An Article Submitted to

Journal of Quantitative Analysis in Sports

Manuscript 1115

A New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings

Brady T. West* Madhur Lamsal†

*University of Michigan-Ann Arbor, bwest@umich.edu

†University of Michigan-Flint, mlamsal@umflint.edu

Copyright ©2008 The Berkeley Electronic Press. All rights reserved.

A New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings*

Brady T. West and Madhur Lamsal

Abstract

This paper begins with a thorough review of previous quantitative literature dedicated to the development of ratings for college and professional football teams, and also considers various methods that have been proposed for predicting the outcomes of future football games. Building on this literature, the paper then presents a straightforward application of linear modeling in the development of a predictive model for the outcomes of college football bowl games, and identifies important team-level predictors of actual bowl outcomes in 2007-2008 using real Football Bowl Subdivision (FBS) data from the recently completed 2004-2006 college football seasons. Given that Bowl Championship Series (BCS) ratings are still being used to determine the teams most eligible to play for a national championship and a playoff system for determining a national champion is not yet a reality, the predictive model is then applied in a novel method for the calculation of ratings for selected teams, based on a round-robin playoff scenario. The paper also considers additional possible applications of the proposed methods, and concludes with current limitations and directions for future work in this area.

KEYWORDS: College Football Ratings, Prediction, Linear Modeling, College Football Bowls, NCAA Football

---

*The authors wish to acknowledge the Undergraduate Research Opportunity Program (UROP) at the University of Michigan-Flint for providing funding to assist with this research.

1. INTRODUCTION

The college football bowl season is an extremely popular time for fans of American college football, and an extremely important time for the colleges involved, both financially and in terms of recruiting athletes for their football programs. Casual college football fans enter into recreational bowl pools for fun, putting their knowledge of the game on the line, and the more serious fans exchange a great deal of money via various betting enterprises. College football teams competing in a given bowl game split a purse, the size of which depends on the bowl; because more prestigious bowls have higher payouts for the two teams, securing an invitation to a bowl has serious financial implications for the schools involved. The directors of the bowl games are charged with selecting and inviting bowl-eligible (i.e., achieving six-wins against NCAA Division I-A opponents) teams that will play in competitive games, attracting fans and keeping fans tuned in across different media formats (making bowl games important for advertising).

Sixty-four bowl-eligible teams were invited to participate in 32 bowl games in the 2007-2008 bowl season. Five of the bowl games were a part of the Bowl Championship Series (BCS), featuring 10 of the strongest college football teams as determined by a variety of ratings and polls. One of the BCS games featured the two teams with the highest BCS ratings, with the winner (Louisiana State University) deemed the national champion of the Football Bowl Subdivision (FBS, formerly Division I-A) of college football. Each year, the directors of the BCS bowl games are responsible for making use of a great deal of quantitative information in an effort to match up the strongest teams in the nation, and identifying the two teams most eligible to compete for the national championship.

Given the importance of the college football bowl season, a number of statisticians and quantitative analysts have explored the possibility that statistical methods can be used to rate college football teams and predict the outcomes of future games (allowing for the possibility of selecting evenly matched teams and identifying the two “best” teams under the current BCS system). These methods could potentially have an impact on the development of rating systems for the teams, in addition to the determination of the betting line (or spread) for college football games; the need to have reliable and accurate team ratings under the current BCS system indeed provides statisticians with a variety of interesting challenges (Stern 2004). Unfortunately, the BCS places restrictions on the inputs that can be used to develop team ratings and suffers from an ill-defined objective for the development of team ratings, leading Stern (2006) to suggest a “quantitative boycott” of the BCS.

This paper presents a thorough examination of previous quantitative efforts exploring these problems, and presents a new application of statistical

modeling that can be used to directly predict the outcomes of college football bowl games, given team-level information that is available prior to the onset of bowl season. Further, given the ongoing controversy surrounding the BCS ratings that are calculated each year to determine the two teams most eligible to compete for the national championship (especially in 2007), this paper considers an application of the proposed prediction model in the development of ratings for college football teams based on a round-robin playoff scenario. Results based on the recently completed 2007-2008 FBS season are presented and discussed.

2. LITERATURE REVIEW

A large body of recent quantitative literature has been dedicated to the development of ratings and rankings for American football teams, considering both the professional and college games. The rating methods proposed in these papers and articles can be evaluated by their ability to predict the outcomes of future games, and several papers have in fact evaluated proposed rating systems in that manner. Many papers have considered methods based on various forms of least squares estimation for the development of team ratings, where ratings are formulated as parameters in linear models predicting game outcomes. These papers include work by Stefani (1980), who incorporated home field advantage into least squares ratings; Harlow (1984), who developed a computer program for calculating ratings; Stefani (1987), who discussed additional applications of least squares in the prediction of future outcomes; Stern (1995), who used ratings based on past performance to predict the outcomes of future NFL games; Bassett (1997), who proposed the use of least absolute errors rather than least squares estimation to reduce the influence of outliers; and Harville (2003), who proposed a modified least squares approach incorporating home field advantage and removing the influence of margin of victory on ratings (per BCS requirements), identified seven key attributes of any ranking system, and showed that the ratings based on the modified least squares approach had better predictive accuracy for future games than the Las Vegas betting line.

