Using the Weibull count distribution for predicting the results of football matches

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

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

December 14, 2015

In our previous article, we described a new discrete distribution that appears to provide a better fit to goals data than the Poisson distribution. On its own, this finding isn't earth shattering - just interesting (if you are in to that sort of thing, and we are ;-)). In this article, we 'up-the-ante' and use the Weibull count distribution as the basis for a model that we use for betting.

1 The *default* model for football

The 'go-to' model for football is based on Maher's (Maher, 1982) original specification in which teams are assigned attack and defence 'strengths'. Just in case you don't know how the model works, we will give our take on it.

First, consider the Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

which gives the probability of x goals in a match by a team which scores goals at a rate of λ per game. An obvious thing to do is to say each team has a different scoring rate (some teams are better than others and score at a higher rate than others). Thus, we might say that each team has its own λ, i.e. the ith team has scoring rate λ_i. But this can be improved further, because surely it matters who the opposition is - Manchester City can be expected to score at different rates when playing against Chelsea or Sunderland. This is where Maher's model is really neat: the scoring rate for team i playing against team j is given by

$$\lambda_{ij} = \alpha_i \times \beta_j,$$

where α_i is the attack strength of team i (as α_i increases, the scoring rate of team i increases) and β_j is the defence strength of the opposition team, team j (as β_j increases, the opposition defence must be worse since it corresponds to an increase in the scoring rate of team i).

Lastly, surely the scoring rate depends on where the match is taking place: if team i is at home, we might expect them to score at a higher rate than if they were playing away to team j. To allow for this possibility, we let the scoring rate for team i playing at home to team j be

$$\lambda_{ij} = \alpha_i \times \beta_j \times \gamma,$$

where if γ is estimated to be bigger than 1, the scoring rate increases so that there is a *home advantage*. If $\gamma = 1$, there is no home advantage, and if $\gamma < 1$ the scoring rate at home decreases below that which would be expected away from home.

For those unfamiliar with statistical modelling, you might be thinking, "but what actual values should we use for the α's, β's and γ?". It is a good question. The short answer is that in what follows, we let the data decide by using maximum likelihood estimation - "we choose the values of the parameters (the α's, the β's and γ) which maximise the probability of observing the data we have observed". This is deep water and having dipped our toe in, we leave the topic of maximum likelihood estimation there! However, if you don't want to use maximum likelihood, you can get pretty good estimates of the λ for each team in each match using averages:

1. Calculate the mean goals scored at home ($= \bar{X}$), and the mean goals scored by away teams ($= \bar{Y}$), for the whole data set. Then do the same for each team (call these $\bar{X}_i$ and $\bar{Y}_i$ respectively).

2. Similarly, calculate the mean goals conceded by all teams at home ($= \bar{P}$), and the mean goals conceded by all away teams ($= \bar{Q}$), for the whole data set. At the individual team level, let the average goals conceded at home and away be $\bar{P}_i$ and $\bar{Q}_i$ respectively.

3. Then for two teams i and j playing a match in which team i is at home, a sensible estimate for the rate at which team i might score goals is

$$\lambda_i^{(j)} = \bar{X} \times \frac{\bar{X}_i}{\bar{X}} \times \frac{\bar{Q}_j}{\bar{Q}},$$

   where the first term represents the baseline scoring rate of teams at home, the second term adjusts this for how much more (or less) than the average for all teams the home team in question scores goals, and the third term represents how much more (or less) the away team concedes goals than the average for all teams.

4. Similarly for the away team's scoring rate, we would use

$$\lambda_j^{(i)} = \bar{Y} \times \frac{\bar{Y}_j}{\bar{Y}} \times \frac{\bar{P}_i}{\bar{P}}.$$

Anyway, back to the Maher model.... Maher's original model was later adapted by Dixon and Coles (Dixon and Coles, 1997) and ultimately used for betting. Dixon and Coles made two modifications to Maher's original model. First, they allowed for dependence by inflating/deflating the probabilities of some scorelines. Second, they allowed more recent matches to effect the estimated attack and defence strengths more than matches played in the past.

Like Dixon and Coles, our model is based on the Maher specification. However, instead of letting the rate parameter in the Poisson distribution be a function of the attack and defence abilities of the two competing teams, we let the rate parameter of the Weibull count distribution vary with the attack and defence abilities. We follow Dixon and Coles and allow more recent matches to effect the estimated parameters more than matches in the past, but instead of using Dixon and Coles' method for allowing for dependency between the number of goals scored by the two teams, we use a Frank copula to generate a bivariate distribution with Weibull count marginals. Full details of our model can be found in our paper.

