Analysis of NBA player statistics to predict MVP voting

Louis Nass

Department of Mathematics

Tulane University

New Orleans, LA 70118

November 25, 2020

Abstract

Every year, sports reporters and analysts make cases for players they deem worthy of receiving the regular season MVP award. Yet, there is no clear definition of the term "most valuable". The goal of this study is to add statistical insight to this definition. Here we analyze the statistics of NBA players who received MVP votes between the 1980 to 2018 seasons. We use these statistics to create generalized linear models, where the response variable is the expected number of MVP votes one will receive. We create models representative of four decades (80's, 90's, 00's, and 10's) as well as a model from the data set as a whole. For each model, we observe the significance of each independent variable to establish which player statistics were most influential for voters in each decade. Finally, we apply our models to player statistics from the 2018-19 season to see which model most accurately decides the MVP.

1 Introduction

The regular season most valuable player award is arguably the most coveted title in the NBA. The award presumably goes to the best player in the NBA for that season. However, the definition of 'most valuable' has been argued year in and out by sport pundits trying to predict which player is most deserving. In many cases, there is no clear candidate. This ultimately leads to contention when a player is inevitably crowned. Additionally, styles of play have evolved throughout decades of play. Different players have different abilities and lead in various statistical categories. Because of this, the definition of 'most valuable' has predictably evolved with the players.

Our goal is to clarify the definition of 'most valuable'. We will analyze season statistics from players who received MVP votes throughout the 1980 to 2018 seasons.

We will use player statistics to create generalized linear models that predict the amount MVP votes a player will receive. We will formulate models from each decade (the 80's, 90's, 00's, and 10's) as well as the data set as a whole. Additionally, we will evaluate the significance of our predictor variables in each model. This will provide insight to which statistics were most and least influential for voters in each decade.

Finally, we will apply our models to individual statistics from the 2018-19 NBA season and compare our results to the actual MVP voting breakdown. We will compare each of the models' results to determine which model was best representative of the most recent NBA season.

2 Pre-analysis

Before we can begin formulating and analyzing our models, we first will describe our data set. Then, we will analyze our output variable and decide on the proper assumed distribution.

2.1 Data description

The data set used in this analysis comes from Kaggle.

We limit ourselves to player statistics from 1980 until the present because journalists began voting for the NBA MVP in 1980. Additionally, we limit the data set to only include players that received MVP votes in their respective season. We do this because we want to focus on players that received votes, and we want to analyze their performances accordingly. If a player did not receive a vote, then we are not particularly interested in their statistics.

Our predictor variables will be comprised on individual player statistics that are normally observed in a box-score, as well as some advanced statistics. We will use a total of 20 predictors, which are listed in table (1).

In this study, for simplicity, we will ignore potential collinearity among predictors.

2.2 Output variable analysis

Before we can formulate models, we need to analyze our response variable. Our data-set observes various player statistics as well as the total number of MVP votes received at the end of their respective seasons. However, the maximum number of votes received varies from year to year. Therefore, we first need to standardize the amount of MVP votes received.

Label	Description
fga	Field goal (shot) attempts per game
fg3a	3-point field goal (shot) attempts per game
fta	Free-throw attempts per game
per	Player efficiency rating (advanced stat)
ts_per	True shooting percentage (advanced stat)
usg_pct	Usage percent (advanced stat)
bpm	Box score plus/minus
win_pct	Win percent
g	Games played
mp	Minutes played per game
pts	Points per game
trb	Rebounds per game
ast	Assists per game
stl	Steals per game
blk	Blocks per game
fg_pct	Field goal percent per game
fg3_pct	3-point field goal percent per game
ft_pct	Free-throw percent per game
ws	Win-share (advanced stat)
ws_per_48	Win-share per 48 games (advanced stat)

Table 1: List of predictor statistics used in model formulation. 15 of our predictors are tallied each game, the other 5 are calculated advanced statistics.

We standardize our response variable by dividing the number of votes received by the maximum number of possible votes received per season. We define this category as the *adjusted votes received*. We then multiply the award share by the maximum quantity of the maximum votes received. The calculation of the *adjusted votes received* is as follows

$$\textit{adjusted votes received} = \left(\frac{\textit{votes received in season } X}{\textit{max votes available in season } X}\right)(\textit{max of max votes available in all seasons})\,.$$

From figure (1), we observe that our response variable is heavily right skewed. This suggests that our response variable appears to have a Poisson distribution. Hence, we observe in figure (2) that once we take the log of our *adjusted votes received* that the distribution is no longer skewed. Hence, for our model construction, we assume that our output is Poisson distributed.