Other recent papers have proposed alternative rating methods that provide alternatives to applications of least squares estimation. Mease (2003) introduced a model based on a penalized maximum likelihood approach that incorporated win/loss information only, and produced rankings for college football teams which were shown to have a higher correlation with expert rankings than BCS models. Fainmesser et al. (2003) used a parametric model based on wins and losses and the relative importance of home versus away games to develop rankings based on regular season performance, and estimated the parameters of the model using historical data and bowl game outcomes from 1999-2003. These rankings were then evaluated by assessing their predictive ability for bowl game

outcomes in 2004, and shown to do a better job of predicting bowl outcomes than BCS rankings. Annis and Craig (2005) showed the effectiveness of incorporating additional information into paired comparison models that can be used to develop rankings for teams. Park and Newman (2005) used a somewhat simple network analysis based on linear algebra and "common sense" to develop rankings for teams that gave more weight to wins over stronger teams, and produced rankings very similar to the final BCS rankings from 2004; further, they suggested that additional variables should be used to refine their rankings. Martinich (2003) evaluated the performance of 10 ranking schemes used by the BCS in 1999 and 2000 in selecting teams, and found that all were equally accurate, in terms of correctly predicting the outcomes of games in the immediate future (i.e., one week after the ratings are released).

Several recent papers have been dedicated to somewhat more direct methods of predicting the outcomes of future games, utilizing past information to predict future outcomes. Harville (1980) included results from previous NFL seasons and information other than the point spread to develop ratings for teams in future seasons and predict the outcomes of future games using linear mixed models. Trono (1988) proposed a probabilistic model based on the simulated outcomes of individual plays (where plays were based on a deterministic play-calling strategy) to predict the outcomes of games, where the probabilities of certain events occurring were based on past performance; using this model, Trono correctly predicted the outcomes of 58.7% of bowl games over eight seasons. Some researchers (Ong and Flitman 1997; Pardee 1999) have considered applications of neural networks in an effort to predict future outcomes of football games, building networks based on past information to predict future outcomes, and demonstrated improved prediction accuracy (as high as 76.2% of future games, as reported by Pardee).

Steinmetz (2000) obtained a United States patent for a statistical model (similar to a regression tree) that can be used for the prediction of future outcomes based on quantitative measures only, using historical parameters related to past performance, experience of team personnel, time of the season at which a game occurs, and the Las Vegas betting line. In a Masters Thesis completed at the University of Utah, Reid (2003) introduced a prediction approach for future games based on least squares estimation. In applications of Reid's method, a team's score in a given game against an opposing team could be predicted based on an estimated model predicting individual team scores as a function of home field advantage, conference status, voting points based on the Associated Press and ESPN voting polls, and indicators for team offenses and team defenses (i.e., a team's score in a given game is a function of their estimated offensive contribution, and their opponent's estimated defensive contribution, in addition to the other controls). Reid showed that the "best" model in terms of prediction

accuracy used all predictive factors discussed. Boulier and Stekler (2003) compared power scores published by the *New York Times* with betting market scores and opinions of the sports editor in terms of their ability to correctly predict the outcomes of NFL games from 1994-2000, based on probit models, and found that the betting market was the best predictor. Harville's (2003) method produced predictions with better accuracy than the betting market. Finally, Fair and Oster (2007) examined nine college football ranking systems, including several used by the BCS, and considered them in addition to an indicator of home field advantage and betting spreads as predictors in regression models predicting the outcomes (point spreads) of 1,582 games from 1998 to 2001. The optimal model including betting spread information explained 44.5% of the variance in point spreads, and predicted outcomes in 74.7% of games correctly. Fair and Oster argued that there was no information in the rankings not in the Las Vegas spread, but that there was information in the Las Vegas spread not common to the rankings.

This paper builds on this fairly extensive literature by examining whether a direct linear modeling approach capable of incorporating a variety of team-level inputs reflecting past performance (Stern 2004) can be used to accurately predict the actual outcomes of future college bowl games. The proposed approach uses a simple regression-based method similar to that proposed by West (2006) for predicting future success in the NCAA basketball tournament. The fundamental idea behind the method is to determine the most important team-level predictors of actual bowl game outcomes, given the pairs of teams selected to play in the bowl games, and develop a prediction model that can be used in practice to predict future outcomes given a variety of team-level information. This paper also incorporates suggestions from Morris (1978) that were discussed further by Stern (2004), using the aforementioned predictive linear model to calculate ratings for the teams that are based on predictions of all possible outcomes when a given team faces all other bowl-eligible teams in a round-robin playoff scenario.

