A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament

Brady T. West[1], M.A.

[1]University of Michigan Center for Statistical Consultation and Research (CSCAR), Ann Arbor, MI

Background

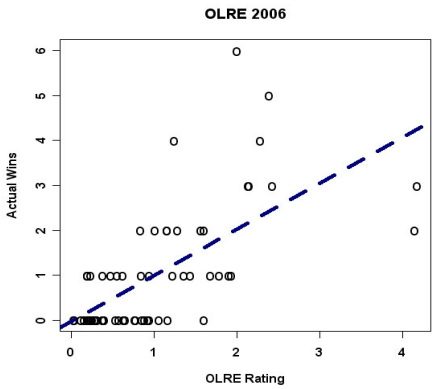
- Each year, a small NCAA-appointed committee is responsible for selecting 34 Division I basketball teams (31 conference champions receive automatic bids) for the wildly popular NCAA Basketball Tournament, and then assigning regions and seeds to the 65 teams
- The seeds represent relative strengths of the teams, and higher seeded teams have a better chance of being assigned to a region close to home
- Numerical ratings used by the committee to seed the teams and assign them to regions need to be strong indicators of expected success in the tournament, resulting in a balanced and competitive tournament

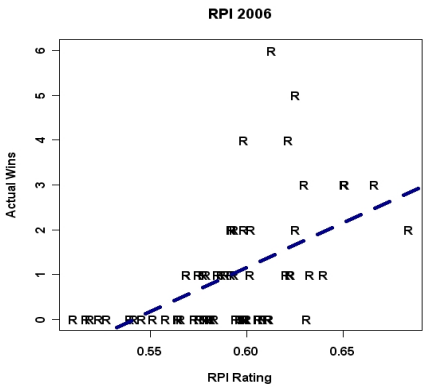
Objective

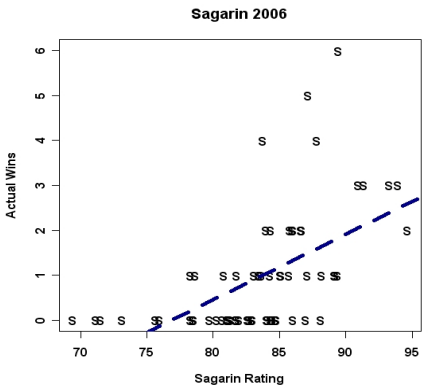
- The primary objective of this work was to highlight the simplicity, flexibility, and effectiveness of a proposed rating method (the OLRE method) for the selected teams, which is based on ordinal logistic regression and the expected value of a discrete random variable

Methods

- Historical data representing team-level variables at the ***ends of the regular seasons*** for the cohorts of teams selected for the tournament were collected from the 2002-2003 season to the 2005-2006 season, using free internet resources (see ***Conclusions*** for the variables)
- The number of wins achieved by each team in the 2003, 2004, and 2005 tournaments were also recorded and used as a dependent variable in an ordinal logistic regression model, where the predictor variables were the team-level regular season variables
- Predicted probabilities of winning 0 through 6 games were then calculated based on the fitted model for the 2006 tournament teams, and adjusted to satisfy known marginal constraints for the expected tournament outcomes (e.g., exactly 32 teams will win 0 games)
- The adjusted predicted probabilities for each team allowed for the calculation of an expected number of wins in the tournament (a rating)







Results

- The results of one million NCAA tournaments for the 2006 teams were simulated in R using the Bradley-Terry model for paired comparisons, where the strength parameters for the teams were the ***ranks*** of the teams based on Jeff Sagarin's regular season-end computer ratings
- The resulting predicted probabilities of winning 0 through 6 games for each team based on the simulation were used to calculate a model-based expected number of wins, for comparison with the OLRE method
- The sums of squared errors between the OLRE and model-based expectations and the actual numbers of wins achieved by the teams in 2006 were calculated and compared to assess predictive power
- The OLRE ratings had the smaller sum of squared errors (63.92, versus 70.02 for the simulation-based expectations), suggesting stronger predictive power
- In addition, the Pearson correlation of the OLRE ratings with the actual number of wins was higher (0.67) than both the final regular season RPI ratings (0.52) and Sagarin ratings (0.55); see the plots above

Conclusions

- The ordinal logistic regression model was re-fitted in 2007, where the 2005-2006 results were considered as additional historical data to strengthen the model, and the simulation was also repeated
- The simulation-based expectations had the lower sum of squared errors (43.92, versus 54.33 for the OLRE ratings)
- The OLRE ratings once again had a higher correlation with actual success in 2007 (0.72) than the final regular season RPI ratings (0.68) and Sagarin ratings (0.67)
- A major limitation of this method is the apparent lack of freely available information on team-level variables at the ***end of the regular season***; additional predictors aside from winning percentage, point differential, strength of schedule, and number of wins against Top 30 opponents would undoubtedly improve the predictive power of the model
- Additional historical data prior to 2002 could also be collected to improve the fit of the model (there were only four outcomes of five and six wins available when fitting the 2007 model)
- Applications in other tournaments are also possible

For additional information, please visit http://www.umich.edu/~bwest

Comments are welcome! Please email me at bwest@umich.edu