CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: http://www.tandfonline.com/loi/ucha20

Who's Number 1 in College Football? ... And How Might We Decide?

Hal S. Stern

Nebraska beat out Penn State for number 1 in the two major polls. Not so, says this author.

Who's Number 1 in College Football? . . . And How Might We Decide?

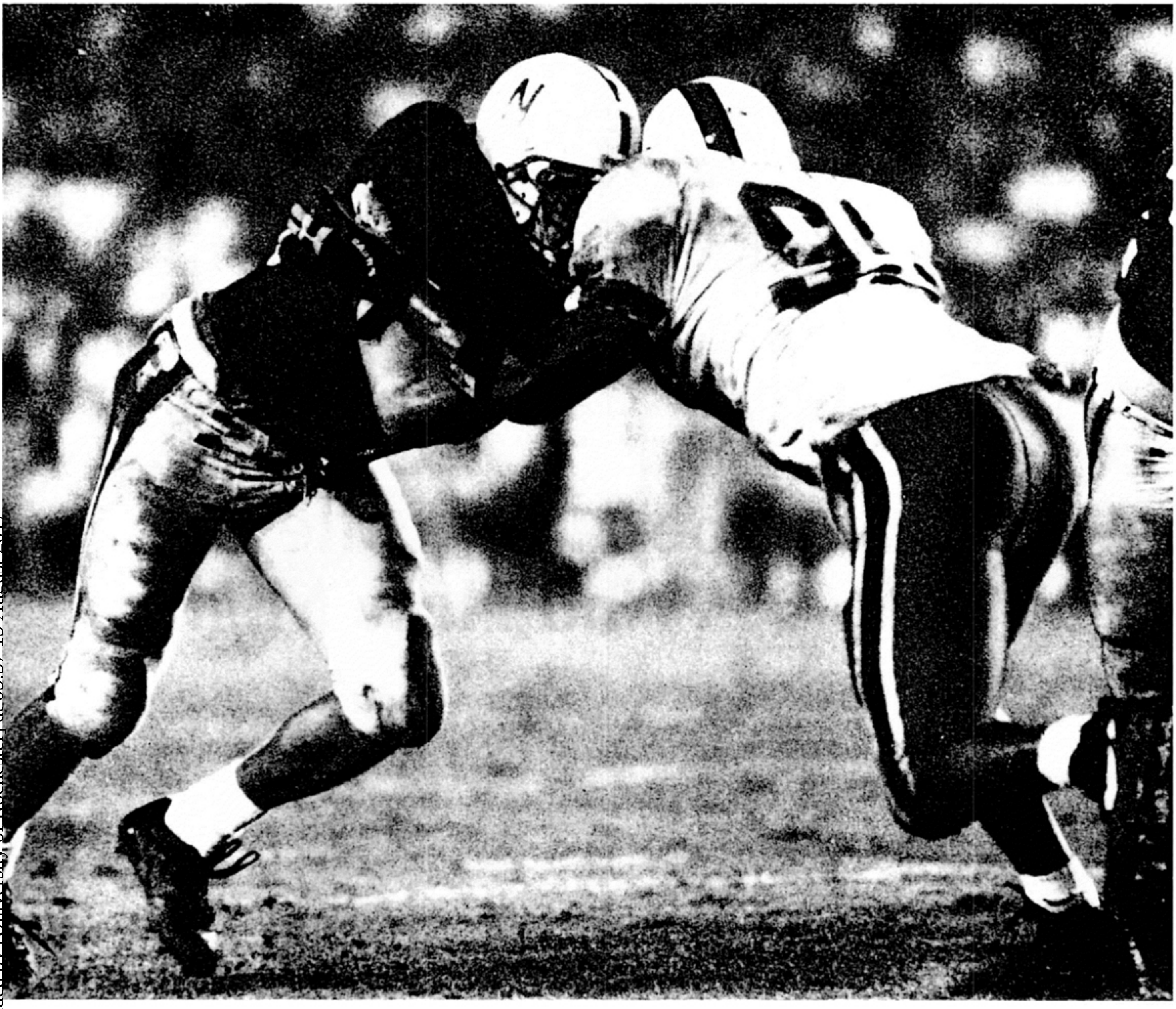
Hal S. Stern

College football is perhaps the only major team sport that does not decide a unique champion via competition each year. College football teams are rated during the season by two polls, one a survey of journalists (The Associated Press) and one a survey of coaches (*USA Today*/CNN). Last season, both groups preferred undefeated Nebraska to undefeated Penn State in their final polls, but there was substantial debate among experts and fans about whether this was the correct decision. In both 1990 and 1991 the two polls disagreed. Statistics is a field concerned with the collection, analysis, interpretation, and presentation of numerical data. Certainly, college football provides abundant numerical data that are amenable to statistical analysis. What can statistics say about choosing the number 1 team? The *New York Times* and Jeff Sagarin (published in *USA Today*) provide ratings based on statistical methods; these methods do not always agree with the polls.

In this article, I explore ratings based on the familiar statistical idea of least squares. Although questions concerning the rating of teams are of special interest in college football, they may also be of interest in other sports. For example, ratings play an important role in selecting the participants for college basketball's annual championship tournament.

Asking the Right Question

Part of the difficulty in applying statistical methods to a problem like rating teams is that there is no unanimity about what the rating should measure. Should the goal be to determine the best team in college football in the sense that it would be most likely to defeat the others or should the goal be to reward the team with the most impressive season-long performance? For example, in 1984 Brigham Young University (BYU) was undefeated but did not play a very difficult schedule. The two polls agreed that as the only undefeated team, BYU was the national champion. It is certainly possible, however, that once-defeated Washington (one defeat against a more difficult schedule) was the better team. Even if we agree that we would like to choose the best team, there can be problematic scenarios. Suppose a team loses its first two games when it is forced



1995 Orange Bowl, Nebraska vs. Miami, by Action Sports. Photographer M. Shaw.

