Final Project: MLB MVP Voting Analysis

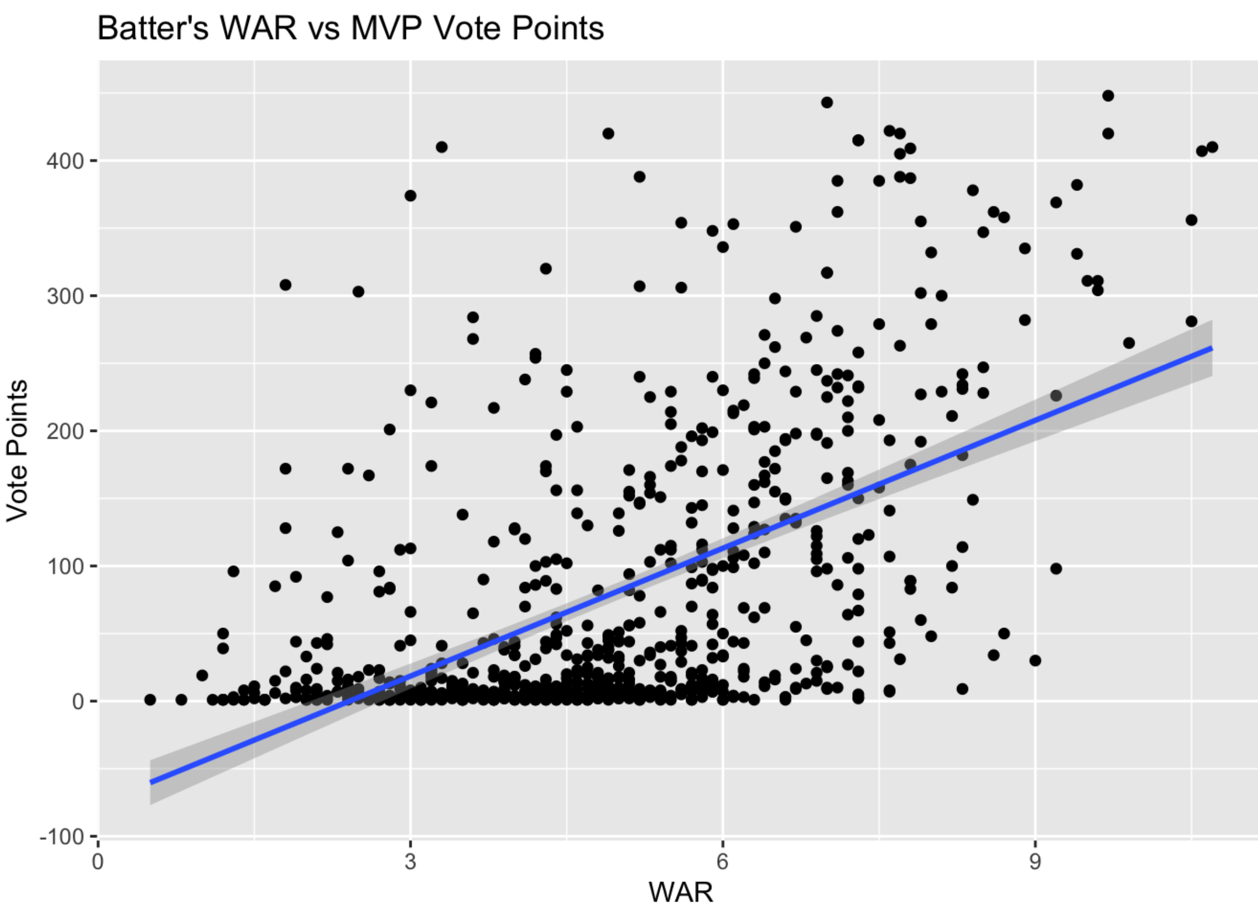
Caleb Woo

4/27/2022

1: Introduction

1.1: Background Information

MLB's most valuable player award is bestowed upon one player in the American League and one player in the National League each year. Every year after the regular season, a group of sports writers are entrusted to evaluate some of the best players in each league and thus nominate MVPs. Given the rather opaque voting process, there are certain years where there is confusion, if not controversy, over which players receive MVP votes. For a while, sports writers could only reference traditional batting and pitching statistics to inform their votes. This all changed with the invention of the statistic wins above replacement (WAR) in 2004 which measures the total offensive and defensive value a player contributed that season. Since WAR takes into account a player's offensive and defensive contributions over a full season, it is now often the go to statistic for quickly evaluating the best players in the league. Perhaps WAR has strongly informed MLB MVP voters since 2004 as we see a moderate positive correlation between WAR and MVP vote points from 2004 to 2021 below.