3. DATA COLLECTION / MEASURES

Building and diagnosing the linear model proposed in this paper involved the collection of a variety of team-level variables for the 240 Football Bowl Subdivision (FBS, or Division I-A) teams selected to play in the 120 bowl games in 2004, 2005, 2006, and 2007. These variables were all publicly available at the conclusion of the each regular season (including conference championships), and were collected *prior* to the onset of the bowl games. Data were collected using free online resources (Yahoo! Sports, ESPN.com, Jeff Sagarin's *USA Today* computer ratings, the NCAA, and cfbstats.com; see References), prior to the onset of the bowl games. Specific team-level variables collected for each of the 120 teams selected to participate in the FBS bowl games included the following:

- Number of games played
- Scoring margin (average points scored per game minus average points yielded per game, despite the BCS decree that margin of victory *not* be used for computer ratings, just to gauge the importance of scoring margin as a predictor)
- Offensive yardage accumulated per game
- Offensive first downs per game
- Defensive yardage yielded per game
- Defensive first downs yielded per game
- Defensive touchdowns yielded per game
- Turnover margin (take-aways minus give-aways)
- Strength of schedule (as computed by Jeff Sagarin for *USA Today*)

Values on these variables were standardized within each year across the teams selected for bowl competition (by subtracting the mean for a given variable in a given year from each team's value, and dividing by the standard deviation for that year), so that all measures would be on the same scale. The standardized variables therefore indicate how much better or worse than the average bowl team a given team is (in standard deviations) on each team-level measure, for a given year.

For each of the 120 bowl games, differences in standardized values on the team-level variables were computed, measuring the difference between the arbitrarily selected "home team" (because all games are played at neutral sites) and the "away team" (H – A). These team differences in standardized values from 2004, 2005, and 2006 were then considered as potential predictors of the actual bowl game outcomes in these years in multiple linear regression models. The difference in the score between the "home team" and the "away team" was recorded for each of the 88 games in the first three years (H – A), and these values defined the continuous outcome variable in the regression models. Data from the 2007 bowl season (n = 32 games), including the outcomes of the bowl games, were only used to examine the predictive ability of the historical model in this paper; future applications of this method would use the 2007 data when fitting a regression model to be used for prediction in future years.

4. MODEL FITTING

Prior to fitting the multiple regression models, pair-wise Pearson correlation coefficients, scatterplots, and Lowess smoothers were used to determine whether any of the team difference predictors had unusually high correlations, and to examine whether the simple bivariate relationships of any of the predictors measured with the actual bowl outcomes were non-linear in nature. Most

relationships appeared to be linear in nature (which was later confirmed using partial regression plots), and two high pair-wise correlations were observed between the team difference predictors (not unexpectedly): difference in offensive first downs per game and difference in offensive yardage per game (r = 0.871), and difference in defensive yards per game and difference in defensive first downs yielded per game (r = 0.787). As a result, only the team differences in (standardized) offensive yards accumulated per game and defensive yards yielded per game were retained as potential predictor variables of the actual bowl outcomes, to minimize potential problems in the regression model due to multi-collinearity.

The remaining six predictor variables measuring "home minus away" differences in standardized values between the teams were then considered in a multiple linear regression model for the actual score difference outcomes. Higher-order interactions between the predictors were not considered in this application due to the small sample size (n = 88), although future applications using more seasons of data to develop a predictive model might consider such interactions (see Conclusions). An intercept term was omitted from the regression models to ensure that the arbitrarily selected home team would not be given an advantage or disadvantage when model-based predictions were calculated (all bowl games are played at neutral sites, so no "home advantage" is expected).

Standard diagnostics for linear models were thoroughly examined at each step of the model fitting process, to assess statistical assumptions of normality in the residual errors, constant variance for the errors, linearity of the relationships, and influence of unusual cases. The SPSS statistical software (Version 16.0.1) was used for all analyses, and ordinary least squares (OLS) estimation was used to fit all models.

5. MODELING RESULTS

The fit of the initial "full" model considering all six predictors did not present evidence of any non-linear relationships of the predictors with the actual outcomes, based on standardized residual diagnostics and partial regression plots. Assumptions of normality and constant variance for the residual errors were justified, but two bowl games appeared to have an unusually strong influence on the fit of the model based on an examination of Cook distances: the Las Vegas bowl between Brigham Young University (BYU) and Oregon in 2006 (won 38-8 by BYU), and the Music City bowl between Clemson and Kentucky in 2006 (won 28-20 by Kentucky, despite the fact that Kentucky had a much smaller scoring margin and significantly worse defensive statistics than Clemson). The R-squared value of the initial model was 0.164, and multi-collinearity was not an issue (largest condition index = 3.968). The predictors measuring difference in scoring

margin and difference in turnover margin were found to have significant ($p < 0.05$) positive and negative relationships with the actual outcomes, respectively.