to play without a top player and wins the remainder of its games easily. It is doubtful that such a team would be voted the best team, although at the end of the season it may be the best team.

We make a distinction between rating teams and ranking teams. Ratings are assumed to incorporate a continuous scale such that differences or ratios of team ratings describe the relative abilities of two teams. The term *ranking* is used here to refer to an ordering of the teams without any associated measure of difference. Thus, in a rating, the difference between the teams rated first and second may be larger or smaller than the difference between the ninth and tenth rated teams. The polls provide a ranking of the teams but no rating. It seems that there is sufficient information available to provide team ratings.

Least Squares Ratings

The statistical idea of least squares can be used to obtain college football ratings. Suppose that we take θ_i to be the unknown rating of team i, and let H be the advantage of a team playing at home (thought to be about 3 points). When team i plays team j at team i's home field, the expected outcome is $\theta_i - \theta_j + H$. The actual outcome will almost certainly differ from this expected outcome because there is considerable variability in football games. (It would not be a very interesting sport if the better team always won!) Suppose that we use the available data to estimate the ratings σ and the home-field advantage so that the actual outcomes are close to the expected outcomes. We choose to minimize the sum of the squared differences between the actual outcomes and the expected outcomes (see sidebar). There are a few technical problems remaining. First, because only differences in the ratings are important for this approach, we can add 10 or 100 or 1,000 to each team's rating without changing the picture. A second problem is that if we focus our attention on Division 1-A college football teams ("the big boys"), we must choose a method for dealing with teams from lower levels that occasionally play against 1-A teams. An easy approach that solves both of these problems is to let the rating of all lower level teams be set to 0 by definition. With this value fixed, the ratings of all the 1-A teams will be determined relative to this baseline. Of course this approach is not correct because there is considerable variability in the strength of teams from this lower level. Jeff Sagarin's ratings in *USA Today* include all Division 1-A and Division 1-AA teams, which means that only Division II and lower teams are grouped together and assigned the 0 rating.

