Winning Games in Canadian Football: A Logistic Regression Analysis

Keith A. Willoughby

Keith A. Willoughby (kwilloug@bucknell.edu) is an Assistant Professor of Decision Sciences at Bucknell University. He received his Ph.D. in Operations Management from the University of Calgary in 1999. His research interests include public transit system analysis, procurement and logistics models, and quantitative applications in sports. At Bucknell, he teaches, among other courses, a first-year seminar entitled "The Science of Sports".

Introduction

The use of statistics is an integral component of sports. The number of home runs hit in baseball, three-point shots made in basketball, and touchdowns scored in football are means by which analysts and fans alike determine "success" for specific players or teams.

Which (if any) of the varied statistics produced in a football game are vital in determining a game's outcome? Using 1985 data, Wagner [**8**] studied 90 U.S. Division I collegiate football games and 98 National Football League (NFL) games. He produced a multiple regression model, with the margin of victory as the dependent variable. Some of the independent variables in the model included per-game differences in passing yardage, rushing yardage, number of turnovers, penalty yardage, and the number of first downs. His results showed close similarity between the parameter estimates in collegiate and professional football.

Hurley [**5**] used a multiple regression model to analyze NFL playoff games from 1970 to 1993. Again using margin of victory as the dependent variable, he was able to show that turnovers and rushing yardage are two extremely important determinants of success in a football game.

Goode [**2**] used factor analysis to determine those variables that were significant predictors of the outcomes of football games. Examining NFL games played between 1969 and 1973, his model successfully predicted the outcomes of 75% of the regular season games and 86% of the playoff contests.

Investigating Ivy League (collegiate) football games from 1964 to 1966, Haberman [**3**] developed a linear regression model to determine those statistics important for winning football games. He found that, among others, rushing yardage and completed passes were critical in determining the outcome of games.

Besides professional and collegiate football, other sports have been examined by researchers. Lo, Bacon-Shone, and Busche [**6**] created a logistic regression model for analyzing the outcomes of horse races. Using actual data from race tracks in Hong Kong, New Jersey, and Japan, they determined the likelihood of horses finishing in particular positions for specific races.

Crowe and Middledorp [**1**] formulated a logistic regression model for the sport of cricket. They analysed results from cricket matches played in Australia from 1977 to 1994, to determine if Australian umpires were biased against visiting teams. They

discovered that leg before wicket decisions, a highly controversial judgement call, were more frequently assessed to visiting countries.

The purpose of this paper is to present a logistic regression model for analyzing those game statistics important in determining the outcome of football games in the Canadian Football League (CFL). Many of the empirical analyses cited in this paper have used multiple linear regression. If one is strictly concerned with the *final outcome* of the game, rather than point differential, logistic regression (with its use of a dichotomous dependent variable) ought to be well suited for analyzing football game results. Ties are extremely rare, so the dependent variable could concentrate on either wins or losses.

Whereas previous analyses have lumped all teams together (by using complete seasons of data), this analysis will focus on three specific teams (categorized as very good, average, and poor), and their performance over a seven-year period. In this way, one may learn if differences exist in those statistics that are vital for a very good team versus those that are important for a poor team. If differences do result and should "winning mean everything", then football coaches and general managers ought to emphasize those game strategies essential to victory for the very good team.

This analysis does not rest on a "snapshot in time", but consists of data gathered for an individual team over a long duration. Table 1 shows the cumulative record of CFL teams between 1989 and 1995. The team with the greatest number of wins over this period (Calgary) was chosen as the "very good" team in the study, while the squad which earned the least number of wins (Ottawa) represented the "poor" team. The "average" team was the one whose number of victories most nearly equalled its number of losses (Saskatchewan).

Table 1. Cumulative Records of CFL Teams, 1989–1995

Team	Record (Wins-Losses-Ties)
Calgary	**90-35-1**
Edmonton	86-40-0
Winnipeg	73-53-0
Saskatchewan	**61-65-0**
British Columbia	58-66-2
Hamilton	50-76-0
Toronto	50-76-0
Ottawa	**38-88-0**

Note: This table only includes those teams which participated in the CFL between 1989 and 1995. It does not include various American-based CFL teams which joined the league beginning in 1993.

Model

Logistic regression is a powerful technique for analyzing those situations in which the dependent variable is discrete (Hosmer and Lemeshow [**4**]). Some its applications include medical research (presence or absence of heart disease) or consumer studies (several different response categories for a new product development).

