Building Better Brackets: An Introductory Analysis of the Impact of Rebounds on NCAA Tournament Progress

Nathan A. Wiggins

Abstract Every year, millions of people fill out brackets for the NCAA March Madness Tournament. Strategies for picking a potential winning bracket range from strenuously analyzing regular season statistics to comparing college mascots. These strategies naturally make the curious mind consider whether one comparison in particular is effective at predicting which teams will do well in the tournament. In particular, this paper investigates whether or not the number of rebounds per game that a team averages has a measurable impact on how far a basketball team progresses in the NCAA College Basketball Tournament. Using data from the 2013 – 2019 College Basketball Tournaments, we construct a series of linear probability models using the ordinary least squares method to analyze this impact and find that each additional rebound per game averaged by a team results in a 2.62 percentage point increase in the probability that the team will progress to the Round of 32 and a 1.99 percentage point increase in the probability that the team will progress to the Sweet 16 when controlling for other performance metrics. Further, we fail to find that rebounding has a measurable impact on the probability that a team will progress past the Sweet 16. These results are rigorously tested for robustness and applications of the results are discussed.

I. Introduction

The March Madness NCAA Basketball Tournament is a unique event in the sporting world that draws the attention of both sports-loving and sports-neutral individuals alike. Perhaps even more unique than the games themselves are the lengths that individuals go to in predicting the outcome of the tournament. Even the novice student of probability understands the unlikeliness of predicting 63 consecutive coin tosses with complete accuracy. Yet, every March, millions of people across the United States complete brackets consisting of 63 College Basketball match-ups in an effort to become the first person to ever flawlessly predict the comprehensive outcome of the tournament. Each year, the frenzy grows larger and larger with an increasing number of high-status individuals and industries offering multi-million dollar (and, in some cases, even billion dollar) prizes.

To the curious statistical mind, it is natural to wonder if there exists a mathematical way to approach the issue of designing a perfect bracket. While correctly predicting upsets and expected wins seems almost as ambiguous as flipping a coin, past research has shown that if one focuses on the most pertinent statistics, correctly choosing the winning team is not as random an event as we might originally think (Lopez et al, 2018). One in-depth study of the NCAA Basketball Tournament observed that we often inaccurately denote games where higher seeded teams lose to lower seeded teams as upsets, noting that these "seeding" rankings are assigned by a committee and are based off an obscure number of observables (Kvam & Sokol, 2006). This analysis make it clear that in order to accurately predict events within the bracket, we have to be focusing on the correct performance metrics.

One performance metric that is often overlooked by fans but heavily emphasized by coaches to their players is rebounds. Players are often told to focus their efforts in games on pulling down as many rebounds as they can under the assertion that this will increase the team's likelihood of winning the game.

Our focus in this project will be to extend this implication to the annual NCAA basketball tournament and analyze whether or not the average number of rebounds per game that a team achieves during the regular season has an impact on how far a team progresses in the tournament. If teams with higher rebounding averages do indeed tend to be more successful in the tournament, fans wishing to fill out more accurate brackets could observe the number of rebounds per game that each team has and use that information in determining which teams to predict for wins, and which to predict for losses. Additionally, teams hoping to improve their chances of success in the NCAA tournament could do so by focusing on rebounding during practice.

To many, choosing rebounds as the main focus of this analysis may seem mysterious. After all, grabbing a rebound does nothing to change the number of points showing on the scoreboard. However, from a coach's perspective, every single rebound is crucial to the eventual outcome of the game because it is associated with another key performance metric: possession time. Using data from the 1994 NCAA Basketball Tournament, Bradley P. Carlin (1996) was able to create a simple model based off of standard performance metrics and data from the Las Vegas point spread predictions that accurately predicted the outcome of a game that resulted in an upset. Subsequent research showed that one of the most critical factors in predicting the results of a basketball game are metrics related to possession time (Lopez & Matthew, 2015, p. 5). Models that focus heavily on these metrics have been shown to be strikingly accurate, though not perfect, at predicting game results. While there is some continued debate about whether models based off of Markov-Chain probabilities are more accurate than Logit/Probit models (Štrumbelj & Vračar, 2012), the fact that possession time is a powerful ingredient in creating an accurate model is clear.

