Modeling NBA MVP Shares using Bayesian Beta Regression

CSSS 564 Final Project

Andrea Boskovic, Harshil Desai, Rebecca Lopez

University of Washington

June 10, 2022

Contents

1 Introduction & Research Question

The National Basketball Association, or NBA, gives out a litany of awards every season. From awarding outstanding defense, to acknowledging tremendous improvement in a player, no award is more prestigious than the Most Valuable Player (MVP) award. Each year, the NBA media votes on which player will receive the award, and the player with the highest vote share wins. The MVP has gone to a wide variety of player archetypes throughout the history of the league, from unstoppable juggernauts like the 7"1' Shaquille O'Neal to the sharpshooting point guard, Stephen Curry. As statisticians, this begs the question: How can we predict who will receive this award? In this paper we propose a Bayesian approach to modeling, creating a Bayesian beta regression model to predict the MVP vote share a given NBA player will receive at the end of the season.

1.1 Background

The MVP award has been given out since the 1955-1956 NBA season (denoted as the 1956 season) [4]. However, we choose to only consider data from the 1980 NBA season onward. The primary reason for this truncation was due to a fundamental rule change that happened in the 1980's–the introduction of the 3 point line. This is widely considered the beginning of the "Modern NBA"[1]. With the emergence of MVP shooters such as Stephen Curry and James Harden, the necessity for this metric to be included is incredibly important.

For our modeling we will predict the MVP vote share a given player will receive in an NBA season. The MVP is awarded by tallying the votes of a wide variety of NBA reporters throughout the nation. Since the amount of voters and total vote points have changed throughout the years, it is natural to predict the share of votes received as this will normalize the data across the 40 seasons. The data used will be scraped from Basketball Reference, a website with vast amounts of easily downloadable NBA data. We give a special thanks to Kobe Sarausad for providing us with his webscraper. While Bayesian modeling for the NBA has become more popularized by statisticians such as Nate Silver at 538, there have not been major efforts to do so in order to predict MVP awards. As such, we will utilize basic box score stats to evaluate the effectiveness of a simpler model to analyze the foundational effectiveness of such methods.

2 Methods

To predict the MVP vote share for a given year, we use a set of predictors from data scraped from Basketball Reference to create a Bayesian beta regression model. These predictors are given below:

- Age (years)
- Points per Game (average taken over all games of the regular season)
- Games Played
- Minutes Played (minutes)
- Total Rebounds
- Assists
- Steals
- Blocks
- Field Goal Percentage (average percentage of shots made from two and three point range during the regular season)
- Three Point Percentage (average percentage of three point shots made from three point range over the course of a season)
- Free Throw Percentage (average percentage of free throws made over the course of a season)

Intuitively, these features are highly predictive of MVP Status because often, players who score points efficiently, i.e., have higher field goal, three point, and free throw percentages, are more likely to be better players. Note that since the scale of these predictors are not all equivalent and that each predictor is numeric, we standardize them to improve our model's performance.

Note that any unanimous MVPs, or players with vote shares of exactly 1, will be problematic in our model, so we adjust these vote shares by subtracting a small value $\varepsilon = 0.0001$ from their vote share. Only one player in NBA history, Stephen Curry, has been a unanimous MVP, however, so this adjustment was only made for one observation out of a total of $n = 685$ observations.

2.1 Model

The model we choose to explore is a Bayesian beta regression model with a Beta posterior distribution with predictors as specified in Section 2. In our model selection process, we also tested models that

included several quadratic terms, but they didn't improve our model. The model we use and discuss in this paper is as follows:

$$
\begin{aligned}
Y_i|X_i, \beta, \phi &\sim \text{Beta}(a_i, b_i) \\
\text{logit}(\mu_i) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} \\
a_i &= \mu_i \times \phi \\
b_i &= (1 - \mu_i) \times \phi,
\end{aligned}
$$

where $k = 11$ represents the total number of predictors in the model. We implement this model in JAGS using the R package `rjags`. In our implementation, we choose the following uninformative priors to regularize our parameters and to avoid overfitting:

$$
\begin{aligned}
\beta_i &\sim \mathcal{N}(0, 1) \\
\phi &\sim \text{Uniform}(0, 100)
\end{aligned}
$$

Based on our prior predictive checks, shown in section 3.1, these priors are indeed uninformative, as they do not allow us to model the vote share distribution accurately.

