

The Success of College Basketball Teams can be predicted using several advanced team statistics*

A Linear Model Approach

Thomas D’Onofrio

21 April 2022

Abstract

The March Madness college basketball tournament is one of the most exciting sports events in the world due to the intense games, unpredictable results, and the creation of brackets by millions. In this paper, data displaying each college basketball schools’ advanced team statistics was used to create a linear model with team winning percentage as the response variable, where the explanatory variables are the team statistics that play the largest factor in leading to team success. It was found that [insert statistics here once model has been created] are the most important team statistics that affect the amount of wins a team gets the most, meaning fans can use these statistics to fill out their brackets more successfully. Unfortunately, the data is slightly biased due to inconsistent scorekeeping across teams and arenas, and the racial biases of referees when calling fouls, causing some implications to the findings.

1 Introduction


March is a very exciting time for sports fans in the US and many other countries, as it is when the NCAA Men’s Division 1 Basketball Tournament, more commonly known as March Madness, is played. 68 college basketball teams go head to head, playing single elimination games until there is only one team remaining, making for some of the most intense and unpredictable sports games in the world. As a result of this, March Madness draws in a lot of viewers. 16.9 million people tuned in to the 2021 March Madness finals, making it the 2nd most viewed North American sports event only behind the Superbowl (Baker 2021). A big reason for the massive following is that many people create brackets for the tournament, trying to predict the results. With 9.2 quintillion different bracket possibilities, all the insane upsets that nobody sees coming, and the overall chaotic nature of the tournament, creating a bracket and seeing how well you do is very appealing to both sports fans and regular people, causing all the hype around March Madness (Geiling 2014). People put a lot of work into their brackets, comparing teams to one another and trying to figure out who has the best odds of winning. With this in mind, is there a way fans can more successfully predict which teams will have more success than others in the tournament?

In this paper, we attempt to answer this intriguing question by analyzing a variety datasets from the college basketball section of Sports Reference (Sports Reference 2022), using R (R Core Team 2020). Datasets containing college basketball statistics from 2017 to 2020 were used, where each dataset contains many different advanced team statistics for each NCAA Division 1 school with a Men’s basketball team in a specific season. The data was first graphically analyzed to determine which team statistics have a strong or noticeable relationship with how many games a team wins. These statistics were then used to create a linear model that attempts to predict winning percentage, the response variable that is being used to measure team success, where only the best and most necessary team statistics were included. As a result of this, it was determined that the following statistics [Insert statistics here once model is completed], play the most important factor in leading to team success. Thus, when filling out their brackets, fans can compare these statistics across teams

*Code and data are available at: <https://github.com/TDonofrio62/Paper-On-NCAAB-Stats>.

Table 1: 10 Observations from dataset of NCAA Basketball team statistics

Year	School	Over .500	SRS	Pace	O-rating	3pa Rate	TS %	TRB %	AST %	STL %	BLK %	TOV %	ORB %
2017-18	Austin Peay	TRUE	-3.28	70.1	104.7	0.294	0.532	52.2	49.5	9.4	8.9	16.8	35.7
2017-18	Green Bay	FALSE	-8.51	74.1	102.2	0.401	0.537	48.7	53.4	9.2	7.8	16.2	26.3
2018-19	South Carolina Upstate	FALSE	-15.30	69.4	96.9	0.436	0.515	46.7	56.0	9.2	8.1	17.3	26.2
2018-19	Xavier	TRUE	9.61	66.5	107.0	0.374	0.553	53.4	56.3	8.1	10.6	16.5	32.2
2019-20	Duke	TRUE	22.55	72.9	111.4	0.315	0.559	53.8	52.1	11.2	13.6	15.2	34.8
2019-20	Eastern Michigan	TRUE	-3.12	68.1	95.4	0.378	0.509	48.6	42.9	14.4	11.6	18.3	27.8
2020-21	Central Connecticut State	FALSE	-15.12	72.0	95.2	0.376	0.516	44.7	51.7	9.6	7.1	17.6	23.4
2020-21	Marshall	TRUE	5.70	73.1	108.7	0.410	0.571	48.7	53.2	10.6	13.2	14.6	23.8
2021-22	Eastern Kentucky	FALSE	-6.75	73.0	106.5	0.484	0.525	47.5	54.7	13.7	10.1	13.3	29.7
2021-22	Lipscomb	FALSE	-8.42	71.2	104.0	0.417	0.563	49.8	59.1	6.2	7.0	16.4	23.1

