An Investigation of Basic and Sabermetric Statistics for Individual Players as Predictors for MVP Candidates in Major League Baseball

Micheas Yimam
Math Major Class of 2019
College of Wooster
Wooster, OH
myimam19@wooster.edu

Kory Sansom
Math Major Class of 2019
College of Wooster
Wooster, OH
ksansom19@wooster.edu

Abstract

The premise of the paper is to analyze baseball player's worthiness with a combination of sabermetric and basic stats. By comparing sabermetric and basic stats of MLB players since 1975, we will be able to infer how the voter committee decides most valuable player points. Additionally, a step-wise regression will be performed to remove insignificant factors, thus enhancing our model to accurately predict MVP winners from 2010-2017 Major League Baseball seasons. **Keywords:** step wise regression, basic, sabermetric, generalized linear model;

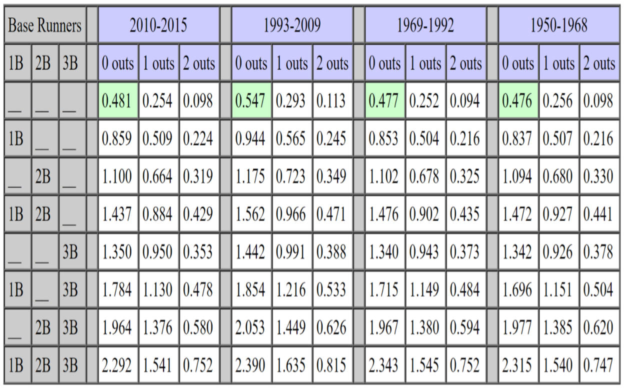
Background

In sports, statisticians, general managers, and fans all use statistics to summarize or explain a players contribution to the team and how he stacks up against other players. In baseball specifically many fans look at batting average, RBIs (runs batted in), and home runs to compare position players among other teams. These numbers are constantly plastered on ESPN, FOX Sports, and other sports media to rank potential MVP (most valuable players) candidates. Statisticians have invented an all-in-one statistic to determine MVP worthiness. Most fans of the MLB use WAR (Wins Above Replacement) and others use WPA (Wins Probability Added) to determine whether a player is the most valuable player for their team.

This leaves us with three questions that we plan to address with our project. First we need to separate our two statistics which can be classified as sabermetric or basic. Sabermetric stats, which are statistics calculated with a multitude of variables; examples of sabremetric statistics are WAR and WPA. Basic statistics are statistics that are recorded and accumulated for every instance. Since we are only looking at position players the statistics that are in this category are hits, plate appearances, at bats, walks, home runs, RBIs, strikeouts, stolen bases, and caught stealing. Thus once the model is created we are looking to find which predictor basic or sabermetric stats is more successful. Second question would be, when running a regression with all of these attributes which statistic affects the model the most, and what combination of sabermetric and individual statistics will have the greater impact. The final question is, by crafting what specific data set we use by using win expectancy methodology will that allow for better predictors of writers voting tendencies in the MVP races? The MVP race is decided by the baseball writers association, similarly our data will be taken from *baseball-reference.com* where we will gather both individual and sabermetric data from 1975 - 2017.

Process and Data

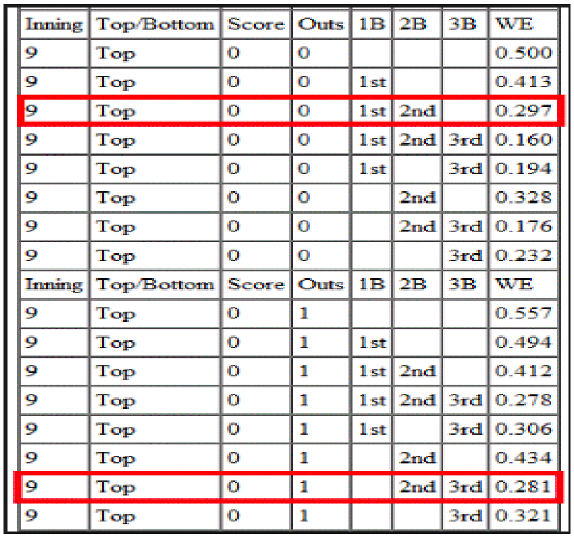
When deciding what data would be collected and what parameters would be considered all of this was based on the articles and books on this topic. For example, when considering the basic individual statistics was required we only looked at stats that contributed to runs and wins. Baseball statistician Tom Tango highlights all the possible combination of run expectancy for each possible scenario man on first no outs, man on first one out, etc. [7]