As with least squares regression, least squares ratings are sensitive to outlying values. Such values may have an excessive impact on the ratings. There is a large difference between a team winning by 50 points and a team winning by 25 points in terms of the least squares calculation. Most sports fans, however, would think of 25- and 50-point margins as being indicative of similarly superior performances. It seems desirable to reduce the effect of wide margins of victory; both the *New York Times* computer rankings and Jeff Sagarin's ratings do so. In this article, if the margin of victory in a game is greater than 20 points, we replace the outcome by a "modified" outcome equal to 20 plus the square root of the additional margin of victory beyond 20 points. Thus, a 29-point outcome results in a 23-point modified outcome, and a 56-point outcome results in a 26-point modified outcome. The ratings computed from these adjusted outcomes are called "modified least squares ratings."

Table 1 gives the top 25 teams according to the modified least squares ratings along with the ratings of several other teams that were given high rankings by the news organizations' polls. In all, there were 107 Division 1-A teams during the 1994 college football season. The modified least squares ratings can be interpreted easily—the difference between the ratings of two teams is an estimate of the number of points by which one team would beat the other at a neutral site. The record of each team, along with the points scored and allowed, is also provided in the table. Nebraska, the number 1 team in both polls, is rated third best here. Penn State is number 1 by a wide margin. They would be expected to defeat Nebraska by about 5 points on a neutral field. The least squares ratings implicitly reward teams for playing dif-

Using Least Squares to Rate Teams

We briefly provide a few details about rating football teams using least squares. Suppose that games are indexed by k and Y_k is the outcome of the kth game. Y_k may be the difference between the points scored by the home team, labeled i_k, and those scored by the visiting team, labeled j_k, or some other outcome measure, possibly incorporating a modification to minimize the effect of blowouts or a bonus for winning the game. If there is no home team, then choose one arbitrarily and denote it as i_k. The least squares ratings are obtained by determining the ratings σ that minimize

$$\sum_k (Y_k - (H + \theta_{i_k} - \theta_{j_k}))^2$$

where H is a parameter measuring the advantage of playing at home and the rating for any non-Division 1-A team is set to 0. For games played at neutral sites, H is omitted. It turns out that estimates of the ratings and home-field advantage parameter can be obtained using any regression program as follows. The regression model assuming that there are n Division 1-A teams is

$$Y_K = HZ_{k0} + \sum_{m=1}^{n} \theta_m Z_{km} + e$$

where $Z_{km} = 1$ if team m is the home team in game k, $Z_{km} = -1$ if team m is the visiting team in game k, and $Z_{km} = 0$ if team m is not a participant in game k. Z_{k0} is 0 if the game was played at a neutral site and 1 otherwise. There is no intercept. If one of the teams in game k is not from Division 1-A, then only a single Z_{km} would be nonzero for that game.

Some insight into the ratings is obtained by examining the conditions required to minimize the sum of squared errors. Let us focus attention on a single team, team m, and assume that it plays n_m games. For any game s in which team m is involved, let opp_s represent team m's opponent, and let S_s be 1, 0, or −1 depending on whether that game was played at team m's home site, a neutral site, or the opponent's site, respectively. By differentiating the sum of squares criterion with respect to σ_m and setting the result equal to 0, we find that the least squares ratings satisfy

$$\theta_m = \frac{1}{n_m} \sum_s (Y_s - \hat{H}S_s) + \frac{1}{n_m} \sum_s \theta_{\text{opp}}$$

where the sums are over games involving team m. The first term is a measure of team m's average performance (adjusted for the site of the game), and the second term is the measure of schedule strength. For Penn State in 1994, the average performance was 18.5, and the average opponents strength was 19.2; for Nebraska, the average performance was 18.0, and the average opponents strength was 15.0.

