A Penalized Maximum Likelihood Approach for the Ranking of College Football Teams Independent of Victory Margins

David Mease
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1092

Abstract

We introduce a penalized maximum likelihood approach for ranking all NCAA Division 1-A college football teams. The model does not consider margin of victory and is based solely on win/loss data. Despite the simplicity, the model leads to rankings which exhibit greater agreement with expert opinion (i.e. AP and Coaches polls) than the models historically and currently used by the Bowl Championship Series (BCS). The model can be implemented using standard statistical software packages.

Key Words: Paired Comparisons; Football; Rankings; Bayesian.

1. INTRODUCTION

Prior to 1998, the national championship of US Division 1-A college football was based solely on two opinion polls, commonly referred to as the Coaches and the AP (Associated Press) polls. The ranking of the teams in these two polls is determined by coaches and sports writers respectively who vote weekly for the top 25 teams based on the teams' performances in all games played. Up until 1998, the team that finished the year in the number one position in both polls was deemed to be the national champion. In cases in which the two polls did not agree on the choice of the top team, the two teams involved shared the title of national champion, which occurred most recently in 1997 with Michigan and Nebraska.

With the 1998 season came the inception of a new system under which the top two teams at the end of the season would play one final game for the championship, thereby eliminating the possibility of a shared title. This new system was called the Bowl Championship Series, or BCS. The BCS system employed rankings produced by a number of computer models in addition to the rankings of the AP and Coaches polls in order to determine the top two teams.

The purpose of these computer models was to lessen dependence on the AP and Coaches polls which, although historically trusted as representing expert opinion, have often been criticized on the following two accounts. First, the human pollsters are not objective observers and may have biases toward certain schools based on regional loyalty, historical perception, etc. Secondly, it is impossible for a human pollster to recall all outcomes of all games involving the 117 Division 1-A teams over the course of an 11 to 14 week season, even if he or she had witnessed or read about every game.

Although the computer models employed by the BCS do not have any such bias and are able to consider the outcomes of all games played, they have also proven to be extremely controversial as a result of many instances in which they produced nonintuitive rankings which differed significantly from the AP and Coaches polls. For instance, in 2001 the University of Oregon which finished 2nd in both the AP and Coaches polls, finished 8th in one of the eight BCS computer models and 7th in two of the others. The low ranking of Oregon in these three computer polls was thought to be attributable to their many victories by narrow point margins, since the four BCS computer models that did not use margin of victory ranked Oregon no lower than third. This was not the first time that controversy resulted from the computers polls weighing margin of victory much more heavily than public opinion, and as a result it was mandated that all computer polls either ignore margin of victory or be excluded from the BCS system beginning in 2002. The idea was that by forcing computers to ignore margin of victory, the resulting rankings would be more consistent with the public's opinion, which tends to be more a function of a team's winning percentage and quality of opposition than a function of the point margins. Furthermore, this would remove any incentive for a team to "run up the score" in a game that is a foregone conclusion, which is universally considered to be bad sportsmanship.

It should be noted that the belief that the BCS computer models are more influenced by margin of victory than public opinion is not shared by everyone. Some people, including the creators of some of these computer models, would argue the human pollsters themselves can be highly influenced by large victory margins, citing examples in which a team that wins by a large margin climbs higher in the polls than a team that wins by a small margin. While it is in fact possible for human pollsters to be influenced in such a way, a number of instances similar to the one involving Oregon described above were enough to convince the BCS that the computer polls tended to weigh margin of victory too heavily (even after a restriction was made limiting the maximum margin to 21 points). Further evidence that human pollsters tend

to put relatively small weight on margin of victory is given by Harville (2003), who noted that the version of his model which ignores margin of victory agreed more strongly with the human pollsters for the football season he considered. Sentiment for eliminating margin of victory from the computer polls also arose as a result of the fact that a computer poll cannot discern between a large margin of victory resulting from one team dominating another for an entire game versus a large victory margin as a result of a large amount of scoring occurring after the game has already effectively been decided. A human pollster, on the other hand, can differentiate these two situations.

In this article we examine the problem of constructing a computer model to rank the (currently) 117 Division 1-A football teams without using victory margins. A new computer model is proposed that is shown to produce rankings which on average agree more strongly with the AP and Coaches polls than the models used by the BCS as well as other competing models. The proposed model uses a penalized likelihood approach which results in a ranking process that attempts to mimic the thought processes of the human pollsters of the AP and Coaches polls. The paper is organized as follows. Section 2 gives a simple example illustrating the complexities involved in ranking teams without using victory margins. Section 3 describes some of the models used in the statistics literature for ranking football teams. Section 4 discusses the models that are or have been used by the BCS. Sections 5 and 6 present the proposed model and discuss its implementation using statistical software. Section 7 compares the rankings produced by the proposed model to those of some competing models, including the BCS computer models. Section 8 describes some possible modifications to the model.

