Josh Sellers

MATH.421.01 - FYW: Mathematical Modeling

Dr. Cherry

21 November, 2017

Most Valuable Player

SUMMARY

Major League Baseball (MLB) is a highly analytical sport. This has led to a disconnect between the teams and their fans, most of whom do not have the resources or background to replicate or understand how their favorite teams operate.

In order to bridge this gap, I formulated a model that is able to analyze a player's previous statistical performance and salary to extrapolate his future value. The model uses discrete difference equations to estimate a player's salary, statistical performance and injury potential over time. These equations are combined to output a value for the player. In order to test the equations, they were tested against historical data for three of the scenarios that the model covers: adding a free agent, retaining a currently rostered player and bringing up a minor league player.

The model is used by calculating multiple players' expected value over a given time period and comparing their yearly statistical values to ascertain which player would contribute the most value to a given baseball organization. Based on the test cases, the model is most accurate for rookies and least accurate for rostered players. This is due to the fact that rostered players tend to be older and older athletes are usually experiencing a performance decline that translates over to their statistical performance; this decline was difficult to model due to how each player experiences decline a little bit differently. Additionally, the unpredictable nature of injuries was difficult to factor and the model did not satisfactorily mitigate that variability. Until the decline and injury issue are corrected, it is better to only use the model for younger baseball players and older baseball players with short contracts.

INTRODUCTION

The popular movie *"Money Ball"*, follows Billy Bean, the then general manager for the Oakland Athletics, an American League professional baseball team. The movie shows how Bean was an early innovator in the use of mathematics and statistics to identify underrated baseball players that had been overlooked by other baseball organizations.

That film is emblematic of the continued trend in baseball; finding the best possible player within

the constraints of a team's budget. This can be seen among the composition of currently successful teams like the Cleveland Indians and the Chicago Cubs. Through some combination of good drafting, the initial step in selecting players, then player training and development, smart trades and free agent pick-ups, these teams have created rosters that consistently win. These choices are informed by an analytical strategy.

Each team's group of executives have their own unique and protected strategies for choosing players. Dedicated staff, in particular, perform critical mathematical analysis on potential and current players in order to select and pursue the best option. For the average fan, there is mystery behind their favorite team's choices. Due to their non-technical background, it only possible for most fans to make educated guesses as to whom the team will choose. However, baseball diehards, wanting as much intelligence as their team's general manager, do not want to settle for this mystery. Regular fans, however, do not have the same access to the huge sets of data, the numerous analysts and the computational resources.

While it is not possible to completely recreate the research capabilities that baseball teams have, it is possible to make a simple model for choosing players emulating the analytical models of the teams.

The model put forward in this paper allows for some insights regarding what a baseball team will decide concerning a given positional need without using a prohibitive amount of time and money. I conceived, created and tested the suggested model for this specialized player analysis.

MODEL

The model computes the impact, over time, of professional baseball players. Data generated by the model will be used to compare the three most popular options towards player selection: trade, free agency, retaining whoever is the currently rostered player for the team and bringing up an up-and-coming minor league player. This comparison will help fans answer the question of which player the team will probably chose.

The impact of these three options to the team is based on Wins Above Replacement (WAR) per fifty thousand contracted dollars. WAR quantifies the number of game wins a player adds to a team compared to an average player in that same position. It is an aggregate of offensive statistics from a given season. A player's WAR will converge over time to their expected skill level, oscillating around that value to factor in better and worse seasons. As the player ages, his WAR usually declines. Additionally, each year that a player is on the team, there is a probability that he will be injured. If injured, he will produce less than his expected WAR for the team that year. Hence, the

probability for injury and the amount of WAR lost due to an injury are incorporated; this is done by subtracting an amount based on the potential for the player to get injured combined with the percentage of the season an injured player misses, on average, due to injury.

The player's salary will follow four scenarios over time: 1) If the player is a minor leaguer, his salary will first follow the rookie salary standards for six years, gradually rising from its initial value to account for increases in the Major League Baseball minimum salary and rookie salary arbitration. Once the rookie contract is complete, the contract will follow a pattern similar to a rostered player's contract, giving the player a significant raise. 2) If a player is a free agent, he will receive an increase in his salary compared to what he earned on his previous team. 3) In the case of trading the player from one team to another, his salary will be whatever it was on his previous team. 4) For rostered players, the team will also give some sort of salary raise, similar to that of a free agent's, in order to retain his talents.