Table 1—NCAA Football Ratings, 1994 Season

			Least squares ratings							
			Modified		Unmodified		Rank with bonus B		Polls	
Team	Record	Points off-def	Rating	Rank	Rating	Rank	B=50	B=9999	AP	CNN
Penn State	12-0-0	564-252	37.7	1	55.5	1	1	1	2	2
Florida	10-2-1	538-228	33.7	2	51.9	2	5	6	7	7
Nebraska	13-0-0	459-162	33.1	3	47.3	4	2	2	1	1
Florida St.	10-1-1	428-200	32.9	4	48.4	3	3	3	4	5
Colorado	11-1-0	439-235	30.2	5	43.4	5	4	4	3	3
Miami (Fl)	10-2-0	365-143	30.0	6	43.4	6	7	8	6	6
Michigan	8-4-0	330-268	28.8	7	40.7	10	9	9	12	12
Illinois	7-5-0	309-156	28.4	8	42.0	7	20	31	29	27
Ohio St.	9-4-0	336-211	27.5	9	39.0	12	12	14	14	9
Auburn	9-1-1	359-199	26.9	10	38.3	13	10	10	9	*[a]
Texas A&M	10-0-1	319-147	26.7	11	36.0	17	8	7	8	*
Tennessee	8-4-0	363-208	25.7	12	40.8	9	14	18	22	18
Alabama	12-1-0	305-190	25.6	13	36.1	16	6	5	5	4
Southern Cal.	8-3-1	356-243	25.4	14	41.0	8	11	11	13	15
Utah	10-2-0	426-210	25.1	15	39.1	11	13	13	10	8
Virginia	9-3-0	370-188	23.9	16	35.3	19	16	16	15	13
Wash. St.	8-4-0	192-136	23.8	17	35.3	20	17	17	21	19
Notre Dame	6-5-1	342-280	23.7	18	35.6	18	23	24	nr	nr
Boston College	7-4-1	271-169	23.6	19	33.8	22	24	27	23	22
Wisconsin	7-4-1	357-238	23.4	20	37.5	14	25	28	28	26
Oregon	9-4-0	349-250	23.4	21	36.2	15	18	19	11	11
Kansas St.	9-3-0	312-168	23.1	22	34.2	21	19	20	19	16
Washington	7-4-0	295-233	23.0	23	33.7	23	21	22	30	*
Arizona	8-4-0	278-190	21.1	24	33.0	24	27	23	20	20
Michigan St.	5-6-0	280-267	20.9	25	32.4	25	35	38	nr[b]	nr
Colorado St.	10-2-0	400-269	20.9	26	32.3	26	15	12	16	14
Virginia Tech	8-4-0	327-247	19.4	28	30.6	29	29	25	26	24
Texas	8-4-0	366-291	19.1	29	29.0	33	30	29	25	23
Miss. St.	8-4-0	373-262	18.9	31	31.9	27	32	32	24	25
BYU	10-3-0	385-300	18.6	32	27.7	37	26	21	18	10
N. Carolina	8-4-0	374-267	17.8	36	30.1	30	31	30	27	21
N.C. State	9-3-0	305-275	16.0	44	24.7	48	22	15	17	17

[a]An asterisk indicates ineligibility for this poll.

[b]nr indicates did not receive any top 25 votes from any voter.

ficult opponents. The ratings represent a combination of each team's performance and its opponent's ratings (see sidebar). The least squares ratings are apparently quite impressed with Penn State's athletic conference, the Big Ten (consisting of 11 teams). Penn State, Michigan, Illinois, Ohio State, Wisconsin, and Michigan State all appear in the least squares top 25. Some Division 1-A teams are assigned ratings lower than 0; for example, Akron's rating is –16.2. This means that they would be expected to lose to teams from lower divisions, and, in fact, Akron did lose one such game during the season. The estimated home-field advantage is 1.9 points based on this one season. Estimates of the home-field advantage in previous seasons are somewhat higher, suggesting that the 1994 season was a bit anomalous in this regard. It is also possible to estimate the standard deviation of the differences between the observed modified outcomes and the expected outcomes. This standard deviation is 11 points, suggesting that, even with ratings treated as known quantities, there is considerable uncertainty in football game outcomes. In actuality, there is probably even more uncertainty because these ratings are not estimated perfectly.