In a linear regression model, the mean value of the outcome variable (given the value of the independent variable) is termed the conditional mean. It may be expressed as an equation linear in x, such as

$$E(Y \mid x) = \beta_0 + \beta_1 x.$$

In a logistic regression model, the conditional mean may also be represented by the expression $E(Y \mid x)$. However, the equation is no longer linear in x. In fact, we have

$$E(Y \mid X) = \frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}}.$$

The dependent variable used in this paper's logistic regression model is $Y \in \{0, 1\}$, with 0(1) indicating a loss (win) for the team (very good, average, or poor) under consideration.

The actual logistic regression model employed is

$$E(Y \mid X) = \frac{e^{(Z)}}{1 + e^{(Z)}}$$

where Z is

$$\beta_o + \beta_1 \text{DIFF\_RUSH} + \beta_2 \text{DIFF\_PASS} + \beta_3 \text{DIFF\_INT} + \beta_4 \text{DIFF\_FUMB} + \beta_5 \text{DIFF\_SACK}.$$

The independent variables, similar to those illustrated in Wagner [**8**], represent various football game statistics. They are:

DIFF_RUSH: The difference in rushing yardage
DIFF_PASS: The difference in passing yardage
DIFF_INT: The difference in the number of interceptions
DIFF_FUMB: The difference in the number of fumble recoveries
DIFF_SACK: The difference in the number of quarterback sacks

Thus, if Calgary rushed for 150 yards in a game while its opponent rushed for 100 yards, the DIFF_RUSH variable would have a value of 50 for Calgary.

An interesting statistical feature of these data is that, in games involving any two of the three teams, the outcomes of the games are not independent. That is, if Calgary played Saskatchewan and Calgary won, then Saskatchewan would have lost. In order to guard against dependencies in the data, we removed all games between any two of these three football teams. This resulted in a sample size of 92 for both Calgary and Saskatchewan, and 99 for Ottawa.

Results

The SAS/STAT® [**7**] statistical software package (using PROC LOGISTIC) was employed to obtain all model results. Table 2 indicates the results of the logistic regression analysis. Variable coefficients are shown in the table along with the corresponding p-values in parentheses. We note that the $\beta_1, \beta_2, \ldots, \beta_5$ coefficients indicate the change in the log odds of team victory for a one unit increase in the particular explanatory variable.

Table 2. Model Results

	Team		
Variable	Calgary	Saskatchewan	Ottawa
β_o	0.9948*	−0.8210	−0.8340***
	(0.0106)	(0.1248)	(0.0046)
DIFF_RUSH	0.0173***	0.0529***	0.00743*
	(0.0032)	(0.0011)	(0.0353)
DIFF_PASS	0.0148***	0.0173***	0.00427
	(0.0015)	(0.0030)	(0.0828)
DIFF_INT	0.8136***	1.4184***	0.3423*
	(0.0085)	(0.0011)	(0.0397)
DIFF_FUMB	0.3451	0.9074***	0.5303***
	(0.2912)	(0.0066)	(0.0098)
DIFF_SACK	0.4199*	1.1287***	0.4638***
	(0.0151)	(0.0027)	(0.0004)

Note:
***Significant at the 1% level
*Significant at the 5% level

An analysis of these findings reveals differences between those factors that are important in predicting a game's outcome, given the team under consideration. For instance, differences in rushing yardage and the number of interceptions were significant at the 1% level for the very good and average teams (Calgary and Saskatchewan, respectively) in our study. These game statistics were significant at the 5% level for the poor team (Ottawa). In addition, differences in passing yardage were highly significant for both Calgary and Saskatchewan while not significant at all for Ottawa.

On the other hand, differences in the number of fumble recoveries were not significant for Calgary, while they were highly significant for Saskatchewan and Ottawa. Finally, differences in the number of quarterback sacks were significant at the 5% level for Calgary, while they were significant at the 1% level for the average and poor teams. Further, we note that the estimated coefficient of the quarterback sacks variable for the average team (Saskatchewan) is roughly 2.5 times larger than those coefficients for the very good or poor teams. This would indicate that this variable has a larger effect on the log odds of the team's victory than the corresponding variable for the other two teams.

What does this mean? Quite simply, the better the team in the CFL, the more likely it is to rely on rushing and passing yardage as well as the number of interceptions in winning football games. A poor team is less likely to depend on these factors. Consequently, football coaches, in an effort to build championship squads, should try and control the line of scrimmage (through rushing and passing yardage) and record more interceptions than their opponents.