Base Runners			2010-2015			1993-2009			1969-1992			1950-1968		
1B	2B	3B	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs
_	_	_	0.481	0.254	0.098	0.547	0.293	0.113	0.477	0.252	0.094	0.476	0.256	0.098
1B	_	_	0.859	0.509	0.224	0.944	0.565	0.245	0.853	0.504	0.216	0.837	0.507	0.216
_	2B	_	1.100	0.664	0.319	1.175	0.723	0.349	1.102	0.678	0.325	1.094	0.680	0.330
1B	2B	_	1.437	0.884	0.429	1.562	0.966	0.471	1.476	0.902	0.435	1.472	0.927	0.441
_	_	3B	1.350	0.950	0.353	1.442	0.991	0.388	1.340	0.943	0.373	1.342	0.926	0.378
1B	_	3B	1.784	1.130	0.478	1.854	1.216	0.533	1.715	1.149	0.484	1.696	1.151	0.504
_	2B	3B	1.964	1.376	0.580	2.053	1.449	0.626	1.967	1.380	0.594	1.977	1.385	0.620
1B	2B	3B	2.292	1.541	0.752	2.390	1.635	0.815	2.343	1.545	0.752	2.315	1.540	0.747

Figure 1: Run Expectancy Matrix

Tango then uses these run probabilities to create a 'win probability matrix' to illustrate the game state change with the amount of runs already scored, how many base runners are on, how many outs, and what inning the game is in. [7] When looking at win expectancy data common beliefs in baseball are sometimes proven less significant. For example, when looking at figure 2 we see with no outs and runners on first and second in a tie game it is widely accepted that using a sacrifice bunt to move the runners to second and third for the cost of a out is worth it. However according to the highlighted win expectancy as a team you are only slightly at an advantage by bunting to the tune of 0.016.



Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
9	Top	0	0				0.500
9	Top	0	0	1st			0.413
9	Top	0	0	1st	2nd		0.297
9	Top	0	0	1st	2nd	3rd	0.160
9	Top	0	0	1st		3rd	0.194
9	Top	0	0		2nd		0.328
9	Top	0	0		2nd	3rd	0.176
9	Top	0	0			3rd	0.232
Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
9	Top	0	1				0.557
9	Top	0	1	1st			0.494
9	Top	0	1	1st	2nd		0.412
9	Top	0	1	1st	2nd	3rd	0.278
9	Top	0	1	1st		3rd	0.306
9	Top	0	1		2nd		0.434
9	Top	0	1		2nd	3rd	0.281
9	Top	0	1			3rd	0.321

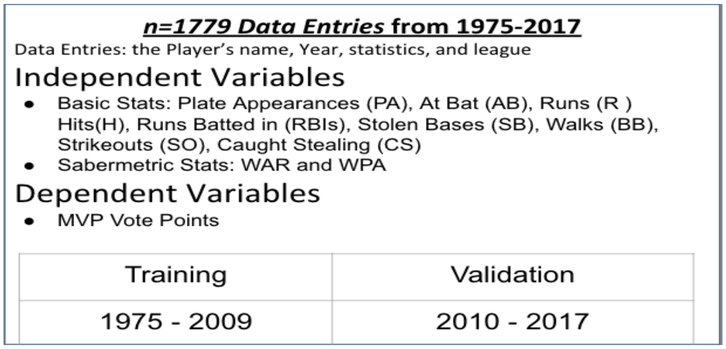
Figure 2: Win Expectancy Matrix (From the Home Team Standpoint)

Why all of this is relevant in deciding what data points to include in our analysis relates back to the original question; how do you determine a player's value. In baseball, the goal is to win the game, and the player that best puts you in that position to win is the most valuable. Thus by using Tango's game state methodology where when a hitter gets on base he positively impacts the game for his team by increasing the run expectancy and win expectancy. However, when that hitter strikes out he negatively impacts the win and run expectancy. That is why plate appearances and at-bats will be used as a measure of total chances; hits, runs, runs batted in, stolen bases, walks all are a measure of positive impact; while strikeouts and caught stealing are measures of negative impacts. For the first case only these statistics will be considered.