Pursuant to this crucial idea of possession time, in this paper, we explore whether rebounds per game have a significant impact on the progress a team makes in the NCAA Basketball Tournament. Using rebounds as our key performance metric is based on the importance of possession time, since more rebounds results in more time with the ball. Rather than focusing on one particular NCAA tournament, our study uses data from 7 such tournaments, spanning the years 2013-2019. Using these data we construct a linear probability model and find that rebounding is correlated with success in the first two rounds of the tournament, and find no significant relation for the later rounds.

2. Data

To explore our question we collected data from the NCAA on tournament results from the 2013 - 2019 NCAA Basketball Tournaments. For each tournament year, each team and the seed that they were assigned along with how many games they won were recorded. Performance metrics at the per game level were then downloaded from Basketball-Reference and matched with each tournament team from these years. Together, these sources provided us with a data set that contains detailed information on the population of NCAA qualifying tournament teams spanning the years 2013 - 2019.

Among the performance metrics included in our data set is the number of rebounds per game achieved by each team during the season. Each game, a team achieves a certain amount of offensive rebounds and defensive rebounds. When added together, they produce a statistic called, “total rebounds,” or more simply, just, “rebounds.” Reducing this statistic to a per-game-seasonal-average produces our statistic of interest: “rebounds per game.” Analyzing this metric is crucial to answering the question posed by this paper, and hence our independent variable for this study will be the number of rebounds per game.

In addition to rebounding and NCAA tournament data, our data set also includes other team metrics all measured at the per game level. These include points, field goal percentage, steals, blocks, turnovers, personal fouls, and each team’s assigned seed number. Having these data readily available is very useful in controlling for other factors besides rebounds in our analysis of tournament progression.

Defining our outcome variable, however, is somewhat more complicated since the progress a team makes in the NCAA tournament has several different interpretations. For this purpose, our study includes multiple, similar, outcome variables, described in the table below.

TABLE 1. *List of Included Outcome Variables*

Outcome	Wins Required	Variable
A team progresses to the Round of 32	1 Win	R_{32}
A team progresses to the Sweet 16	2 Wins	R_{16}
A team progresses to the Elite 8	3 Wins	R_8
A team progresses to the Final 4	4 Wins	R_4
A team progresses to the Championship Game	5 Wins	R_2
A team wins the NCAA Tournament	6 Wins	R_1

Each of our outcome variables R_n functions as a dummy variable, with $R_n = 1$ if the associated outcome statement is true, and $R_n = 0$ otherwise. Thus, for a team that successfully makes it to the Elite Eight, but fails to progress to the Final Four, we have:

$$R_{32} = 1, R_{16} = 1, R_8 = 1, R_4 = 0, R_2 = 0, R_1 = 0$$

Before we investigate a model, an analysis of our key variables and statistics proves useful in directing our research. We begin with a statistical summary of the aforementioned performance metrics included in our data set.

TABLE 2. *Per Game Summary Statistics for NCAA Qualifying Teams from 2013-2019*

Performance Metric	Mean	Standard Deviation	[Min, Max]
Rebounds Per Game	**35.841**	**2.804**	**[26.0, 44.1]**
Points Per Game	74.612	5.130	[61.5, 89.8]
Field Goal Percentage	0.459	0.022	[0.403, 0.526]
Steals Per Game	6.571	1.239	[3.5, 11.7]
Blocks Per Game	3.962	1.026	[1.3, 7.8]
Turnovers Per Game	11.859	1.326	[7.4, 15.7]
Personal Fouls Per Game	17.577	1.833	[12.5, 24.2]

Table 2 shows that the season average number of rebounds per game that NCAA qualifying teams achieved between the years 2013 and 2019 is 35.841. If there does in fact exist a correlation between rebounding more and progressing further in the NCAA tournament, then we would expect that if we limit our sample to only the teams who win at least one game, we would find a higher per game rebounding average when compared to the full sample. Table 3 explores this and other filter statistics.