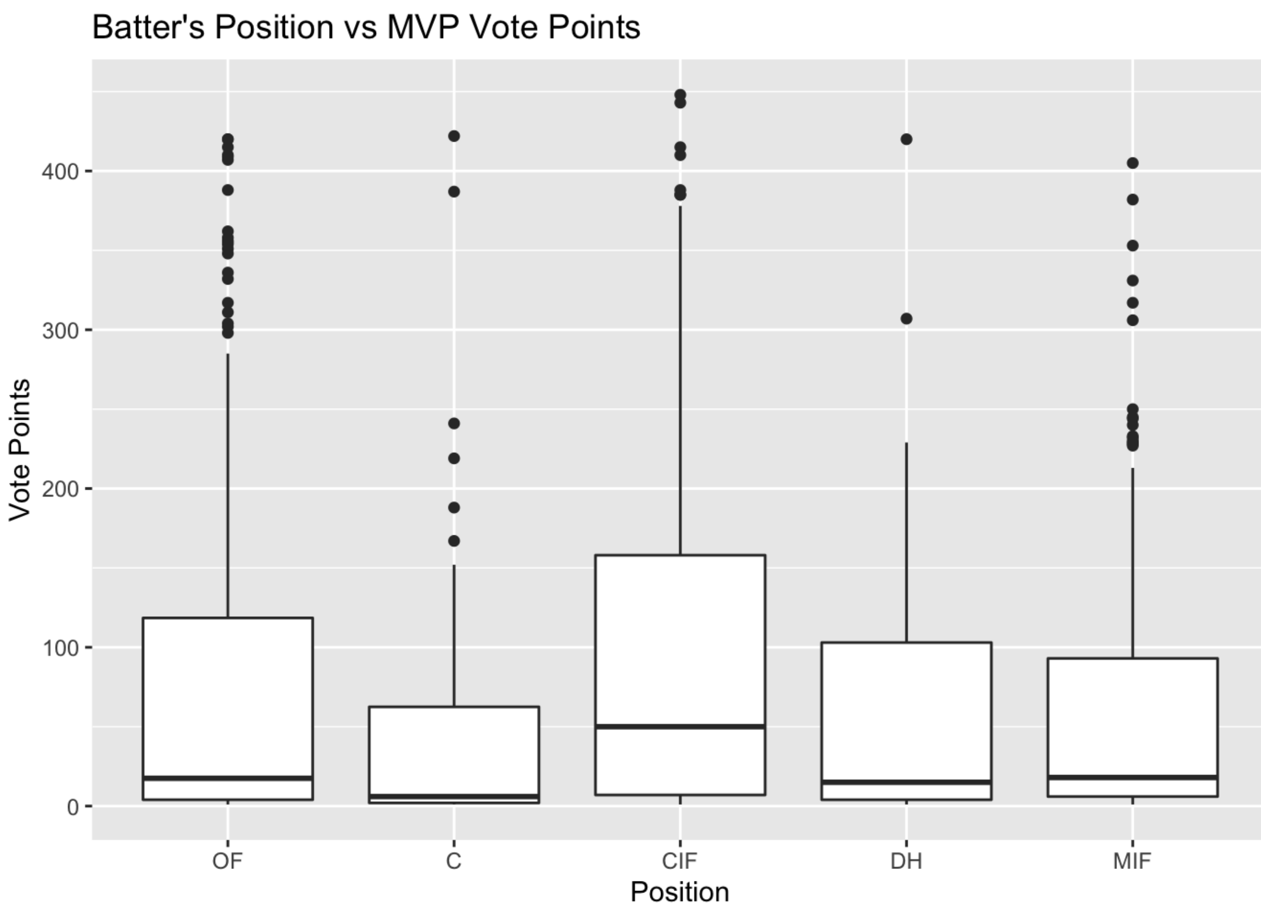


While WAR appears to correlate with MVP vote points, I became interested in other factors and considerations that correlate with MVP voting performance as well. Voters appear to value health and availability. Even if a certain player was elite for half a season, the player will have a difficult time garnering high nominations on MVP ballots. In addition, voters tend to nominate players on good, if not winning, teams. While the MVP is an individual award, many voters seem to believe that the best player on a great team is more valuable than the best player in the league who happens to play on a mediocre team. Furthermore, how voters treat players of different positions will be interesting to investigate. In general, few pitchers have ever won the MVP award over the past century with voters tending to favor elite offensive players.