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
srs	1418	0	-0.6	10.3	-38.2	-1.6	27.8	
o_rtg	295	0	102.9	6.1	76.9	103.0	122.3	
ts_percent	154	0	0.5	0.0	0.4	0.5	0.6	
trb_percent	146	0	50.3	2.6	38.0	50.3	57.6	
tov_percent	108	0	16.2	1.8	10.8	16.2	23.3	

to determine which teams have a good chance of going far or even winning the tournament, as opposed to just looking at wins or their seeding like many people do.

Unfortunately, there are some implications with these findings. There appears to be several aspects of bias in the datasets being used. For one, there is a lot of human error in the tracking of statistics by scorekeepers, causing certain ambiguously defined statistics in the data to be incorrectly tracked and thus skewed (Williams 2017). This means that data is tracked differently for each team in the league in their respective arenas, made evident by Madison Square Garden’s scorekeepers overstating the quality of shots in the NHL (McCurdy 2020). Thus, the data for some teams may be biased, making them look either better or worse relative to the other teams in the league. There is also unfortunate racial biases in basketball, as referees tend to show racism by calling more fouls on players with different races than them, leading to these teams scoring less points in games where this racial injustice occurs (Wolfers and Price 2012). With more white referees in the league, teams with more white players may have team statistics skewed to make them seem better than they truly are, where teams with more players of colour will look worse because of the bias. We must be aware that our findings have been affected by these biases and as a result may not be perfectly correct. Nevertheless, a lot can be learned from the datasets which are still quite informative.

The rest of this paper will be as follows: Section 2 goes over the datasets being used in the paper, and analyzes the data with the use of tables and graphs to see which team statistics have a relationship with winning percentage. Section 3 includes the creation of a linear regression model that uses only the most important team statistics as explanatory variables to predict winning percentage, the response variable that determines team success. Section 4 interprets the coefficients and provides a statistical analysis of the model in order to understand what is being done. Finally, Section 5 discusses what exactly the model and its outcome is telling us, what can be learned from it, and what implications come along with the findings.