ASSUMPTIONS

To use this model, certain restrictions are incorporated. First, it is impossible to predict what happens after a player's current contract runs out. Therefore, all compared players are assumed to have both of their contracts beginning at the same time. This mitigates any uncertainty between a player with a contract that ends sooner, who gets more or less money in a subsequent contract, and a player who has a longer contract. Additionally, with the exception of the linearly increasing rookie contracts, contracts are assumed to be constant over time (ignoring incentives and bonuses).

Second, scouting intelligence is used to determine how well a player will do during his time on the team. Scouting data is based on previous statistical performance for a starter. For a minor league player, the data is based on the ranking that he is given by the sports magazine *Baseball America*, generally considered to be a good benchmark for predicting professional success. Since *Baseball America* only lists the top 100 player prospects, the top 50 of those will be given the WAR of an All-Star, one of the best players in baseball, and the bottom 50 will be given the mean WAR for starters. Any other minor leaguer will be given the minimum WAR for starters. The cutoff at starter is due to the assumption that the team's management will not promote a player for a crucial position unless they believe that he is capable of successfully competing at a professional level. Once in the model, the players will oscillate around their expected values and not diverge; this assumes that players do not dramatically improve or regress in performance.

Third, the only identified epidemiology on player injuries is associated to pitcher- and fielder-only information. Accordingly, all fielders will be assumed to have the same risk of injury. Since

catastrophic injuries are rare among baseball players, the players are assumed to never miss more than a full season because of an injury. The number by which their WAR is decreased is an estimate based on the incidence rates of injuries for pitchers and for fielders per 1,000 players combined with the mean time that injured pitchers and fielders miss due to injuries (Posner et al. 1678). Finally, with the exception of minor leaguers' rankings, all players' initial parameter values are assumed to come from Baseball-Reference.com, an online repository of current and historical baseball team and player statistics.

Beyond the initial assumptions, several additional restrictions were made in the creation of the equations and selection of the parameters and constants for this model. Players are assumed to follow typical trends for their positions and not deviate from their contemporaries. This is seen in the injury equations, where it is assumed that players will follow the average rates of injuries and missed time. For the statistical decline equations, the same assumption of following normal player patterns is made, since most players start to decline around age 28. Additionally, the salary equations assume that players will not receive huge pay raises or cuts compared to the norm. Furthermore, players are not expected to have a sudden dramatic improvement in their performance. Essentially, the model works best if players have careers in line with their projections. Another assumption made was that players will play out their contracts and not retire or demand trades. This expected loyalty permits the model's users to obtain a better idea of how different players will compare over time. One last assumption: teams will be willing to pay players the amount the model suggests; they will not be constrained by budgets. In conclusion, the combined assumptions allow for a simple and easily utilized model.

EQUATIONS

The model will be comprised of the following equations, parameters and variables.

$$V_n = \left(\frac{W_n - p_n}{log_{10}(S_n/50,000)}\right)$$

$$S_n = \begin{cases} b_f S_0 & \text{for free agents} \\ b_r S_{n-1} & \text{if } n \leq 6,\ n \neq 3, \text{ rookie and } n \geq 1 \\ b_3 S_{n-1} & \text{if } n = 3 \text{ and rookie} \\ b_c S_6 & \text{if } n > 6 \text{ and rookie} \\ b_c S_0 & \text{for rostered players} \\ S_0 & \text{for trades} \end{cases}$$

$$W_n = \begin{cases} \left(\frac{W_{expected} - W_{n-1}}{|W_{expected} - W_{n-1}|}\right) r_p + W_{n-1} & \text{if } A_n \leq 26 \text{ and pitcher} \\ \left(\frac{W_{expected} - W_{n-1}}{|W_{expected} - W_{n-1}|}\right) r_h + W_{n-1} & \text{if } A_n \leq 28 \text{ and hitter} \\ \frac{1}{2}(A_n - 26) d_p + W_{n-1} & \text{if } A_n > 26 \text{ and pitcher} \\ \frac{2}{9}(A_n - 28) d_h + W_{n-1} & \text{if } A_n > 28 \text{ and hitter} \end{cases} \quad (1)$$

$$p_n = \begin{cases} (s_p)(i_p)(W_n) & \text{if pitcher} \\ (s_h)(i_h)(W_n) & \text{if fielder} \end{cases}$$

Variables List			
Variable	Summary	Units	Range
V_n	Player value	WAR	No range for WAR
W_n	Player WAR	WAR	No range for WAR
S_n	Player salary	Dollars	$S_n \geq$ League minimum
p_n	Injury loss potential	WAR	No range for WAR
$W_{expected}$	Expected WAR	WAR	Value based on WAR table
A_n	Age during contract	Years	Must not be a student
S_0	Initial salary	Dollars	$S_0 \geq$ League minimum
W_0	Initial WAR	WAR	No range for WAR
n	Number of years since first year of contract	Years	$n \geq 0$