1.2: Data Description

I gathered all data from the baseball reference website. Baseball reference has a page for the full MVP voting results for every year from 1911 to 2021. The MVP voting results include the number of vote points each candidate secured along with batting and pitching statistics. Each voter has 10 spots on their ballot and a 1st place nomination from a voter gives the player 14 vote points. 2nd to 10th place nominations are worth 9, 8, 7, 6, 5, 4, 3, 2, 1 points, respectively. Scraping the MVP results pages provided me with most of the data I needed.

However there is some crucial extra information that I thought would be necessary to supplement my data. One important piece of information is the player's team's winning percentage for that season. While voters are supposed to nominate the most valuable player regardless of how well a player's team is performing, there have been some noticeable biases towards good players on winning teams as opposed to great players on losing teams. As Mike Trout dominated the 2010s on subpar LA Angels teams, many voters acknowledged how superior he was to his peers and nominated him multiple times despite this bias. I think quantifying the correlation between a player's team's winning percentage and how well a player does in MVP voting will reveal some insights into this bias.



Another important piece of information I added was the primary position of every player. As shown in the boxplot of batter positions vs MVP vote points above, there are some noticeable differences in MVP voting outcomes between different positions. I have narrowed down the different categories of positions to OF, C, CIF, MIF, and DH which refer to outfielder, catcher, corner infielder, middle infielder, and designated hitter respectively. The outfielder category corresponds to all center fielders, left fielders, and right fielders. The corner infielder category corresponds to first basemen and third basemen. The middle infielder category corresponds to shortstops and second basemen. Some of the best hitters in baseball are often first basemen which may explain why the average vote points of corner infielders is so much higher than other positions, especially given the context that MVP voting tends to favor offensive performance over defensive performance. On the other hand, the middle infield positions and catcher are considered the most demanding defensive positions and as a result, most middle infielders and catchers are poor hitters on average when compared to corner infielders. Perhaps voters were more forgiving of lower offensive performances among middle infielders and catchers which led to their nominations in the first place. However, catchers and middle infielders have much lower average vote points than corner infielders which suggests that the overall lack of offensive starpower in these positions may contribute to low nominations and few vote points awarded to these players.

1.3: Variables

There are several baseball statistics that I will use in this analysis. WAR is wins above replacement and measures a player's overall value by calculating how many more wins a player is worth than a replacement-level player. A replacement-level player would be a Minor League player or an unsigned free agent. This means that players who man more defensively demanding positions such as catcher, shortstop, second base, and center field will have a higher WAR than a first baseman, third baseman, left fielder, or right fielder who offers the same overall production because the more demanding defensive positions offer a lower level of production from replacement-level players.

At bats are the number of times a batter faces pitchers without including plate appearances where the batter walked, was hit by a pitch, or hit a sacrifice fly ball. On base percentage is how often a batter reaches first base or further which includes hits, walks, and hit by pitches but does not include errors or fielder's choices. Slugging percentage is the total number of bases a player averages per at bat. Slugging percentage assigns a value of one to a single, two to a double, three to a triple, and four to a home run. By doing so, slugging percentage combines the information of batting average with how much damage, in terms of total bases, that player contributes per hit.

Innings pitched are the total number of innings throughout the season a pitcher pitched. Earned run average is the average number of earned runs a pitcher allows per nine innings. Earned runs are runs that are scored without the intervention of some error in the field. Hits and Walks allowed, which I will refer to as H&BB, are the number of hits and walks allowed by the pitcher that season. This was derived from WHIP which is the walks and hits per inning pitched. I multiplied WHIP by innings pitched to derive H&BB which I believe to be a more interpretable variable in my MVP pitchers model below.

2: Methodology

In order to understand and quantify some of the factors that correlate with MVP voting, I fit 2 separate Bayesian generalized linear models on the MVP vote points of batters and pitchers respectively. Since batters and pitchers have different statistics correlating with their level of success, I felt it to be most logical and interpretable to fit separate models rather than try to use 1 model with many indicator variables.

Below is the equation for the MVP batters model:

$$\log(\text{Vote Points}) = \beta_0 + \beta_1 * \text{I(League = AL)} + \beta_2 * \text{I(Position = C)} + \beta_3 * \text{I(Position = CIF)} + \beta_4 * \text{I(Position = DH)}$$

$$+\beta_5 * \text{I(Position = MIF)} + \beta_6 * \text{I(COVID = True)} + \beta_7 * \text{Team's Win \%} + \beta_8 * \text{WAR} + \beta_9 * \text{AB} + \beta_{10} * \text{OBP} + \beta_{11} * \text{SLG}$$

There are three categorical variables in league, position, and COVID. League refers to whether a batter played in the American League or National League for the given season. Position corresponds to the primary position that the batter played on the field. As mentioned above, I have narrowed down the different categories of positions to OF, C, CIF, MIF, and DH in order to consolidate 9 potential categories into 5 categories with enough observations in each category. COVID refers to whether the player and his stats correspond to the 2020 pandemic shortened season or not.

