Variable	Intercept	MajorsOPS	MajorsAV	MinorsOPS	MinorsAV	Int'l	Top3	ServiceTime	Age
Coef.	-2.5504	**.0266**	-31.749	.0066	12.677	-.38	1.1387	**2.2203**	**-.6301**
P-value	N/A	.001***	.1594	.497	.709	.6737	.1603	.0001***	.006***

After performing the regression task of adjusted annual value of hitters' contracts against their characteristics and performance metrics, three variables emerged as significant predictors of how much a player earns per year in his extension. First, as expected, the control variables ServiceTime and Age are both significantly positively correlated with AdjAnnVal. The variable ServiceTime's direction follows our hypothesis—the closer a player is to exiting the pre-arbitration stage, the more he is able to negotiate for in these contracts. For every additional year of service time accrued, a player's annual deal value increases by about $2.2M. Meanwhile, age moves in the opposite direction than what the hypothesis expected—holding all else constant, for every year older a player becomes before signing a pre-arbitration extension, he loses an estimated $630,100 per annum on his deal. There are two potential causes for this result. Because pre-arbitration extensions often buy out years during a player's free-agency eligibility, and the industry is moving away from signing older free agents (players in early-to-mid 30s) to lucrative contracts, those who are still pre-arbitration eligible in their late 20s would not cost as much to retain beyond their current years of control. The other potential explanation is that players who reach the majors at younger ages are typically better than those who reach the majors at more advanced ages, and are therefore more valuable assets that command more money.

Above and beyond the control variables described above, the main performance metric that forecasts a hitter's extension AdjAnnVal is his OPS in his most recent MLB season prior to signing the deal. OPS, as mentioned, represents the sum of a hitter's ability to get on base (OBP) and to hit for power (SLG). Because OPS can range anywhere from 0 to 5.000 (although typically values range between 0 and 1.200), multiplying each value by 1000 in the dataset made it easier to interpret regression coefficients in this step. Holding all other regressors constant, the model suggests that a one-point increase in a player's Major League OPS in the season before his deal is associated with an increase in annual contract value of about $26,600. While this may not seem like a large sum on the surface, the magnitude of this marginal effect is quite large in practice. Of 121 qualified major league hitters in 2019, the median OPS was .825, courtesy of Arizona's Christian Walker. If we were to compare the difference in projected annual earnings to the 75th percentile-OPS (.745, for Washington's Victor Robles), this would predict an estimated AdjAnnVal gap of $2.128M between the players' extensions, assuming the players are otherwise identical.

Above all, the significantly positive coefficient on OPS shows that, when a team negotiates a pre-arbitration contract extension with a hitter, the average annual value of the deal is determined more by his raw power and plate discipline (as measured by OPS) than his ability to merely collect base hits (as seen by the p-value of ~.16 on MajorsAV). This seems to fall in line with the increasing prevalence of "three true outcomes" hitters in baseball—those who post high frequencies of walks, home runs, and strikeouts—and the fact that a player is better at creating runs when he attempts to hit home runs at a lower success rate, than to hit singles at a higher success rate. While this is excellent news for new-wave hitters who embrace this "three true outcomes" plate approach, this does not bode well for speed and contact hitters' ability to sign lucrative pre-arbitration extensions.

Another result from this study is the lack of significance on the coefficient for career minors OPS and AV. Rather than prioritizing track record in the minor leagues (where most players in the dataset had more career at-bats than they did in the majors), the value of pre-arbitration contracts appears tied solely in very recent performance in the Major Leagues, perhaps signaling teams' distrust of minor league performance as a robust measure of future success over the length of a long contract.

