Modelling the Time to the First Goal in Football

Georgi Boshnakov[1], Tarak Kharrat[1,2], and Ian G. McHale[2]

[1]School of Mathematics, University of Manchester, UK.
[2]Centre for Sports Business, Salford Business School, University of Salford, UK.

December 14, 2015

> *If you are first, you are first.*
> *If you are second, you are nothing.*
>
> Bill Shankly, *Liverpool F.C. Manager from 1959 to 1974*

In this article we look at what is the most important event in determining the outcome of a football match: the first goal. Bill Shankly was saying that being first mattered, well in football it is pretty intuitive that scoring the first goal helps a team come first. Indeed, in matches in which there are goals, the team that scores first, goes on to win 68% of matches[1]. Our objective here is not simply to show that Bill Shankly was right, but rather, by investigating the distribution of the time to first goal, we should gain insight into the distribution of time between goals, and ultimately derive improved models for in-play predictions.

1 A Competing Risks Model for Time to First Goal

To investigate the distribution of the time to the first goal in a match we use a 'competing risks' model. In some sense, this exercise is a generalisation of the simplified situation studied in Nevo and Ritov (2013) who examine the time to the first goal irrespective of which team scores. Here we consider the time to the first goal but acknowledge that the event (the first goal) can happen as a consequence of one of two 'competing' causes: either the home team *or* the away team can score the first goal in the match. A simple analogy of this situation is that of a study in which an individual can die (the event) from one of two possible causes, e.g. cancer or heart attack, or for us, a first goal (the event) can be scored by one of two possible 'causes' (the home team scores, or the away team scores).

Let X_t be the 'state' of the match at time t where $t \in [0, \tau]$ and τ can be thought of as the end of the match (90 minutes plus some injury time). The competing risks process begins when the match starts (kick-off) in an initial state, X_0, with the scoreline at 0-0. The process moves out of the initial state X_0 when the first goal is scored at time T and hence T is the period of time in the match when the scoreline is 0-0.

In survival analysis T is known as the survival time or failure time. If no goal is scored by the final whistle, the time to the first goal is 'censored'. We assume here random (right) censoring and that the censoring is independent of the distribution of T. This is a reasonable assumption in football as we do

[1]Premier League from 2010/2011 to 2014/2015

not expect the referee to compute the additional time based on the timing of any goals scored, or on the current score. In some sense, it is equivalent to assuming a 'fair' (not corrupted) referee!

The key quantities when analysing competing risks data are the 'cause-specific' hazards which are $\alpha_1(t)$, the home team's intensity of scoring, and $\alpha_2(t)$, the away team's intensity of scoring. The hazards are defined by

$$\alpha_k(t) = \lim_{\Delta_t \to \infty} \frac{\mathsf{P}(T \in [t + \Delta_t] \, , D = k \mid T \geq t)}{\Delta_t} \qquad k = 1, 2, \tag{1}$$

where D is a random variable representing the type of 'failure', with $D = 1$ being a home team goal, and $D = 2$ being an away team goal. In the case of football, one can think about these cause-specifc hazards as the probability that the next attack is successful (i.e. that it results in a goal) for that team when the score is still 0-0.

The hazard function, defined by (1), and the cumulative cause-specific hazard function (given by $A_k(t) = \int_0^t \alpha_k(u)du$ for $k = 1, 2$) are analogous to the probability density function and cumulative distribution function. To model the hazards, we look for the most appropriate functions for α_1 and α_2, and just like when fitting other types of models in statistics, the task is to estimate the parameters of the chosen parametric hazard function. In the same way that one might use the histogram of the observed data to give a clue as to which probability to use as a model, we can use the empirical hazard function to give a clue as to the form of the model we should use for the time to first goal.

The cumulative hazard function can be estimated from the observed data (using counting process theory and the Nelson-Aalen estimator, see, for example, Andersen et al. (2012)). The hazard function can then be recovered using kernel-based methods as explained in Muller and Wang (1994). When estimating the hazard for the home (away) team, goals scored by the opposition result in a censored event for the home (away) team.

2 Results

We now estimate the empirical hazard functions and fit two models to the time to first goal - the exponential distribution, and the Weibull distribution. Our data are the times to the first goal in matches in the English Premier League for the five seasons from 2010-11 to 2014-15. Out of the 1900 games, there are 1755 matches (92.4% of matches) with a 'first' goal and 145 (7.6%) matches with censored observations (this is the same as the number of 0-0 scorelines).

Figure 1 shows the non-parametric estimated scoring intensities for the first goal event, together with two parametric models: an exponential and a Weibull model. The Weibull model has hazard function given by

$$\alpha(t) = \lambda c t^{c-1},$$

with $\lambda \in (0, +\infty)$ being a rate parameter and $c \in (0, \infty)$ a shape parameter. The hazard is monotonically increasing for $c > 1$, monotically decreasing for $c < 1$, and constant (and equal to λ) when $c = 1$ (the exponential case). The parametrisation adopted here is the one we used in our previous articles (add link to them).