TABLE 3. *Summary Statistics for Rebounds Per Game Filtered by Round Achieved*

Filter Condition	Observations	Rebounds per Game Mean*	Compared to All Qualifiers' Mean	Compared to Previous Filter's Mean
-	448	35.841 (2.804)	-	-
$R_{32} = 1$	224	36.087 (2.771)	+0.246	+0.246
$R_{16} = 1$	112	36.297 (2.862)	+0.456	+0.210
$R_8 = 1$	56	36.107 (2.965)	+0.266	-0.190
$R_4 = 1$	28	36.157 (2.871)	+0.316	+0.050
$R_2 = 1$	14	36.393 (3.502)	+0.552	+0.236
$R_1 = 1$	7	36.457 (3.299)	+0.616	+0.064

*Standard Deviation in Parentheses

A brief analysis of Table 3 shows that with very few exceptions, the number of rebounds per game increases as the tournament round progresses. This increase adds validity to our research question and leads to the natural hypothesis that the average number of rebounds per game that a team achieves does have an impact on how far a team progresses in the tournament. In order to really investigate this, however, we will need a solid mathematical model and method.

It also bears noting that as our data set contains data on all tournament teams from 2013-2019, we can be assured that there are no flaws in our sampling method. Additionally, our analysis of the impact of the number of rebounds per game on how far a team makes it in the NCAA tournament will not be hindered by selection bias related to our independent variable since rebounds per game is a competitive metric and cannot be completely controlled by any one single team, adding a somewhat "random" factor to them in addition to the skill it takes to obtain them.

3. Method

In our analysis of the effect that then number of rebounds per game has on the progress a team makes in the NCAA tournament, we use a linear probability model. As mentioned previously, our key independent variable is the seasonal average number of rebounds that a team achieves on a per game basis (rebounds per game). We also include other team measures: points per game, field goal percentage, steals per game, blocks per game, turnovers per game, and personal fouls per game. These team measures also include a dummy variable that takes on the value of 1 if the team is assigned a high (1-8) seed and 0 if the team is assigned a low (9-16) seed. Finally, we also control for year fixed effects (which will be represented by γ, with 2013 as the omitted category).

The outcome variable of our model (R_n) is the probability that a team reaches the round of n, where n represents the number of teams not yet eliminated (See Table 1). Since there are six possible rounds that a team could potentially progress to, we will use this model to analyze all six of these rounds, each with a separate regression.

Summarizing our choice of outcome variables, our independent variable, and our control variables, our model is represented by the following equation:

$$Prob(R_n = 1) = \beta_0 + \beta_1(\textit{rebounds per game}) + \theta(\textit{team measures}) + \gamma + u$$

Our choice of control variables is deliberate and calculated. We are particularly careful not to include any variables that exhibit multicollinearity with other variables in the model. Additionally, since our star variable is a common team statistic, it is essential that we include other team statistics that could reasonably have an effect on how far a team progresses in the yearly championship. It seems natural to suppose that being a high scoring team or a team that is extremely accurate would have a favorable effect on the probability that a team progresses to a given round in the tournament. Both of these statistics also interact with the rebounding aspect of the game. For these reasons, we control for points per game and field goal percentage in our model design. A similar argument can be made for turnovers, steals, blocks, and personal fouls.

Our "high seed" dummy variable is also interesting and important. The difficulty level of the pathway that a team takes in the NCAA tournament is largely determined by the seeding it receives at the beginning of the tournament. For instance, a team that receives a high seed (1-8) has a very realistic chance of playing 2 or 3 games before it encounters a team seeded higher than it, while a team that receives a low seed (9-16) may never play a team seeded lower than it. Introducing this dummy variable controls for this effect due to seeding and allow us to use our model without fear of bias due to differences in tournament pathway difficulty.

Controlling for fixed effects is also an essential part of this model. Year fixed effects eliminate the possibility of bias being introduced into our model in the case of fundamentally different tournaments due to characteristics associated with the calendar year.

Yet, even with all these controls, the possibility of bias in the model still exists due to potential omitted factors. One such factor could be the level of confidence that a team has coming into the tournament. Although a somewhat subjective concept, this could be measured by variables associated with the experience of the coach (years of coaching), experience of the players (number of seniors on the team), and whether or not the team has won a NCAA tournament game before. Although variables such as these have not been included in this model (mostly due to the complexity of collecting the data), we can determine the bias associated with omitting them. As these confidence variables would likely be positively correlated with the probability of progressing in the tournament and also positively associated with the average number of rebounds per game a team achieves during a season, there is a positive bias associated with this omission. Thus, as we review our results associated with β_1, we must consider these an upper bound rather than a true unbiased estimator.