Table 1 also includes the ratings that would be obtained if the actual outcomes were used in place of the modified outcomes. The changes are mainly minor (e.g., Nebraska falls behind Florida State) but occasionally major (e.g., Southern California jumps from 14th to 8th). Southern California won twice during the year by big margins (61 and 41 points) against respectable opponents.

The final two columns of Table 1 give the final sportswriters' and coaches' polls. The polls do not assign ranks to teams beyond 25, but for this table, other teams receiving votes have been assigned ranks above 25. There are some major differences, none more obvious than the position of Illinois. Illinois finished just outside the top 25 in the two polls (27th and 29th) but 8th in the modified least squares ratings. Other teams favored by least squares include Wisconsin, Michigan State, and Notre Dame, each of which was excluded from the top 25 by the two polls. In each of these cases, the explanation seems to be a discrepancy between the poll participants' opinions of the Big Ten and the computer's impression of the Big Ten. Notre Dame is not a member of the Big Ten but did play four Big Ten opponents. On the other hand, the polls include Colorado State (#26 based on the modified least squares), Virginia Tech (#28), Texas (#29), Mississippi State (#31), Brigham Young (#32), North Carolina (#36), and North Carolina State (#44). Most of these omissions would be considered minor because there is only a 4-point difference between #24 and #41 in the least squares ratings. Still, North Carolina State was rated 17th in both polls and considerably lower by least squares, and Brigham Young was rated 10th by the coaches' poll.

Some Refinements

The ratings in Table 1 represent one answer to the question of who is number 1. One can easily imagine the objections from around the country. Could Michigan State—with five wins and six losses—really be one of the top 25 teams in the nation? Could Alabama—with only a single loss by 1 point to second-rated Florida—actually be the 13th best team? Certainly computers do not make mistakes, but computers are only as good as the rules they are given. We, therefore, point out some of the issues that seem important and possible modifications to the least squares ratings that might be used to address them.

Our approach to reducing the effect of blowouts is somewhat arbitrary. Any number of rules might be tried or a different criterion than least squares might be applied. Ratings chosen to minimize the absolute value of the differences between actual and expected outcomes are one possibility. Alternatively, rating methods that use only win/loss outcomes and are not affected by wide margins of victory might be used (such as paired comparisons methods).

Even assuming that an acceptable method for dealing with blowouts is agreed upon, some problems remain. How does it come to pass that Alabama (12 wins, 1 loss) is rated 13th, far below Illinois (7 wins, 5 losses)? This occurs in part because the difference between a 1-point victory and a 1-point defeat is relatively small in terms of the least squares criterion. For most football observers, the difference between an outcome of 1 and an outcome of −1, indicating a change from victory to defeat, is quite significant. For the least squares criterion, the difference between 1 and −1 is no larger or smaller than the difference between 9 and 11 or 19 and 21. One approach to solving this problem is to redefine the outcome measure yet again. Perhaps a bonus of B points should be added for winning a game so that the outcome used in creating ratings would be the actual outcome (or modified outcome) plus the bonus. As the bonus B becomes larger, the score of the game is less important. Rankings (but not ratings) for least squares ratings with $B = 50$ and $B = 9999$ are included in Table 1. Note that Alabama moves up as the size of the bonus is increased, because their small margins of victory become less important and only their large number of wins is important. Unfortunately, if B is greater than 0, the resulting ratings are no longer interpretable as differences in points, and that is why the ratings are not included in Table 1. The rankings corresponding to $B = 9999$ match the rankings obtained by the two polls more closely than those with no bonus; schedule strength is still sufficient to rank Notre Dame ahead of many other teams with superior records.

All games are treated equally in the least squares ratings of Table 1. It is easy to introduce weights so that games played more recently are given greater weight than games played in the past. Bowl games played after the regular season might be given two or three times as much weight as regular season games. To add weights to the analysis, the sum of squared errors is replaced by a weighted sum of squared errors. Moreover, each team's rating is assumed to be a constant for the entire year. An alternative would be to allow the teams' ratings to evolve during the year.