There are five numerical variables in team's winning percentage, WAR, AB, OBP, and SLG. The team's winning percentage corresponds to the percentage of games the player's team won that year. For the few players in the dataset who were traded midseason, I averaged the winning percentage of the two teams they played on during that season. WAR is wins above replacement, AB is at bats, OBP is on base percentage, and SLG is slugging percentage. I have multiplied team's winning percentage, on base percentage, and slugging percentage by 100 so that model interpretations can be in terms of a unit increase in the percentage.

The five different categories for position will help me quantify how MVP voting correlates differently among different positions. The player's team's winning percentage will help me uncover potential biases towards players on good teams. Since at bats are cumulative across a whole season, it is a measure of a player's availability and will help me understand the correlation between health and MVP voting outcomes. WAR is a measure of overall success and OBP and SLG are measures of offensive success which will help me quantify how overall success and offensive success correlates with MVP voting among batters. League is a control variable to uncover potential differences in MVP voting outcomes between the 2 leagues. Finally, COVID is also a control variable to help account for the pandemic shortened season in 2020 and the lower overall at bats and WAR numbers because of that.

Below is the equation for the MVP pitchers model:

$$\log(\text{Vote Points}) = \beta_0 + \beta_1 * \text{I}(\text{League} = \text{AL}) + \beta_2 * \text{I}(\text{Position} = \text{RP}) + \beta_3 * \text{I}(\text{COVID} = \text{True})$$

$$+\beta_4 * \text{Team's Win \%} + \beta_5 * \text{WAR} + \beta_6 * \text{IP} + \beta_7 * \text{ERA} + \beta_8 * \text{H\&BB}$$

Similar to the batters model, the pitchers model also includes the same three categorical variables, two of the same numerical variables, and three other numerical variables. League and COVID in the pitchers model is identical to the same two variables in the batters model. There are two categories for pitcher position: SP and RP which refer to starting pitcher and relief pitcher respectively. The team's winning percentage and WAR in the pitchers model is identical to the same two variables in the batters model. IP is innings pitched, ERA is earned run average, and H&BB is the number of hits and walks allowed by the pitcher that season. Just like the MVP batters model, I multiplied team's winning percentage by 100 so that model interpretations can be in terms of a unit increase in the percentage.

Although pitchers are only classified as starters or relievers, including position in the MVP pitchers model will help me quantify differences in MVP voting outcomes between the 2 types of pitchers. Innings pitched are cumulative so similar to at bats in the MVP batters model, it will help me understand the correlation between health and MVP voting. I also utilize WAR as a measure of overall success in the MVP pitcher model in combination with ERA and H&BB, which are measures of pitching success, to help me quantify the correlation between overall success and pitching success and MVP voting among pitchers. Just like the MVP batters model, I include league and COVID as control variables.

3: Results

Below are the coefficient and 95% Bayesian credible interval results for the MVP batters model:

MVP Batters Model Results

	Mean	**Lower Bound**	**Upper Bound**
Intercept	-14.866	-16.600	-13.106
American League	-0.101	-0.265	0.062
Catcher	0.502	0.168	0.860
Corner Infielder	0.095	-0.094	0.286
Designated Hitter	-0.014	-0.341	0.338
Middle Infielder	-0.064	-0.288	0.155
COVID Season	3.169	2.624	3.731
Team's Win %	0.069	0.056	0.082
WAR	0.278	0.229	0.329
At Bats	0.007	0.006	0.009
On Base Percentage	0.091	0.061	0.120
Slugging Percentage	0.114	0.098	0.130
shape	0.900	0.819	0.981
mean_PPD	117.082	84.456	184.867
log-posterior	-3547.999	-3553.863	-3544.040

Among positions, catcher is the only position among the 4 non baseline positions to be significant at the 95% Bayesian credible interval. With an exponentiated coefficient of 1.65, catchers are expected to have 1.65 times the MVP vote points of the baseline outfielders, holding all else constant. Meanwhile, all 5 numerical variables are found to be significant at the 95% Bayesian credible interval. A 1% increase in the team's winning percentage is expected to increase MVP vote points by a multiplicative factor of 1.07, holding all else constant. A 1 unit increase in WAR is expected to increase MVP vote points by a multiplicative factor of 1.32, holding all else constant. A 1% increase in slugging percentage is expected to increase MVP vote points by a multiplicative factor of 1.12.