That leads to the question: is there a single statistic that best represents a players worth? Which in turn answers the question who is the most valuable player. Wins above Replacement (WAR), is a statistic that generates the number of wins a player adds to a team compared to a replacement level player [2]. WAR is used commonly in baseball conversations to talk about who deserves to win MVP; it even is used as a metric to judge potential Hall of Fame candidates. WAR has become almost industry standard for evaluation players worth even when looking at the past five years in the American league three times the Most Valuable player award has been given to the league leader in WAR. Yet, there are shortcomings of WAR such as in the calculation for WAR there is an adjustment for what is considered ”league average” and ”positional adjustment”. These values are constants across the MLB for each position not specific by team [6]. Where this is a problem for example is the Red Sox where if Dustin Pedroia gets hurt a player like Eduardo Nunez would come off the bench to replace him with a season average WAR of 1.5 the past three seasons. If 0 is the WAR of a replacement level player and the Red Sox would replace an injured player with a WAR of 1.5 then Dustin Pedroia true Wins above replacement would be his WAR - Eduardo Nunez WAR. Also looking at the other side of the spectrum Mike Trout who has a career average WAR of 6.95 would be replaced with Jefery Marte if he was injured and his career average WAR is 0.2.

Another stat like WAR is Wins Player Added (WPA). WPA is not as well known as WAR, but still garners some respect in the sense of explaining a player's worth. WPA is based purely on cumulative assortment of all the positive and negative impacts of game states an individual player had throughout the season. What is unique about this statistic is that most sabermetric statistics do not consider the situation of a particular event or how some plays are more crucial to a win than others [3]. Even though WPA is very useful in determining a players contribution in an individual games WPA fails in the specificity of it's own equation. The way WPA is calculated now, there are only positive and negative impacts on pitchers and hitters, but that is not how baseball is really played. If there is a ground ball to second with two outs, and a runner on third in a tie game and the second basemen makes an error and the run scores. In the current form of WPA the batter would be awarded a positive WPA for getting that run and the pitcher would be penalized for giving up the go ahead run; when in reality the fielder made the costly error, and his WPA is unaffected because WPA does not take in account fielding.

Below in figure 3 you will see how we broke apart the data into different sub groups.



***n=1779 Data Entries* from 1975-2017**

Data Entries: the Player's name, Year, statistics, and league

Independent Variables

- Basic Stats: Plate Appearances (PA), At Bat (AB), Runs (R) Hits(H), Runs Batted in (RBIs), Stolen Bases (SB), Walks (BB), Strikeouts (SO), Caught Stealing (CS)
- Sabermetric Stats: WAR and WPA

Dependent Variables

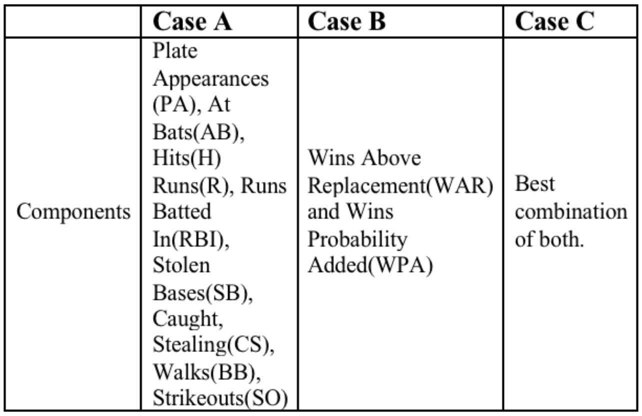
- MVP Vote Points

Training	Validation
1975 - 2009	2010 - 2017

Figure 3: Data breakdown

Methodology

The methodology for how all the statistics that were collected are then broken down into their respective cases. Each case will use the statistical components of each player for a given year to predict the total number of MVP votes they get.