2. A SIMPLE EXAMPLE

To illustrate the complexity in fairly ranking teams based solely on win/loss data, consider the following season in which 5 teams (A, B, C, D, and E) played a total of 8 games with the following outcomes:

Game 1: Team A defeated Team C	Game 2: Team A defeated Team E
Game 3: Team B defeated Team A	Game 4: Team B defeated Team E
Game 5: Team C defeated Team D	Game 6: Team C defeated Team E
Game 7: Team D defeated Team E	Game 8: Team D defeated Team E

Team B finishes the season with a win/loss record of 2 wins and zero losses (denoted 2-0), Team E finishes with a record of 0-5, and Teams A, C, and D all finish with records of 2-1.

Now suppose no scores are available and we seek to fairly rank the five teams based only on the above outcomes. We can judge the strength of each team based on its win/loss record as well as on the strength of its opponents, who in turn are judged based on their win/loss records and the strength of their opponents, and so on. An argument for ranking these five teams may go as follows.

Team B should be ranked as the first place team, having no losses and having defeated team A, who is a strong opponent since if not for being defeated by team B, would have no losses. Team E should be ranked as the worst team, having no wins and having loss to team D twice, who is a weak opponent since if not for defeating team E (twice) would have no wins. Teams A, C, and D remain to be ranked from 2nd to 4th. Although all three have the same record of 2-1, Team A should be ranked 2nd since their only loss came to team B, who is a strong opponent since they are the number one team. Finally, Team C should be ranked above team D since team C defeated team D while team D defeated only team E, who is a weak opponent since they are ranked as the worst team. Thus the ranking of the teams from first to last should be B, A, C, D, and E.

While such an argument is tractable for the five teams playing eight games in the example, for the 117 Division 1-A teams who each play 11 to 14 games on average the problem becomes extremely difficult for a human pollster to handle. The voters of the AP and Coaches polls attempt to reason in a similar fashion as above, but it is impossible for them to simultaneously consider all games played. This is the motivation for developing a computer model that can simultaneously judge the strength of all 117 teams based on winning percentage and strength of opposition.

3. FOOTBALL RANKING MODELS IN THE STATISTICAL LITERATURE

A number of articles exist in the statistics literature dealing with the ranking of college football teams. Many of these use a linear model approach in which margin of victory is modelled as a function of the two competitors in each game as well as a number of covariates. For instance, Harville (1977) proposed a linear model which considered as covariates the location of the game (i.e. which team was the home team), the date of the game, and the division membership of the two teams involved in each game. In order to downplay victories by large margins he also considered two variations on this model, one in which the margin of victory was truncated at 15 points and one in which the margin of victory was truncated at 1 point. Note that latter model in essence ignores margin of victory and as such offers a possible solution

to the problem of interest in this paper. Results from Harville's models will be considered in Section 7. In addition to truncating margin of victory at a certain value as Harville did, other approaches to downplaying the effect of large victory margins in linear models are offered by Stern (1992) and Bassett (1997). In Stern's (1992) model a smaller weight was given to games in which the victory margin was large, while Bassett (1997) proposed using absolute error as opposed to least squares.

In addition to linear model approaches, likelihood based approaches can also be found in the statistics literature. Keener (1993) considered a number of models for ranking college teams, including a version of what is often called the Bradley-Terry model after a paper dealing with paired comparisons for experimental design (Bradley and Terry 1952). In this model, each Team i is assigned a parameter θ_i that can be thought of as representing its strength. The probability that team X defeats team Y is then given by $\frac{\theta_X}{\theta_X+\theta_Y}$, and the estimates for the θ_i values can then be obtained by maximizing the likelihood over all the game outcomes for the entire season. Ordering the MLE estimates for the θ_i values then determines a ranking of the teams. While in his model Keener used a weighting factor to weigh large margins of victory more than close victories, it is not necessary.

It is important to note that a problem with the Bradley-Terry model is that if team i has no losses, the likelihood becomes larger as θ_i becomes larger. It follows that the maximum likelihood estimator for the θ_i value of any undefeated team is infinity. As a consequence, the model will produce a tie for first among all undefeated teams. This is problematic due to the substantial variability in the strength of opponents among the different Division 1-A teams. It is undesirable to restrict all teams with at least one loss to be ranked below all undefeated teams. Moreover, a two- or three-way tie for first is a rather unsatisfactory answer to the question of who is number one. Keener applies the model to a season in which all teams have at least one loss so that this is not a problem; however, for general use the model would need to be modified since it is not uncommon to have one or more undefeated team in Division 1-A football.