4. Results

As discussed previously, our model analyzes the probability of a team progressing to a specific round in the NCAA tournament (i.e. The Round of 32, The Sweet 16, The Elite 8, The Final 4, The Championship Round, Winning the Championship) based on key performance metrics, measured at the team level. These results are included in Table 4.

The regression of this model yields some interesting results. Each additional rebound that a team averages on a per game basis leads to a 0.0262, 0.0199 increase in the probability that it will progress to the Round of 32, and the Sweet 16 respectively. We find that the estimated coefficient for rebounds in the Round of 32 model and the Sweet 16 model are statistically significant at the 95% level and the 90% level respectively. Additionally, we find no evidence to support the claim that rebounds per game has an impact on the probability that a team will progress to the Elite 8, Final 4, the Championship Round, or eventually become the Champion.

If robust, our results indicate that every additional rebound a team averages is correlated with a 2.62 and a 1.99 percentage point increase in the probability that the team will make it to the Round of 32 and to the Sweet 16 respectively. The idea that each additional rebound averaged per game will lead to a full 2 percentage point increase in the probability of success in the first two rounds of the tournament, all else constant, is mind-blowing.

Since our model estimates that rebounds per game have this measurable impact on the likelihood of a team progressing to the Round of 32, and the Sweet 16, we wish to test the robustness of these results. To do this, we examine the results of a “naive model” that includes only rebounds per game as an independent variable. We then add our control variables related to other team performance metrics (e.g. points per game, steals per game, high seed, etc.) to that model to create a “partial model” and reexamine the results. Finally, we compare the results of those two models to our “original model." The results of this robustness check are reported in Table 5.

TABLE 4. *Linear Probability Regression on the Outcome $R_n = 1$*

	Outcome Variable					
	$P(R_{32} = 1)$	$P(R_{16} = 1)$	$P(R_8 = 1)$	$P(R_4 = 1)$	$P(R_2 = 1)$	$P(R_1 = 1)$
Rebounds Per Game	**0.0262****	**0.0199***	**0.0102**	**0.0051**	**0.0075**	**0.0033**
Points Per Game	-0.0129*	-0.0054	-0.0050	-0.0013	-0.0013	0.0000
Field Goal Pct	0.0511***	0.0406***	0.0342***	0.0101	0.0110*	0.0045
Steals Per Game	0.0700***	0.0501***	0.0366***	0.0226**	0.0126	0.0120*
Blocks Per Game	0.0468**	0.0220	0.0267	0.0175	-0.0011	-0.0022
Turnovers Per Game	-0.0832***	-0.0347**	-0.0348**	-0.0173	-0.0200***	-0.0066
Pers. Fouls Per Game	-0.0080	-0.0272**	-0.0178	-0.0187**	-0.0072	-0.0044
High Seed (1-8)	0.3110***	0.2640***	0.1130***	0.0558**	0.0317*	0.0179
Year FE	*Included*	*Included*	*Included*	*Included*	*Included*	*Included*
Constant	-1.421**	-1.544***	-1.090**	-0.194	-0.354	-0.233

***p < 0.01 **p < 0.05 *p<0.10

TABLE 5. *Robustness of Linear Probability Model Verification*

	Outcome Variable					
		$P(R_{32} = 1)$			$P(R_{16} = 1)$	
	Naive	*Partial*	*Original*	*Naive*	*Partial*	*Original*
Rebounds Per Game	**0.0157***	**0.0327*****	**0.0262****	**0.0145****	**0.0281*****	**0.0199***
($\triangle$)	(-0.0105)	(+0.0065)	-	(-0.0054)	(+0.0082)	-
Team Measures	-	*Included*	*Included*	-	*Included*	*Included*
Year Fixed Effects	-	-	*Included*	-	-	*Included*

***p < 0.01 **p < 0.05 *p<0.10

In all cases, the estimated coefficient associated with rebounds per game is extremely similar to the coefficient derived from our original model. When using the likelihood of progressing to the Round of 32 as our outcome variable, we see just a 0.0105 difference when regressing the naive model. Similarly, we see a 0.0054 difference between the naive and original estimations with the Sweet 16 round. Controlling for other team measures, but excluding fixed effects due to year and seeding, results in a more inflated coefficient in both cases.