Our JAGS model creates simulated observations, used in our posterior predictive check, and calculates each players expected average vote share as predicted by our model, defined as:

$$\bar{y_i} = \frac{a_i}{a_i + b_i}.$$

We can then compare the observed player vote share to expected player vote share, $\bar{y_i}$ for each player i.

2.2 Markov Chain Diagnostic Checks

Before analyzing our model, we must check that our Markov Chains are converging properly to ensure the model's validity. We choose the following parameters:

- `n.chains = 4` - Run 4 Markov Chains,
- `n.adapt = 1000` - Set a burn-in period of 1,000 iterations, and
- `n.iter = 3000` - Run 3,000 iterations for each Markov Chain.

To assess our chain, we evaluate the traceplots, the Gelman-Rubin Statistic $\hat{R}$, and the effective sample size of each parameter in our model. The summary of these results is given in Table 1.

Parameter	**Effective Sample Size**	$\hat{\mathbf{R}}$
β_0	2781	1
β_1	5277	1
β_2	2864	1
β_3	4599	1
β_4	2697	1
β_5	2391	1
β_6	2689	1
β_7	3413	1
β_8	2663	1
β_9	4129	1
β_{10}	5028	1
β_{11}	3510	1
ϕ	2552	1

Table 1: We show the point estimate for the Gelman-Rubin statistic $\hat{R}$ and the effective sample size for each parameter $\beta_0, \ldots, \beta_{11}$, and ϕ in our model.

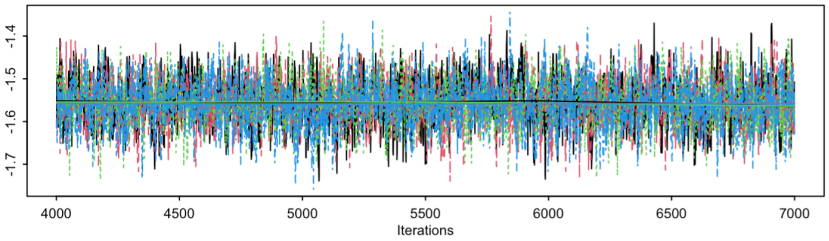

Figure 1: Traceplot for β_0 showing convergence and adequate mixing.

In accordance with our number of iterations, we have a large effective sample sizes, and our traceplots show good mixing and convergence. Figure 1 shows the traceplot for β_0, but we found all traceplots to exhibit similar behavior and convergence patterns. Based on these diagnostics, our Markov Chains are converging, and we can proceed with our analysis.

3 Results

3.1 Prior Predictive Check

In order to asses whether our chosen prior is appropriate for our data set, we run a prior predictive check where we simulate data using simulated parameters according to our sampling distribution.