Parameters List			
Parameter	Summary	Units	Value
s_p	Percentage of season missed for pitcher	N/A	0.41
s_h	Percentage of season missed for hitter	N/A	0.23
d_p	Rate of decline for pitchers	WAR	-0.187
d_h	Rate of decline for hitters	WAR	-0.576
i_p	Injury incidence rates for pitchers	N/A	0.00416
i_h	Injury incidence rates for hitters	N/A	0.00210
r_p	Rate of increase in WAR for pitchers	WAR	0.695
r_h	Rate of increase in WAR for hitters	WAR	0.651
b_f	Boost in salary for free agent	N/A	1.81
b_r	Boost in salary for a rookie	N/A	1.33
b_c	Boost in salary for rostered player	N/A	1.66
b_3	Boost in salary for rookie arbitration	N/A	4.5

For a player's initial WAR, the estimation will be based on the explanation given by Baseball-Reference.com, which is summarized in the table below.

WAR Explanation	
WAR Range	Equivalent
8+	MVP
[5,8)	All-Star
[2,5)	Starter
(0,2)	Reserve
$0 \leq$	Replacement

The estimation for WAR for major leaguers, denoted in the model as $W_{expected}$, will be based on if they won a Cy Young Award, Most Valuable Player (MVP) award or had been previously voted an All-Star. The Cy Young Award recognizes each baseball league?s - American and National - best pitcher, while the MVP more broadly honors each league?s top-performing player. Baseball, as does other types of team sports, annually creates an middle-of-season super team of All-Star players, another widely recognized badge of top performance.

The number of these received honors will dictate how high their estimated WAR will reach within its range. For each Cy Young award bestowed to and All-Star appearance made for non-MVP players, the WAR will be increased by 0.5, capping off at 8. For each MVP, the WAR will be increased by 0.2. If they had not made an All-Star appearance or received an MVP, then the WAR

estimation will be based solely on the upper bound of whichever bracket their W_0 belongs. For the case of rookie hitters, their expected WAR will be the mean amount for whatever category they are projected to attain, and their initial WAR will be half of their expected WAR. For rookie pitchers, their expected WAR will be the lower bound of their projected category.

EXPLANATION OF EQUATIONS

The model's three main equations are fairly straightforward. The W_n and p_n equations are the easiest to summarize. For the injury loss potential equation, it is the probability of getting injured and the average amount of time missed multiplied by the player's WAR (Posner et al. 1678). That loss in WAR is split into two cases: pitchers and everyone else. For the WAR equation, there are four cases; 1) before age 26 for pitchers; 2) after age 26 for pitchers; 3) before age 28 for hitters; 4) after age 28 for hitters (Fair 26). These two ages are the approximate time that pitchers and hitters respectively start to decline statistically. The fractions $\frac{2}{9}$ and $\frac{1}{2}$ for the pitcher and hitter decline equations simulates how pitchers decline faster than hitters. The rates of improvement and decline for pitchers and hitters were calculated by sampling 152 and 183 seasons, from a set of players, and finding the average change in WAR before and after the two sets' decline cutoff. The bounding function of $\left(\frac{W_{expected}-W_{n-1}}{|W_{expected}-W_{n-1}|}\right)$ uses the player's expected WAR, whose calculation was discussed in the previous section, as an upper bound for the player's WAR. Besides the WAR and injury equations, the most complicated equation was for player salary (S_n).

For the salary equation, there are six possible cases. The most basic of these involves when a player is traded: in this scenario, his salary would remain the same as the one he received from his previous team (S_0). If the player is a free agent, his new salary is calculated by multiplying his previous salary by the free agency boosting constant, b_f. The constant was calculated by utilizing the data concerning the salary increases that 50 players received via free agency and averaged those salary increases. The free agency boosting rate was determined to be 0.807985381 $\pm$ 0.11 on a 90% confidence interval. Another 90% confidence interval was used to calculate the amount of salary increase for currently rostered players; for a sample of 62 players, the interval was found to be 0.661645563 $\pm$ 0.096.

Finally, there are three cases for rookies: before , during and after salary arbitration. The rookies' S_0 is expected to be the league minimum, since only a small number of rookies receive greater starting contracts. Based on a sample of 110 players, the increase in their salaries, b_r, was identified as 0.330120501 $\pm$ 0.13 on a 90% confidence interval. Rookies are eligible for salary arbitration after their third season (mlb.com). Based on a sample size of 23, the 90% confidence interval for this increase was shown to be 3.500610892 0.97. After a rookie's sixth season, he will follow a similar

renumeration pattern as a rostered player, receiving a raise based on his sixth year and b_c. The salary equation, the WAR equation and the injury equations were combined into the first equation, which calculates player value.

