The Rosenthal Fit:

A Statistical Ranking of NCAA Men's Basketball Teams

by

Jeffrey S. Rosenthal
Professor, Department of Statistics, University of Toronto
Author of "*Struck by Lightning: The Curious World of Probabilities*"
(March 17, 2013)

Introduction:

I was asked by the TSN sports television network to make predictions for the 2013 NCAA Men's Basketball "March Madness" tournament bracket, based solely on a statistical analysis, without using any specific knowledge of NCAA teams (which is just as well since, although I like sports and watch them sometimes and even play a bit of neighbourhood pick-up basketball myself, I haven't closely followed any spectator sports in years).

I proceeded by:

(a) Gathering lots of different data variables for each team, for each of the past four regular[1] seasons.

(b) Separately gathering the results of each game of each of the past three years' March Madness tournaments.

(c) Combining all of that data together for my computer programs to read (which turned out to be very time-consuming, since different data are available on different web sites in different formats with different team name abbreviations, so I had to "teach" my computer to match them all up).

(d) Exploring different "non-negative linear combinations" of the data, i.e. formulas which use the data from a given regular season, to give an overall

[1] I use the phrase "regular season" to include *all* games from that season prior to the NCAA March Madness tournament, including conference tournament games.

score to each team.

(e) Developing computer programs to "fit" the formula based on previous seasons, i.e. to do an extensive search to figure out which of those formulas did the best job of predicting the winners for each game in that year's tournament, using data from the corresponding regular season.

(f) Eventually coming up with a single best formula for this, which I call the "Rosenthal Fit".

(g) Then, filling in the actual bracket simply by picking, for each game, whichever team has a larger value of their Rosenthal Fit.

The formula for the Rosenthal Fit, plus an evaluation of how well it performed when applied to data from the previous three years' tournaments, is provided below. Corresponding values for all teams for the 2012–2013 regular season (to be used to predict the 2013 tournament bracket) are listed in the attached file "RosenthalFitValues.txt".

General Observations:

The NCAA tournament is inherently hard to predict. Indeed, the total number of different ways of filling in your bracket predictions is 2^{63} (i.e., 63 different 2's all multiplied together), which works out to about 9×10^{18}, i.e. a 9 followed by 18 zeros, which equals nine billion billion, or nine million million million. That's a lot of possibilities!

In fact, even the experts find it challenging. For example, in past tournament games, the higher-seeded team only won about 70% of the games. This means that even when many of the most knowledgeable people get together to seed the teams, they can still only correctly predict the winner about 70% of the time. Individual expert basketball predictors (e.g. Kem Pomeroy at KenPom.com) tend to perform similarly, accurately predicting the winner in only about 70% of the tournament games. Part of the reason is that each

matchup is a single-elimination game, rather than e.g. a seven-game series, so there is lots of inherent day-to-day randomness, and it is quite possible for a weaker team to beat a "better" team in any one game, making predictions that much more difficult.

So, despite my extensive computer programming and statistical modeling, I do not expect to do better than calling about 70% of the games correctly. Indeed, I would say that anyone who does much better than 70% would have to get fairly lucky (in addition to perhaps having a good predictive model and/or good knowledge of the basketball teams).

Statistical Data Considered:

To perform my statistical analysis, I downloaded and considered lots of different statistics, including the following (listed with sources):

- **WinFrac:** The team's overall game-winning fraction for the entire regular (pre-March Madness) season. (teamrankings.com)
- **WinFrac3:** The team's game-winning fraction in their final three regular season games. (teamrankings.com)
- **CWinFrac:** The team's game-winning fraction for games within their own conference. (realtimerpi.com)
- **NCWinFrac:** The team's game-winning fraction for games outside of their own conference. (realtimerpi.com)
- **AdOff:** The team's "adjusted" offensive efficiency rating. (KenPom.com)
- **AdDef:** The team's "adjusted" defensive efficiency rating. (KenPom.com)
- **OffEff:** The team's unadjusted offensive efficiency rating. (teamrankings.com)
- **DefEff:** The team's unadjusted offensive efficiency rating. (teamrankings.com)

- **SOS:** The team's "Strength of Schedule", a measure of the average strength of the opponents they played. (realtimerpi.com)
- **RPI:** The team's "Ratings Percentage Index". (realtimerpi.com)
- **PntPG:** The team's average number of points scored per game. (teamrankings.com)
- **OpPnt:** The team's average number of points scored against them per game. (teamrankings.com)
- I also examined the team statistics provided at ncaa.com and at espn.go.com, but they largely overlapped with the above statistics, so in the end I did not need to use them directly.