A different likelihood based approach was proposed by Thompson (1975) for ranking the professional teams of the National Football League (NFL). Similar to the model above, each team i was assigned a strength parameter θ_i, but in this model the probability that team X defeated team Y was given by $\Phi(k(\theta_X - \theta_Y))$ where θ_X and θ_Y are the θ_i parameters for teams X and Y respectively and Φ is the standard normal cumulative distribution function. The reason for the inclusion of the parameter k in the model is that Thompson restricted

the values for the θ_i's to be the integers $1, ..., n$ where n is the number of teams. While this restriction eliminates the problems associated with undefeated teams described in the above paragraph, it also makes maximization of the likelihood extremely difficult, requiring a search over $n!$ possibilities. Thompson avoids this problem by further restricting that no team can be ranked below a team having a lower winning percentage. While such a restriction may be reasonable for the teams of the NFL who play schedules that are relatively comparable in difficulty, such a restriction is undesirable when ranking college football teams due to the variability in the schedule strengths mentioned above. Note that the use of the normal CDF for paired comparisons can be traced back to Thurstone (1927).

4. COMPETING MODELS USED BY THE BCS

In the five years of existence up to the time of the writing of this article, ten different models have been used by the BCS for ranking the Division 1-A college football teams. Moreover, some of these ten models have been modified during this period. Due to the large number of models and the fact that in some cases the technical details regarding the models are not public, we will not give descriptions of each specific model. Instead, we will evaluate these models based on the rankings they produced for the specific football seasons during which each model was included in the BCS. This will be done in Section 7. While we do not provide a description of each BCS model, it should be noted that the BCS models cover a wide range of complexity. Some models take into account only win/loss data while others consider such factors as location, date, rankings from previous seasons, and margin of victory (prior to 2002).

5. THE PROPOSED MODEL

To begin, let us suppose the intrinsic performance level of each team i varies from day to day according to a normal distribution with mean θ_i and variance of $1/2$ (for simplicity). Treating the performance level as random is consistent with the fact that even very good teams can be be "upset" by weaker teams. The use of the normal distribution was used largely because it results in a model that produces reasonable rankings, but it can also be motivated by supposing that the performance level is a sum of a large number of independent factors and invoking the central limit theorem. Further suppose that this intrinsic performance level of each team is independent of its opponent's level and that every game is won by the team that has the greater performance level on that day. Under these assumptions the probability that

team X defeats team Y is $\Phi(\theta_X - \theta_Y)$, similar to the Thompson model described earlier. Note that the same expression also arises from the model of Harville (2003) by supposing that the difference in the score when team X plays team Y is normal with mean $\theta_X - \theta_Y$ and variance 1 and that only win/loss data is available. However, we prefer the interpretation of the θ_i parameter as a mean performance level rather than a mean score for modelling win/loss data. This is the basic idea behind our likelihood which is given in its entirety for the n teams by

$$l(\theta) = \underbrace{\prod_{(i,j)\in S} [\Phi(\theta_i - \theta_j)]^{n_{ij}}}_{\text{Part 1}} \times \overbrace{\prod_{i=1}^{n} \Phi(\theta_i)\Phi(-\theta_i)}^{\text{Part 2}} \times \underbrace{\Phi(\theta_{n+1})\Phi(-\theta_{n+1}) \times \prod_{(i,j)\in S^*} [\Phi(\theta_i - \theta_j)]^{n_{ij}}}_{\text{Part 3}}$$

and discussed in detail in the following paragraphs. This likelihood function $l(\theta)$ is maximized as a function of $\theta = (\theta_1, ..., \theta_{n+1})$ where θ_{n+1} is a nuisance parameter. We will rank the teams according to the MLE's of their respective θ_i values such that the team with the largest value of θ_i is ranked as the top team.

Part 1 of the likelihood captures the motivation from the first paragraph in this section where we proposed that the probability that team X defeats team Y is $\Phi(\theta_X - \theta_Y)$. For Part 1 we define the set S to consist of all ordered pairs (i, j) for which team i defeated team j and both team i and team j are members of Division 1-A. The value n_{ij} is the number of times team i defeated team j. Tie games are counted as half a win and half a loss.

In the maximization of Part 1, a team earns a large value of θ_i by defeating other teams who themselves have earned relatively large θ_i values. In this way, maximization of Part 1 simultaneously judges the strength of each team based on its win/loss record and based on the strength of its opponents, which is in turn determined based on their win/loss records and the strength of their opponents. Thus, the process of maximizing Part 1 alone mimics the thought processes a human would attempt to use to rank the teams as described in Section 2. However, there are a couple of problems with using Part 1 alone.

The first problem is the difficulty mentioned earlier in conjunction with the Bradley-Terry model. That is, if a team i has no losses, Part 1 increases without bound as θ_i becomes larger so that the MLE for any undefeated team is infinite. To solve this problem we penalize the likelihood by multiplying by $\Pi_{i=1}^{n}\Phi(\theta_i)\Phi(-\theta_i)$ as given by Part 2. For more discussion on estimation based on penalized likelihood the reader is directed to Green (1998) and Silverman (1985).