The model was re-fitted by excluding these two unusual bowl games ($n = 86$), and there were notable changes in the fit. All assumptions underlying the model were once again justified, but the R-squared value was now 0.216 and the positive effect of the SOS difference predictor was now significant at the 0.10 level ($p = 0.058$). Table 1 presents unstandardized estimates of the regression parameters associated with these predictors, along with standardized estimates (which assume that the actual bowl outcomes were standardized as well). These standardized coefficients reflect the relative impacts of changes in the team-level difference predictors on the expected bowl outcomes.

Table 1. Estimated regression parameters in the final predictive model, based on data collected from 2004 to 2006.

Difference Predictor	Unstandardized Estimate	Standard Error	Standardized Estimate	t-value (80 d.f.)	p-value
Scoring Margin	9.593	2.982	0.577	3.217	0.002
Offensive Yds./Game	-2.995	2.171	-0.245	-1.379	0.172
Defensive Yds./Game	-2.669	2.262	-0.195	-1.180	0.242
SOS	3.229	1.679	0.213	1.923	0.058
Defensive TDs/Game	3.881	2.368	0.276	1.639	0.105
Turnover Margin	-2.923	1.186	-0.273	-2.463	0.016

$n = 86$, $R^2 = 0.216$; Overall test of all parameters against 0: $F(6,80) = 3.678$, $p = 0.003$.

Examining the parameter estimates for the predictors in Table 1, the differences in standardized values for scoring margin, strength of schedule, defensive touchdowns per game and turnover margin were the strongest predictors of the actual bowl outcomes. With every additional standard deviation difference in scoring margin in favor of the "home team," the expected outcome was roughly 9.6 points higher in favor of the home team. Further, with every additional standard deviation difference in strength of schedule in favor of the "home team," the expected outcome was roughly 3.2 points higher in favor of the home team. Interestingly, the difference in standardized values of turnover margin also had a significant *negative* relationship with the actual outcomes, suggesting that larger differences in favor of the home team would result in larger margins of

victory for the *away* team (and vice versa for the away teams, given the linear relationship). Hypothesis tests for the regression parameters associated with the team-level differences in offensive yards per game, defensive yards per game, and defensive touchdowns yielded per game suggested that these variables were not as important as the other difference predictors.

The R-squared value for the final model was 0.216, suggesting that about 22% of the variance in the 86 bowl outcomes from 2004 to 2006 was explained by these six team-level difference predictors. In practice, these estimated parameters would be used to write a prediction equation for the outcomes of future bowl games:

$$\hat{out}_{H-A} = 9.593 \times MARG_{H-A} - 2.995 \times OYDS_{H-A} - 2.669 \times DYDS_{H-A} + 3.229 \times SOS_{H-A}$$
$$+3.881 \times DTD_{H-A} - 2.923 \times TOM_{H-A}$$

To clarify the notation used in this equation, $TOM_{H\text{-}A}$ represents the difference between the home team and the away team in standardized values of turnover margin. Consider a hypothetical example, where the home team was one standard deviation higher than the away team in terms of scoring margin, one standard deviation lower than the away team in offensive yards per game, one standard deviation lower in terms of defensive yards per game, zero standard deviations higher in terms of SOS (equal schedule strength), one standard deviation lower in defensive touchdowns yielded per game, and one standard deviation higher in turnover margin. The predicted outcome of this bowl game would be calculated as follows: H – A = 9.593 + 2.995 + 2.669 - 3.881 - 2.923 = 8.453, suggesting that the home team would be expected to win the game by approximately eight points.

To measure the amount of uncertainty in the predictions based on the estimated regression parameters in Table 1, a 95% confidence interval for mean predictions can be calculated as follows:

$$\hat{out}_{H-A} \pm t_{n-p}^{(0.025)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

In this equation, n refers to the number of games used to estimate the model, p refers to the number of parameters in the final model, the value of 0.025 refers to the 0.025 critical value of the t distribution with $n - p$ degrees of freedom, and x_0 refers to a vector of values on the team-level difference predictors retained in the final model. The matrix X refers to the model matrix containing values on the predictor variables for all n cases. These confidence intervals can be used to reflect uncertainty in the expected game outcomes, and determine whether the prediction (or expectation) would suggest a clear-cut winner (i.e., the confidence interval does not include a value of 0 for the predicted difference in scores).

The "historical" predictive model estimated using data from 2004 to 2006 was applied using the data collected before and after the 2007 bowl season to examine the predictive ability of the model. Table 2 compares predictions of the outcomes in the 32 bowl games in 2007-2008 based on the historical 2004-2006 model (with 95% confidence intervals for the predicted outcomes) to the actual outcomes.

Table 2. Predicted bowl outcomes in 2007-2008 according to the historical (2004-2006) predictive model vs. actual outcomes.