3 Models

Next, we will establish generalized linear models for each decade as well as the data set altogether. To generate our models, we will use the generalized linear model **R** function, *glm()*.

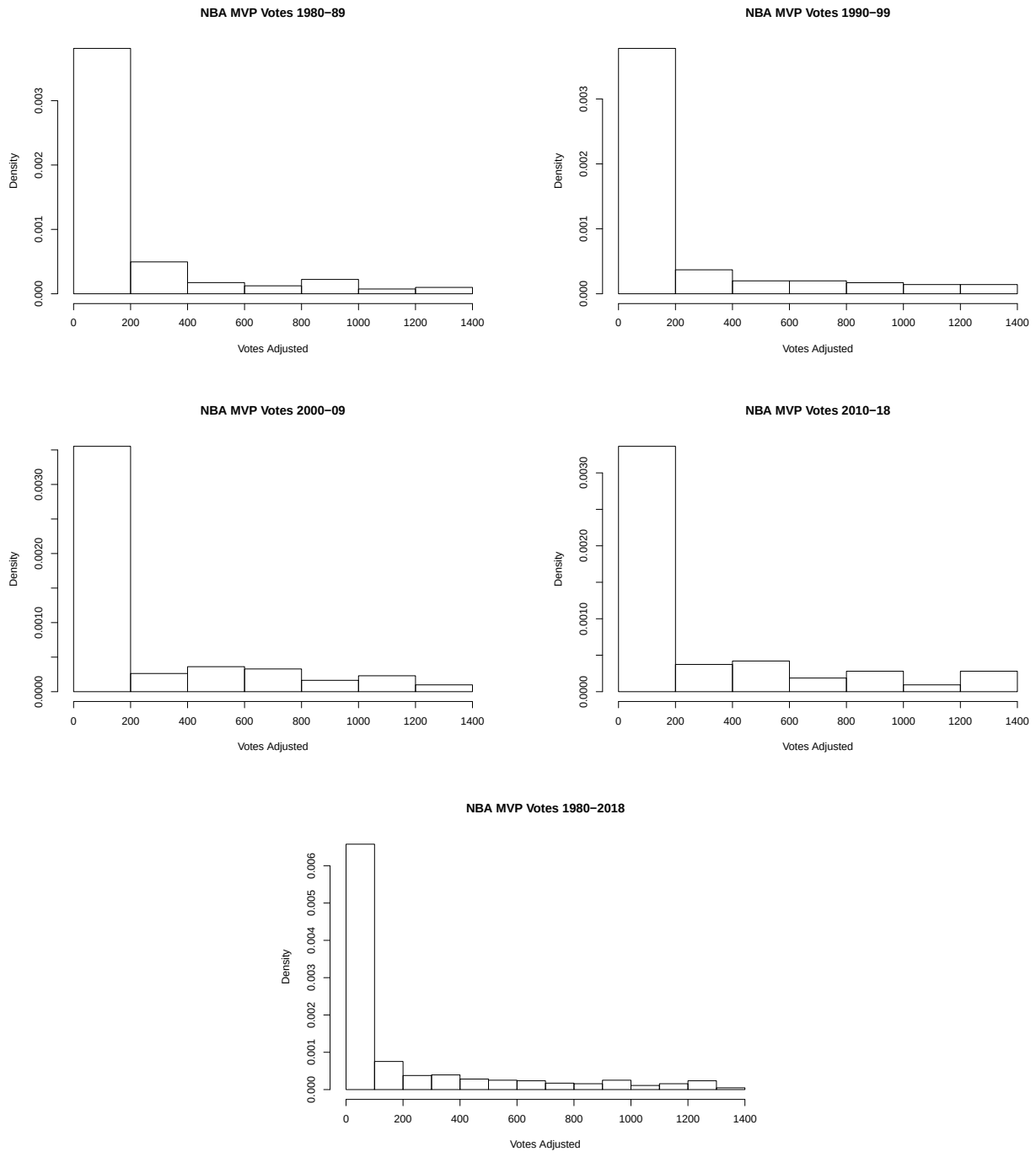

Figure 1: The distribution of the adjusted MVP votes received from each decade (top two rows) and from all the data(bottom row). We observe that the distribution is heavily right skewed. This shows that a majority of the players receive fewer votes and the few winners receive larger vote counts.