While the main reason for using the penalty term above is that it restricts the MLE estimates of the θ_i's to be finite, it will also be shown that it leads to rankings which consistently

agree with human polls (i.e. AP and Coaches polls). Furthermore, the use of this penalty term is easily justified. Two justifications for the penalty term are given in the following two paragraphs.

One way to view the penalty term is as Bayesian prior distribution. To illustrate this we will associate with a team X the parameter ξ_X defined by $\xi_X = \Phi(\theta_X)$. Note that these ξ parameters can be thought of as the probability of defeating a team with a θ_i value of zero. Now if we treat the unpenalized likelihood described before as the conditional distribution of the game outcomes given the ξ's and assign independent Beta distributions with parameters α and β as the prior for the ξ of each team, the posterior of the ξ's is proportional to the unpenalized likelihood and

$$\Pi_{i=1}^{n} \xi_i^{\alpha-1}(1-\xi_i)^{\beta-1} = \Pi_{i=1}^{n} \Phi(\theta_i)^{\alpha-1}(1-\Phi(\theta_i))^{\beta-1}.$$

Taking $\alpha = \beta = 2$ gives the likelihood proposed. The motivation for taking $\alpha = \beta$ is to give an equal prior probability of winning and losing for each team, while the choice of the value 2 for these two parameters was selected after testing the model on actual football seasons.

A second way to view the penalty term is by considering one additional "virtual" team with a θ_i value of zero. Giving every team exactly one win and one loss to this virtual team also results in a likelihood with the penalty term proposed. (Note that in general one could give $\alpha - 1$ wins and $\beta - 1$ losses to this virtual team.) Viewing the penalty term in this way makes it clear how the problems presented by undefeated teams are solved, since when the virtual team is considered each team actually has at least one loss. Furthermore, this view of the penalty term also suggests why it is simple to fit the model using existing binary response regression procedures to be described in Section 6.

Finally, maximization of the product of Parts 1 and 2 will produce an adequate ranking of the teams, but without using Part 3 we would be ignoring all games in which one opponent was a member of Division 1-A but the other was not (recall the set S does not include such pairings). Since such games are not extremely common, it is not unreasonable to ignore such games. In fact, the BCS computer model of Wes Colley does exactly that. However, it is foreseeable that this could lead to considerable controversy if, say, a certain team's only loss was to a non-Division 1-A opponent. In such a case, ignoring this loss would make this team undefeated and ranked as one of the best teams, despite having a loss to an extremely weak opponent. In place of ignoring these games, another solution would be to expand the set S to include all games of any team who has played a Division 1-A team. However, these extra games would involve yet another set of teams for which we would be ignoring games against

opponents not in the set. Thus, unless we are willing to rank almost all college football teams in the nation, a different solution must be found.

The solution that was chosen is to combine all teams not in Division 1-A who play a Division 1-A opponent into one generic $n+1st$ team. The set S^* is then taken to be all ordered pairs (i,j) for which team i defeated team j and one of i or j is equal to $n+1$ and the other is in $\{1,..,n\}$. The term $\prod_{(i,j)\in S^*}[\Phi(\theta_i-\theta_j)]^{n_{ij}}$ in Part 3 of the likelihood then accounts for all games involving Division 1-A teams that were not included in the set S. As before, the n_{ij} are the number of times team i defeated team j. Note that most of the games in the set S^* are losses for the generic non-Division 1-A team, and as such θ_{n+1} invariably has a very small MLE value. Thus, the reward for a Division 1-A team beating a team not in Division 1-A is very small while a loss to a non-Division 1-A has a strong negative effect on the ranking, as should be the case. Finally, the term $\Phi(\theta_{n+1})\Phi(-\theta_{n+1})$ is included in Part 3 to penalize the likelihood of this generic $n+1st$ team just as Part 2 did for the other n teams.

6. IMPLEMENTATION OF THE MODEL USING STATISTICAL SOFTWARE

Standard statistical software packages can be used to maximize the likelihood function for the model by fitting a binary regression model employing the probit link function. The procedure for doing this is similar to the method used by Fienberg and Larntz (1976) for fitting the Bradley-Terry model. The data matrix should consist of one column for each team in Division 1-A as well as one additional column to represent the generic non-Division 1-A team. For each game played involving at least one Division 1-A team, a row is included in the data matrix in which the value 1 is placed in the column for the winner of the game and a -1 is placed in the column for the loser of the game while the values for the other columns are all set to zero. In addition to these rows, the data matrix should also contain two additional rows for each team. For any given team these two rows are as follows. The first row should contain a 1 in the column for that team with the rest of the entries being zero, while the second row should contain a -1 for that team with the rest of the values being zero. These two additional rows provide the penalty term for the likelihood. Finally, the response vector should be set to a vector of all 1's equal in length to the number of rows in the data matrix.