The player value equation, V_n, is a player's WAR subtracted by the injury potential divided by the logarithm of his salary divided by \$50,000. The league minimum salary is currently \$507,500; the division by \$50,000 means that the minimum logarithm for the denominator around one (the divisor was adjusted by testing against historical data with lesser minimums) (Associated Press). The logarithm was used because of the wide range of salaries for players; it corrects for that disparity in the values. This is the model's output, and what will be used to compare different players.

The main issue with the equation is in the constants, some of which have significant variability. This variability is most likely due to the wide range of abilities found among Major League Baseball players. The constant represents the middle ground of abilities; this is what led to the assumption that players' numbers will be in line with their projections. If they follow the average career growth and decline, the model will have a reasonable amount of accuracy.

VALIDATION

To check the model's validity, historical data for various players was compared to the model's output. To guarantee a fair test, players who experienced catastrophic injuries were not selected. Additionally, players with a lot of variability in their statistical performance over the course of their careers were not chosen, since it already know that the model would not be able to accurately match their historical data. The first test was run on three players' rookie years. Due to the known variability of rookie performance compared to expectations, only rookies that had a ranking from *Baseball America* were picked for the tests. The three players chosen were Craig Kimbrel (pitcher), Jason Heyward (hitter) and Madison Bumgarner (pitcher). Their rookie rankings from *Baseball America* were respectively 84, 1 and 14. The Table below shows their projected values compared to their projected values; a divisor of \$40,000 was used against their league minimum.

Craig Kimbrel Test	
Expected V_n	Actual V_n
1.712501393	2.352585909
2.13358191	2.8234279
2.472395263	2.717880745
1.99287864	1.114559783
1.532892321	0.5526782
1.35927698	0.367483053
1.15588934	1.428483063
Totals	
12.35941585	11.35709865
Error	
0.088254688	

Jason Heyward Test	
Expected V_n	Actual V_n
3.24843025	6.4
3.469422186	2.285485974
3.646541169	5.043528887
2.735779756	1.887530839
2.889799792	3.022690869
3.026063648	2.838385048
2.470175799	-0.109740088
Totals	
21.4862126	21.36788153
Error	
0.005537801	

Madison Bumgarner Test	
Expected V_n	Actual V_n
2.374285317	2.283208144
2.714501537	1.832256026
2.989835517	3.063626161
2.344784648	2.028427111
2.539112555	2.155130049
2.080752303	2.094728109
1.814345892	1.17951491
Totals	
16.85761777	14.63689051
Error	
0.151721246	

The model did a good job of predicting the rookies' values. The most divergent output was for Madison Bumgarner. He had an injury in the final year of the time period tested, which accounts for this error.

Another three players were tested under the rostered scenario: Brandon Phillips (hitter), Joe Mauer (hitter) and Felix Hernandez (pitcher). The actual raises in their respective salaries were \$12,000,000, \$23,000,000 and \$25,000,000 (spotrack.com). The predicted and actual outputs are shown below.

Brandon Phillips Test	
Expected V_n	Actual V_n
1.611847605	1.680523111
1.372834085	0.7562354
1.038215156	0.714222322
0.607990819	1.470457722
0.082161073	0.336104622
-0.53927408	0.336104622
Totals	
4.173774657	5.293647799
Error	
0.211550368	

Joe Mauer Test	
Expected V_n	Actual V_n
2.411763129	0.56332573
2.317516243	1.614867093
2.129022469	1.99041758
1.846281809	0.788656022
1.469294262	0.56332573
0.998059829	0.86376612
0.432578508	1.276871655
Totals	
11.60451625	7.661229931
Error	
0.514706692	

Felix Hernandez Test	
Expected V_n	Actual V_n
1.83566295	1.926660908
1.791752497	2.519479649
1.660021138	1.630251538
1.440468873	0.592818741
1.133095702	0.296409371
Totals	
7.861001159	6.965620207
Error	
0.12854289	

Error was identified for this test, and is due to the variability in when and how much a player declines. Since players do not receive these types of contracts until the middle or the end of their

careers, a high likelihood exists that they will begin to decline during their contract timeframes. Thus, the variability speaks to the inadvisability of handing out these contract types.

The final test was on free agency. For this test, the three players were Robinson Cano (hitter), Albert Pujols (hitter) and Darren O'Day (pitcher). Their salaries respectively were $24,000,000, $24,000,000 and $2,900,000 (spotrack.com). The predicted and actual outputs are shown below.