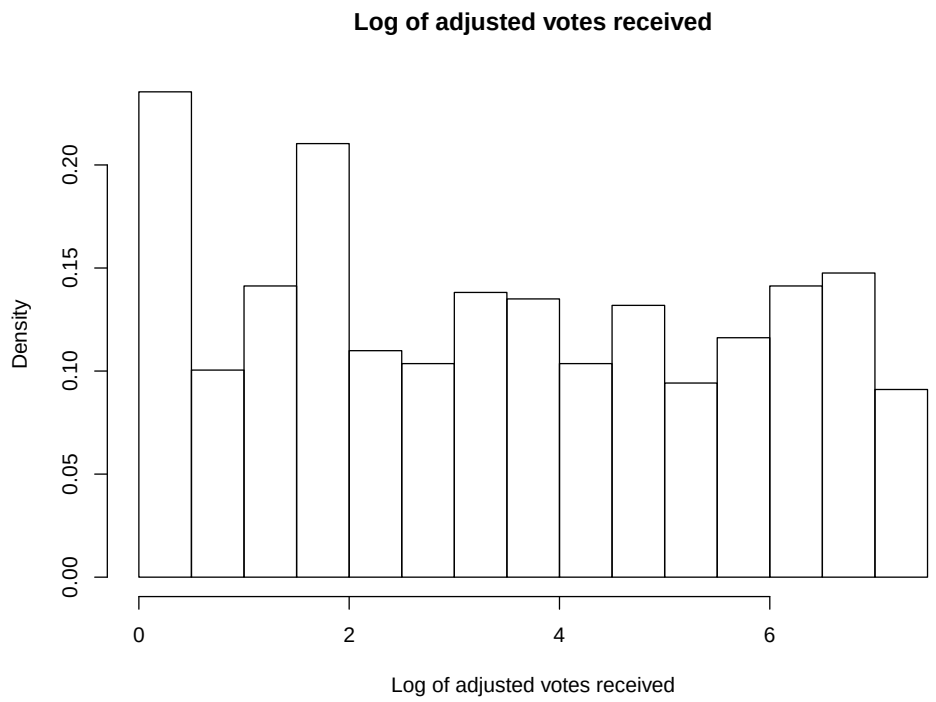

Figure 2: The distribution of the log of the adjusted MVP votes received from the entire data set. We observe that the distribution is no longer skewed.

3.1 Most significant predictors

First, we will observe the most significant predictors in each model. We determine predictor significance by observing corresponding p-values. The p-value tests our model against the null-hypothesis. Hence, small p-values indicate significance while larger p-values indicate little to no significance from our predictor variables.

We observe the list of the four most significant predictors of each model in table (2).

Model	**Most significant predictors (ranked 1-4)**
1980-89 model	Win percent, points, field goal attempts, rebounds
1990-99 model	Usage percent, win percent, minutes, win share per 48 games
2000-09 model	3-point attempts, BPM, win percent, assists
2010-2018 model	Minutes, win percent, games played, win share per 48 games
1980-2018 model	Usage percent, win percent, games played, minutes

Table 2: Most significant predictors from each model, ranked 1-4.

We observe that our decade specific models have some intuitive indicators of the style of play in that era. We observe that the 1980-89 model has rebounds as a highly significant predictor, which makes sense in that time period. In the 80's, generally the 'best' players were larger individuals who primarily used strength and physicality.

On the contrary, we observe that in the 2000-09 model, we observe that 3-point attempts and assists are among the most significant predictors. This indicates a shift from a physical style of play to more of a skill-based style.

Lastly, in the model from the entire data set, we observe that usage and win percent and games and minutes played are the most significant predictors. Intuitively, this makes sense because this identifies players who are on the winningest teams (win-percent) and play the most on these teams (games and minutes). We can think about these indicators fleshing out the common definition of the MVP, which is "the best player of the best team".

3.2 Least significant predictors

Next, we will consider which predictors are least significant. Here we will apply the Akaike information criterion (AIC) in order to eliminate insignificant predictors from our models.

We observe the eliminated predictors from each model in table (3).

Model	**Removed predictors**
1980-89 model	Free-throw percent
1990-99 model	BPM
2000-09 model	No terms dropped
2010-18 model	True-shooting percent, BPM, rebounds
1980-2018 model	True-shooting percent

Table 3: Least significant predictors from each model according to the AIC.

