ELSEVIER

International Journal of Forecasting 19 (2003) 313–317

international journal of forecasting

www.elsevier.com/locate/ijforecast

Predicting discrete outcomes with the maximum score estimator: the case of the NCAA men's basketball tournament

Steven B. Caudill*

Department of Economics, Auburn University, Room 203, 415 W. Magnolia, Alabama, AL 36849-5242, USA

Abstract

Seedings as a predictor of winning in the men's NCAA basketball tournament have recently been examined by Boulier and Stekler [*Int. J. Forecast.* 15 (1999) 83]. BS estimate a probit model to establish a relationship between seedings and the probability of winning. This study discusses the merits of a maximum score estimator for the prediction of discrete outcomes. Unlike the probit model, the maximum score estimator maximizes the number of correct predictions. The maximum score estimator is applied to updated data on the men's NCAA basketball tournament. The score estimator has better in-sample performance than the probit model used by BS. When out-of-sample predictions are examined using a series of rolling or recursive regressions, the maximum score estimator performs slightly better than the probit/maximum likelihood models. These results illustrate the potential advantages of using the maximum score estimator when predicting discrete outcomes.

Keywords: Maximum score; Discrete choice; Forecasting

1. Introduction

Several authors have recently examined the predictive power of seedings in athletic events. Recent studies by Schwertman, McCready and Howard (1991), Carlin (1996), Schwertman, Schenk and Holbrook (1996), and Smith and Schwertman (1999) have used various measures to predict the probability of winning and the margin of victory. As part of their study, Boulier and Stekler (1999) examine the usefulness of seedings as a predictor of winning in the men's NCAA basketball tournament. Using data from 1985–95, BS estimate probit models to establish a relationship between seedings and game outcome. Probit models are not calibrated to maximize the number of correct predictions. As an alternative, the maximum score estimator is presented in this study. The maximum score estimator maximizes the number of correct predictions. We apply the maximum score estimator to updated data on the men's NCAA basketball tournament. Based on in-sample predictions, the score estimator performs better than probit/maximum likelihood models. Comparisons are also made between the maximum score estimator and probit/maximum likelihood using out-of-sample predictions based on a series of rolling regressions. In these predictions the maximum score estimator demonstrates some slight advantages compared to probit. Together, these

*Tel.: +1-334-844-2907; fax: +1-334-844-4615.
E-mail address: scaudill@business.auburn.edu (S.B. Caudill).

0169-2070/02/$ – see front matter
doi:10.1016/S0169-2070(02)00008-0

results illustrate the virtue, in some cases, of using the maximum score estimator to predict discrete outcomes.

2. Data

The data for this study comes from the men's NCAA basketball tournament for the period 1985 to 1998. The field consisted of 64 teams for this entire period. There are sixteen seeded teams in each of four regions. Fifteen games are played in each region to determine a single team to send to the Final Four. Like BS, we do not consider the final four games so our data consists of sixty games per year for fourteen years for a total of 840 games.

Our data set contains information on the winner of each game and the seed numbers of the competing teams. Of the 840 games played the higher seeded team won 619, or nearly seventy-four percent of the time. This result suggests that seeding has an impact on winning, but a more careful analysis using regression methods is needed.

In order to determine the effects of seeding on game outcome and to demonstrate the usefulness of the score estimator, several models are estimated. In each case the dependent variable is equal to one if the higher seeded team wins, zero otherwise. Like BS, the independent variable used is DSEED which is the difference in seed numbers for the two teams. DSEED is calculated as high seed (HSEED) minus low seed (LSEED). For example, if a one seed plays a sixteen seed, DSEED will equal $1-16$ or -15. If models are estimated in this way, the effect of lowering the high seed by one position is constrained to be the same as increasing the low seed by one position. This means that the probability that a one seed defeats a four seed is constrained to be the same as the probability that a thirteen seed defeats a sixteen seed. We also relax this constraint by estimating models with high seed and low seed included as separate independent variables.

2.1. Maximum score estimation

The problem with using probit models to predict outcomes is that the probit model is not estimated so as to maximize the number of correct predictions (see Greene, 2000, pp. 833). If the prediction of a discrete outcome is the goal, a more appropriate estimator is the maximum score estimator. This estimator is based on maximizing the number of correct predictions so better (at least, no worse) in-sample performance compared to probit/maximum likelihood is assured.

