Analysis of Predictions for NFL Game Outcomes from Burke's Model

Ryan K. Brown, Undergraduate; Dr. Thomas Short, Professor, Department of Mathematics and Computer Science; Christopher A. Wetherill, Undergraduate, John Carroll University

Abstract

Burke provided a binary logistic model for assessing the likelihood of a given NFL team winning against any other team, and has used the model to predict game results, in particular, over the four-year period from 2009 through 2012. We demonstrate how such predictions may be evaluated using the Hosmer-Lemeshow and the Pigeon-Heyse goodness-of-fit tests.

Burke's Multivariate Logistic Model

Burke's model for predicting game outcomes takes the general form:

$$GameLogit = X_0 + 0.72X_1 + (Logit_A + Logit_{A.Opp}) - (Logit_B + Logit_{B.Opp})$$

where X_0 is the equation constant (-0.36), X_1 is the home field coefficient (1 for a team playing at home; 0 otherwise), $Logit_T$ is the log of the odds ratio for a given team (A or B), and $Logit_{T.Opp}$ is the log of the odds ratio for that team against an average opponent. The log of the odds ratio for a given team is given as

$$Logit_T = \beta_0 X_0 + \cdots + \beta_n X_n$$

where $X_0, \cdots, X_n$ represent a set of independent variables selected by Burke.

Hosmer-Lemeshow and Pigeon-Heyse Tests for Goodness of Fit

Both the Hosmer-Lemeshow and the Pigeon-Heyse tests for goodness of fit were employed to evaluate Burke's model. The Hosmer-Lemeshow test statistic takes the form:

$$\hat{C} = \sum_{g=1}^{G} \sum_{k=1}^{2} \frac{(O_{gk} - E_{gk})^2}{E_{gk}}$$

where G is the number of groupings (here, the probabilities of home team wins predicted by Burke's model are sorted and then divided into 10 bins of approximately equal sizes), k signifies a win (k = 1) or loss (k = 2), O_{gk} is the number of observed k outcomes for a group, and E_{gk} is the number of expected k outcomes. This statistic follows a χ^2 distribution on $G - 2$ degrees of freedom.

The Pigeon-Heyse test takes a similar form

$$\mathcal{J}^2 = \sum_{g=1}^{G} \sum_{k=1}^{2} \frac{(O_{gk} - E_{gk})^2}{\Phi_g E_{gk}},$$

$$\text{with } \Phi_g = \frac{\sum_{n=1}^{n_g} \hat{\pi}_{i1}(1-\hat{\pi}_{i1})}{n_g \bar{\pi}_{i1}(1-\bar{\pi}_{i1})},$$

where n_g is the group size, $\hat{\pi}_{i1}$ the individual home team win probability for all teams in a group, and $\bar{\pi}_{i1}$ the average groupwise home team win probability.

