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

Christopher Wetherill, Ryan Brown, and Thomas Short

John Carroll University

Author Note

Christopher Wetherill, Department of Psychology; Ryan Brown and Thomas H. Short, Department of Mathematics and Computer Science, John Carroll University

Correspondence concerning this article should be addressed to Thomas H. Short, Department of Mathematics and Computer Science, John Carroll University, University Heights, OH 44118. Email: tshort@jcu.edu

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 over the four-year period from 2009 to 2012. Here we demonstrate how such predictions may be evaluated using the Hosmer-Lemeshow and Pigeon-Heyse goodness-of-fit tests and assess the appropriateness of each test as a diagnostic mechanism.

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

Since 2009, Burke has offered outcomes predictions for each NFL game played using a multivariate logit model with eight contributing coefficients based on team efficiency statistics.

Burke's model for predicting game outcomes then 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, 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

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

where $X_0, \ldots, X_n$ represent a set of independent variables selected by Burke (2009a; 2009b).

However, to date, there has been little examination of the goodness of fit of such models. Typically, for such logistic models, the Hosmer-Lemeshow goodness-of-fit test is employed. This test takes the form of a χ^2 test where the n observations are sorted by their estimated probabilities of success and then divided into g equal-sized groups. These are then evaluated by:

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

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 (Hosmer & Lemeshow, 1980).

Evaluating Burke's predicted win probabilities for each NFL game ($n = 876$) from 2009 through 2012, we see that the predicted rates of home team wins do not differ significantly from a χ^2

distribution ($\hat{C}(8) = 3.351$, p-value $= 0.910$ using IBM SPSS 21). On this basis, we might assume that Burke's model is an adequate descriptor of the available data. However, in practice, we see that the results of a Hosmer-Lemeshow test may vary significantly depending on the sparseness of the data being sampled and the statistical software used to perform the analysis. Specifically, Hosmer et al. in subsequent research reported six different p-values given by various statistical packages, with probabilities ranging from 0.02 to 0.16; Pigeon and Heyse gave a more impressive example, with p-values for a single data set ranging from 0.02 to 0.45 (1997; 1999b). As Kuss notes, this wide variance results from statistical packages using different calculation algorithms without offering the user a clear idea of what calculations are actually being performed.

Indeed, Bertolini et al. have suggested that this may result simply from how the data are ordered before being binned into deciles. As Pigeon and Heyse note, the sorting may result in the collective low expected probabilities for an event of the observations in the first deciles and low expected probabilities for a non-event of the observations in the final deciles, challenging the validity of the χ^2 distribution (2000; 1999b). Notably, however, recent work has attempted to address some of these concerns and to standardize the power of the Hosmer-Lemeshow test across different sample sizes (cf. Paul, Pennell, & Lemeshow, 2013). Such work does, however, rely on sample sizes of $n = 500$ or larger with significant decreases in power at samples fewer than $n = 1000$.

Given these concerns, we also evaluated Burke's model using the Pigeon-Heyse goodness-of-fit test, created in part to remedy the noted irregularities with the Hosmer-Lemeshow test. The Pigeon-Heyse test takes a similar form to Hosmer and Lemeshow's:

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

This test differs from the Hosmer-Lemeshow test in its inclusion of Φ_g, the ratio of the proper

variance of O_k and the multinomial variance of O_k, defined as

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

where n_g is the group size, $\hat{\pi}_{i1}$ (in the context of Burke's model) is the individual home team win probability for all teams in a group g, and $\bar{\pi}_{i1}$ the average groupwise home team win probability. $\mathcal{J}^2$ is approximately distributed as a χ^2 distribution on $G - 1$ degrees of freedom (Pigeon & Heyse, 1999a).

If we evaluate Burke's model using this measure, we again see that as a group the observed rates of home team wins do not significantly differ from expected rates of home team wins ($\mathcal{J}^2(9) = 7.148$, p-value $= 0.622$).

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:

Year	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 (1999b). Moreover, before altering his model, using the Pigeon-Heyse test, we see evidence to question the model's fit to the data, although neither p-value for 2009 or 2010 reaches a level of statistical significance.

Indeed, observing the densities of predicted probabilities of team wins by year, we can see that the 2009 and 2010 seasons' predictions depart (if not to a level of significance) from an expected

χ^2 distribution. Specifically, we can see that for 2009 and 2010, before Burke changed two of the model's coefficients, the predicted probabilities of team wins do not obviously follow an expected χ^2 distribution. That is, we would expect a noncentral χ^2 distribution to be a unimodal curve with a peak about the mean: that is, similar to the PDF curves for the 2011 and 2012 predicted probabilities. Rather, both the 2009 and 2010 predicted probability curves are, to some degree, platykurtic and lacking a distinct peak about the mean predicted probability.

Given these results, we conclude that, although the Hosmer-Lemeshow test for goodness-of-fit does provide an overall adequate description of the fit of the model, in cases where the number of observations is small, it will often fail to accurately assess the model's goodness-of-fit and the $\hat{C}$ statistic returned may vary substantially among statistical packages. Alternately, the Pigeon-Heyse test provides a more robust analysis of goodness-of-fit even when sample sizes are smaller and is more resistant to sorting effects than is the Hosmer-Lemeshow test. Moreover, we find that Burke's model not only fits the available data well, but also improves with respect to fit over the four years in which it has been applied.

In the context of Burke's model, we sought to use these goodness-of-fit tests to assess the correctness of the model's predicted probabilities of team wins. Here, we found that, although no test revealed any individual year's predicted probabilities to be significantly incorrect, those for 2009 and 2010 fail to visually approximate a χ^2 distribution and trend towards significantly departing from the distribution. Nevertheless, assessing Burke's model by the Pigeon-Heyse test, we see that the predicted probabilities of team wins better approximate the distribution as a function of year.

References

Bertolini, G., DAMico, R., Nardi, D., Tinazzi, A., & Aplone, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics, 5*, 251:253.

Burke, B. (2009a). How the model worksa detailed example part 1. *Advanced NFL Stats.* Web.

Burke, B. (2009b). How the model worksa detailed example part 2. *Advanced NFL Stats.* Web.

Hosmer, D.W., Hosmer, T., Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine, 16*, 965-980.

Hosmer D.W., & Lemeshow S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics, A9*, 1043-1069.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine, 21*, 3789-3801.

Paul, P., Pennell, M., & Lemeshow, S. (2013). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in Medicine, 32*, 67-80.

Pigeon, J. G., & Heyse, J. F. (1999a). An improved goodness of fit statistic for probability prediction models. *Biometrical Journal, 41*, 71-82.

Pigeon, J.G., & Heyse, J.F. (1999b). A cautionary note about assessing the fit of logistic regression models. *Journal of Applied Statistics, 26*, 847-853.