Pitchers

Variable	Intercept	MajorsERA	MajorsKBB	MinorsERA	MinorsKBB	Int'l	Top3	ServiceTime	Age
Coef.	2.252	0.109	.659	-2.468	-1.187	1.612	1.842	1.330	0.404
P-value	N/A	0.869	.164	.001***	.013**	.081*	.033**	.041**	.076*

When performing the same analysis on the pitcher group (n=40), the predictive value of minor league performance is the opposite. Pitchers' minor-league statistics (both ERA and K/BB) significantly correspond with the average annual value of pitcher extensions. More specifically, a one-run decrease in pitcher minor league ERA is associated with an estimated increase in the AdjAnnVal of a deal by $2,468,000. An explanation for minor league statistics' effect on pitchers' contract value may reside in the difference between what it means for a pitcher, versus a hitter, to succeed in the minor leagues. In fact, certain minor league affiliate leagues, such as the Pacific Coast League (Triple-A) and the California League (High-A), locate the majority of their teams in cities with extremely high altitudes, such as Las Vegas, Albuquerque, and Colorado Springs. As is seen with the major league equivalent city, Denver, and its team, the Colorado Rockies, runs are scored at a higher rate at their stadium than anywhere else in the country, because thinner air at higher altitudes decreases air resistance, causing pitchers' pitches to break less (decreasing effectiveness) and the ball to travel further off of the bat. Thus, while offensive competition is relatively weaker in the minor leagues than in the major leagues, executives might determine that the ability to limit runs (low ERA) in thin-air minor league affiliate leagues is a robust measurement of a pitcher's true skill, by virtue of the sheer difficulty to limit runs in such hitter-friendly environments. Given the fact that some, but not all minor league affiliate leagues are "hitter-friendly," this explanation would best be confirmed through a comparative study of pitchers' performances across different affiliate minor leagues.

Not only are the performance metrics that contribute to AdjAnnVal different for hitters and pitchers, but so are the personal characteristics that affect the regression. Particularly, the coefficient on Top3 is significantly positive for pitchers, such that a pitcher who qualified for "elite prospect status" (either a Top 3-round pick or an international signing bonus of at least $100,000) is estimated to average $1.84M more per year in a pre-arbitration extension than a pitcher who was not an elite prospect. Eighteen out of 40 (45%) of the pitchers in the sample fall into this category, while 35 out of 64 hitters qualify (54.7%). Finally, service time, as is the case in the hitter analysis, is significantly positively correlated with extension AdjAnnVal, and likely for the same reasons.

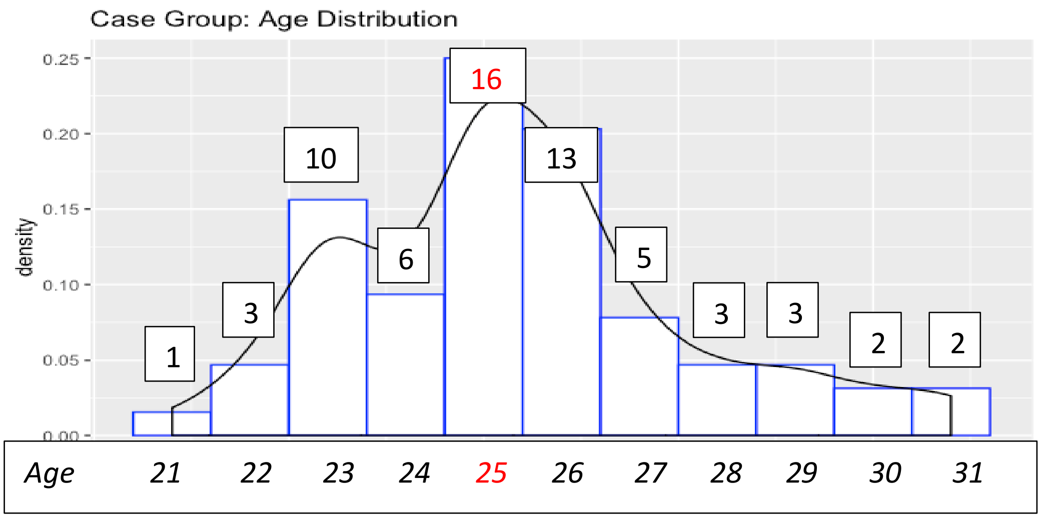
In summary, this pair of linear regression analyses has identified factors that teams value in extending their players in the pre-arbitration stage, and furthermore, how those factors differ between hitters and pitchers. However, aside from the issue of a relatively small sample size in each regression, the main limitation of these analyses are their inability to predict whether or not a specific player is likely to receive an extension. As presently constituted, the data only describes those who *have* received an extension, thus providing no context to the universe of players that have opted not to sign extensions prior to arbitration. In Analysis II, with the infusion of new data and the use of machine learning techniques, a case-control is performed that attempts to assess this more complex question.