	Case A	Case B	Case C
Components	Plate Appearances (PA), At Bats(AB), Hits(H) Runs(R), Runs Batted In(RBI), Stolen Bases(SB), Caught, Stealing(CS), Walks(BB), Strikeouts(SO)	Wins Above Replacement(WAR) and Wins Probability Added(WPA)	Best combination of both.

Figure 4: Cases

Limitations with Data

We understand that our model has limitations that may affect the results and predictions. Listed below are potential limitations of our models.

- **Data does not consist of pitchers:** Pitchers have won three MVPs between both leagues since 1975. Two of them in the past six years. However we felt since the MVP is traditionally a position player award because the average pitcher only plays in a fifth of the season compared to a position player barring injury will play the entire season. However that isn't necessarily how all the voters see it. When looking at the data even though in most circumstances you won't find pitchers in the top three of vote getters, you see them spread out across the top 25. These instances of voters granting even a small percentage of the vote total for pitchers will hurt our model because we did not consider them at all.
- **Data does not take into account defensive statistics** Since defensive stats are very difficult to be used in comparisons across positions. For example, comparing the errors of a shortstop and a first basemen would not reflect who is the better defender because the shortstop plays a more demanding defensive position. Without traditional statistics we would have been obligated to use only sabermetric data like Defensive Uniform Zone Rating which is the standard in evaluating players across all positions. Thus since WAR already has an inclusion of an adjustment of for position an additional defensive metric would be redundant, and not necessarily add any substantive information [1].
- **Limited offensive statistics** When selecting data we had to be realistic in the amount of data we could collect. The first step was limiting the amount of data we collected and we did that by doing research on the most influential statistics [4]. The most influential stats are categorized by the stats that affect the run expectancy of the game which in turn affects the win expectancy. However including all the basic and sabermetric statistics could have only improved our model not made it worse by preforming a step wise regression.
- **Steve Garvey Effect:** The Steve Garvey Effect also known as the Joey Votto effect is more of a critique of the writers in the MVP selection then the actual statistics the player garners throughout the season. As Bill James explains [5] when a player wins his first MVP he his held to a higher standard in the seasons following. For both Steve Garvey and Joey Votto following their initial MVP season. Also the Steve Garvey effect can work in reverse where players who won an MVP or have had many MVP caliber seasons earlier in their career will warrant sympathy votes from the baseball writers and win an MVP late in their career like Barry Larkin in 1995 or Miguel Tejada in 2002 after a 16th and 19th finish.
- **Coors effect:** The Coors effect is another voter trend that Bill James mentions in [5]. This is a voter bias that has effected multiple players throughout the years in MVP selections. In the past players like Jason Giambi and Todd Helton who had great numbers were not highly considered for the MVP because they played for the Rockies. The Rockies home stadium is a mile higher than sea level, so because they play at a higher altitude the ball "jumps" off the bat differently. The writers feel because these players are at an unfair advantage because they play so many games here that there numbers are actually inflated. Even though a simple solution as Bill James mentions is by looking at home and road splits meaning separate the statistics into two categories one for there home games and one for there away games and see if these players still stack up with other players who do not have this advantage.
- **Teams success:** This seems to be the strangest limitation to include because it has nothing to do with the individual player,but the number of wins a team has is probably the largest [5] bias or impact besides individual numbers on a player winning the MVP. Some of the baseball writers feel that if a player is not on a playoff team they shouldn't even be considered for the award because their impact wasn't enough to land there team a top ten finish. After our analysis of our data it will be interesting to look at the players who failed our model, if there teams wins led to a bias against them that led to less MVP votes.

Analysis and Results

For our analysis, we performed a comparison analysis to determine whether sabermetric statistics or individualistic statistics were better predictors for MVP votes in their

Accuracy (%)	GLM (Saber)	GLM (Individual)
American League	66.7%	66.7%
National League	70.8%	70.8%

Table 1: Comparison of Model Accuracy

respective years and league. Major League Baseball rewards most valuable player in both the American League and the National League, thus when constructing our model, we had to separate our data frame across the leagues. With the total data entries at 1779, the data frame was split with 909 data entries of the years for National League and 870 data entries for American League.