In addition to using these comparative models to check for robustness, we also use a logit model to verify the results of our linear probability model for the Round of 32 and the Sweet 16, controlling for all of the same variables. Table 6 reflects marginal effects of the results of this logit estimation in comparison to our original regression.

TABLE 6. *Logit Model and Linear Probability Model Comparison*

	Outcome Variable			
	$P(R_{32} = 1)$		$P(R_{16} = 1)$	
	Logit	*LPM (Original)*	*Logit*	*LPM (Original)*
Rebounds Per Game	**0.0273****	**0.0262****	**0.0191******	**0.0199***
(△)	(+0.0011)	-	(-0.0008)	-
Team Measures	*Included*	*Included*	*Included*	*Included*
Year FE	*Included*	*Included*	*Included*	*Included*

***p < 0.01 **p < 0.05 *p<0.10

A comparison between the logit and linear probability models a result that is even more similar than before. For the Round of 32, our coefficients on rebounds per game differ by only 0.0011. We see only a slight variation in the rebounding coefficients in our Sweet 16 iteration, with a 0.0008 difference. The similarity between these coefficients indicates that our model and the results it has produced are sound. So in all cases our robustness checks have agreed with our original model.

5. Conclusion

The results of the additional regressions reported in Table 5 and Table 6 show that our original results are robust, and we conclude that the more rebounds per game that a team averages, the higher the likelihood is that it will progress to the Round of 32, and the Sweet 16. In fact, every additional rebound that a team averages during the regular season is expected to lead to a 2.62 percentage point increase in the probability that it will win at least one round in the tournament, and a 1.99 percentage point increase in the probability that the team will progress to the Sweet 16 when controlling for the factors mentioned in this paper.

For each rebound to have roughly a 2 percentage point impact on the probability of a team progressing through the first two rounds of the tournament is an astonishing result, and based upon these findings, we would expect high rebounding teams to be very successful in the early rounds of the tournament.

Some discussion on why we see this result in the early rounds of the tournament, but not in the later rounds is also necessary. On the surface, our results seem to suggest that in the final rounds of the tournament, season-average rebounding performance doesn't make a huge difference in the result of the game. It could be argued that the high-energy environment of the final rounds of the tournament create a unique situation for teams where other factors such as mental toughness, handling pressure, and so on, play bigger roles in success than simply season consistency.

A more likely culprit, however, is the small sample size of data that this study included for the final rounds of the tournament. Because this paper was based on the results of 7 NCAA tournaments, it includes data for just 7 champion-winning teams, 14 championship-contending teams, 28 Final 4 teams, and 56 Elite 8 teams. Smaller sample sizes make it extremely difficult to obtain statistically significant results, and there is the definite possibility that rebounding still has a measurable impact on performance in the final rounds of the tournament. Additional research including large sample sizes of teams from the final rounds of the tournament is the next step in the process of determining the impact that rebounding has on all rounds of the tournament, and not just the early rounds.

The results of this study have far-reaching implications for both basketball teams, and the fans of the game. A team wanting to prepare itself for the end-of-year NCAA tournament can increase its odds of success in the early rounds of the tournament by focusing on rebounding during the regular season, as our results have shown that effectiveness in this area of the game is correlated with success in the tournament. Meanwhile, for fans filling out their brackets at the beginning of March, our results show that it would be prudent to examine how many rebounds a team averages per game before making a decision on whether to include that team in his or her bracket. Comparing rebounds per game in every matchup may even lead to more accurate predictions of upset wins.

In any case, this discovered correlation between rebounds per game and progress in the NCAA tournament is yet another tool in the process of building a better bracket.

Works Cited

Carlin, B. P. (1996). Improved NCAA basketball tournament modeling via point spread and team strength information. *The American Statistician*, *50*(1), 39-43.

Lopez, M. & Matthews, G. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, 11(1), 5-12.

Lopez, M. J., Matthews, G. J., & Baumer, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in North American sport. *The Annals of Applied Statistics*, 12(4), 2483-2516.

Kvam, P., & Sokol, J. S. (2006). A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NrL)*, 53(8), 788-803.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, 28(2), 532-542.