VIII. Analysis II: A case-control study predicting probability a hitter receives pre-arb extension

After studying the properties of players who received pre-arbitration extensions, a further goal was established: to be able to predict, based on certain performance data and personal characteristics, the probability that a player would receive a pre-arbitration extension. As discussed previously, such an analysis would deliver insights to both team management and player representatives about the market dynamics surrounding pre-arbitration extensions. For teams, it would give them insight into which characteristics and abilities make it most likely that a player will accept a pre-arbitration extension. On the flip side, because of the heavy discounts that teams get when extending players early in their careers, such an analysis would inform the MLBPA and player representatives of the types of players that are most susceptible to such deals.

In order to identify the profiles of players who receive such extensions, it was necessary to gather data from a control group (those with similar characteristics, but who have not received a pre-arbitration extension in their careers). Focusing on hitters alone, a case-control study was conducted, with a 1:1 ratio in each group (n=64 in each). Data from every individual hitters' season statistics from 2008-2019 (n=2,725) was scraped from Fangraphs.com, with the following columns extracted: MajorsOPS, MajorsAV, and Age. Because the two performance variables of interest (OPS and AV) are rate statistics, and therefore small samples of plate appearances poorly represent a player's true skill, a minimum of 300 plate appearances—about one half of a full season—was required for a player to be included in the set.

Caution was taken to ensure that the case and control groups were sufficiently similar—if this were not the case, variation across the two groups would be a product of the differences in the their baseline characteristics, thus reducing any predictive power of the probability of a player receiving an extension. Because of the relatively small sample of cases, similarity between groups was achieved by coercing the control group into the exact age distribution as observed in the case sample. The age distribution for the case group is shown in the figure below.



With the above information, the Fangraphs data was sliced into smaller datasets by age in years. Before sampling from these datasets, all rows of seasons belonging to players currently in the *case* set were removed using filtering techniques. Then, from each age dataset created, individual seasons were randomly sampled *without replacement* using the sample() function in R. The number of rows sampled from each age group followed the graph above exactly (e.g. sixteen 25-year-olds, two 31-year-olds). When sampling was complete, the samples were merged, creating the study's 64-row control group.

Upon creation of the control group, the variables International and Top3Round were manually inserted as new columns. Then, after merging the case and control groups into one dataset (n=128), the binary dependent variable, "Case", was created—those in the case group (having received extensions) were assigned a 1, while those in the control group (not having received extensions) were assigned a 0.

Finally, before the construction of models, training and testing sets were created, using the conventional 80%/20% split, ensuring that equal proportions in each set belonged to the case and control groups. The table below explains this symmetrical breakdown

	Case	Control	Total
Train	51	51	102
Test	13	13	26
Total	64	64	128

IX. Analysis II: Models and Methods

Modeling on the training data, logistic regression was performed using the glm function in R. The binary dependent variable, Case, was regressed against the following four variables: MajorsOPS, MajorsAV, International, and Top3Round. The model is below, where p is the probability that Case=1:

$$Log(p/(1-p)) = B_0 + B_1\textbf{MajorsOPS} + B_2\textbf{MajorsAV} + B_3\textbf{International} + B_4\textbf{Top3Round} + Error$$

After modeling on the training data and receiving coefficients and their corresponding significance levels, the model was tested on data in the test set, generating predicted probabilities of each hitter receiving a pre-arbitration extension based on the model's interaction with their characteristics. Additionally, predicted probabilities for players in the test set were compared to the observed values of "Case," and measurements of model fit and prediction accuracy were conducted. Finally, the fit of the logistic regression was compared to the fit given by performing random forest classification, a machine learning technique that uses a "voting" system to classify players as either having a value of 0 or 1 for "Case."