Bowl	Home (H)	Away (A)	Predicted Outcome, H-A (95% CI)	Actual Outcome, H-A
Poinsettia	Navy	Utah	-3.79 (-15.39, 7.19)	-3
New Orleans	Memphis	Florida Atlantic	-2.66 (-9.71, 4.38)	-17
Papajohns.com	Cincinnati	Southern Miss	7.96 (-3.69, 19.60)	10
New Mexico	New Mexico	Nevada	4.04 (-3.75, 11.83)	23
Las Vegas	BYU	UCLA	4.07 (-3.96, 12.11)	1
Hawaii	East Carolina	Boise State	-28.61 (-41.86,-15.34)	3
Motor City	Central Michigan	Purdue	-11.38 (-19.32,-3.43)	-3
Holiday	Texas	Arizona State	-1.60 (-6.18, 2.98)	18
Texas	Houston	TCU	-7.08 (-17.81, 3.66)	-7
Champs Sports	Michigan State	Boston College	1.49 (-4.52, 7.51)	-3
Emerald	Oregon State	Maryland	7.38 (2.52, 12.24)	7
Meineke Car Care	Wake Forest	UCONN	1.77 (-2.60, 6.15)	14
Liberty	Mississippi State	UCF	-0.82 (-8.66, 7.00)	7
Alamo	Texas A&M	Penn State	-13.82 (-21.22,-6.41)	-7

Independence	Colorado	Alabama	-2.57 (-7.00, 1.85)	-6
Armed Forces	California	Air Force	2.08 (-4.64, 8.79)	6
Humanitarian	Fresno State	Georgia Tech	-5.48 (-13.52, 2.57)	12
Sun	South Florida	Oregon	1.33 (-3.24, 5.90)	-35
Music City	Florida State	Kentucky	-6.74 (-12.41,-1.06)	-7
Chick-fil-A	Auburn	Clemson	-2.04 (-7.98, 3.90)	3
Insight	Oklahoma State	Indiana	-0.38 (-8.78, 8.03)	16
Outback	Tennessee	Wisconsin	-4.28 (-9.44, 0.90)	4
Cotton	Arkansas	Missouri	0.83 (-3.65, 5.30)	-31
Gator	Virginia	Texas Tech	-5.77 (-15.62, 4.08)	-3
Capital One	Florida	Michigan	17.65 (7.92, 27.36)	-6
Rose	USC	Illinois	12.98 (5.66, 20.29)	32
Sugar	Georgia	Hawaii	1.97 (-9.01, 12.95)	31
Fiesta	West Virginia	Oklahoma	-3.09 (-8.39, 2.20)	20
Orange	Kansas	Virginia Tech	2.90 (-5.29, 11.09)	3
International	Rutgers	Ball State	15.54 (4.89, 26.18)	22
GMAC	Bowling Green	Tulsa	-1.57 (-9.14, 6.00)	-56
Championship	Ohio State	LSU	6.92 (-2.14, 15.97)	-14

We note in Table 2 that the expected outcomes based on the predictive model agreed with the actual winners of the games in 19 out of the 32 bowl games (59.4%), and that of the 13 games where the expected outcome was not in the

same direction as the actual outcome, three of the 95% confidence intervals for the expected outcomes covered the actual outcomes. The model appeared to do fairly well at predicting close games (defined by a less than seven point difference in final scores), with eight of the 12 'close' games (67%) having the 95% confidence interval for the expected outcome including the actual score. However, the model performed poorly at predicting 'blowouts,' with only one difference of 20 or more correctly covered by the 95% confidence interval for the expected outcome (Rutgers vs. Ball State).

6. APPLICATIONS OF THE PREDICTIVE MODEL

Given additional resources, the linear model proposed in this paper could be fitted using additional historical data (to further assess the importance of the predictors currently being investigated, in addition to other potential predictors). This makes a wide variety of applications of the model-based predictions possible. For example, odds-makers in Las Vegas or elsewhere could use the predictions as one possible quantitative tool for determining the lines for future college football bowl games, because the predictions represent expected outcomes based on previous predictors of bowl success. More importantly, an interesting application of the proposed prediction method could be the development of team ratings, based on a round-robin playoff scenario.

The calculation of team ratings for the current bowl season would start with a predictive model estimated using previous seasons of team-level data and bowl outcomes. Each bowl-eligible (six-win) team would be set as the "home team," and matched up against *all other* bowl-eligible teams in a round-robin playoff scenario. Based on the predictive model, 63 predicted (or expected) outcomes would be calculated, considering that team as the home team and all 63 opponents as "away" teams (predicted outcome = home – away). Each expected outcome would have an associated 95% confidence interval, reflecting statistical uncertainty in the expectation. If the 95% confidence interval includes 0, the expected outcome would be declared 'uncertain,' and the team would receive a single (1) point. If the 95% confidence interval does not include 0 and only includes positive values (i.e., the team is expected to be a clear winner), the team would receive 2 points. If the 95% confidence interval does not include 0 and only includes negative values (i.e., the team is expected to be a clear loser), the team would receive 0 points. The rating for a team would then be calculated as the sum of these points across all 63 hypothetical games. If the BCS is adamant that a playoff format for determining a national champion will not be adapted at any time in the near future (and may never be a reality, due to the financial incentives for schools to participate in bowl games), these ratings based on a hypothetical

round-robin playoff scenario could be used in part to help determine the two highest-rated (or strongest) teams.