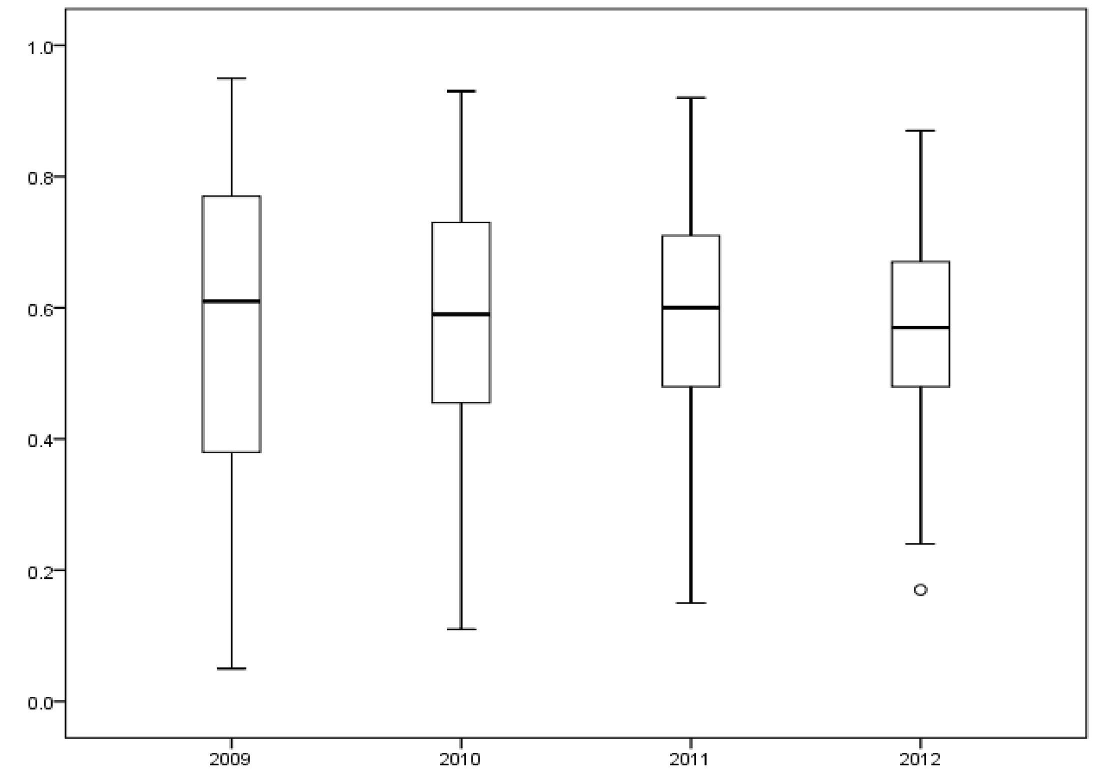
Evaluating Burke's Goodness of Fit

If we evaluate Burke's model across the four years that he has used it, we see that as a group the observed rates of home team wins do not significantly differ from expected rates of home team wins ($\hat{C}$ (8) =3.351, p-value = 0.910 and $\mathcal{J}^2(9)$ =7.148, p-value =0.6217). However, Burke between the 2010 and 2011 seasons altered his model to better fit the data. With this in mind, the year-to-year p-values for each of the two tests (on 8 and 9 degrees of freedom, respectively) become:

	H-L	P-H
2009	0.169	0.111
2010	0.622	0.070
2011	0.797	0.604
2012	0.397	0.868

Notably we see occasional stark differences between the p-values of the Hosmer-Lemeshow and Pigeon-Heyse tests: likely this is attributable to the differences in sorting strategies of the covariates outlined by Pigeon and Heyse (1999).

However, we nevertheless see that the observed rates of home team wins predicted by Burke's model do not differ significantly from the expected rates. Moreover, we may also see that the model's variability in predicted probabilities of home team wins has consistently decreased as a function of time (although the accuracy of his predictions remains at an average of 65%):



Probability of home team winning by year

Conclusion

In sports, predictive models proliferate, often without any robust test of their objective validity. Here, we illustrate a method for assessing the goodness of fit of these models and illustrate the reliability of the less well-known Pigeon-Heyse goodness-of-fit test (relative to the Hosmer-Lemeshow test). Notably, this test is less dependent on the calculating algorithm that sorts the coefficients of a linear model than is its Hosmer-Lemeshow counterpart (Kuss, 2002; Pigeon & Heyse, 1999). Finally, we demonstrate the validity of the present multivariate logistic regression model.

References

Burke, B. Advanced NFL stats: How coaches think: Run success rate. 04 Oct. 2010. Web. 05 May 2013.

Burke, B. Advanced NFL stats: how the model works–A detailed example. 23 Jan. 2009. Web. 01 May 2013.

"Chi-Square Calculator." Chi-Square Calculator. Web. 05 May 2013.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. Statistics in Medicine, 21. 3789–801.

Pigeon, J. G. & Heyse, J. F. (1999). An improved goodness of fit statistic for probability prediction models. Biom. J., 41: 71–82.