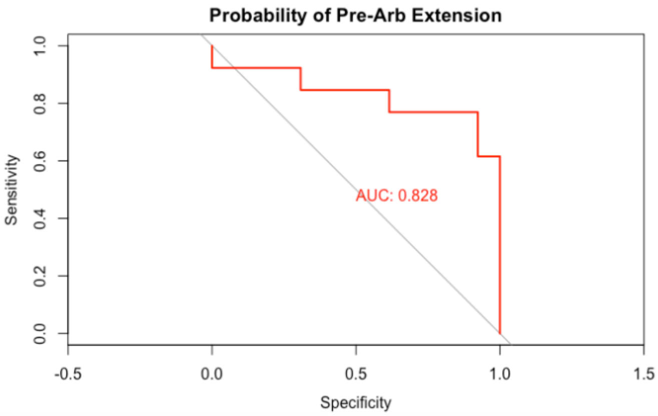
X. Results, Analysis II

Variable	Intercept	MajorsOPS	MajorsAV	International	Top3Round
Coef.	-9.096	**.009382**	5.925	**.9236**	-.0155
P-value	N/A	.014**	.580	.062*	.972

In logistic regression output, only the direction and significance of the coefficients can be immediately interpreted. Most apparently, an increase in a player's major league OPS in the previous season is significantly associated with an increase in the probability that he then receives a pre-arbitration extension. To better understand the magnitude of these effects, estimated probabilities were calculated for two theoretical players: player A has an OPS of .900, and player B has an OPS of .800—as for the other metrics, each of them were assigned the sample mean value in the training set. By using the formula $\frac{e^{Log(\frac{p}{1-p})}}{1+e^{Log(\frac{p}{1-p})}}$—a full breakdown of the coded calculations can be found in the appendix—there is a 77.4 percent chance that player A would have received a pre-arbitration extension following the season, and just a 57.3 percent chance that player B would have. Thus, that increase in 100 OPS points raises the probability of a pre-arbitration extension by roughly 35%.

There is also, to a less certain degree (p-value =.062), a positive effect of a player being internationally born on the probability that he receives a pre-arbitration extension. Utilizing the same formula as above for estimating probabilities from logistic regression, and using the training set averages for OPS, AV, and Top3Round for both, probabilities of receiving an extension were assigned to players C and D—C being internationally born, and D being born and trained in the United States. Following this calculation, the estimated probability that a player like player C (average, and internationally born) would have received an extension is 65.4%, while a player with the same characteristics as C except domestically born (player D) would have only a 42.8% chance of receiving an extension in the following season. Thus, the previously stated hypothesis—that international players, holding all else constant, are more likely to accept pre-arbitration extensions—is supported by the output of the logistic model.

While the interpretations of the coefficients provide insights as explained above, predicted probabilities of "Case=1" were generated on the test set, and compared to the actual values of each test member, in order to assess the ability of the regression to classify members of each group. To measure this, a ROC (receiver operating characteristic) plot was generated, which graphically displays the ability of the regression to correctly classify the test group as either "case" or "control." As seen below, the AUC (area under the curve) is .828—A higher AUC implies a greater ability of the model to correctly classify the data into the correct group, and a value above 0.8 is indicative of a good classifier.



Finally, in order to determine if other models classified case and control groups more effectively than logistic regression, random forest classification was performed using the caret package in R. Like in the logistic regression, the variable Case was regressed against the same four variables as in the logistic regression. Ultimately, it was determined that the random forest algorithm was slightly worse than logistic regression at classifying members of the test set, generating an AUC of .799.

While random forest is an inferior classifier to logistic regression in this study, it sparks reconsideration of the variable International as a strong predictor of whether or not a player receives a pre-arbitration extension. Using the "variable importance" feature in R, which determines which explanatory variables had the most influence in swaying the random forest model to vote either "case" or "control," it is determined that OPS (importance=100) and AV (importance=74.96) are much more influential than is "International" (importance=2.37). Therefore, while there might be a greater chance that a Venezuelan-born player will accept a pre-arbitration extension than a player from the U.S., random forest suggests that this classification can be performed, more or less, by almost completely ignoring "International" and basing the classification on OPS and AV alone.

XI: Conclusions and Discussion: Analysis I, Analysis II

Overall, many insights can be gained from this pair of analyses, from an understanding of what teams value in pre-arbitration extensions, to the profiles of players that are most likely to sign such extensions.