Below are the coefficient and 95% Bayesian credible interval results for the MVP pitchers model:

MVP Pitchers Model Results

	Mean	**Lower Bound**	**Upper Bound**
Intercept	3.502	0.483	6.571
American League	0.114	-0.307	0.515
Reliever	-1.672	-2.877	-0.435
COVID Season	-1.366	-2.668	0.067
Team's Win %	0.025	-0.021	0.069
WAR	0.329	0.088	0.573
Innings Pitched	0.000	-0.018	0.018
Earned Run Average	-0.102	-0.688	0.487
Hits and Walks Allowed	-0.016	-0.029	-0.003
shape	0.649	0.533	0.774
mean_PPD	31.950	22.397	45.532
log-posterior	-716.996	-722.815	-713.422

The reliever position is found to be significant at the 95% Bayesian credible interval. With an exponentiated coefficient of 0.19, relievers are expected to have 0.19 times the MVP vote points of the baseline starting pitchers, holding all else constant. While a player's team's winning percentage is not found to be significant at the 95% Bayesian credible interval for the MVP pitchers model, WAR and hits and walked allowed are found to be significant. A 1 unit increase in WAR is expected to increase MVP vote points by a multiplicative factor of 1.39, holding all else constant. Each additional hit or walk allowed by a pitcher is expected to decrease MVP vote points by a multiplicative factor of 0.98.

4: Discussion

4.1: Conclusions

Using a 2 model approach, I have identified some of the factors that are correlated with the MVP voting outcomes of batters and pitchers respectively. Given the high positive values of the coefficient and 95% credible interval bounds of the catcher position in the MVP batters model, my model suggests that if a catcher and outfielder had the same offensive and defensive contributions throughout a season, the catcher would garner more MVP vote points. This makes sense because as mentioned above, catcher is one of the most demanding defensive positions in baseball. So even if the catcher and outfielder provided the same defensive production at their respective positions, the catcher's defensive production is much more valuable because of how difficult the position is to begin with. In addition, if a catcher were able to provide the same elite level offensive production as an outfielder, the catcher's offensive production is much more valuable because few catchers have the offensive prowess to offer that level of production out of the catching position. Contrary to what my boxplots in the introduction may suggest, corner infielders do not have an advantage in MVP voting outcomes. Instead, corner infielders likely had better average MVP voting outcomes because of their high average offensive production relative to other more defensively demanding positions. But if there were players in more demanding defensive positions such as catcher who were able to provide the same level of offensive and defensive production as a corner infielder, my model suggests that voters would favor the catcher because of how difficult it is provide that level of production from such a demanding position.

The negative values of the coefficient and 95% credible interval bounds for the reliever position in the MVP pitchers model suggests that if a starter and reliever provided the same level of pitching production throughout a season, the starter would garner more MVP votes. This positional correlation among pitchers also makes sense because of the high value put on starting pitchers in the first place. Front offices and fans alike often lament about how difficult it is to acquire and/or develop elite starting pitching. While elite relievers are certainly valuable in their own respect, especially in high leverage situations in the late innings, most relievers are failed starters who did not quite have the stamina, pitch mix, or control to make it as a quality starter in the major leagues. Furthermore, elite starters more than double the innings workload of relievers which typically makes starting pitchers more valuable over a full season. Nonetheless, the coefficient is to be interpreted while holding all else constant so it is interesting to note that even if a starter and reliever had the same exact pitching contributions throughout a season, the starter would still garner more MVP vote points.

As expected, WAR had a high correlation with MVP vote points in both the batters and pitchers models. This does not come as much of a surprise because WAR is such an all encompassing statistic that provides accurate and informative values of the true overall production of players. Since the models only utilize data from 2004 to 2021 which is the period of time in which WAR became available as a statistic, my model also suggests WAR is likely a major factor in the decisions of MLB MVP voters. Since slugging percentage is more or less a measure of hitting ability and power, the significance of slugging percentage also suggests the voters' bias towards offensive production among MVP batters.

Team's winning percentage was only significant in the MVP batters model but not the MVP pitchers model. Furthermore, the multiplicative effect of team's winning percentage in the MVP batters model is subtle which suggests that voters are less influenced by the success of a player's team than I originally thought. In addition, innings pitched was not significant in the MVP pitchers model and at bats was significant in the MVP batters model but also only had a subtle multiplicative effect on MVP vote points. These results suggest that while health and availability may be important for a player to be considered for MVP, slight differences in availability will not outweigh the overall production a player provides in a season.

