The Probability of Winning a High School Football Game

Jennifer L. Bell

Auburn University

This paper, which is part of a series, was intended to demonstrate the application of nonparametric statistics while appealing to the general interest in football.

Non-parametric Statistics Workshop
EERA 2008 (Hilton Head, SC)
J. Bell

The Probability of Winning a High School Football Game

After a weekend filled with football games, the media presents a wide range of statistics to explain the game's ultimate result, win or lose. During a high school, college, or professional football game, the sports broadcasters will discuss probabilities, frequencies, and overall season statistics regarding offensive or defensive plays (e.g., third down and four, field goal from fifty yards, or goal line stand by the defense). Even though the general game format is the same for high school, college, and professional football, rules and regulations differ depending on the level, conference, and division.

Wagner (1985) wanted to explain the margin of victory by the significant statistics according to three multiple regression models and eliminate inferior or insignificant statistics that do not affect the margin of victory or were highly correlated to the other standard statistics. The participants were 90 Top 20 Division 1 schools and 98 professional games, who were randomly selected, from the 1985 season. The author hypothesized that the margin of victory can be explained based on these nine independent variables: (a) difference in the number of first downs; (b) difference in rushing yardage; (c) difference in passing yardage; (d) difference in return yardage; (e) difference in penalty yardage; (f) difference in the number of turnovers; (g) difference in the number of quarterback sacks; (h) difference in the time of possession; and (i) home field designation.

For college football games, Wagner (1985) found passing and rushing yards increased the margin of victory. Moreover, rushing yards increased the margin of victory more than passing yards, but turnovers decreased the margin of victory. His results revealed that time of possession, number of first downs, penalty yards, and home field advantage were not significant. Actually,

time of possession had a negative effect on the margin of victory. Comparable to the college results, the results for the professional football games suggested that rushing yards had more of an effect on the margin of victory than passing yards. Two differences existed between the college and professional games: number of first downs and quarterback sacks. These variables were significant for the professional games but not for the college games. According to Wagner, the margin of victory was more easily explained for college games than professional games (R^2 for the college model was 72.35%, and R^2 for the professional model was 56.87%).

Similar to the Wagner (1985) study, Willoughby (2002) examined 191 Canadian Football League (CFL) games between 1989 and 1995 using logistic regression. He predicted the outcome of the CFL game by the differences in rushing yardage, passing yardage, interceptions, fumbles recoveries, and quarterback sacks. The results suggested that the better teams in the CFL were more likely to win the game based on rushing yardage, passing yardage, and the number of interceptions, which was parallel to the findings of Wagner. By using the logistic regression model, the average prediction probability of the model was 85%, which exceeded the R^2 values in the Wagner study.

Research Question

In the Southeast, on any given Friday night during the months of August through November, the local high school's stadium will be filled with football fans of all ages. Often, the high stakes involved with these high school games are overlooked compared to college and professional games. There have been a number of studies (Wagner, 1985; Willoughby, 2002) about predicting the outcome of professional and college football games. Are the same variables used to determine the probability of winning a game with college and professional teams applicable to high school football teams?

Methods

Participants

The participants for this study were four Division 5-A high school football teams from a southeastern state. They were selected from a total population of 69 high schools. Using the school's enrollment from the previous school year, the state's high school association determined division classification. Division 5-A included the top 20% of schools based on large student populations and membership in their organization. These schools are divided into eight different geographic regions. Participants were selected based on regional designation and availability of game box statistics. A total of 46 individual game box statistics were collected from the 2006 regular and postseason seasons with 31 overall wins (67.4%) and 15 overall losses (32.6%). If 1 of the 4 participants played another team within the same set of selected participants, that game was entered for the winner of the football game and was not entered for the loser of the game.

Measures

During the football games, a certified teacher from the high school or a member of the booster club stood on the sidelines and recorded the results of each offensive and defensive play. After the game, the team's statistician analyzed the recorded yardage and attempts. The amount of information available for each team varied depending on the team statistician's preferred game box statistics.

Procedures

The individual game box statistics were collected by the researcher from the team's webmaster or head coach. The predictor variables includes difference between the team's and the opponent's (a) total passing yards; (b) rushing yards; (c) penalty yards; (d) fumbles; and (e) first downs.

Results

Predictors of Win/Lose

Descriptives. Descriptives for each predictor variable were assessed by the Win/Lose dependent variable. For the entire sample, the difference in total passing yards ranged from -199 to 89 with a mean of -37.83 and a standard deviation of 74.64. The mean score for the difference in total rushing yards was 86.91 with a standard deviation of 134.05. The range of scores was from -171 to 457. The range for the difference in total penalty yards was -54 to 62 with a mean of -0.20 and a standard deviation of 26.40. For the difference in the number of fumbles, the mean was -0.54 with a standard deviation of 1.59, and the values ranged from -6 to 4. The difference in the number of first downs had a range of -12 to 17 and a mean of 2.39 with a standard deviation of 7.25. For the five predictors, an independent sample *t* test was conducted. Table 1 displays the means and standard deviations for each predictor by Win/Lose in addition to the *t* test results.