In analysis I, our hypotheses anticipated that career minor league statistics, for pitchers and hitters alike, would be most effective at predicting how much a player would earn per year on his pre-arbitration extension. Following the analysis, it was determined that this hypothesis was only half true. While it was supported for pitchers (minor league ERA and K/BB were both significantly negatively correlated with extension AdjAnnVal), this was not so for hitters. Instead, a players Majors OPS from the previous season had a significantly positive relationship with the AdjAnnVal he earned on his extension. As aforementioned in VI., one possible explanation of these differences might be that teams believe minor league pitching stats to be more robust measurements of future success than they do than minor league hitting statistics. While opposing hitting and pitching alike are determinants of any hitter or pitcher's minor league success, it appears from this study that teams consider the advantage in minor league games to be held by offensive players, thus discounting their statistical output compared to what pitchers accomplish.

Furthermore, Analysis I provides insight into which types of players are valued most highly in pre-arbitration deals. For hitters, it is clear that teams are willing to spend more on players that display power and patience rather than contact and hit ability (as seen by the positive and significant coefficient on OPS instead of AV). For pitchers, teams value those that purely limit runs (negative correlation between AdjAnnVal and ERA) rather than those that have the ability to control the strike zone well (negative correlation between AdjAnnVal and K/BB). Therefore, players will be able to command more money per year if they belong to low ERA and high OPS groups, and may be more advised to take an extension than a player who is either a contact hitter or a "control" pitcher, and likely undervalued by the pre-arbitration extension valuation system.

Analysis II adds a predictive dimension to the insights developed in Analysis I. After inclusion of a control group (players never having received pre-arbitration extensions), it was determined that a hitter's Majors OPS in his previous season has a significantly positive effect on the probability that he receives a pre-arbitration extension in the next year. This can mean one of two things. For one, perhaps players that break out early in their careers are more likely to take security in the form of an early, long-term deal, to protect against their previous seasons being nonreplicable. If this were the case, those who represent players and negotiate their contracts should run analyses on each of their players to determine whether or not such output is sustainable. If an agent determines that such output is sustainable for the particular player, he might advise the player to hold off on negotiating an extension, and instead bank on the likelihood of performing well again in subsequent seasons, in order to achieve a lucrative free-agent payday. However, another explanation for the positive association between OPS and the probability of an extension may lie in hidden differences in properties between the case and control groups, which may have not been captured by merely using the same age distribution. While the same number of players at each age were present in the case and control groups, perhaps the case group contained players that displayed power and patience relatively earlier in their careers, while the rest of the baseball population developed these skills later on average. While the average age of hitters in the case group was 25.2, various analyses, such as one conducted by J.C. Bradbury of Baseball Prospectus, suggest that players typically peak closer to their age-29 season. Therefore, those in the control group were likely represented before their peak performance age (since their average age was also 25.2), thus showing that pre-arbitration extensions more frequently go to those that develop power and patience at a relatively younger age.

Finally, a positive association was found between whether a player was born internationally and the probability that he signed a pre-arbitration extension. As presented in the hypothesis, this can possibly be attributed to players’ desire for financial security, if they are coming from a country with relatively lesser financial prosperity than the United States. Another explanation, however, could be that players who enter professional baseball through international signing bonuses are paid relatively less than players of equal caliber who enter through the draft. For reference, in 2019, the highest-paid draft pick was awarded an $8.1M bonus, while the highest-paid internationally signed amateur was paid a $5.1M bonus. Therefore, the continued need for financial security even in a second large contract is more likely for international players, who on average have earned less than American players by the time they are pre-arbitration players. In a further study, placing an interaction term between Top3Round and International might have confirmed this intuition if the coefficient was found to be significantly negative. Ultimately, this last insight has far reaching implications. Not only can teams more easily convince their pre-arbitration, internationally born players to secure to long term deals, but also might decide to target more international amateurs in the future, knowing that they can be bought out throughout their primes for somewhere below the market rate. While this represents an opportunity for teams to accrue surplus value from its players, this should also raise concerns in the MLBPA, as international players appear to be less willing to bet on themselves on the market and accept deals below what they are worth. While players should be allowed to opt for financial security of their families, a new collective bargaining agreement could be created that raises the major league minimum salary to the point at which players, such as Ozhaino Albies, feel they have already made a comfortable amount of money, and do not have to sacrifice maximum career earnings to get incrementally more in the present.