Our first model constructed was the sabermetric model for American League and National League. The sabermetric model was conducted via MatLab using a generalized linear model. Although the model only consisted of two explanatory variables, the linear model proven to predict top 3 MVP vote getters from year 2010-2017 in the National and American League 70.8% and 66.7% accuracy, respectively. We understand that this model is not using the significant factors which may cause a decrease in our accuracy; however, we are fascinated to see the strength of sabermetric statistics against individual player statistics.

Our second model was the individual statistic model for American League and National League. The individualistic model was conducted via Matlab using a generalized linear model. The reason for determining these specific individual statistic can be found in section . Once again, we understand factors that may not be significant are included in our model; however the purpose of included all factors is to compare strength of models. Our predictions using the individual statistic model yielded 66.7% and 70.8% for American and Nation League respectively.

From initial results, the accuracy of the models seem incorrect due to the similarities; however further investigation revealed the coincidence. Although the accuracy is identical for both models, the specific validation data entries that weren't captured are different. Typical faults in the sabermetric model are players that have extraordinary individual statistics, but have low WAR and WPA. These players could be found on lower performing teams perhaps. It is only coincidental that the accuracy for sabermetric and individual are identical for both American and National League. Based on our comparison analysis, both models fail to make accurate predictions for the seasons 2010-2017. In our next analysis, we hope that the step-wise linear regression will remove any insignificant factors which will result in a higher accuracy for the model.

For our third model, we wanted to run a step-wise linear regression to determine which predictors are significant in determining MVP votes. The table below breaks down the p-values for each predictors. From Table 1, the insignificant factors in the step-wise regression for the American League are plate appearance, stolen bases and strikeouts. These factors had P-values greater than 0.05, thus resulting in the removal of them from our final model. With only significant factors in the final linear regression, our model's accuracy increased to 83.3% for the American League. Our model significantly increased its strength from 66.6% to 83.3%. Next we built a different step-wise model for the National League because we found that different factors were significant. The table below depicts the p-values for all factors in the National League.

	WPA	WAR	PA	AB
P-value	5.69e-6	0.0012	0.7913	7.65e-9
	RBI	SB	CS	BB
P-value	0.0299	0.7403	1.29e-4	0.0032
	R	H	SO	
P-value	0.0377	2.53e-7	0.37122	

Table 2: American League Factors

	WPA	WAR	PA	AB
P-value	9.37e-5	0.0157	0.6915	1.86e-7
	RBI	SB	CS	BB
P-value	8.07e-5	0.0044	7.46e-4	0.4748
	R	H	SO	
P-value	0.5369	2.57e-5	0.001	

Table 3: National League Factors

From Table 2, we can see that plate appearance, runs and walks are insignificant factors in our step-wise regression model for the National League. After removing these factors, the final model yielded 70.8% accuracy. Refer below to figure 5 to see the overall comparisons among the models.

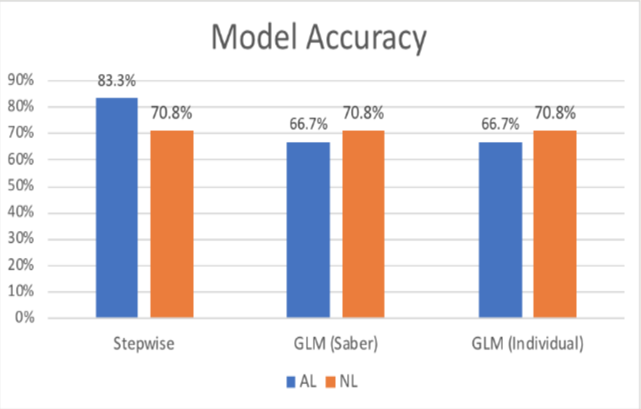


Figure 5: Model Accuracy Comparison

We understand that our model has many limitations that could affect our accuracy; however, we wanted to investigate the players that our final model failed.