If a probit regression model with no intercept is fit to the data matrix and response vector constructed in this way it can be verified that the likelihood being maximized is exactly the likelihood for the proposed model. As such, the parameter estimates output from the software corresponding to each of the different columns of the data matrix will be the maximum likelihood

estimates for the θ_i's in the proposed model. Thus, the single parameter estimate for the generic non-Division 1-A team can be discarded and the remaining values can be used to rank all the Division 1-A teams. Code for implementing this procedure in SAS for the example presented in Section 2 can be found at http://members.accesstoledo.com/measefam/SAScode.html.

Note that if tie games exist in the data the above procedure must be modified slightly since not all n_{ij} will be integers. In these cases, each row in the data matrix described above that does not correspond to a tie game should be entered twice, while rows corresponding to tie games should still be entered only once. In this way the likelihood being maximized is proportional to the square of the desired likelihood, and the ranking of the teams produced will be correct.

7. RESULTS AND COMPARISONS WITH COMPETING MODELS

In this section we will examine the rankings resulting from the competing models introduced earlier. We will compare these rankings to those of the proposed model in terms of similarity to the AP and Coaches Polls. This similarity will be quantified in terms of the average absolute difference between the rankings of each model and the average ranking of the AP and Coaches Polls, although one may argue that a measure which gives large differences smaller weight (or larger weight) than does the absolute value function may be more appropriate.

Harville (1977) gave his rankings for the 1975 college football season. These rankings are included in Table 1 along with the rankings for the model proposed in this paper ("Mease"). The teams listed are the top 15 teams in terms of average rankings of the AP and Coaches Polls ("Poll Avg."). The three columns listed for Harville's model correspond to the original model ("no cap") as well as the two models in which the margin of victory is limited to 15 points and 1 point. The final row in the table gives the average absolute difference between the rankings for each model and the AP and Coaches average rankings for the 15 teams. From these values it can be seen that for the 1975 season the proposed model agrees more closely with the AP and Coaches Polls than the "15 point" and "no cap" Harville models, while the "1 point" Harville performs better than the proposed model with regard to average absolute difference from the AP and Coaches Polls.

Keener (1993) gave rankings based on his model for the 1989 college football season. Table 2 gives these rankings for the top 15 teams as determined by the average of the AP and Coaches Polls for 1989 ("Poll Avg."). From this table it can be seen that the model proposed in this paper has an average absolute ranking difference of 1.60 from the AP and Coaches average

rankings. By comparison, the Keener model has an average difference of 2.87 and as such did not perform as well in terms of similarity to the AP and Coaches Polls as the model proposed in this paper. It is also interesting to note that Houston, the second ranked team in the Keener model, is not included in the table since it was not ranked by the Coaches Poll, which only ranked 20 teams at that time. The model proposed in this paper ranked Houston as 15th, which is similar to their AP ranking of 14th.

Table 1: Comparison to Harville's Models (1975 Season)

Poll Avg.	Team	Record (Wins-Losses-Ties)	Mease	Harville 1 point	Harville 15 point	Harville no cap
1	Oklahoma	11-1-0	2	1	6	3
2	Arizona State	12-0-0	1	2	9	14
3	Alabama	11-1-0	5	3	4	1
4	Ohio State	11-1-0	3	4	2	2
5	UCLA	9-2-1	16	16	16	15
6.5	Texas	10-2-0	7	7	3	5
6.5	Arkansas	10-2-0	9	10	5	6
8	Michigan	8-2-2	13	19	14	8
9	Nebraska	10-2-0	6	5	7	4
10	Penn State	9-3-0	12	9	10	11
11.5	Texas A&M	10-2-0	10	6	12	12
12	Maryland	9-2-1	27	23	15	17
14	Miami(Ohio)	11-1-0	8	8	25	26
14	Pitt	8-4-0	21	13	13	7
14.5	California	8-3-0	15	18	22	22
0	Avg. Abs. Diff.		3.93	3.67	4.13	4.53

Table 2: Comparison to Keener's Model (1989 Season)

Poll Avg.	Team	Record	Mease	Keener
1	Miami(Florida)	11-1-0	2	1
2.5	Notre Dame	12-1-0	1	3
2.5	Florida State	10-2-0	5	5
4	Colorado	11-1-0	3	7
5	Tennessee	11-1-0	4	11
6	Auburn	10-2-0	9	8
7.5	Michigan	10-2-0	8	9
8	Alabama	10-2-0	6	12
8.5	Southern California	9-2-1	10	4
10	Illinois	10-2-0	7	14
11.5	Nebraska	10-2-0	14	13
11.5	Clemson	10-2-0	11	6
13	Arkansas	10-2-0	13	17
15	Penn State	8-3-1	16	18
16	Michigan State	8-4-0	19	15
0	Avg. Abs. Diff.		1.60	2.87