XII: Limitations, Further Areas of Study

While this study develops insights about the types of players that are amenable to pre-arbitration extensions, as well as the factors that influence these extensions' valuations, there are certain judgments it cannot deliver, due to limitations in the scope of data and in its approach.

For one, the sample size is quite small in the study of pitchers (n=40). Typically, it is advised to utilize a 10-to-1 rule of observations to predictors, but determining which predictors to omit for the study was difficult. While the insights from the pitching regression can be taken seriously, they must therefore be taken with a bit of caution. Furthermore, due to unavailability of particular data, this study does not effectively compare how adjusted annual value of contracts change in types of contracts *other* than pre-arbitration extensions. Analysis I found that every additional OPS point represents a $26,600 increase in expected AdjAnnVal, however comparing this to players who have just received free agent or arbitration contracts would have given context into whether or not $26,600 per point was a relatively large or small incremental reward. This can be accomplished in a further study with data that includes both a players statistical output, personal characteristics, *and* annual salary in the following season.

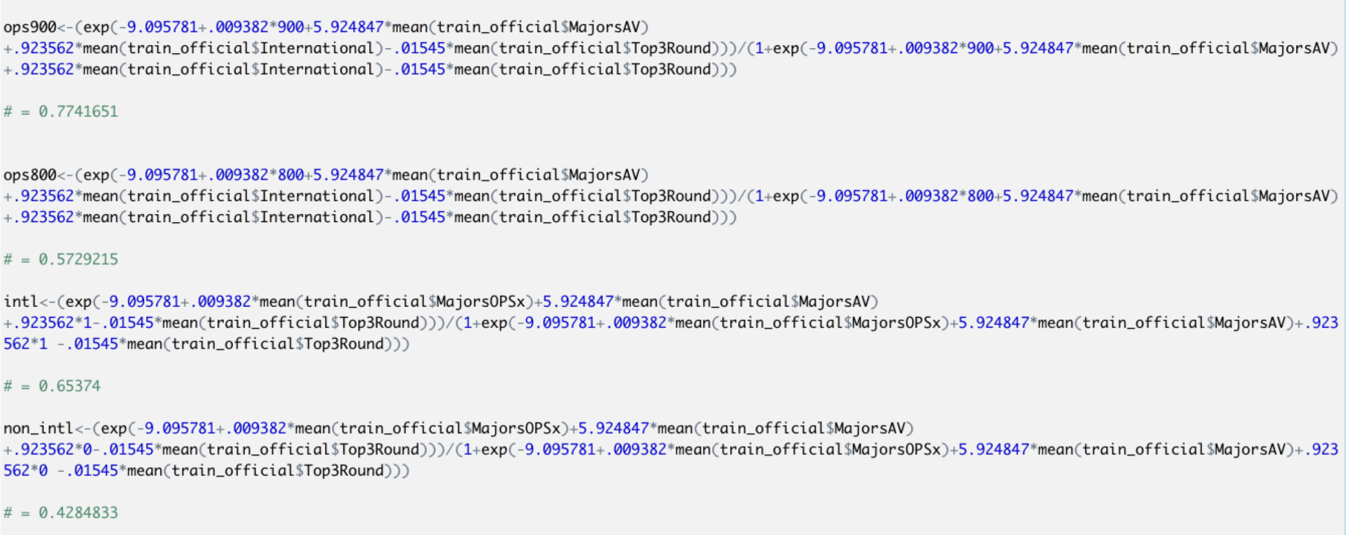
Secondly, while the four performance statistics studied—OPS, AV, ERA, and K/BB—are considered fairly robust measures of player ability, there are other, more robust measures (wRC+ for hitters, xFIP for pitchers) that would have more accurately depicted a player's true value to a team. However, these statistics are considerably more difficult to collect for minor leaguers, and in order for the major and minor league statistics used to be comparable, it was determined that wRC+ and xFIP would not be used for this study. A further analysis, replacing the statistics used with the these and other more “modern” metrics, and perhaps even Statcast batted ball data, might identify deeper connections between how pre-arbitration players are valued in extensions, and their true abilities.