Robinson Cano Test	
Expected V_n	Actual V_n
2.617647472	2.386954188
2.391744801	1.268069412
2.075481062	2.72261962
1.668856255	1.268069412
Totals	
8.75372959	7.645712633
Error	
0.144920037	

Albert Pujols Test	
Expected V_n	Actual V_n
1.698716092	1.790215641
1.426240358	0.559442388
1.062939379	1.454550208
0.608813156	1.156180935
0.063861688	0.522146229
-0.571915024	-0.671330865
Totals	
4.288655648	4.811204535
Error	
0.108610823	

Darren O'Day Test	
Expected V_n	Actual V_n
1.244267985	1.134154617
0.930685622	1.304277809
0.512575805	1.587816463
Totals	
2.687529412	4.026248889
Error	
0.332497944	

For this test, the same decline-based error occurred. Due to the shorter contract time period, that error was mitigated. Hence, free agency will lead to some error, but not as much as giving a raise to a rostered player.

ANALYSIS AND APPLICATION

Overall, the model works well with rookie contracts and free agency negotiations. The potential for error is higher when applied to raises for rostered players.

The model's limitation is due to the variability in performance as a player ages. Since rookies are least likely to experience performance decline, due to their youth, the model is most accurate for them. For free agency candidates, the players are either younger or the contracts are not long enough to be affected by performance decline. Therefore, rostered players, who tend to be older, are the most sensitive to decline, which is tied to age.

Due to the difficulty in encompassing the effects of age-related decline, the majority of sensitivity analysis for the model was done on the decline case for pitchers and hitters in the WAR equation. The fractions used for their equations, $\frac{2}{9}$ and $\frac{1}{2}$, were the main target of the analysis. They were found to cause a great deal of variation in the model's output from slight positive and negative

changes in their value. As such, in the future, those two values will be the first target for updates when improving the model, with attempts made to give them a wider range of accuracy.

This model's best application would be either for younger players or shorter contracts. In general, it is unwise to award a long contract to any player; pitchers, in particular, tend not to receive long contracts, due to the unique wear on their bodies that pitching inflicts. This restriction, however, does not have a major impact on the model's efficacy.

An example for using the model can be seen in the test data. Craig Kimbrel and Darren O'Day are both relief pitchers. Even though Kimbrel wound up earning more, via arbitration, at the end of his contract, he demonstrates greater value, based of the model's output, as a targeted player over the same period of time as O'Day. That quantitative analysis can also be confirmed by looking at the historical data the test was conducted upon, Kimbrel had a much higher WAR. Furthermore, the model can also provide deeper insights. Felix Hernandez is a well-known pitcher with a track record of winning games. A team's leadership would not be faulted for picking Hernandez over most pitchers. The model demonstrates though that he is beginning to decline in the amount of WAR he produces for his expected salary. Based on that information, it would then be better to pick a younger pitcher, like Madison Bumgarner.

CONCLUSION

The best scenario for this model's application is for players that follow expected trends and do not exceed or fall below performance expectations. Since most players fall into this category, this model should be useful for many situations.

This model is not perfect: its limitation lies in the fact that it does not sufficiently factor in wild shifts in player skill-level, decline in skills and injury probability. Since those factors have the most influence in longer contracts for older players, analyzing those situations should be avoided unless the model is refined further.

Overall, this is an intriguing model. It will do a good job of describing player value for different team positions. By employing the model, armchair general managers can understand a portion of the analysis done behind closed doors, and gain a greater sense of connection with their teams.

Works Cited

BaseballAmerica.com, www.baseballamerica.com/.

Baseball-Reference.com, www.baseball-reference.com/.

Fair, Ray. "Estimated Age Effects in Baseball." Journal of Quantitative Analysis in Sports, vol. 4, no. 1, 2008, doi:10.2202/1559-0410.1074.

Posner, Matthew, et al. "Epidemiology of Major League Baseball Injuries." The American Journal of Sports Medicine, 27 June 2011, journals.sagepub.com/doi/abs/10.1177/0363546511411700.

Nowak, Joey. "Everything You Need to Know about Service Time." Major League Baseball, 1 Apr. 2015, m.mlb.com/news/article/115853014/everything-you-need-to-know-about-service-time/.

Press, Associated. "MLB Minimum Salary Remains at $507,500 for 2016." ESPN, ESPN Internet Ventures, 18 Nov. 2015, www.espn.com/mlb/story/_/id/14161690/mlb-minimum-salary-remains-507500-2016.

Spotrac.com, www.spotrac.com/.