Extended Analysis

From our step wise regression we still failed 17% of the time in our best case scenario in the American League. Here is a collection of the list of players that failed to meet our model in both generalized linear models. Meaning these are the players that received votes for MVP, but we did not have them finishing top three in MVP votes in our model.

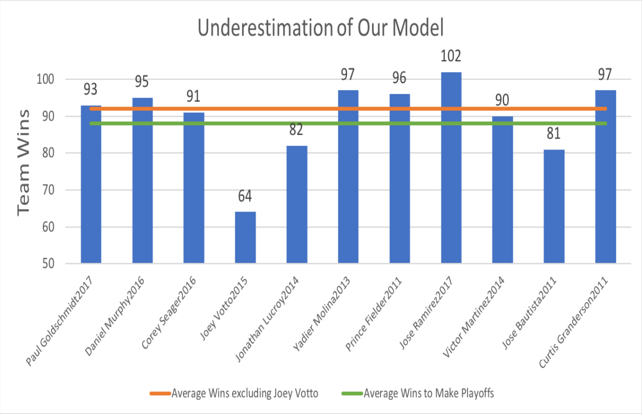


Figure 6: Underestimated Players

As we can see the first thing that jumps out from figure 6 is the fact that most of the players competed on playoff teams. As stated in our limitations we expected players who failed our model to play on good teams because the writers acknowledge there bias in relation to the team's success. As you can see in the graph the average number of wins for these players was four wins higher than the wins required to make the playoffs. The fact that we can justify for eight of the eleven players our model failed directly with team success allows us to focus on that limitation as a major cause is a lower accuracy.

However it is important to not stray from the original question that was proposed which was when looking at sabermetric stats in relation to individual stats which of these stats would build a better model in predicting the MVP. Identifying this limitation as a large reason why these players were not represented can impact future work done in this topic.

In addition looking at players who are in the top three that our model failed for 13 out of the 16 MVP races the second and third place vote getters did not attribute for 25% of the vote. Meaning that these MVP races were very definitive. Predicting the second and third place vote getters gets even more difficult with a smaller differential between vote percentage. For example when we look at the 2011 season Joey Votto who was sixth on the list of total votes and included in our model in comparison to Prince Fielder who finished third on the list and was not. The difference in vote percentage was 15% even with the Steve Garvey effect hurting Joey Votto's total votes.

Now when looking at the players we overestimated it was not as clear as looking at their teams success so we had to look at a multitude of limitations. As you look below in figure seven you will see the players we included in our model that did not receive enough votes to be in the top three for that year. The color coordination represents the reasons we believe they were left off the ballot mostly due to voter bias referenced earlier in the limitations.

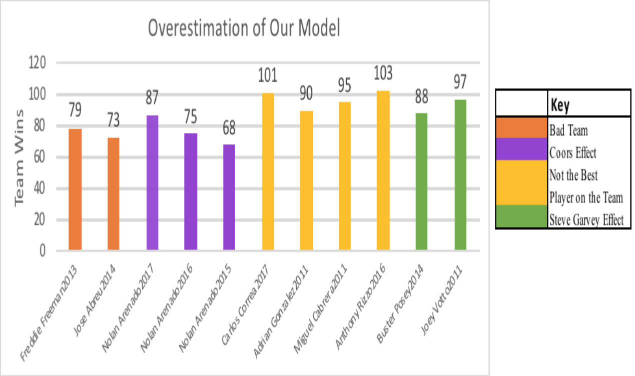


Figure 7: Overestimated Players

If we look at this graph we can point out the Coors effect and the Steve Garvey effect almost immediately. Nolan Arenado who plays for a usually under preforming Rockies team would be penalized when it came to MVP voting because of the Coors effect. Arenado who had a WAR and the offensive stats to top almost every year's candidates of MVP's was consistently left off the top three in votes and this limitation was cited as the cause [5]. Then looking at Joey Votto in 2011 who had similar numbers to his MVP year in 2010 was an obvious victim of the Steve Garvey effect. The other players had great seasons in there own right, but there teammate had won the MVP that year, and for the baseball writers to award one player from a team the lesser quality player will be affected by most of their potential votes going to there more deserving teammate. We see this with Carlos Correa in 2017 and Anthony Rizzo in 2016 whose teammates both won the award.

