

# Finding the True MVP\*

## An NBA Analysis Through Linear Regression

Yan B. Mezhiborsky

December 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

## Table of contents

### 1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section ??....

### 2 Data

#### 2.1 Overview

We use the statistical programming language R (R Core Team 2023), as well as the following packages to help clean and manipulate our data, (Wickham et al. 2023), (Goodrich et al. 2022), (Wickham, François, et al. 2023), (Firke 2023), (Grolemund and Wickham 2023), (Arel-Bundock 2023), (Xie 2023), (Carpenter et al. 2023), (Zhu 2023). Using these packages we cleaned our data, added years for the data sets that only had seasons, referring to the

---

\*Code and data are available at: [https://github.com/RohanAlexander/starter\\_folder](https://github.com/RohanAlexander/starter_folder).

basketball season, as well as helped create some nice tables, unlike the traditional R software can create.

Our data was collected from (Sports Reference LLC 2024), we collected data from every player that received a vote for any All-NBA teams since 1986, this includes a vote cast for the First All-NBA Team, the Second All-NBA Team, and the Third All-NBA Team. As we believed that taking the players that received votes for the MVP awards would result in a smaller sample. We also created a new statistic call `MVP_Index` that will be defined in Section ?? . In addition, we collected a data set for all NBA MVPs since 1986, so we could compare to the ones we have chosen with the `MVP_Index`.

Following Alexander (2023), we use techniques and methods from this text we have analyzed the data using models, graphs, tables and other methods of data analysis to conduct our research to find players that could have been MVPs and what the MVP for the following years will look like(statistically).

## 2.2 Variables

### 2.2.1 All NBA Variables

As we collected data from each year since 1986 individually, we have utilized the following variables:

- **Team:** Whether the player was in the 1st, 2nd, or 3rd All NBA Teams or OVR if they did not make the top three teams but still received a vote.
- **Pos:** The player's position including the generalized positions G, F, C, for Guard, Forward, and Center respectively.
- **Player:** The name of the NBA player.
- **Age:** The age of the NBA player.
- **Tm:** The three letter designation for the NBA team the player played for the year.
- **G:** Games played by the player.
- **MP:** Minutes played per game.
- **PTS:** Points per game.
- **TRB:** Total Rebounds per game, including offensive and defensive rebounds.
- **BLK:** Blocks per game
- **STL:** Steals Per game
- **WS:** The amount of wins allocated to the individual player

As each year has an individual data set we took PTS, TRB, AST, BLK, STL, and WS took the average for that year and normalized each player's statistics by dividing them by the average of that year with this we have created the MVP Index which we have defined as:

\$\$

$$\text{MVP\_}\{\text{Index}\} = 0.8\text{PTS\_}\{\text{Norm}\} + 0.25\text{AST\_}\{\text{Norm}\} + 0.25\text{TRB\_}\{\text{Norm}\} + 0.1\text{STL\_}\{\text{Norm}\} + 0.1\text{BLK\_}\{\text{Norm}\} + 0.15\text{WS\_}\{\text{Norm}\}$$

\$\$

With this MVP Index we have created our own list of who could have been MVP base on statistics for that year, essentially comparing how good the player was to othe rplayers in the basketball season.

### 2.2.2 Historical MVP Variable

For our second set of date we have all MVPs since 1986 also sourced from (Sports Reference LLC 2024), and the variables that have been used are the following:

- **Player:** The name of the NBA player.
- **Year:** The year in which the player was selected as MVP.
- **Tm:** The three letter designation for the NBA team the player played for the year.
- **G:** Games played by the player.
- **MP:** Minutes played per game.
- **PTS:** Points per game.
- **TRB:** Total Rebounds per game, including offensive and defensive rebounds.
- **BLK:** Blocks per game
- **STL:** Steals Per game
- **WS:** The amount of wins allocated to the individual player

## 2.3 Measurement

Data collected in sports and more specifically basketball is rather simple. When an NBA game is being played there are people who’s entire job is to keep track of a players statistics throughout the game, whenever a player makes a shot, or steals they ball it is being logged by that specific person. After the standard 82 games, there could be morte or less under different circumstances such ast COVID-19, injuries, or even trades all regular season statistics have been collected. As most of our in game statistics are based on the players ‘per game’ average, the total number of points blocks, etc., are recorded it is divided by the number of games the player played that season.