The maximum score estimator is developed by Manski (1975, 1985, 1986) and Manski and Thompson (1989). This estimator of the discrete choice model is chosen as the solution of the following problem,

$$\text{Maximize } S_{N\alpha}(\beta) = \frac{1}{N}\sum_{i=1}^{N} [\tau_i - (1-2\alpha)]\,\text{sgn}(\beta' x_i), \tag{1}$$

where N is the sample size, α is a preset quantile, and $\tau_i = 2y_i - 1$, where y_i is the dependent variable, taking on values 0 or 1. If α is set equal to 0.5, the expression above simplifies to

$$\text{Maximize } S_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} [2y_i - 1]\,\text{sgn}(\beta' x_i), \tag{2}$$

and the parameter vector, β, is chosen so as to maximize the number of correct predictions, that is, the number of times the sign of $\beta' x_i$ matches the sign of $2y_i - 1$. Being based on the sign, the maximum score estimator can only identify β up to a positive scale constant. Consequently, the constraint $\beta'\beta = 1$ is imposed in the estimation to identify the parameter vector. This constraint has the effect of forcing β to be of unit length.

3. Results

The prediction results are based on the estimation of four models: probit/ML including DSEED as the independent variable, probit/ML including HSEED and LSEED, maximum score including DSEED, and maximum score including HSEED and LSEED. For our sample, the maximum score estimator gives the largest number of correct predictions (623), following next by a tie between the probit/ML model including DSEED and the maximum score model including DSEED (619). The fewest correct predic-

tions are obtained from the probit/ML model including HSEED and LSEED (617).

The superior in-sample performance of the maximum score estimator including HSEED and LSEED is due to two factors. First, unlike either probit model, the maximum score estimator is chosen to maximize the number of correct predictions. Second, the added flexibility from including HSEED and LSEED leads to a model capable of predicting victory for a lower seeded team. The maximum score estimator including DSEED always predicts the higher seed will win. Many times, improved predictions can be obtained by deviating from this simple rule. For example, in our sample of 840 games the team with the higher seed won 619. In this sample the eighth seed played the ninth seed fifty-six times with the ninth seed winning thirty-one games and the eighth seed winning twenty-five. On four occasions a twelfth seed played a thirteenth seed, with the twelfth seed winning three times. Our maximum score estimator for this sample predicts the lower seeded team will win if the difference in seed is one position and no team is seeded higher than number seven. This accounts for the four additional correct predictions for the maximum score estimator. The details vary from sample to sample but the potential for improvements with the maximum score estimator is due to the fact that, when needed, a model can be produced to sometimes predict that the lower seeded team will win. This accounts for the superior in-sample performance of the maximum score estimator including HSEED and LSEED. Next, out-of-sample performance of the models is compared.

To compare out-of-sample predictive performance, a series of rolling or recursive regressions is estimated. Predictions in each year are based on estimation results using data from previous years. With our sample, these recursive regressions yield predictions for thirteen years, 1986 to 1998.

Out-of-sample predictions for the models previously discussed are given in Table 1. The prediction results from the probit/ML model including only DSEED are given in column two, the probit/ML results including HSEED and LSEED are given in column three, the maximum score results including DSEED are given in column four, and the maximum score results including HSEED and LSEED are given in column five.

The second row of the table indicates the number of years in our sample for which each model yields the strictly largest (no ties) number of correct predictions. The maximum score model including HSEED and LSEED is first with two years, followed by the maximum score including DSEED and the probit/ML including HSEED and LSEED with one year each. The probit/ML including only DSEED is last with zero years.

Row three indicates the number of years for which each model yields the largest number of correct predictions with ties included. First is maximum score including DSEED (10 years), next is probit/ML including DSEED (9 years), followed by maxi-

Table 1
Annual predictions: relative performance of various models

	Probit/ML DSEED	Probit/ML HSEED, LSEED	MSCORE DSEED	MSCORE HSEED, LSEED
Number of years with highest number of correct predictions (excluding ties)	0	1	1	2
Number of years with highest number of correct predictions (including ties)	9	6	10	7
Number of years with lowest number of correct predictions (excluding ties)	0	1	0	4

mum score including HSEED and LSEED (7 years), and finally probit/ML including HSEED and LSEED (6 years).