2 Data

```
ncaab_stats %>% select(srs, o_rtg, ts_percent, trb_percent, tov_percent) %>% datasummary_skim()
```

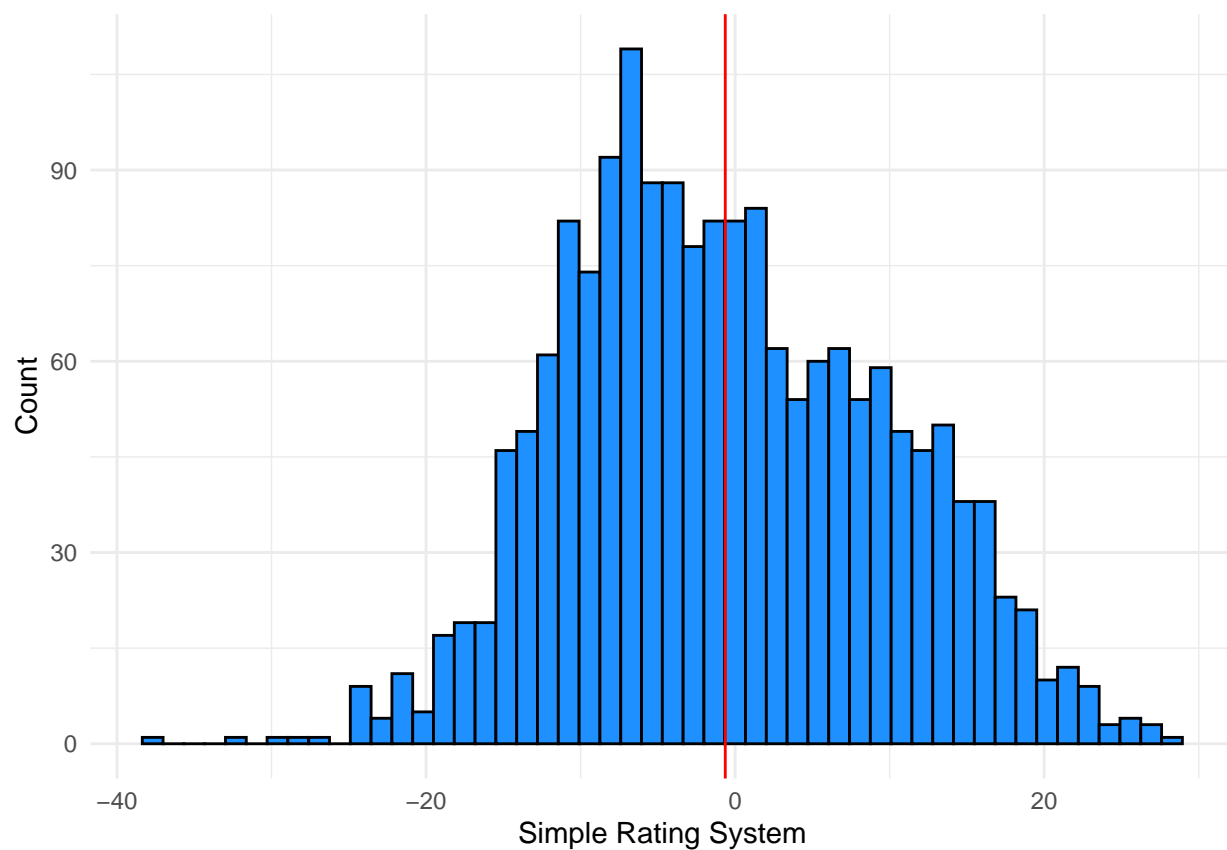


Figure 1: Distribution of Simple Rating System for NCAA Basketball teams

	srs	o_rtg	ts_percent	trb_percent	tov_percent
srs	1
o_rtg	0.65	1	.	.	.
ts_percent	0.48	0.86	1	.	.
trb_percent	0.51	0.51	0.33	1	.
tov_percent	-0.46	-0.62	-0.29	-0.08	1