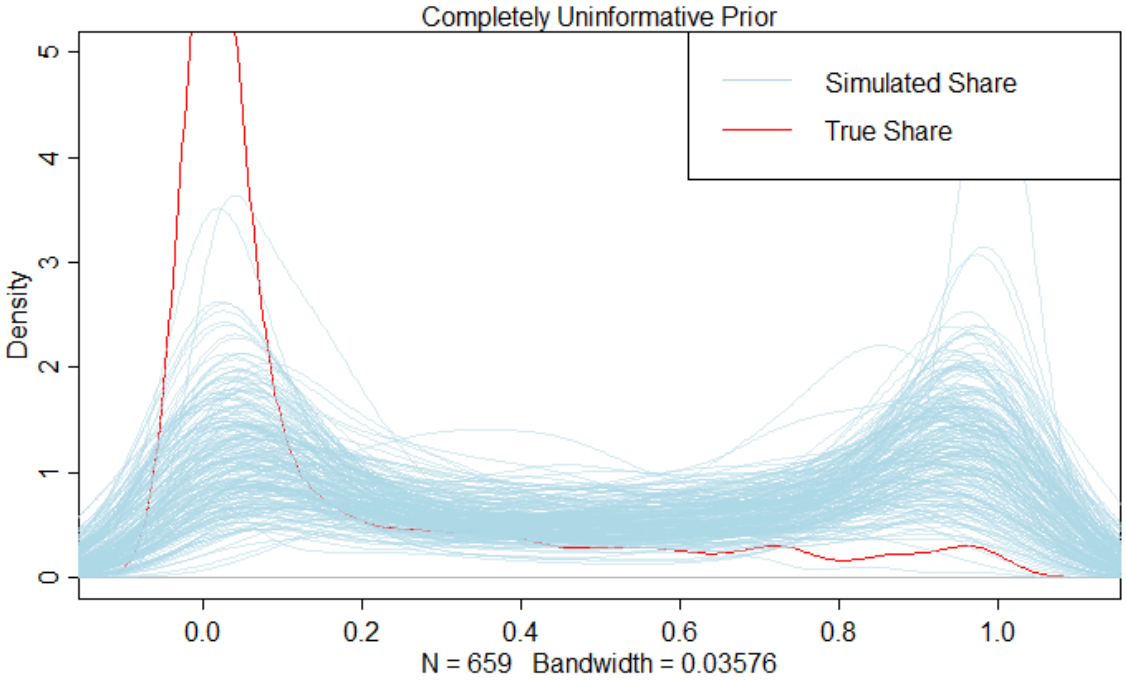

Figure 2: Prior predictive check in which we overlay the prior generated densities over the density of our observed data.

We choose weakly informative priors to allow for the gain in the model's robustness in terms of the parameter space. We can clearly see the density of our simulated vote shares stray from the true vote shares in our data. However, given the fact that we would like to predict a proportion, which ranges from 0 to 1, we see that a beta distribution will be the most appropriate representation of our outcome.

3.2 Posterior Predictive Check

After running our model and performing the diagnostic checks on our Markov Chains shown in section 2.2, we focus on the analysis of our model and its results by sampling from our posterior distribution. By simulating data from our model, we can get a good representation of our model's performance and how well it fits our observed data.

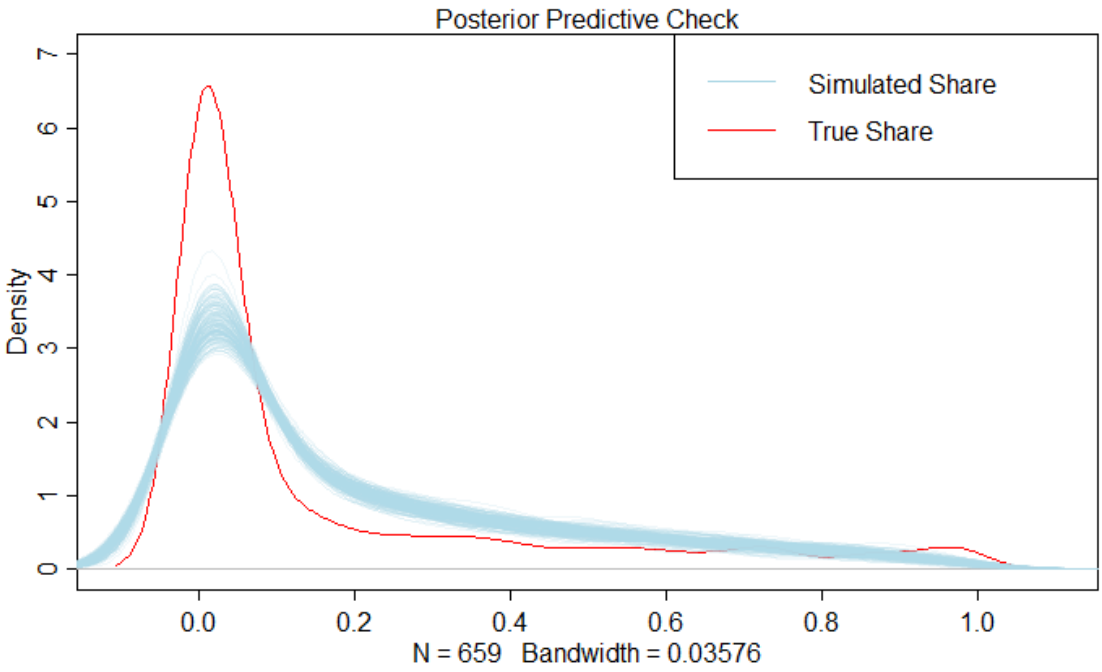

Figure 3: A posterior predictive check in which we overlay the densities of our simulated posterior predictions (blue) over the density of our observed data (red).