4.2: Limitations and Future Work

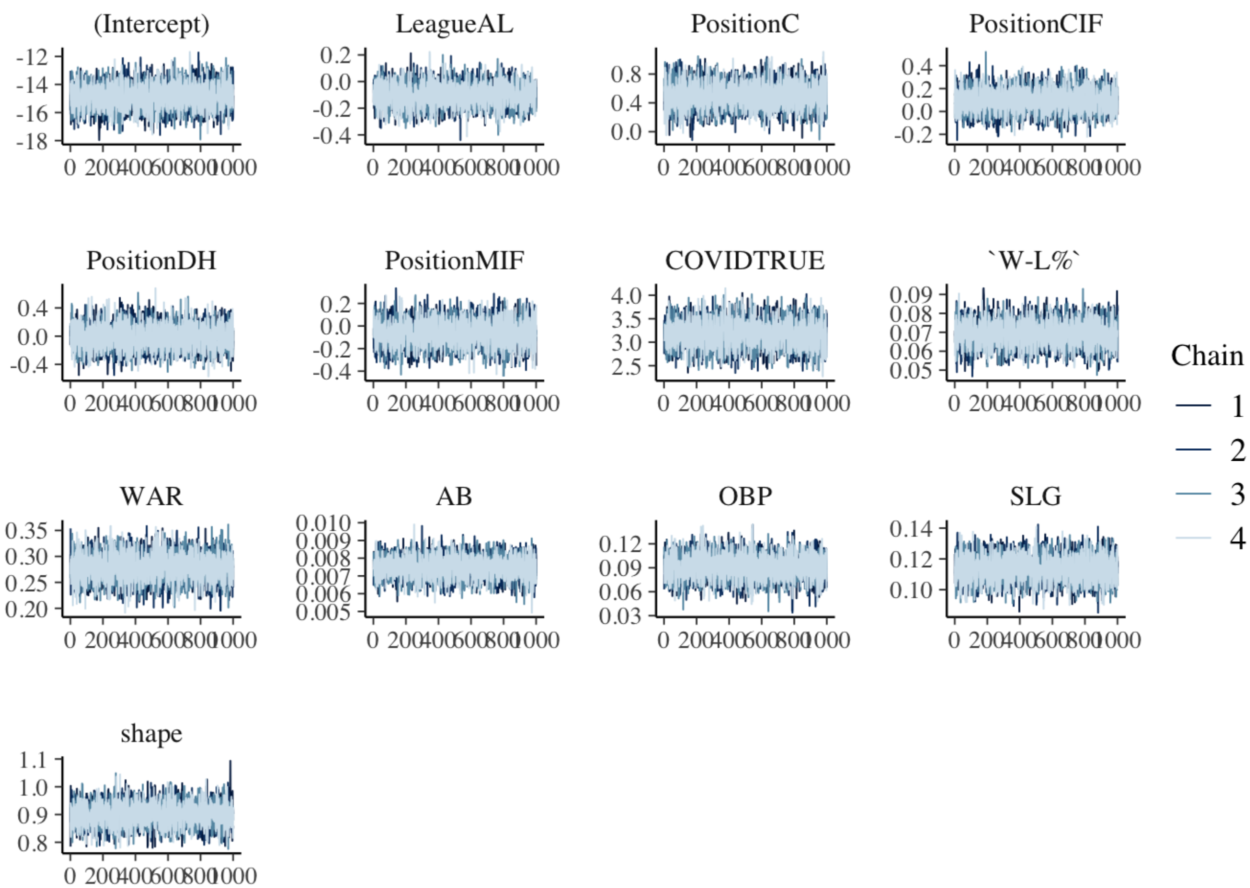
While the 2 model approach allowed me to model both batters and pitchers in the MVP voting process, it does not really allow for direct comparison between the MVP voting considerations of batters and pitchers. I believe that pitchers in general have a disadvantage in the MVP voting process but this cannot be quantified with the 2 model approach. Furthermore, the 2 model approach is insufficient for modeling the newest face of baseball: Shohei Ohtani. As the first 2 way player since Babe Ruth, Ohtani shattered all kinds of MLB records in 2021 on his way to a strong MVP season. Ohtani is unique in that he was one of the best hitters in baseball throughout the season as a designated hitter while also being a top level starting pitcher. The 2 model approach forced me to include Ohtani's batting and pitching contributions separately in my models. Therefore, my models would be very poor at predicting his MVP vote points because it cannot combine the information of his batting and pitching production. Finding a single model to be able to analyze all MVP candidates, regardless of position or multi-positionality would be the next step for this project.