Table 2—Comparing Rating Systems on Professional Football Data

Rating method	Percent correct prediction	Root mean squared prediction error	Mean absolute prediction error	Mean absolute error versus point spread
Least squares	63.6	14.52	11.45	3.63
Least squares Modified to reduce blowouts	63.5	14.35	11.29	2.94
Least squares Modified to reduce blowouts 10-Point bonus for winning	62.9	—	—	—
Least squares Modified to reduce blowouts 50-Point bonus for winning	61.7	—	—	—
Least squares Modified to reduce blowouts More weight to recent games	63.6	14.43	11.40	3.13
Least squares Modified to reduce blowouts 10-Point bonus for winning More weight to recent games	62.5	—	—	—
Bayesian approach Uses info from >1 season	64.9	14.11	11.02	2.35
Point spread	66.3	13.94	10.88	0.00

An Evaluation of Rating Methods

The previous section suggests that there are many possible rating systems obtained by making different choices about the various features of the system—using actual or modified outcomes, incorporating a bonus for winning or not, weighting games evenly or unevenly, using least squares or some other criterion, and so forth. The number of permutations is large, and even then, the list would exclude entirely distinct methodologies—paired comparisons methods that ignore the scores and evaluate teams based only on wins and losses, or perhaps a Bayesian approach that incorporates information from previous seasons. Given this large pool of possible rating approaches, we now consider how rating systems might be evaluated.

At this point, a clear definition of the goals and aims of the ratings is required. If ratings are to be used to choose teams for participation in a championship playoff, then it might be inappropriate to use information from prior seasons. If prediction of future games is the ultimate goal, then the information from prior seasons may be extremely helpful. Professional sports typically reward teams that have performed sufficiently well over the entire season with the opportunity to compete for the championship. This suggests ratings that give all games equal weight. For prediction, however, it is of greater interest to estimate the team that is performing best at a particular time.

We suggest evaluating rating systems based on their predictive ability. After all, even those not inclined to bet would tend to interpret the ratings as indicating which team should be expected to win a game. At this point, we take a brief detour from college football to professional football. College football includes many teams (Jeff Sagarin rates more than 300 teams in *USA Today*) so that a large dataset covering several seasons is difficult to acquire. The National Football League consists of 28 teams playing 16 games each season so that each season is only 224 games in total. To compare rating systems, the ratings based on the data prior to week W (for W in the second half of the professional season) are used to predict the games that are played during week W. We restrict attention to the second half of the season, so that enough data have accumulated to obtain meaningful ratings. Table 2 provides the results of a comparison over the second half of nine seasons (1981–1991 minus two seasons affected by labor problems), totaling 1,027 games (in some recent years, the games were spread unevenly over 17 or 18 weeks). Rating systems are compared in terms of the percentage of games for which the outcome was correctly predicted and, where relevant, the standard deviation and median absolute value of the prediction errors.

There are many possible rating systems based on the least squares ideas of the previous section and many other possible systems beyond that. We restrict attention to the six possibilities included in Table 2. Each system is characterized by whether or not scores are truncated, the size of the bonus for winning a game, and whether later games were given additional weight (in this case, each week's games from the first two-thirds of the elapsed season received weight 1, whereas games from the most recent one-third received weight 2, except for the last week's games, which received weight 3). The basic least squares ratings correctly predicted the winner in 63.6% of the 1,024 games that did not end in a tie. Adding a bonus for winning or increasing the weight of more recent games seems to decrease or not affect predictive ability. Only a few bonus values and only a single choice of weights are considered in Table 2, so the results are not conclusive, but the results do suggest that the basic least squares ratings are quite effective in predicting game outcomes. The final line of Table 2 is included as a benchmark. The point spread of each professional game is set in Las Vegas. It is not a predicted outcome, rather it is a number set to entice equal numbers of bettors on both sides of the point spread. It turns out that it is a very effective predictor even though this is not its principal goal. The team favored by Las Vegas wins 66.3% of the time. This may seem surprisingly low, but it is indicative of the variability in sports.