We observe that predictors such as true-shooting percent and BPM are eliminated from some of our models. In the 1980-89 model, we remove free-throw percent. This again mimics the dominant style of the 80's in the NBA. Generally, the best players of that time-period were bad at shooting free-throws.

Another interesting takeaway is that we eliminate rebounds from the 2010-18 model, even though it was one of the most significant predictors for the 1980-89 model. Again, this is suggesting an evolution of the style of play.

3.3 Goodness of fit

Lastly, we will evaluate the goodness of fit of our models. Here we will use the Pearson-χ^2 value to determine how well our models fit the data.

We observe the Peason-χ^2 values of each of our models in table (4).

We observe that the values are in the tens of thousands for each of our decade models and greater than one hundred thousand for the model from all the data. This indicates a great lack of fit to our data. This lack of fit can be attributed to potential collinearity among our predictor variables, as well as limitations we put on our initial data set. Although numerically we observe a lack of fit, we can still gain intuition from our models, as observed in sections (3.1) and (3.2).

Model	Pearson-χ^2 value
1980-89 model	18355.33
1990-99 model	16502.78
2000-09 model	29589.75
2010-18 model	17661.88
1980-2018 model	108084.1

Table 4: Pearson-χ^2 values from each model.

4 Applying models to 2018-19 season

Now we will apply our models to test data from the 2018-19 NBA season. The test data has individual player statistics from each predictor from the 2018-19 season. To apply our models to this test data, we will use the **R** function *predict()*.

We observe the top 5 vote leaders from each model, with their predicted vote count, as well as the true 2018-19 MVP voting results in table (5).

1980-89 model		**1990-99 model**		**2000-09 model**	
Player	**Votes**	**Player**	**Votes**	**Player**	**Votes**
Anteto.	7.06	Harden	10.03	Harden	8.82
Jokic	5.47	Anteto.	7.99	Anteto.	7.89
Gobert	5.40	Embiid	5.96	West.	6.72
Capela	4.97	Curry	5.90	Embiid	6.26
Aldridge	4.17	George	5.80	Lillard	6.15

2010-2018 model		**1980-2018 model**		**Actual Result**	
Player	**Votes**	**Player**	**Votes**	**Player**	**Votes**
Harden	7.18	Anteto.	7.24	Anteto.	941
Anteto.	6.57	Harden	7.13	Harden	776
Lillard	5.16	Durant	5.35	George	356
Gobert	5.13	Embiid	5.35	Jokic	212
Durant	4.98	Jokic	5.33	Curry	175

Table 5: Top 5 MVP vote receivers and predictions and actual results from the 2018-19 NBA season.

We observe that in 3 models James Harden is awarded the MVP and in 2 models Giannis Antetokounmpo is the MVP. Their margins are close within each of the models, except the 1980-89 model, and they both seem to have a multiple vote leads from the rest of the players in each model. This represents a small success in our model, since we observe that the models replicate the close race between the top MVP front-runners, Giannis Antetokounmpo and James Harden. However, we observe that our models only predicted single digit vote counts, instead of the hundreds of votes received in the actual MVP race.

Another interesting result from this test is the 1980-89 model results. The top 5 candidates in this model all play the same position in the modern NBA, and reflect the physical style of play from the 80's that we discussed earlier.

Hence, although our models did not accurately predict correct vote counts for our MVP candidates, we again are able to extrapolate intuition from our results.

5 Conclusion

In conclusion, we observe that the results of our models are mostly intuitive. We gain insight from the significance of predictors in each model that reflect the styles of play in the NBA throughout various decades. Our significance results indicate an evolution of play, moving from a physical, brutish style to more of a skill-based focus. We observed similar findings among our least significant predictors result. Ultimately, we define the MVP as a player who exhibits the most effective play style for their respective era. Observing the model from the entire data set, we conclude that the MVP generally will be the best player on the best team.

Furthermore, we observed that our models mirrored a close race between the two leading candidates, James Harden and Giannis Antetokounmpo. These two were the clear cut favorites in 2018-19, therefore it is promising that all of our models had either Harden or Antetokounmpo winning the MVP.

However, there are clear issues with our models. The observed lack of fit and the inaccurate final vote counts indicate that these models are far from perfect. We can attribute these issues to potential collinearity among our predictor variables, as well as a restricted data set that does not include all NBA players.

Yet, at the end of the day, the intuition we gain from our models represents NBA MVP voting.