The final tables listed give comparisons with the BCS models. Tables 3-7 give the rankings for the BCS models for the five college football seasons beginning in the years 1998, 1999, 2000, 2001 and 2002 respectively along with the rankings for the model proposed in this paper ("Mease"). The rankings for all models as well as the average of the AP and Coaches Polls ("Poll Avg.") exclude the bowl games played at the end of the season. This was done since the primary function of the BCS is to select teams to participate in these bowl games. The teams included in Tables 3-7 are the top 15 teams as computed by the overall BCS ranking system prior to the bowl games and are listed in the order of these overall BCS rankings.

From Table 3 it can be seen that the model proposed in this paper outperformed two of the three models used by the BCS in 1998 in terms of average absolute difference from the AP and Coaches average ranking. In Tables 4-6 the proposed model outperformed 3, 7 and 7 of the eight models used in the years 1999, 2000, and 2001 respectively and Table 7 shows the proposed model outperformed 4 of the seven models used in 2002. Table 8 summarizes the average absolute ranking differences from the AP and Coaches average rankings over the five seasons in which the BCS existed. From this comparison it can be seen that the model proposed in this paper in fact had the smallest average difference from the AP and Coaches average over the five year period among all BCS models.

Table 3: Comparison to 1998 BCS Models

Poll Avg.	Team	Record (Wins-Losses)	Mease	Jeff Sagarin	New York Times	Anderson/ Hester
1	Tennessee	12-0	1	2	2	1
2	Florida State	11-1	2	3	1	2
4	Kansas State	11-1	4	1	5	4
3	Ohio State	10-1	8	6	3	7
5.5	UCLA	10-1	3	4	6	3
8.5	Texas A&M	11-2	5	5	4	6
5.5	Arizona	11-1	6	9	9	5
7	Florida	9-2	10	8	11	10
8.5	Wisconsin	10-1	9	10	10	9
10	Tulane	11-0	7	14	23	8
15	Nebraska	9-3	11	7	15	11
12.5	Virginia	9-2	15	18	17	13
11	Arkansas	9-2	14	12	22	17
13	Georgia Tech	9-2	16	20	12	16
17.5	Syracuse	8-3	24	16	7	24
0	Avg. Abs. Diff.		2.47	3.07	3.80	2.33

Table 4: Comparison to 1999 BCS Models

Poll Avg.	Team	Record	Mease	Jeff Sagarin	New York Times	Anderson/ Hester	Richard Billingsley	Dunkel	Kenneth Massey	Herman Mathews	David Rothman
1	Florida State	11-0	1	1	1	1	1	1	1	1	1
2	Virginia Tech	11-0	2	2	2	3	2	2	2	2	2
3	Nebraska	11-1	3	3	4	2	3	3	3	3	3
5.5	Alabama	10-2	4	6	3	4	5	7	6	4	4
5.5	Tennessee	9-2	9	5	5	8	7	6	5	5	6
7	Kansas State	10-1	6	4	6	5	4	5	4	6	5
4	Wisconsin	9-2	13	7	8	12	8	4	7	11	9
8	Michigan	9-2	8	9	7	6	10	9	8	7	10
9	Michigan State	9-2	7	8	10	7	6	8	9	8	8
10	Florida	9-3	10	11	16	9	9	12	12	9	7
15	Penn State	9-3	11	10	20	11	11	10	10	10	11
11	Marshall	12-0	5	13	11	15	33	31	11	22	12
12	Minnesota	8-3	24	15	21	21	14	19	17	15	15
15.5	Texas A&M	8-3	16	17	15	14	13	16	15	19	16
16	Texas	9-4	14	14	21	13	17	13	16	14	13
0	Avg. Abs. Diff.		2.77	1.57	2.43	2.83	3.10	2.90	1.30	2.50	1.77

Table 5: Comparison to 2000 BCS Models

Poll Avg.	Team	Record	Mease	Jeff Sagarin	New York Times	Anderson/ Hester	Richard Billingsley	Dunkel	Kenneth Massey	Herman Mathews	David Rothman
1	Oklahoma	12-0	1	3	3	1	1	3	2	2	1
3	Florida State	11-1	2	1	1	3	2	1	1	1	2
2	Miami(Florida)	10-1	3	2	2	4	3	2	3	3	3
4	Washington	10-1	4	8	5	2	10	11	5	4	4
5.5	Virginia Tech	10-1	6	5	4	6	5	5	4	7	7
5.5	Oregon State	10-1	5	7	8	5	7	9	8	5	5
7	Florida	10-2	7	6	6	7	4	4	7	6	9
8.5	Nebraska	9-2	9	4	10	9	6	13	6	8	6
10	Kansas State	10-3	11	9	12	12	8	12	11	11	8
9.5	Oregon	9-2	8	14	15	8	12	17	14	9	11
10	Notre Dame	9-2	10	16	8	10	14	15	15	10	12
12	Texas	9-2	15	10	11	15	11	6	9	12	10
16	Georgia Tech	9-2	12	11	7	11	9	8	10	13	14
14.5	TCU	10-1	14	12	20	20	16	7	12	14	15
14.5	Clemson	9-2	13	15	19	13	13	21	13	15	13
0	Avg. Abs. Diff.		1.00	2.47	2.73	1.60	2.33	4.33	2.33	0.87	1.33