Another major limitation of my approach are the batting and pitching statistics I utilized. On base percentage, slugging percentage, ERA, and WHIP are some of the most traditional batting and pitching statistics that have been tracked since the beginning of baseball. While they are great for quantifying the actual results of the player, it does not quantify the expected performance of a player. Baseball is such a random luck driven sport that there are inevitably players that get luckuy or unlucky throughout a whole season. Statistics such as wOBA and FIP are performance independent statistics that would help identify the true expected performance of a player without the influence of luck or chance. I think utilizing these perforamcne independent statistics would help me further analyze the best and most valuable players in a season without the added variability of luck.

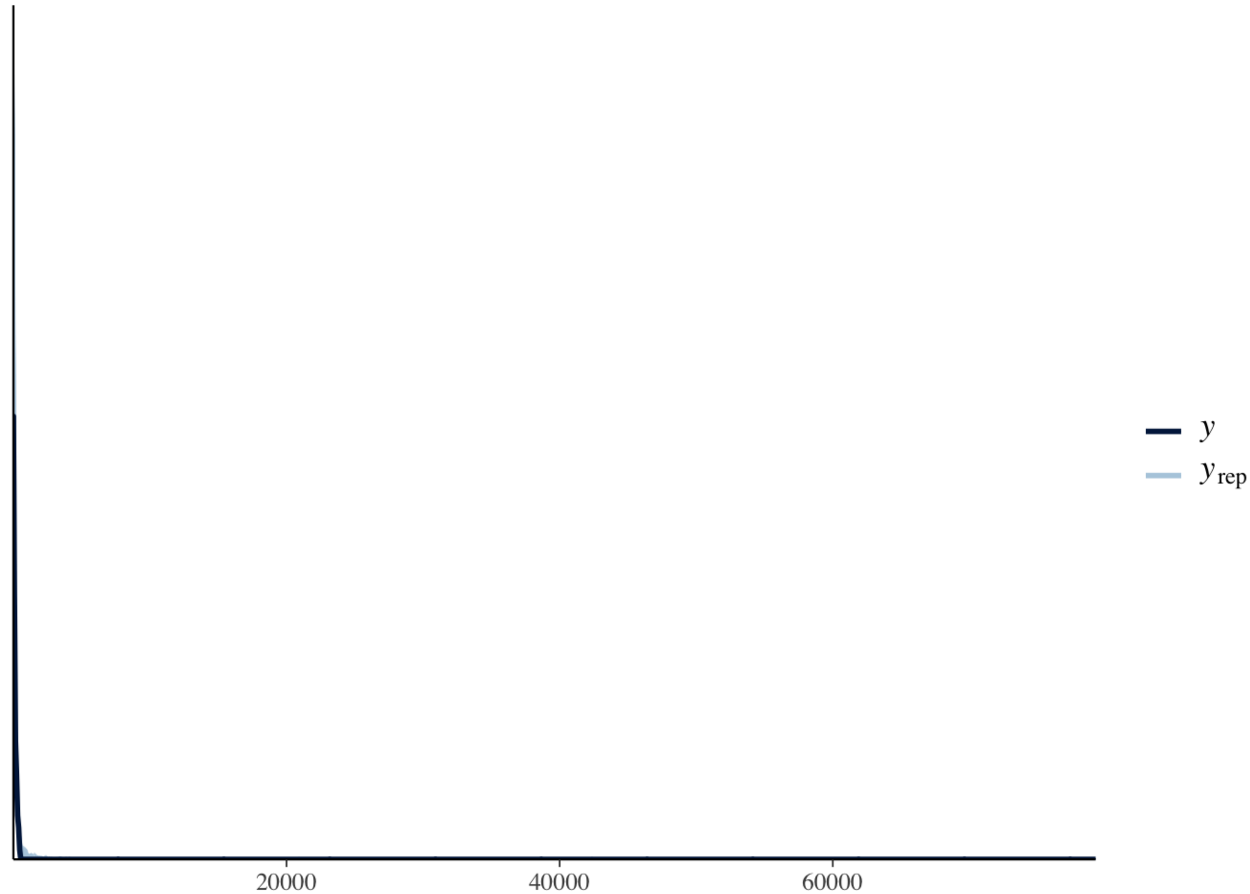
5: Appendix

Appendix A: MVP Batters Model Diagnostics

A.1: Trace Plot Convergence



A.2: P-P Plot



A.3: Effective Sample Size and R hat