## 2.4 Outcome variable: MVP Index

Our outcome variable is the same as our estimand in Section ???. As MVP Index is dependent on our other statistics, Points, Assists, Rebounds, Steals, Blocks, and Win Shares. As we are studying MVPs in our paper this is our Variable of interest ast it represents how good a player

was in comparison to the other best players in the NBA that year, from the sample of players that have received any ALL NBA Team vote.

### 3 Model

The goal of our modelling strategy is twofold. Firstly, we will be using a linear regression model to predict the MVP Index of the mvp in the following year regardless of the player. This will tell us how much better we can expect the 2025 MVP to be compared to his fellow NBA players. Secondly, We will use another data set we have created by selecting players that had the highest MVP index in their respective year whether they had won MVP or not. So, Using both these predictions we will be able to compare whether we believe the player with the highest MVP index will in fact be the next NBA MVP.

Here we briefly describe the Linear analysis model used to investigate what the MVP Index of the next NBA MVP will be. For our model we will use collected data describe in Section ??.

Background details and diagnostics are included in Appendix ??.

#### 3.1 Model set-up

Define  $y_i$  as the number of seconds that the plane remained aloft. Then  $\beta_i$  is the wing width and  $\gamma_i$  is the wing length, both measured in millimeters.

$$y_i | \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \quad (1)$$

$$\mu_i = \pi_i + \alpha_i + \rho_i + \beta_i + \xi_i + \omega_i \quad (2)$$

$$\mu \sim \text{Normal}(0, 2.5) \quad (3)$$

$$\pi \sim \text{Normal}(0, 2.5) \quad (4)$$

$$\alpha \sim \text{Normal}(0, 2.5) \quad (5)$$

$$\rho \sim \text{Normal}(0, 2.5) \quad (6)$$

$$\beta \sim \text{Normal}(0, 2.5) \quad (7)$$

$$\xi \sim \text{Normal}(0, 2.5) \quad (8)$$

$$\omega \sim \text{Normal}(0, 2.5) \quad (9)$$

$$\sigma \sim \text{Exponential}(1) \quad (10)$$

```
# create a linear regression model
lm_model <- lm(MVP_index ~ PTS + AST + TRB + BLK + STL + WS, data = nba_master)

# View the summary of the model
summary(lm_model)
```

Call:

```
lm(formula = MVP_index ~ PTS + AST + TRB + BLK + STL + WS, data = nba_master)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.27967	-0.06232	-0.00013	0.06198	0.33044

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.1545172	0.0136329	11.33	<2e-16 ***
PTS	0.0337247	0.0005182	65.08	<2e-16 ***
AST	0.0445034	0.0012761	34.88	<2e-16 ***
TRB	0.0295472	0.0010966	26.95	<2e-16 ***
BLK	0.1125198	0.0041166	27.33	<2e-16 ***
STL	0.1060966	0.0056427	18.80	<2e-16 ***
WS	0.0199776	0.0009258	21.58	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09672 on 1536 degrees of freedom

Multiple R-squared: 0.8992, Adjusted R-squared: 0.8988

F-statistic: 2284 on 6 and 1536 DF, p-value: < 2.2e-16

```
next_year_mvp_stats <- mvp_table %>%  
  filter(Year == 2024) %>% # Replace with the target year  
  select(PTS, AST, TRB, BLK, STL, WS)  
  
# Predict MVP_index for the next year's MVP  
predicted_index_mvp_table <- predict(lm_model, newdata = next_year_mvp_stats)  
print(paste("Predicted MVP Index (mvp_table):", predicted_index_mvp_table))
```

```
[1] "Predicted MVP Index (mvp_table): 2.45027833674002"
```

```
next_year_real_mvp_stats <- merged_real_mvp %>%  
  filter(Year == 2024) %>% # Replace with the target year  
  select(PTS, AST, TRB, BLK, STL, WS)  
  
# Predict MVP_index for the next year's MVP  
predicted_index_real_mvp <- predict(lm_model, newdata = next_year_real_mvp_stats)
```