Finally, and most importantly, the "outcome" measure was:

- **TourRes:** The game-by-game, line-by-line win/loss results for each game of each of the past three March Madness tournaments. (kusports.com)

Statistical Modeling Approach Taken:

My approach was to try to figure out which linear combination of (i.e., formula using) the above-listed regular-season statistical values would do the best job of ranking the teams from highest to lowest, in terms of who won which games in the corresponding year's tournament. I computed this using regular-season statistical values, and corresponding tournament game results, for each of the three seasons 2009–2010, 2010–2011, and 2011–2012.

To perform this computation, I wrote computer programs in C and in R, which used such techniques as "linear regression", "constrained linear regression", and finally a "Monte Carlo (randomised) search algorithm", to find an optimal formula.

Although my computer programs considered all of the above variables, they ultimately selected just a few of those variables as being most relevant

for prediction, namely: **WinFrac**, **WinFrac3**, **OffEff**, **DefEff**, **SOS**, and **NCWinFrac**.

Final Formula:

Using the above statistical analysis, the resulting best linear combination turned out to be:

$$\textbf{Rosenthal Fit} \;=\; 6.2337 \times \textbf{WinFrac} + 1.7180 \times \textbf{WinFrac3}$$
$$+1.1179 \times \textbf{OffEff} + 1.9189 \times \textbf{DefEff} + 11.9846 \times \textbf{SOS} + 7.3712 \times \textbf{NCWinFrac}\,.$$

I then applied this linear combination formula to the regular-season statistics for the current (2012–2013) season. This provided an overall numerical rating for each team this year, based on their regular-season statistics. These ratings are listed, in order from highest to lowest, in the attached file "RosenthalFitValues.txt".

Then, to fill out this year's tournament bracket using this Rosenthal Fit, simply choose, for each game, whichever team has a higher value of the Rosenthal Fit (i.e., comes first in the file "RosenthalFitValues.txt").

Note: The above rating system is based purely on statistical analysis, without taking any other factors into account. Certain late-breaking events (e.g. Kentucky Wildcats superstar Nerlens Noel's major injury on February 12) could potentially have a large impact on a team's tournament performance despite making only small changes to their regular-season statistics, which could throw off my model's predictions. I did consider making a few post-hoc adjustments to account for such developments, but in the end I decided not to – thus keeping the Rosenthal Fit as a purely statistical measure.

Comparison to Other Predictors:

The following table shows how the Rosenthal Fit, and also the tournament seedings, and also the RPI (Ratings Percentage Index) itself, would have done at predicting tournament games in each of the past three tournaments. (In two of the tournaments, there was one game between two equally-seeded teams; those two games are excluded from the evaluation of the tournament seedings.)

Season	Tournament Seedings	RPI Values	Rosenthal Fit
2009–2010	42/62 (67.74%)	44/63 (69.84%)	48/63 (76.19%)
2010–2011	43/63 (68.25%)	38/63 (60.32%)	43/63 (68.25%)
2011–2012	46/62 (74.19%)	44/63 (69.84%)	45/63 (71.43%)
Total	131/187 (70.05%)	126/189 (66.67%)	136/189 (71.96%)

This table shows that the Rosenthal Fit compares favourably with RPI and with the tournament seedings. This should not be taken as evidence of any particular superiority, since the Rosenthal Fit was developed precisely to try to maximise these predictions. Still, it does suggest that the Rosenthal Fit is at least roughly comparable in predictive power to these expert measures. In a few weeks, we will know how well it performed this year.

Acknowledgements: I thank Ken Volden at TSN for arranging this project, Kate McKenna and her colleagues at TSN for helping me locate the relevant statistical data, and my cousin David Rosenthal – a professor of mathematics at St. John's University in New York City – for discussing basketball issues with me.