For purposes of this paper, we present an example of calculating these ratings using actual data from the 2007-2008 season. Based on the "final" model presented earlier, predicted bowl outcomes were calculated according to a round-robin playoff format (Morris 1978) involving the ten teams that were selected to play in the BCS bowl games: Ohio State (OSU), Louisiana State (LSU), Southern California (USC), Illinois, Georgia, Hawaii, Kansas, Virginia Tech, Oklahoma, and West Virginia. These ten teams were all matched up against each other, and ratings were determined based on the predicted outcomes using the method described above. Table 3 presents results from this example.

Table 3. Round-robin predictions of victories, losses, and statistical "ties" for the ten BCS teams in 2007-2008 based on the estimated prediction model from 2004-2006, with 95% confidence intervals for the outcomes. Ratings are based on a system of two points per win, one point per tie (if the 95% confidence interval includes 0), and zero points per loss. Details are included for OSU and LSU, and the remaining eight teams have their ratings presented.

Team (H)	**Opponent (A)**	**Expected Outcome (H-A, 95% CI)**	**Points**	**Rating**
OSU	LSU	6.9 (-2.1, 16.0)	1	**14**
OSU	GEORGIA	11.2 (4.4, 18.0)	2	
OSU	KANSAS	7.3 (-1.4, 15.9)	1	
OSU	HAWAII	13.2 (2.0, 24.4)	2	
OSU	MISSOURI	14.0 (4.0, 24.0)	2	
OSU	USC	5.0 (0.5, 9.6)	2	
OSU	VA TECH	10.2 (3.1, 17.2)	2	
OSU	OKLAHOMA	1.1 (-6.4, 8.6)	1	
OSU	WVU	4.2 (-2.2, 10.5)	1	
LSU	OSU	-6.9 (-16.0, 2.1)	1	**9**
LSU	GEORGIA	4.3 (-3.8, 12.4)	1	
LSU	KANSAS	0.4 (-8.1, 8.8)	1	
LSU	HAWAII	6.3 (-7.4, 20.0)	1	
LSU	MISSOURI	7.1 (0.0, 14.2)	1	
LSU	USC	-1.9 (-8.7, 5.0)	1	
LSU	VA TECH	3.3 (-5.2, 11.7)	1	
LSU	OKLAHOMA	-5.8 (-14.7, 3.1)	1	
LSU	WVU	-2.7 (-6.6, 1.1)	1	
OKLAHOMA	*Details available upon request*			**14**

WVU		**11**
USC		**10**
KANSAS		**8**
HAWAII		**7**
VA TECH		**7**
GEORGIA		**5**
MISSOURI		**5**

Based on the current prediction model and these 10 opponents, Oklahoma and OSU would be the highest-rated teams. The method illustrated above would be extended to all 64 bowl-eligible teams to develop team ratings. Interestingly, Georgia had the lowest rating despite the fact that it was widely considered to be a very strong team. This is likely due to the fact that Georgia was merely average in terms of scoring margin relative to the other 63 bowl teams in 2007.

7. CONCLUSIONS

Building on the fairly extensive quantitative literature regarding the prediction of outcomes in college football games and the development of ratings for college football teams, this paper presented a new and straightforward application of linear modeling in the prediction of college football bowl game outcomes, and considered an application of the predictions in the development of ratings for teams based on a hypothetical round-robin playoff scenario. Using actual data from the 2004, 2005, 2006, and 2007 college football seasons for the Football Bowl Subdivision (FBS, formerly known as Division I-A), the regression analysis performed in the paper identified six team-level regular season predictors that explained roughly 22% of the variance in the actual bowl game outcomes from 2004 to 2006. These six predictors included the difference between teams in standardized scoring margin, the difference in standardized strength of schedule (based on Jeff Sagarin's computer ratings for *USA Today*), the difference in standardized offensive yardage per game, the difference in standardized defensive yardage yielded per game, the difference in standardized defensive touchdowns yielded per game, and the difference in standardized turnover margin per game. Despite the BCS mandate that rating systems not consider scoring margin for a team in the development of ratings, the team-level difference in standardized scoring margin was found in this paper to be a strong predictor of bowl game outcomes. Expected bowl outcomes in 2007 based on the estimated 2004-2006 regression model were found to be in agreement with actual outcomes in 59.4% (19/32) of the bowl games, and of the 13 games incorrectly predicted, three of the 95% confidence intervals included the actual game outcomes.