Finally, this study is retrospective in nature. While Analysis II uses logistic regression to assign probabilities that a player in the test set received a pre-arbitration extension, it would not necessarily prove accurate if applied to a group of players in 2020 and beyond. However, players and teams can gain insights from this analysis about which player types, ability levels, and backgrounds are most predictive of a player having received a pre-arbitration extension, and then conduct measures of similarity between their situations and those examined in this study.

XIII: Appendix

Calculating Estimated Probabilities of from logistic regression, using:

$$\frac{e^{Log(\frac{p}{1-p})}}{1+e^{Log(\frac{p}{1-p})}}$$



```
ops900<-(exp(-9.095781+.009382*900+5.924847*mean(train_official$MajorsAV)
+.923562*mean(train_official$International)-.01545*mean(train_official$Top3Round)))/(1+exp(-9.095781+.009382*900+5.924847*mean(train_official$MajorsAV)
+.923562*mean(train_official$International)-.01545*mean(train_official$Top3Round)))

# = 0.7741651


ops800<-(exp(-9.095781+.009382*800+5.924847*mean(train_official$MajorsAV)
+.923562*mean(train_official$International)-.01545*mean(train_official$Top3Round)))/(1+exp(-9.095781+.009382*800+5.924847*mean(train_official$MajorsAV)
+.923562*mean(train_official$International)-.01545*mean(train_official$Top3Round)))

# = 0.5729215

intl<-(exp(-9.095781+.009382*mean(train_official$MajorsOPSx)+5.924847*mean(train_official$MajorsAV)
+.923562*1-.01545*mean(train_official$Top3Round)))/(1+exp(-9.095781+.009382*mean(train_official$MajorsOPSx)+5.924847*mean(train_official$MajorsAV)+.923
562*1 -.01545*mean(train_official$Top3Round)))

# = 0.65374

non_intl<-(exp(-9.095781+.009382*mean(train_official$MajorsOPSx)+5.924847*mean(train_official$MajorsAV)
+.923562*0-.01545*mean(train_official$Top3Round)))/(1+exp(-9.095781+.009382*mean(train_official$MajorsOPSx)+5.924847*mean(train_official$MajorsAV)+.923
562*0 -.01545*mean(train_official$Top3Round)))

# = 0.4284833
```

XIV: Works Cited

Bradbury, J.C. “How Do Baseball Players Age?: Investigating the Age-27 Theory.” *Baseball Prospectus*, 11 Jan. 2010, www.baseballprospectus.com/news/article/9933/how-do-baseball-players-age-investigating-the-age-27-theory/.

Brown, Maury. “Breaking Down MLB's New 2017-21 Collective Bargaining Agreement.” *Forbes*, Forbes Magazine, 3 Dec. 2016, www.forbes.com/sites/maurybrown/2016/11/30/breaking-down-mlbs-new-2017-21-collective-bargaining-agreement/.

Cameron, Dave. “The Possible Extinction of the Early-Career Superstar Extension.” *FanGraphs Baseball*, 6 Apr. 2017, blogs.fangraphs.com/the-possible-extinction-of-the-early-career-superstar-extension/.

Humphreys, B. R., and Pyun, H. (2017) Monopsony Exploitation in Professional Sport: Evidence from Major League Baseball Position Players, 2000–2011. *Manage. Decis. Econ.*, 38: 676– 688. doi: 10.1002/mde.2793.

Turvey, Jim. “The Future of Baseball Contracts: A Look at the Growing Trend in Long-Term Contracts.” *The Future of Baseball Contracts: A Look at the Growing Trend in Long-Term Contracts | Society for American Baseball Research*, 2010, sabr.org/research/future-baseball-contracts-look-growing-trend-long-term-contracts.

XV: Data Sources

Fangraphs.com

MLBTradeRumors.com

BaseballReference.com

TheBaseballCube.com

Statista.com