Table 1

Means and Standard Deviations for each Predictor by Win/Lose

	Win		Lose		
Predictor	*M*	*SD*	*M*	*SD*	*t*
Difference in rushing yards	126.06	124.88	6.00	118.02	3.11**
Difference in passing yards	-40.26	72.23	-32.80	81.80	-0.31
Difference in penalty yards	0.77	26.66	-2.20	26.68	0.36
Difference in number of fumbles	-0.74	1.61	-0.13	1.51	-1.23
Difference in number of first downs	3.87	6.58	-0.67	7.83	2.06*

Note. *p < .05. **p < .01.

To determine the relationship between the predictor variable, Win/Lose, and the five predicting variables, bivariate correlations were conducted. The difference in total rushing yards and the difference in the number of first downs were significantly and positively correlated (r = .73; p < .001). The negative relationship between the difference in total rushing yards and the difference in total passing yards was significant (r = -.50; p < .001), meaning as the difference increased for rushing yards the difference decreased for passing yards. In addition, the difference in total rushing yards had a significant and negative relationship with the dependent variable, Win/Lose, (r = -.43; p < .01). The intercorrelations for each predictor variable are presented in Table 2.

Table 2

Intercorrelations for Win/Lose and Predictor Variables

Predictor	Win/Lose	Difference in passing yards	Difference in rushing yards	Difference in penalty yards	Difference in number of fumbles	Difference in number of first downs
Win/Lose	--					
Difference in rushing yards	-.43**	--				
Difference in passing yards	.05	-.50**	--			
Difference in penalty yards	-.05	.37**	-.17	--		
Difference in number of fumbles	.18	-.13	.19	-.30*	--	
Difference in number of first downs	-.30*	.73**	-.00	.09	.05	--

Note. *p < .05. **p < .01.

Prediction of Win/Lose. After the initial descriptives and intercorrelations, a logistic regression analysis was conducted using Win/Lose as the dichotomous dependent variable. The Hosmer and Lemeshow Goodness of Fit Test revealed no significant difference between the observed and expected values ($\Pi^2 = 2.74$; $p = .91$). The overall predicting percentage for the logistic regression model was 80.4%. The most significant predictors for a win were the difference in total rushing yards, the difference in total passing yards, and the difference in the number of first downs. Table 3 displays the summary of the full regression analysis including the unstandardized coefficients, Odds ratio, and Wald statistics for each predictor.

Table 3

Summary of Logistic Regression Analysis for Variables Predicting Win/Lose (N=46)

Predictor	*B*	*SE B*	Odds ratio	Wald statistic
Difference in rushing yards	-0.04	0.01	0.97	8.42**
Difference in passing yards	-0.03	0.01	0.97	6.33**
Difference in penalty yards	0.04	0.02	1.05	3.66
Difference in number of fumbles	0.74	0.43	2.09	2.87
Difference in number of first downs	0.32	0.15	1.37	4.41*
Constant	0.27	0.50	1.31	0.29

Note. Overall Predicting Percentage 80.4%. *p < .05. **p < .01.

Discussion

Previous studies (Wagner, 1985; Willoughby, 2002) found the number of first downs was not significant for college football games and was a significant predictor for professional football games. The results of this study indicate that the number of first downs was a significant predicting variable for high school football games. In addition, these findings support the previous studies regarding the significant contributions of rushing and passing yardage to winning a football game. Based on the results of this study, the likelihood of winning a high school football game depends upon increased rushing yardage and decreased passing yardage. Thus, these findings suggest that the predictor variables used in determining the probability of winning college and professional games are applicable to high school football games.

One significant outlier was found to have a *z* score of 3.33. In this game, the model predicted a win, but the game was lost. After further inspection of the raw data, the researcher hypothesized that the number of penalties and penalty yards (6 penalties for 50 yards versus 3

penalties for 30 yards) possibly caused the 17 to 20 loss. The participant had 165 total rushing yards and 15 first downs to the opponent's 66 rushing yards and 4 first downs.

The literature (Wagner, 1985; Willoughby, 2002) addressed the possible significance of the difference between the teams' quarterback sacks, number of interceptions, and time of possession, but those individual statistics were not available for all football teams; therefore, those variables were not included in the analysis. Future research should be considered using quarterback sacks, number of interceptions, and time of possession. Another limitation to this study was the highly correlated predictor variables of total rushing yards and number of first downs. Since the sample included various games involving the same team, future research could randomly select one week of individual game box statistics for all high school football games within the state.

References

Wagner, G. O. (1985). College and professional football scores: A multiple regression analysis. *American Economist, 31*(1), 33-37.

Willoughby, K. A. (2002). Winning games in Canadian football: A logistic regression analysis. *The College Mathematics Journal, 33*(3), 215-220.