The line titled Bayesian approach in Table 2 refers to an analysis that uses information from preceding seasons to rate games in the current season. Essentially, a separate least squares rating is applied to each season, except that the changes in ratings from season to season are assumed to follow a normal distribution with mean 0 and standard deviation 3. This means that even when game results early in the season suggest a large change in ability for one team, the team's rating is not modified by a large amount until more time elapses. The approach using information from prior seasons is better than the single-season methods. It comes quite close to the performance of the Las Vegas point spread. The experts in Las Vegas are able to account for previous seasons as well as other information (injuries, weather) that are omitted from the statistical analyses.

Comparing the proportion of correct predictions is useful, but it does not address the extra information contained in a rating (as opposed to a ranking). Due to the different nature of the various rating approachs, alternative measures are difficult to construct. We consider several possibilities. The least squares ratings and the Bayesian ratings each estimate directly the difference in ability among the teams (except for those least squares ratings which award extra points for winning the game). The second column of Table 2 gives the square root of the mean squared error of the game predictions. Note that such measures are not relevant when a bonus is included, because the ratings cannot be interpreted as predicted outcomes in that case. The mean absolute error is also reported. Once again, the rating systems perform almost as well as the point spread. Several authors have found the standard deviation of the difference between professional football game outcomes (the difference between the favored team and the underdog) and a good predictor (the least squares ratings or the point spread) to be about two touchdowns (14 points). Table 2 shows that this is indeed the case. The Bayesian ratings appear to be more accurate than the single-season least squares ratings. As a final point, we can compare the least squares predictions directly to the Las Vegas point spread. Over 1,000 games, the mean absolute difference between the least squares estimates and the point spread is approximately 3 points. The Bayesian model is even more accurate; the mean absolute difference is only 2.35 points.

Although this comparison is performed on professional games, the conclusions seem plausible in the context of college football games as well. For predictive purposes, it appears that the least squares ratings, with no bonus for winning, are the most effective. There does not seem to be any reward for giving some games additional weight. There is some improvement obtained by eliminating the effect of blowouts so that we would recommend some type of modified outcome be used.

Conclusions

Rating football teams provokes much debate because an appropriate rating methodology is in the eye of the beholder. The results here indicate that least squares ratings provide useful information about the relative strength of different teams. Returning to college football, it seems clear that Penn State's 1994 performance was more impressive to the least squares methodology than Nebraska's performance. Of course, it is possible that there is some other statistical rating system that would favor Nebraska (the *New York Times* computer and Jeff Sagarin concur with the result of this article). The claim here is not that a statistical method can objectively decide which team is the best team (after all, the choice of method is subjective) but that a statistical method is more open about the criteria used and more consistent in its application. For predicting games, least squares ratings based on games scores (modified to minimize the effect of blowouts) with no bonus for winning seems best. Some might argue that prediction is not the natural goal when determining a national champion. As Table 1 shows, the poll voters' notion of an appropriate ranking seems to be consistent with assigning a fairly large premium for winning games. Settling the championship on the field would yield a clearly defined champion (although not necessarily a clear best team), but that does not appear to be a realistic alternative in the near future.

Additional Reading

Harville, D. (1980), "Predictions for National Football League Games Via Linear-Model Methodology," *Journal of the American Statistical Association*, 75, 516–524.

Keener, J. P. (1993), "The Perron–Frobenious Theorem and the Ranking of Football Teams," *SIAM Review*, 35, 80–93.

Leake, R. J. (1976), "A Method for Ranking Teams with an Application to College Football," *Management Science in Sports*, eds. R.E. Machol, S.P. Ladany, and D.G. Morrison, Amsterdam: North-Holland, pp. 27–46.

Stefani, R. T. (1980), "Improved Least Squares Football, Basketball, and Soccer Predictions," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-10, 116–123.

Stern, H. (1991), "On the Probability of Winning a Football Game," *The American Statistician*, 45, 179–183.