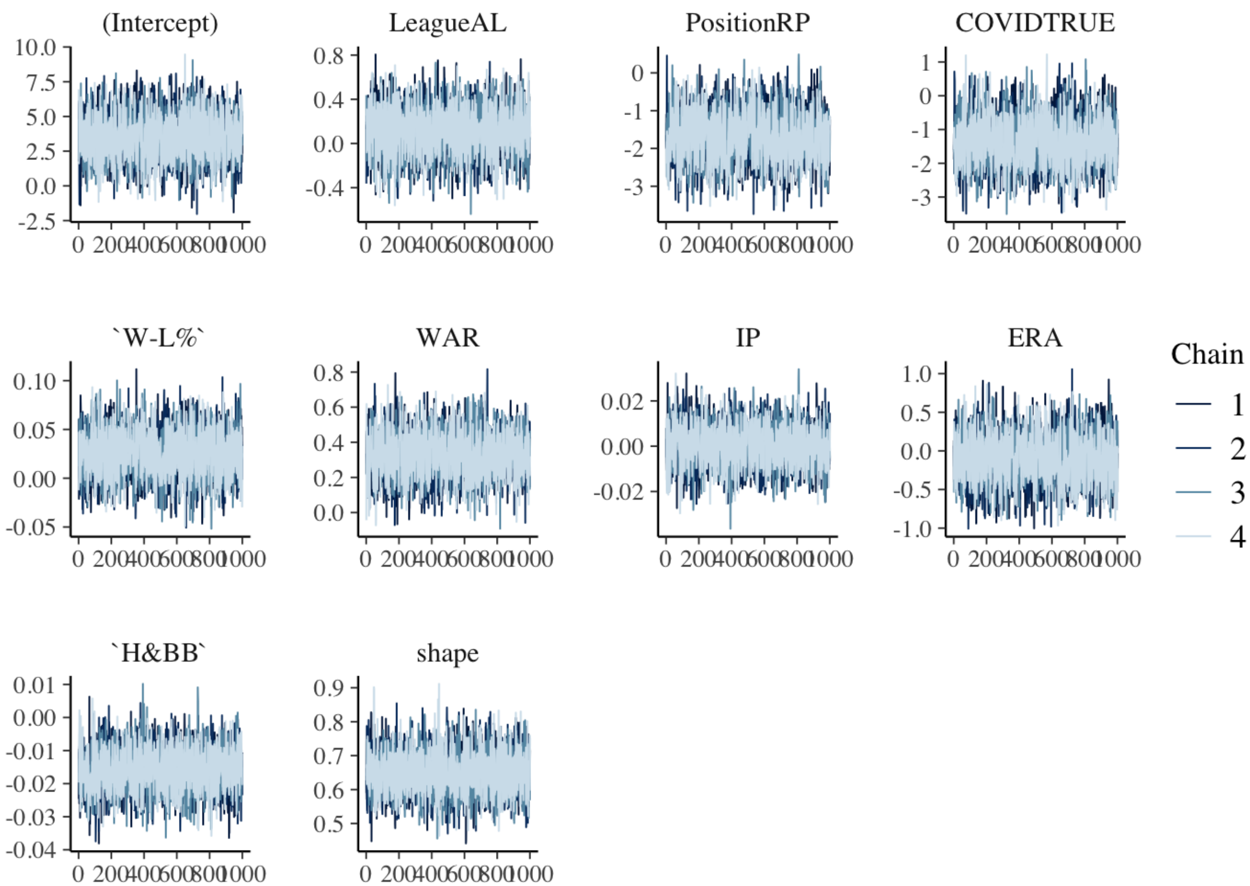
MVP Batters Model Diagnostics

	Effective Sample Size	R hat
Intercept	4196.266	1.000
American League	4887.884	1.000
Catcher	4497.899	1.000
Corner Infielder	4307.317	1.000
Designated Hitter	4016.162	0.999
Middle Infielder	4102.730	1.000
COVID Season	3825.410	1.000
Team's Win %	4994.132	1.000
WAR	4287.711	1.000
At Bats	3647.982	1.000
On Base Percentage	4908.036	1.000
Slugging Percentage	4398.588	1.000
shape	5546.637	0.999

	Effective Sample Size	R hat
mean_PPD	4342.228	1.001
log-posterior	1890.084	1.000

Appendix B: MVP Pitchers Model Diagnostics

B.1: Trace Plot Convergence



B.2: P-P Plot

B.3: Effective Sample Size and R hat

MVP Pitchers Model Diagnostics

	Effective Sample Size	R hat
Intercept	2915.406	1.001
American League	3464.171	1.000
Reliever	1930.641	1.001
COVID Season	2253.579	1.000
Team's Win %	3184.188	1.000
WAR	2453.447	1.003
Innings Pitched	1871.157	1.002
Earned Run Average	2256.265	1.001
Hits and Walks Allowed	2172.497	1.000
shape	3711.253	1.000
mean_PPD	3224.246	1.001
log-posterior	1446.830	1.000