Table 6: Comparison to 2001 BCS Models

Poll Avg.	Team	Record	Mease	Anderson/ Hester	Wes Colley	Richard Billingsley	Kenneth Massey	David Rothman	Jeff Sagarin	Herman Mathews	Peter Wolfe
1	Miami (Florida)	11-0	1	1	1	1	1	1	1	1	1
4	Nebraska	11-1	2	2	2	2	3	2	3	2	2
3	Colorado	10-2	4	4	5	4	4	5	5	5	3
2	Oregon	10-1	3	3	3	3	2	8	7	6	7
5	Florida	9-2	7	9	8	7	8	4	2	3	5
8	Tennessee	10-2	5	5	4	8	6	7	8	7	4
9	Texas	10-2	8	8	9	10	9	3	4	4	6
7	Illinois	10-1	6	7	6	6	12	13	12	10	12
11	Stanford	9-2	10	6	7	11	5	9	9	8	8
6	Maryland	10-1	9	14	10	5	10	11	11	14	11
10	Oklahoma	10-2	11	10	11	9	13	6	6	9	9
13	Washington State	9-2	12	12	12	12	7	10	10	11	10
12	LSU	9-3	15	11	13	14	14	12	18	13	14
14	South Carolina	8-3	18	20	19	19	17	17	23	23	17
20.5	Washington	8-3	16	13	15	15	11	16	25	17	13
0	Avg. Abs. Diff.		1.90	2.70	2.30	1.57	3.03	3.03	3.63	3.10	2.90

Table 7: Comparison to 2002 BCS Models*

Poll Avg.	Team	Record	Mease	Anderson/ Hester	Wes Colley	Richard Billingsley	Kenneth Massey	New York Times	Jeff Sagarin	Peter Wolfe
1	Miami (Florida)	12-0	2	2	1	1	1	1	1	2
2	Ohio State	13-0	1	1	2	2	2	3	2	1
4	Georgia	12-1	3	3	3	3	4	4	3	3
5	USC	10-2	4	5	4	6	3	2	4	4
3	Iowa	11-1	5	4	5	5	8	5	5	5
7	Washington State	10-2	8	8	8	9	5	9	6	6
8	Oklahoma	11-2	6	7	7	4	7	6	8	7
6	Kansas State	10-2	11	14	12	11	10	7	11	10
11.5	Notre Dame	10-2	7	6	6	8	6	13.5	7	8
9	Texas	10-2	9	10	9	7	11	11	9	11
11.5	Michigan	9-3	10	9	10	16	9	8	10	9
10	Penn State	9-3	13	11	13	14	14	10	15	13
13.5	Colorado	9-4	15	13	15	22	13	16	13	15
15.5	Florida State	9-4	14	12	11	23	12	18	12	12
13.5	West Virginia	9-3	18	18	16	15	18	15	18	17
0	Avg. Abs. Diff.		2.03	2.17	2.03	3.10	2.43	1.67	1.97	2.10

*Note: Rankings of all models were adjusted to exclude Alabama which was on probation in 2002 and consequently not ranked by the Coaches Poll or the BCS.

Table 8: Summary of Average Absolute Differences for all BCS Models

Model	1998	1999	2000	2001	2002	Average
Proposed Model (Mease)	2.47	2.77	1.00	1.90	2.03	2.03
David Rothman		1.77	1.33	3.03		2.04
Herman Mathews		2.50	0.87	3.10		2.16
Wes Colley				2.30	2.03	2.17
Kenneth Massey		1.30	2.33	3.03	2.43	2.27
Anderson/Hester	2.33	2.83	1.60	2.70	2.17	2.33
Peter Wolfe				2.90	2.10	2.50
Richard Billingsley		3.10	2.33	1.57	3.10	2.53
Jeff Sagarin	3.07	1.57	2.47	3.63	1.97	2.54
New York Times	3.80	2.43	2.73		1.67	2.66
Dunkel		2.90	4.33			3.62

8. POSSIBLE MODIFICATIONS TO THE MODEL

Although the purpose of this paper is to present a model that performs well using only win/loss data, the model in fact can be easily adapted to incorporate other available information such as the location of the game, the date of the game, and even the margin of victory if so desired.