2 Betting with the model

We have fitted the model to data five seasons of results from the English Premier League from 2010/11 to 2014/15. The values of the estimated α's and β's aren't really the focus of this article. Suffice to say, they are as you would expect: Manchester City has the best attacking and defensive abilities, with the usual suspects (Manchester United, Chelsea etc.) close behind.

In addition to match results, `football-data.co.uk` also provides the average bookmaker odds for the 1X2 and the over-under 2.5 goals markets and we use these in a 'proper' test of the model: out-of-sample betting with the model.

Our "simulated" betting exercise is as follows:

- Team strengths are estimated using results prior to the match to be bet on. Initially, we use the first four and a half seasons results leaving the final 190 games of the 2014–15 season to bet on.
- After each week of matches have been played, we re-estimate the team strengths by dropping the first week's matches and adding the 'new' week's set of results.

The tradional approach in the literature (Dixon and Coles (1997) and Koopman and Lit (2015) for example) when definifing a betting strategy has been to invest a unit stake on every *'quality'* event. A quality event is defined as an event (say A) with an expected value ($EV(A)$) that exceeds some benchmark threshold t :

$$EV(A) = P(A) \times 1/b(A) - 1 > t,$$

where $P(A)$ is the probability of event A occurring and $b(A)$ is the bookmaker's decimal odds for event A.

Our betting strategy is slightly more sophisticated as we use the difference between the model computed probability and the bookmakers' implied probability to define the size of our stake. In order to do that, we use (none other than!) the Kelly Criterion (Kelly, 1956). Just in case you don't know, the Kelly Criterion says that if you want to maximise long-run log-utility you should invest a fraction f of your overall wealth, given by

$$f = \frac{bp - 1}{b - 1},$$

where p is the bettor's estimate of the probability of an event (e.g. the home team winning the game), and b is the decimal odds offered by the bookmaker (where $1/b$ can be interpreted loosely as the bookmaker's implied probability of the event occurring). We also keep the idea of betting on 'quality' events here, as we still only bet if the $EV(A) > t$. It is worth noting here that choice of t is not straightforward. In fact, choosing a small t will allow the placement of large number of bets (with small stakes in the Kelly framework). On the other hand, choosing a large t will result in a small number of bets with big stakes. This paradox highlights one of the weaknesses of the Kelly criterion: it says "the more you disagree with the bookmaker, the more you should bet". This seems worryingly naive: a small difference between your model's probability and the bookmaker's price may mean that you have evaluated the probability more accurately, however, a big difference is more likely to be indicative of your model not knowing something that the bookmaker does. This is likely to be the case in betting with our model - we have no idea of the team lineups, yet the bookmaker's traders will take this information, and more, into account when setting their odds.

In response to such a possibility, people often use fractional Kelly betting, and/or add a little 'protection' to their betting strategy. The approach adopted in our case is twofold. First, we focus on 'quality'

bets with an intermediate value of t. Experimentation led us to use $t = 0.15$ which is a good compromise between betting too much (and losing) and placing a reasonable number of bets. Second, we reset our bank roll after each bet (to avoid risking too much on a single event especially when our probability is far from the market).

Using this protection rule means that of the 190 games we can bet on, we place 57 bets (34 on the 1X2 market and 23 on the over-under market). Not many actual bets, but when we bet, we want to be sure we are investing our hard-earned money wisely ;-). Overall, we end up with returns (based on total stakes placed) of 4.6% on the 1X2 market and 30% on the over-under market. Not bad, especially when you consider that the 1X2 market has an average over-round of 4.9% and the over-under market has an average over-round of 5.7% which needs to be overcome before you start making a profit. For a comparison, using the Poisson model (with a copula to induce dependence), and using the same investment scheme results in returns of -4.4%.

3 What next?

So the model makes some money. What can be done with the model now? We believe the Weibull count distribution better models the distribution of goals in a football match. But crucially, the model needs a good estimate of the scoring rates of the two teams. With bad estimates of the scoring rates, making a profit is going to be rather difficult to say the least. Maher and Dixon and Coles provide one way of estimating the scoring rates, but there are countless others. For example, it is likely that the market itself can estimate good values of the the scoring rates (market efficiency, wisdom of crowds and all that). If you think this may be true, you could 'backward engineer' odds from the total goals and goal supremacy markets to obtain the 'best' estimates of the scoring rates. These rates can in turn be used in our model to generate a matrix of correct score probabilities.

An alternative to estimating scoring rates based on the identity of the two teams playing, would be to use the identity of the players on each team. This would eliminate the need to estimate team strengths and time-decaying results from past matches. However, you replace one set of complications with another: data. For player-based forecasting models to work, you need to know how good all of the players in the whole league are. Building models for player ratings and using the ratings for forecasting is something we are now working on, and hope to write about in the near future.

References

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926.

Koopman, S. J. and Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):167–186.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.