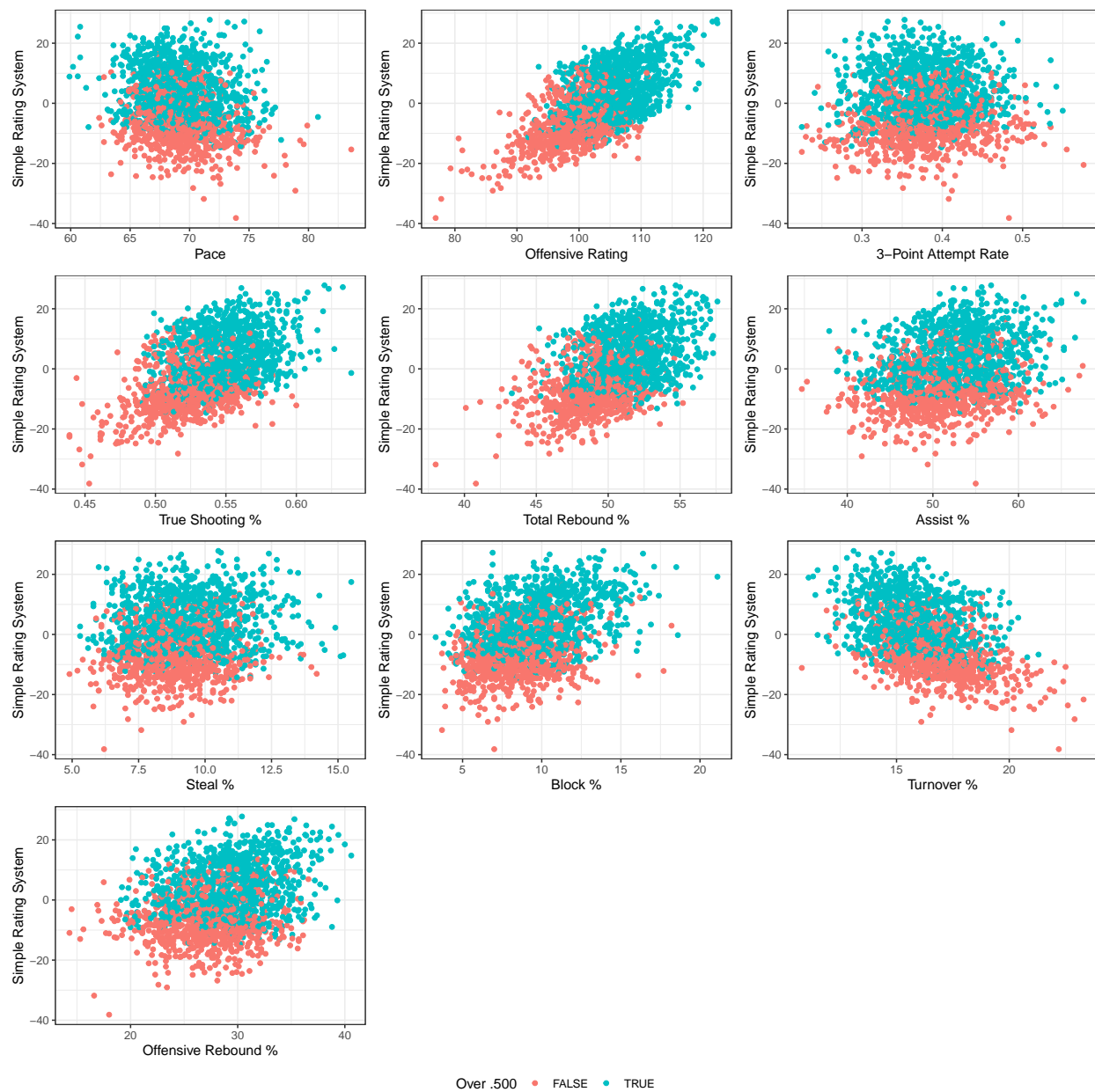


Figure 2: Relationship between Simple Rating System and several advanced team statistics, coloured by whether a school's winning % was above .500 or not

```
ncaab_stats %>% select(srs, o_rtg, ts_percent, trb_percent, tov_percent) %>% datasummary_correlation()
```

3 Model

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \quad (1)$$

Equation (1) seems useful, eh?

Here's a dumb example of how to use some references: In paper we run our analysis in R (R Core Team 2020). We also use the `tidyverse` which was written by (???) If we were interested in baseball data then (???) could be useful.

We can use maths by including latex between dollar signs, for instance θ .

4 Results

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

References

- Baker, Alison. 2021. *The Most Watched Sporting Events in the World*. Road Trips: The Ultimate in Sports Travel. <https://www.roadtrips.com/blog/the-most-watched-sporting-events-in-the-world/>.
- Geiling, Natasha. 2014. *When Did Filling Out a March Madness Bracket Become Popular?* Smithsonian Magazine. <https://www.smithsonianmag.com/history/when-did-filling-out-march-madness-bracket-become-popular-180950162/>.
- McCurdy, Micah Blake. 2020. *Scorer Bias Adjustments*. Hockey Viz. <https://hockeyviz.com/txt/scorerBias>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Sports Reference. 2022. *NCAA Seasons Index*. <https://www.sports-reference.com/cbb/seasons/>.
- Williams, Rob. 2017. *SFU Study Reveals There's Scorekeeper Bias in the Nba*. Daily Hive. <https://dailyhive.com/vancouver/sfu-study-nba-scorekeeper-bias>.
- Wolfers, Justin, and Joseph Price. 2012. *Racial Discrimination Among Nba Referees*. University of Pennsylvania. [https://users.nber.org/~jwolfers/papers/NBARace\(QJE\).pdf](https://users.nber.org/~jwolfers/papers/NBARace(QJE).pdf).