The location of each game played is almost always on the campus of one of the two competing teams. This team is called the "home team" and is thought to generally have an extra advantage due to crowd support and a number of other factors. This "home-field" advantage can be incorporated into the model by replacing $\theta_i - \theta_j$ by $\theta_i - \theta_j + \lambda$ if team i is the home team or by $\theta_i - \theta_j - \lambda$ if team j is the home team. The likelihood can then be maximized over this single home-field advantage parameter λ along with θ. The teams can be ranked based on their respective values of θ_i as before. The model considered by Harville (1977, 2003) uses an analogous single parameter to capture the home-field advantage. Alternatively, one can consider a separate home-field advantage parameter for each team as is done by Harville and Smith (1994) for college basketball teams. While this model is more flexible and may in fact be more realistic, it is not as useful for ranking teams since it will generally result in two different ranks for each team depending on whether the teams are ranked with their home-field advantage included or not.

In addition to the location of the games, it is sometimes suggested that the date of the games should be considered when ranking the teams. Specifically, some feel that games that occur later in the season should be weighed more heavily in the final rankings than games

that occur earlier in the season. One method for doing this in the proposed model would be to incorporate these weights into the values of n_{ij}. As a simple example, if the games in the second half of the season were thought to be twice as important as the games in the first half, one could double the n_{ij} values for all games occurring in the second half of the seasons. Note that in general if one scales the weights for the games such that they are integers, the model can still easily be fit using standard software as described in Section 6. The only necessary change would be to create replicate lines for some entries in the input matrix to reflect these weights, analogous to the method for dealing with tie games described earlier.

Similar to weighting games based on date, games could also be weighted based on some measure of difference in the final score. That is, the n_{ij} values could be used to account for margin of victory. Again, by scaling the weights to be integers, standard software can still be used to fit the model. Careful selection of the n_{ij} may produce a model which agrees even more strongly with the polls by giving more weight to large victories (as some suggest that human pollsters do) but not to the large extent that previous models have done.

A final possible modification to the model is to replace the normal CDF function Φ by some other function. The logit CDF function is one possible alternative that still allows for the model to be fit using standard software. Although the logit CDF and normal CDF have quite different tail behavior, the use of the logit CDF function did not affect the resulting rankings substantially for the football seasons considered.

9. SUMMARY

We have described a penalized maximum likelihood model for ranking college football teams independent of victory margins. By analyzing resulting rankings from actual college football seasons we have shown that the model on average agrees more strongly with the AP and Coaches Polls than many of the more complex models discussed in the statistics literature and used by the BCS. We have shown that the model can be implemented using binary response regression procedures available in most standard statistics software packages.

While no one model for ranking teams is necessarily "better" than any other model, the proposed model has many attractive features. Its strong agreement with the AP and Coaches polls suggests that it is consistent with popular/expert opinion, while at the same is free from the problems of personal bias and limited memory inherent in human polls. Secondly, the model ignores all factors other than wins and losses and is simple in form, suggesting a parsimonious solution to the problem of ranking college football teams. Finally, the model is motivated by statistical theory and can be fit using standard statistical software.

References

Bassett, G. W. (1997), "Robust Sports Ratings Based on Least Absolute Errors," *The American Statistician*, 51(2), 99-105.

Bradley, R. A., and Terry, M. E. (1952), "The Rank Analysis of Incomplete Block Diagrams. I. The Method of Paired Comparisons," *Biometrica*, 39, 324-345.

Green, P. (1998), "Penalized Likelihood," In *Encyclopaedia of Statistical Sciences*, Update Volume 2, New York: Wiley.

Harville, D. (2003), "The Selection or Seeding of College Basketball or Football Teams for Postseason Competition," *Journal of the American Statistical Association*, 98, 17-27.

Harville, D. (1977), "The Use of Linear-Model Methodology to Rate High School or College Football Teams," *Journal of the American Statistical Association*, 72, 278-289.

Harville, D. and Smith, M. (1994), "The Home-Court Advantage: How Large Is It, and Does It Vary From Team to Team?," *The American Statistician*, 48, 22-28.

Keener, J. P. (1993), "The Perron-Frobenius Theorem and the Rating of Football Teams," *SIAM Review*, 35(1), 80-93.

Thompson, M. (1975), "On Any Given Sunday: Fair Competitor Orderings with Maximum Likelihood Methods," *Journal of the American Statistical Association*, 70, 536-41.

Silverman, B. W. (1985) "Penalized Maximum Likelihood Estimation," *Encyclopaedia of Statistical Sciences*, 6, 664-667.

Stern, H. (1992), "Who's Number One? - Rating Football Teams," in *Proceedings of the Section on Statistics in Sports* 1992, pp. 1-6.