Discussion

There were a lot of options for further work with our best accuracy percentage being only 83 percent. Here is a list of three ideas we had that we could improve our model.

- **Including team success:** By including the team's wins for that season we would be differing from original question, but over 70% of the players that failed our model played on playoff teams. Without a doubt including this into our a model would improve our accuracy because we would be introducing a attribute outside of individual performance that the voters obviously weigh heavily. However, for this project going back through data collection to collect the number of wins for each player's team during there MVP caliber season isn't feasible for an extension.
- **Including boolean variables for voter bias:** Identifying the Coors effect and the Steve Garvey effect from research allowed us to expect certain things in the players that were included in our model that didn't win the MVP. Even though these bias aren't associated with the statistics of the individual player by incorporating these voter trends our model will only improve it's accuracy in preventing players who preformed well in our model in doing so in the future. Yet, like the above situation going back and combing through data to add this information would be difficult at this point.
- **Incorporating predictive stats for the season to predict the MVP of future seasons:** Using our model as it stands now and with an understanding of its shortcomings taking the full predictive statistics for 2018 and predicting the winner and runner up both in the American and National League. Meaning take a data set from Fangraphs that has already formulated the predictions of individual performance; run those values through our model; see the results we get, and with some discretion from what we know our model fails at make an informed prediction.

Results of Extension

Although our model showed predictive power for the years 2010 to 2017, baseball enthusiasts are not curious about what has already happened. People want to know who will be in the top three for the current or upcoming season, thus we built a historical model on the years prior to 2017 in order to predict 2018 season. Based on our step-wise linear regression, our projection model will use the significant factors found in Analysis and Results . Additionally, the testing set for the projection model was scraped from Fangraphs.com. The table consists of projection stats for the current year. We understand that these are only projection stats and may increase the error of our model; however, the results may be interesting for the baseball world.

Some limitations to this projection model include restricted projection statistics, inability to account for injury and league changes. Due to the nature of WPA and how the statistic is calculated, projections are not applicable, thus we had to remove this statistic from our model. Deleting this statistic may result in false positive results; however, we decided to go forth with this model. The model also doesnt take into account player injuries and play league changes. Players with serious injuries during the season will jeopardize their contention for top three; however, our model doesnt take this into account. Similarly, if a player happened to change leagues, the result may be catastrophic for top three candidacy. Our model is separated between the leagues, thus factors for each are different. This could affect a players MVP potential by not conforming to the other leagues significant factors.

With all limitations in mind, the model was built to predict the 2018 top three players for each league. The table below highlights the potential top three players for National and American League.

League	Player
American	1. Mike Trout 2. Giancarlo Stanton 3. Manny Machado
National	1. Bryce Harper 2. Kris Bryant 3. Nolan Arenado

Table 4: 2018 Predicted Top Three in MLB

- **Mike Trout:** Arguably the best player in the MLB, who already has a career WAR higher than some Hall of fame inductees.Trout has won two MVP's already, but the Steve Garvey effect doesn't pertain to him because he is a top five vote getter every year he has been in the league.
- **Giancarlo Stanton:** Stanton who won the MVP last year for his obscene power numbers in a huge stadium is projected to shatter those offensive numbers according to fangraphs. Stanton moved to the Yankees which will put him in a significantly better lineup and a much smaller park which all contributes to his highly projected offensive statistics. However as we know due to the Steve Garvey effect Stanton who isn't like Mike Trout will be hindered by this and could hurt his vote total.
- **Manny Machado:** Manny Machado was already an offensive force for his team as a third basemen. This season he is moving to shortstop a more valued position for offensive performance. Manny's increased projections is reasonable, and he has always been a strong candidate for MVP.
- **Bryce Harper:** Bryce Harper who had an amazing 2015 campaign where he won the MVP has not found the same type of success the past two seasons. Bryce was deemed the chosen one of baseball, and has had great numbers to start his career similar to Mike Trout. A perennial MVP candidate we are comfortable with him leading the way as our top NL candidate.
- **Kris Bryant:** Kris Bryant who won the MVP in 2016 is a central piece on the formidable Chicago Cubs team. What could hurt him in the vote count is the Steve Garvey effect, but there is no debate he is a top level talent in the national League.