The exponential model has the property that the scoring intensity is constant throughout the match - hence it is a flat horizontal line in Figure 1. It is clear that the exponential model and the resulting constant hazard is not appropriate. It fails to catch the time-varying effect of the hazard and rather estimates the 'average' effect over time.

The Weibull model describes the data better and captures the increasing intensity of scoring the first

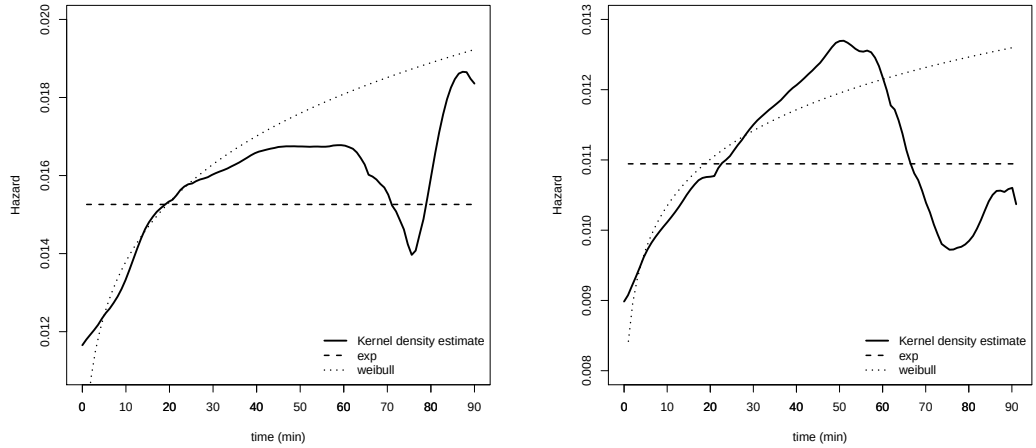

Figure 1: Scoring intensity for time to first goal by (a) the home team and (b) the away team.

goal, though does not mimic the subsequent decline in intensity. The hazard for the Weibull distribution is either increasing or decreasing and hence cannot replicate a non-monotonic hazard. Possible candidate distributions might be, for example, the generalized gamma (Cox et al., 2007) or the generalized F distribution (Cox, 2008). Deriving a closed form for the counting processes using these three-parameters distributions as inter-arrival times could be the topic of future research. Nevertheless, the Weibull provides an improved fit to the data compared to the exponential, as demonstrated by the AICs for the four models (Tables 1 and 2) for the home hazard the AIC for the exponential is 10595 but the the Weibull is 10568, whilst for the away hazard is 8086 for the exponential and 8081 for the Weibull. Likelihood ratio tests suggest the Weibull distribution is a statistically significant better fit in both cases.

Distribution	Deviance	Parameters	AIC
exponential	10593	1	10595
weibull	10564	2	10568

Table 1: Godness of fit summary for hazard of first goal for the home team $\alpha_1(t)$. Likelihood ratio test = 29 (p-value = $7.24 \cdot 10^{-8}$).

Distribution	Deviance	Parameters	AIC
exponential	8084	1	8086
weibull	8077	2	8081

Table 2: Godness of fit summary for hazard of first goal for the away team $\alpha_2(t)$. Likelihood ratio test = 7 (p-value = $8.15 \cdot 10^{-3}$).

3 Future Work

If one uses an exponential model for the scoring rate of teams, Figure 1 hopefully shows that the results (betting results that is) may not be optimal. Further, although our study shows that the Weibull distribution provides an improvement to the exponential, the characteristics of the empirical hazard functions are not fully replicated. We think that a sensible next step would be to look at using the generalised gamma distribution, or other distributions that can mimic the increasing then decreasing hazard function observed in Figure 1.

In this article, we implicitly assumed that our data is homogeneous i.e. all teams score the first goal at the same rate. We didn't take into account the individual teams' identities or the match conditions. As usual in statistics, heterogeneity can be modelled using individual (team specific) covariates. In

a parametric framework, `R` users can model the effect of covariates using techniques described in the `flexsurv` package (Jackson, 2014). We will investigate the effect of covariates in our next article.

The parametric models studied in this post is not the only option. As a matter of fact, the most used model in the survival analysis literature for the hazard function is non/semi-parametric (depending if you include covariates or not) and was proposed by Cox (1972). The model has the advantage of being more 'flexible' as it avoids assuming any parametric distribution for the hazard. The drawback is that deriving a closed formulae for the final counting processes is not possible. `R` users interested in fitting Cox-based hazard models can use the routines available in the `survival` package (Lumley and Therneau, 2004).

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Cox, C. (2008). The generalized f distribution: an umbrella for parametric survival analysis. *Statistics in medicine*, 27(21):4301–4312.

Cox, C., Chu, H., Schneider, M. F., and Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine*, 26(23):4352–4374.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Jackson, C. (2014). *flexsurv: Flexible parametric survival and multi-state models*. R package version 0.5.

Lumley, T. and Therneau, T. (2004). The survival package. *R News*, 4(1):26–28.

Muller, H.-G. and Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, pages 61–76.

Nevo, D. and Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. *Journal of Quantitative Analysis in Sports*, 9(2):165–177.