The 13 games with outcomes poorly predicted by the model were examined in more detail, given that much of what happens in college football bowl games is difficult to predict using objective methods. This examination found evidence of "over-confidence" for certain teams when they were playing against "poorer" teams (objectively defined based on their past ratings). The mean final Sagarin ratings for the teams in these 13 games were calculated using data from 2004 to 2007, and higher mean ratings were found to act as a positive catalyst for being "over-confident." Many of the players on these teams may not know the details regarding the opposing teams' scoring margin, offensive yards per game, defensive yards per game, and values on the other predictors considered in this paper, but they will be well aware of historical ratings of the opposing team and their previous performances in bowl games. Among the 13 unexpected results, four games were good examples of "over-confidence." For example, in the bowl game between Fresno State and Georgia Tech, the mean Sagarin rankings of these teams from the past four years were 78.50 and 27.75, respectively, but Fresno State somewhat surprisingly won the game by 12 points. We observed similar patterns in three of the other games. In other cases, better average rankings were found to have a favorable impact on the outcomes of the games, where teams with higher average historical rankings were found to come out on top. In three of the games, the teams had fairly even average rankings and the games were in fact close. Future research into this area should routinely consider closer examinations of unexpected results based on the historical models, and the impacts of player confidence and historical team ratings on future success would be interesting to examine further.

The estimated historical model in the paper (fitted using data from 2004 to 2006) was also used to demonstrate an application of the predictions in the development of team ratings in 2007, based on a round-robin playoff scenario where bowl-eligible teams are matched up against all other bowl-eligible teams. In each hypothetical game, the expected outcome is computed using the estimated regression model, along with a 95% confidence interval for the outcome. A team's rating is determined based on a scoring system using the 95% confidence intervals, where a team is assigned two points for every clear expected win, one point for expected ties (when the 95% confidence interval for the outcome includes 0), and zero points for every clear expected loss (see Section 6 for details). Considering a hypothetical round-robin playoff scenario involving the 10 BCS teams in 2007, the teams with the highest ratings based on this new method were the Ohio State University and the University of Oklahoma, suggesting that these two teams were expected to fare the best overall when playing the other BCS teams. The paper suggests that similar prediction models (possibly incorporating additional team-level predictors) could be estimated using past seasons of team-level data and bowl outcomes, and then used to calculate

predicted outcomes and team ratings based on the round-robin method in future seasons.

The method presented in this paper certainly has limitations that warrant discussion. First, the model considers *regular season* team statistics in developing the "team difference" predictors of the *post-season* bowl outcomes. Each year, the teams accumulate these statistics by playing against familiar teams in their conference schedules, in addition to a small number of unfamiliar teams in non-conference schedules. The bowl games generally match up teams that are not familiar with each other, so team-level statistics accumulated by playing against a majority of 'familiar' teams may not be the optimal predictors of outcomes when 'unfamiliar' teams are matched up in bowl games. This could be a reason for the relatively large proportion of unexplained variance (78%) in the regression model fitted to the historical data. Second, one could definitely argue that the historical regression model could be improved by using additional team-level difference predictors (e.g., average team experience, coaching experience, change in environmental conditions from home city to bowl city, etc.) and additional historical seasons of data (prior to 2004). The data analyzed in this paper took five months to collect and process using readily available web resources, and substantial additional resources would be required to collect additional predictors and additional years of data (the authors would welcome suggestions regarding additional resources). Third, a great deal of time passes between the last regular season game and the bowl games (generally at least a month), so the team playing in the bowl game may be very different from the team that played in the regular season and accumulated the team-level statistics. A good example of this limitation is the University of Michigan in 2007, which struggled with injuries to two of its top players (quarterback Chad Henne and tailback Mike Hart) in the regular season, and then upset a strong Florida team in the Capital One bowl. The regular season offensive numbers for Michigan may not have been as good as they could have been had these players been healthy all season, and Michigan was expected to lose (clearly) based on the historical regression model because Florida had much better team-level statistics. The expected outcome in the Capital One bowl may not have been as heavily in favor of Florida had the Michigan team in the bowl game been playing all season.

The regression model discussed in this paper was re-fitted including the 2007 data analyzed in the paper, resulting in an updated historical model that could be used to compute expected outcomes and develop ratings for the 2008-2009 bowl season. Based on this updated model, the team-level differences in standardized scoring margin, strength of schedule, and turnover margin remained to be significant ($p < 0.05$) predictors of the bowl outcomes. Further, the estimated coefficients in the updated model suggested that these predictors had similar relationships with the bowl outcomes. Data from the next (2008-2009)

college football season could certainly be used to further assess the predictive ability of this model, obtain better estimates of the parameters in the model, and examine whether predictors that are seemingly unimportant based on the relatively small sample in this paper become more important given more data. Another potential method of validation that warrants discussion was proposed by Stern (1991). Basically, given more seasons of data and applications of the predictions, the empirical distributions of actual outcomes for all games with a given prediction based on the model could be examined, to see if the distributions are in fact centered at the predictions, with small standard deviations. Obviously this method could not yet be applied at this point in time, given the small sample used in this paper.

Despite these limitations, the methods presented in this paper represent straightforward and logical means of applying simple statistical models to important quantitative problems in college football. Potential future applications of these methods provide quantitative researchers with exciting opportunities to explore the possibility that patterns exist in college football enabling the prediction of future bowl game outcomes.