We can see that the density of simulated vote shares matches the shape of the density of the true vote share relatively closely. The peak in the simulated data is half the size of that in the true data, however, likely caused by the fact that there are many players each year with low vote shares because even one vote out of the hundreds placed each year puts them on the list of players with a nonzero vote share. These players, however, don't rank highly in MVP voting due to their low vote shares, so our model is still effective at predicting MVP, since players who win MVP must have the highest vote share.

In an attempt to accurately capture the number of low vote shares, we introduce quadratic and cubic terms into the model, but we find no significant improvement in the simulated density. Future models may consider removing players with very few votes, who virtually aren't even in contention for MVP, i.e., those with vote shares below 5-10%, since historically, the top few candidates each have over 20% of votes. Other than the discrepancy in the peak of the simulated and observed densities, the simulated vote shares seem relatively realistic.

Predictor	Mean	Standard Deviation	2.5 %	97.5%
Intercept	-1.555	0.056	-1.663	-1.438
Age	0.057	0.0461	-0.030	0.144
Points per Game	0.617	0.057	0.504	0.731
Games Played	0.156	0.046	0.068	0.253
Minutes Played	-0.142	0.059	-0.261	-0.025
Total Rebounds	0.353	0.066	0.222	0.485
Assists	0.438	0.060	0.318	0.555
Steals	0.108	0.053	0.001	0.212
Blocks	0.173	0.066	0.045	0.309
Field Goal Percentage	0.145	0.053	0.038	0.252
Three Point Percentage	0.005	0.047	-0.085	0.100
Free Throw Percentage	0.076	0.058	-0.037	0.189

Table 2: Table showing the posterior mean, standard deviation and 95% credible interval, using the 2.5% and 97.5% percentiles to represent the lower and upper endpoints, respectively, for each of our predictors.

Table 2 shows the details of our posterior model. In order to understand the relationships between these predictors and vote share of our model, we conduct another posterior predictive check, shown in fig. 3. This plot models MVP vote share as a function of points per game (standardized), which is our most predictive predictor according to model diagnostics. The line representing the posterior mean is jittery with many sharp peaks and valleys, which can again be explained by the large number of players with low, nearly zero, vote shares in a given year. Note that the width of the 95% credible interval is quite large, but this is graph just displays the relationship between one out of our eleven predictors and MVP vote share, which likely accounts for the high level of uncertainty in the estimate of the posterior mean.

In addition to our visual checks of the posterior distribution, we performed qualitative analysis. We found the mean square error between our observed mean for each player and the predicted value to be approximately 0.044. Overall, we can see that our model does a good job at simulating the vote share, but of course, like all research, always has room for improvement.

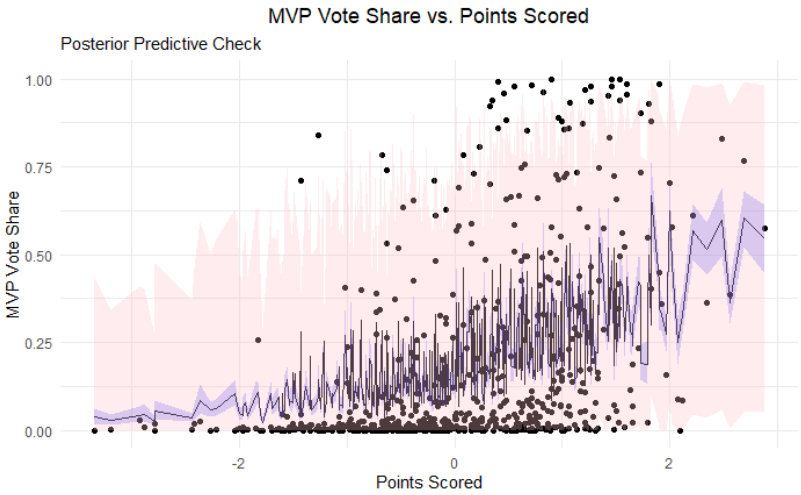

Figure 4: Posterior predictive plot for MVP Vote Share as a function of Points Scored, the main predictor. This illustrates the posterior mean and the 95% prediction interval (blue) and 95% credible interval (red) for MVP Vote Share.