In an effort to determine how well the logistic regression model predicted actual game outcomes, we generated a table of predicted versus actual outcomes. We obtained the number of times the model predicted a probability above 0.5 as well as the number of occurrences in which the probability was under 0.5. (Since the likelihood of a predicted probability exactly equal to 0.5 was extremely rare, we used > 0.5 and

< 0.5 as our cutoff). We then compared these predicted results to the actual outcomes (win or loss) of the respective football games.

Table 3. Predicted vs. Real Outcomes: Calgary

Actual outcome	Predicted Results		Total
	$P > 0.5$	$P < 0.5$	
Win	62	4	66
Loss	9	17	26
Total	71	21	92

Probability of correct model result = 85.9% (79 out of 92)

Table 4. Predicted vs. Real Outcomes: Saskatchewan

Actual outcome	Predicted Results		Total
	$P > 0.5$	$P < 0.5$	
Win	38	5	43
Loss	4	45	49
Total	42	50	92

Probability of correct model result = 90.2% (83 out of 92)

Table 5. Predicted vs. Real Outcomes: Ottawa

Actual outcome	Predicted Results		Total
	$P > 0.5$	$P < 0.5$	
Win	20	12	32
Loss	9	58	67
Total	29	70	99

Probability of correct model result = 78.8% (78 out of 99)

Each of the three models does quite well in predicting actual game outcomes. The probability of a correct model result was 85.9%, 90.2%, and 78.8% for Calgary, Saskatchewan, and Ottawa, respectively.

Tests were done to determine the extent of any multicollinearity in the respective models. A series of linear regressions were run by regressing a given independent variable on the remaining ones. If a certain regression produced a very high value of R^2, then a multicollinearity problem may exist in the data. For our regression models, the highest such R^2 was 0.2294. Multicollinearity does not appear to be a problem in this data set.

Conclusions

These findings appear to suggest that differences exist between those factors that are significant in determining victory for teams of different abilities. Researchers should not lump teams together (as they have previously done) in an attempt to determine critical factors.

In football, the two prime sources of turnovers are interceptions and fumble recoveries. Previous empirical analyses have combined them into one variable, "turnover". However, the model suggests that interceptions are the more important game statistic (they were significant at the 1% level for Calgary and Saskatchewan, while significant at the 5% level for the poorest team, Ottawa). Perhaps this indicates that more emphasis ought to be placed on interceptions, rather than combining the two sources of turnovers.

Some game statistics were purposely not included in these logistic regression models (e.g., time-of-possession statistics and return yardage (from punts and kickoffs)). The desire to keep the model as parsimonious as possible meant that only a handful of the statistics collected in a game could be used.

One of the shortcomings of this type of modelling is that the binary dependent variable fails to differentiate between levels of victory. Some may argue that the game statistics from a 50-point victory would be markedly different from those obtained in a 1-point win. The logistic regression model, however, is not concerned with the margin of victory, only with the specific outcome (win or loss). This concern does have some merit. Nonetheless, it ought to be remembered that football teams earn two points in the standings for a victory *regardless of point differential*. When the season is completed, it is the number of wins that determines the first-place finisher, not the cumulative margins of victory. As a result, logistic regression modelling can be an important tool in determining the critical factors to a team's overall success.

Acknowledgements. The research leading to this paper was partially supported by the Natural Sciences and Engineering Research Council of Canada under Grant No. A1485 and by the Carma Chair at the University of Calgary.

The author gratefully acknowledges the helpful advice obtained from Dr. Wynne Chin (Faculty of Management, University of Calgary), Dr. Tak Fung (University Computing Services, University of Calgary) and an anonymous referee in the completion of this paper.

References

1. S. M. Crowe and J. Middledorp, A comparison of leg before wicket rates between Australians and their visiting teams for test cricket series played in Australia, 1977–94, *The Statistician* **45** (1996) #2, 255–262.
2. M. Goode, Teaching Statistical Concepts with Sports, presentation to the American Statistical Association, Boston, 1976.
3. S. J. Haberman, Analysis of scores of Ivy League football games, in *Optimal Strategies in Sports*, S. P. Ladany and R. E. Machol, editors, North-Holland, 1976.
4. D. W. Hosmer Jr. and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, 2000.
5. B. Hurley, What wins football games in the NFL? presentation at the Canadian Operational Research Society (CORS) Conference, Montreal, Quebec, 1994.
6. V. S. Y. Lo, J. Bacon-Shone, and K. Busche, The application of ranking probability models to racetrack betting, *Management Science* **41** (1995) #6, 1048–1059.
7. SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volumes 1 and 2*, SAS Institute Inc., Cary, North Carolina, 1989.
8. G. O. Wagner V, College and professional football scores: a multiple regression analysis, *American Economist* **31** (1987) #1, 33–37.