REFERENCES

Annis, D.H. and Craig, B.A., "Hybrid paired comparison analysis, with applications to the ranking of college football teams," *Journal of Quantitative Analysis in Sports*, 2005, 1(1), Article 3.

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

Boulier, B.L. and Stekler, H.O., "Predicting the outcomes of NFL Games," *International Journal of Forecasting*, 2003, 19(2), 257-270.

Fainmesser, I., Fershtman, C., and Gandal, N., "A consistent weighted ranking scheme with an application to NCAA college football rankings," CEPR Discussion Papers 5239, 2005. Available online at http://tournamenttheory.org/conference/viewpaper.php?id=61

Fair, R.C. and Oster, J.F., "College football rankings and market efficiency," *Journal of Sports Economics*, 2007, 8(1), 3-18.

Faraway, J.J., *Linear Models with R*, 2005, Chapman and Hall / CRC, London, New York.

Farlow, S.J., "A computer program for ranking athletic teams," *International Journal of Mathematical Education in Science and Technology*, 1984, 15(6), 697 – 702.

Harville, D.A., "Predictions for national football league games via linear-model methodology," *Journal of the American Statistical Association*, 1980, 75(371), 516–524.

Harville, D.A., "The selection or seeding of college basketball or football teams for postseason competition," *Journal of the American Statistical Association*, 2003, 98, 17-27.

Martinich, J., "College football rankings: do the computers know best?", *Interfaces: Experimental Economics in Practice*, 2003, 32(5), 85-94.

Mease, D., "A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins," *The American Statistician*, 2003, 57, 241-248.

Morris, C., Discussion of "Football ratings and predictions via linear models," *American Statistical Association Proceedings of the Social Statistics Section*, 1978, Alexandria, VA: ASA, 87-88.

Ong, E.S. and Flitman, A.M., "Using neural networks to predict binary outcomes," *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems*, Beijing, October 28-31, 1997, IEEE, New Jersey, USA, ISBN/ISSN: 7-80003-410-O/TP-17, 427-431.

Pardee, M., "An artificial neural network approach to college football prediction and ranking," University of Wisconsin – Electrical and Computer Engineering Department, Technical Paper, 1999.

Park, J. and Newman, M.E.J., "A network-based ranking system for US college football," *Journal of Statistical Mechanics: Theory and Experiment*, Oct. 31, 2005. Available online at stacks.iop.org/JSTAT/2005/P10014.

Reid, M.B., "Least squares model for predicting college football scores," Masters Thesis (University of Utah Department of Statistics), 2003.

Stefani, R.T., "Applications of statistical methods to American football," *Journal of Applied Statistics*, 1987, 14(1), 61-73.

Stefani, R.T., "Improved least squares football, basketball, and soccer predictions," *IEEE Trans. Systems, Man and Cybernetics*, 1980, 10, 116-123.

Steinmetz, J.G., *System and Method for Predicting the Outcome of College Football Games*, United States Patent 6112128, issued August 29, 2000.

Stern, H.S., "In favor of a quantitative boycott of the bowl championship series," *Journal of Quantitative Analysis in Sports*, 2006, 2(1), Article 4.

Stern, H.S., "On the probability of winning a football game," *The American Statistician*, 1991, 45, 179-183.

Stern, H.S., "Statistics and the college football championship," *The American Statistician*, 2004, 58(3).

Stern, H.S., "Who's number 1 in college football?...and how might we decide?", *Chance Magazine*, 1995, 8(3), 7-14.

Trono, J.A., "A deterministic prediction model for the American game of football," *ACM SIGSIM Simulation Digest*, 1988, 19(1), 26-53.

West, B.T., "A simple and flexible rating method for predicting success in the NCAA basketball tournament," *Journal of Quantitative Analysis in Sports*, 2006, 2(3), Article 3.

Web Resources:

http://sports.yahoo.com/ncaaf/stats (general statistics)

http://www.usatoday.com/sports/sagarin/fbt07.htm (Jeff Sagarin's strength of schedule information)

http://web1.ncaa.org/d1mfb/natlRank.jsp?year=2007&div=4&rpt=IA_teamturnovermrgn&site=org (turnover margin statistics)

http://www.usatoday.com/sports/sagarin/fbt06.htm
http://sports.espn.go.com/ncf/news/story?id=2473969
http://cfbstats.com/2006/team/index.html

http://www.usatoday.com/sports/sagarin/fbt05.htm
http://sports.espn.go.com/ncf/news/story?id=2054429
http://cfbstats.com/2005/team/index.html

http://www.usatoday.com/sports/sagarin/fbt04.htm
http://scores.espn.go.com/ncf/scoreboard?weekNumber=17&seasonYear=2004&confId=80
http://cfbstats.com/2004/team/index.html

All data files used for the analyses and computations presented in this paper are available upon request (bwest@umich.edu).