3.3 Predictions

In order to test the effectiveness of our model, we used the 2022 NBA MVP voting results to validate our model. We first filtered this dataset for the top 12 MVP candidates, i.e., the twelve highest ranked players in MVP voting. We then calculated their predicted vote share using our Bayesian beta regression model and used these results to find the predicted MVP rankings by ordering the predicted vote shares from highest to lowest. A comparison of the true and predicted rankings, as well as the true and predicted vote shares, is shown in table 3.

	Player	True MVP Rank	Predicted MVP Rank	True Vote Share	Predicted Vote Share
1	Nikola Jokic	1	1	0.875	0.640
2	Giannis Antetokounmpo	3	2	0.595	0.531
3	Joel Embiid	2	3	0.706	0.470
4	LeBron James	12	4	0.001	0.390
5	Luka Doncic	5	5	0.146	0.370
6	Kevin Durant	11	6	0.001	0.349
7	DeMar DeRozan	10	7	0.001	0.240
8	Ja Morant	7	8	0.01	0.221
9	Jayson Tatum	6	9	0.043	0.217
10	Stephen Curry	8	10	0.004	0.196
11	Devin Booker	4	11	0.216	0.174
12	Chris Paul	9	12	0.002	0.145

Table 3: A comparison of true MVP rankings and our model-predicted MVP rankings, ordered according to predicted rank. A comparison of true and predicted vote share for each of the top 12-ranked players is also displayed.

As shown in the table, our model did predict the true winner of the MVP, Nikola Jokic, correctly. It also predicted the both Giannis Antetokounmpo and Joel Embiid would finish in the top 3, although it did switch their order. One of the biggest discrepancies between our model's predictions and the observed MVP rankings was its prediction that LeBron James would finish in 4th place, not 12th. Our model's significant reliance on points per game likely caused this issue, as James finished the 2022 season with the second highest points per game [2]. Similarly our model predicted that, Devin Booker, who finished 4th in MVP voting, would be in 11th place.

Although Booker also had a high points per game average in the 2022 season, he was on the number one team in the league in 2022: the Phoenix Suns [3]. James, on the other hand, played on the Los Angeles Lakers, who won fewer than 50% of their games in the regular season. Many media members who vote on MVP highly value a player's team's success in the regular season, which is not a predictor in our model. Future models may consider adding win-loss percentage as a predictor and a team's seed, which is further discussed in section 4.1.

While our model did predict the placement highly-voted players fairly well, it did not predict the

actual vote shares of players well. This shortcoming is clear in our posterior predictive check, where we saw the simulated distribution of vote shares did not have the same peak as the distribution of observed vote shares. Given that our model treats each vote share as an independent event, when proportions must sum to one in reality, this discrepancy is fairly justified. Nevertheless, our model does predict relative differences between player vote share fairly well, with Nikola Jokic far outpacing the other players in true vote share and predicted vote share.

4 Conclusion

Using basic box score statistics and demographic information, our model predicts NBA MVP relatively well. This approach illustrates the power of using Bayesian analysis, particularly Bayesian beta regression, to analyze sports data. Although our model doesn't predict low vote shares particularly accurately according to our posterior predictive check, shown in fig. 3, the model is able to predict ranks in MVP voting relatively correctly, exemplified in our model validation with the 2022 predictions, shown in section 3.3. Overall, however, our model works well and provides a good proof of concept for the use of Bayesian models with regard to sports data.

4.1 Future Work

While this model was by no means perfect, it did perform quite well considering the limited amount of training data and model features. From here, we would consider adding more complex features to our model, such as the the win-loss percentage of a player for games they have played, prior NBA awards won, and some advanced statistics such as Nate Silver's RAPTOR metric. This model can also be extended to other awards given out, such as predicting who will make the All Star team, or who will make the first, second, or third team in All NBA. With more and more NBA statistics and research being released every day, models such as these are just scratching the surface of basketball analysis.

References

[1] Nba history: The birth and evolution of the 3-point line. https://www.sportingnews.com/ca/nba/news/nba-history-birth-evolution-3-point-line-stephen-curry-reggie-miller-ray-allen/zlqxs2380v7o1pn4oeumjhsmh.

[2] ESPN. 2021-2022 nba season leaders. http://www.espn.com/nba/seasonleaders/_/league/nba/sort/avgPoints.

[3] ESPN. 2021-22 nba standings. https://www.espn.com/nba/standings/_/sort/wins/dir/desc/group/league.

[4] ESPN. Nba history - mvp. http://www.espn.com/nba/history/awards/_/id/33.