- **Nolan Arendao:** Nolan Arenado who consistently in our model was selected to be a top three player once again is selected in the prediction. However due to the Coors effect his offensive numbers no matter how good they are might not be enough to convince the baseball writers to vote for him.

None of these players are surprising picks for the MVP, when considering the Steve Garvey effect, and Coors effect. A good way to see if the players we selected are reasonable selections is by looking at the preseason betting odds for MVP in each league. Below in the table you will see the predicted MVPs table along side the top three candidates according to most favorable betting odds.

League	Predicted Top 3	Betting Odds Top 3
American	1. Mike Trout 2. Giancarlo Stanton 3. Manny Machado	1. Mike Trout +130 2. Jose Altuvie +650 3. Carlos Correa +1000
National	1. Bryce Harper 2. Kris Bryant 3. Nolan Arenado	1. Bryce Harper +325 2. Kris Bryant +450 3. Nolan Arenado +700

Table 5: 2018 Predicted Top Three in MLB Alongside Preseason Betting Odds

From Table five we see that we predicted players for the National league are exactly right and we predicted the top candidate in Mike Trout in the American League. This shows that our model is accurate in being able to take a players statistics and rank them in potential MVP races especially in the National League.

Appendices

Plate Appearance(PA) is recorded when a player completes a batting turn. A player completes a batting turn when he either strikes out from the pitcher or makes a run. Essentially, a plate appearance ends when a player changes from a hitter to a runner.

At-Bats(AB) is a baseball statistics that is a subset of a plate appearance. At-bats are only recorded when the player does not receive credit for a base on balls, hit by pitch, or interference.

Hits(H) is credited to a batter when the batter safely reaches first base after hitting the ball in fair territory without the benefit of an error.

Runs(R) is scored to a player when the player touches all three bases and safely makes it home plate before three outs are recorded.

Runs Batted In(RBI) is a statistic in baseball that credits a batter for making a play that results in a run for their team except when an error occurs.

Stolen Bases(SB) occurs when a runner advances to next base to which he is not entitled to and is credited to the action of the runner.

Caught Stealing(CS) is charged when a runner either attempts to advance to next base or leads off, but is tagged by a fielder before returning safely to a base. The event occurs only when the ball has not been batted.

Walks(BB) also known as base on balls occurs in baseball when a batter has received four balls and is awarded an advancement to first base.

Strikeouts(SO) is recorded when a batter racks up three strikes during a time at bat which results in an out.

Wins Above Replacement (WAR) is an all inclusive reference point that explains the worthiness of a player. WAR offers an estimation on if a player is replaced by an average player.

Win Probability Added (WPA) is a baseball statistic that attempts to measure a player's contribution to the team's win by weighting each specific play made by a player that has altered the game state.

Run Expectancy is the probability at a specific game state where the team currently batting is to score a run.

Win Expectancy is the probability at a specific game state where the home team is expected to win.

Sabermetric is a term coined to classify advanced baseball statistics that have multiple parameters to determine its outcome.

References

[1] Jim Albert and Jay Bennett. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. New York: Copernicus, 2001.

[2] Fangraphs. What is war?

[3] Fangraphs. Wpa. 2017.

[4] David P Gerard. *Baseball GPA: A New Statistical Approach to Performance and Strategy*. Jefferson, North Carolina: McFarland Company, Inc., Publishers, 2013.

[5] Bill James. Mvp followup. 2017.

[6] Bill James. Who spurred baseballs analytics revolution, is waging a mini war on war. 2017.

[7] Mitchel Lichtman Tango, Tom and Andrew Dolphin. *The Book: Playing The Percentages In Baseball*. CreateSpace Independent Publishing Platform, 2014.