Poor prediction performance is examined in row four of the table. This row indicates the number of years for which each model gives the fewest correct predictions, excluding ties. First is the maximum score model including HSEED and LSEED which gave the fewest correct predictions in four years. Next, the probit/ML including HSEED and LSEED gave the fewest correct predictions in one year. For no years in the sample did either the probit/ML including DSEED or the maximum score model including DSEED give the fewest correct predictions.

Another comparison of predictive performance can be made by examining the total number of correct predictions in all thirteen years of recursive regressions. Ranked by the number of correct predictions, the maximum score model including DSEED is first with 575, next is the probit/ML including DSEED with 571, then the probit/ML including HSEED and LSEED with 567, and last is the maximum score model including HSEED and LSEED with 566.

These results establish that the maximum score model including DSEED performed very well. In one case, this model, alone, gave the largest number of correct predictions, for ten years no other model gave more correct predictions, and for no year did this model yield the fewest correct predictions. This model also gave the highest number of total correct predictions (575) in all recursive regressions. The closest competing model is the probit/ML model including DSEED. For nine years no other model gave more correct predictions and for no year did this model give the fewest correct predictions. In the recursive regression results, this model's 571 correct predictions is second only to the maximum score model including DSEED. The maximum score estimator seems to predict better than maximum likelihood for these samples, although the difference in performance is not great.

These results show that if one is interested in predicting discrete outcomes, the maximum score estimator can lead to improved performance over probit/maximum likelihood models. However, the maximum score estimator does have some shortcomings. For example, the estimates may not be unique, although this problem is less of a concern with a large sample. Also, standard errors are unavailable as part of the estimation and are usually obtained by bootstrapping. In addition, the score estimator does not really provide probability predictions but classifies games into 'win' or 'lose.' Without probability predictions, maximum score estimation results provide no information on how strong a prediction of victory is given. Despite these shortcomings, the estimator may be useful when predicting discrete outcomes.

4. Conclusions

We extend the previous work by exploring the use of the semiparametric, maximum score estimator to predict outcomes in the NCAA men's basketball tournament. We compare in-sample performance and out-of-sample performance for probit/maximum likelihood and maximum score estimators. We show that the maximum score estimator has superior in-sample performance, although the gains are not large compared to probit. The maximum score estimator also demonstrates improved out-of-sample performance when compared to probit although, again, the gains are not large in the simple model examined by Boulier and Stekler. Still, as a supplement to probit/maximum likelihood, the maximum score estimator provides information that may be useful to those predicting discrete outcomes.

References

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting, 15*, 83–91.

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

Greene, W. (2000). *Econometric analysis*, 4th ed. New Jersey: Prentice-Hall.

Manski, C. (1975). The maximum score estimator of the stochastic utility model of choice. *Journal of Econometrics, 3*, 205–228.

Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics, 27*, 313–333.

Manski, C. (1986). Operational characteristics of the maximum score estimator. *Journal of Econometrics*, *32*, 85–100.

Manski, C., & Thompson, S. (1989). Estimation of best predictors of binary response. *Journal of Econometrics*, *40*, 97–124.

Schwertman, N. C., McCready, T. A., & Howard, L. (1991). Probability models for the NCAA regional basketball tournament. *The American Statistician*, *45*, 35–38.

Schwertman, N. C., Schenk, K. L., & Holbrook, B. C. (1996). More probability models for the NCAA regional basketball tournaments. *The American Statistician*, *50*, 34–38.

Smith, T., & Schwertman, N. C. (1999). Can the NCAA tournament seedings be used to predict margin of victory? *The American Statistician*, *53*, 94–98.

Biography: Steven B. CAUDILL is Professor of Economics in the College of Business at Auburn University. His research interests include limited–dependent variable models and finite mixture models. Dr. Caudill has articles published in the *Review of Economics and Statistics*, the *Journal of Econometrics*, the *Journal of Applied Econometrics*, and the *Journal of Applied Statistics*.