An introduction to football modelling at Smartodds
Oxford SIAM Conference 2011

Robert Johnson

Smartodds Ltd

February 9, 2011

Introduction

- Introduction to Smartodds
- Practical example: building a football model

What is Smartodds about?

- Smartodds provides statistical research and sports modelling in the betting sector
- Quant team research and implement the sports models
- Primary focus is on Football, however we also model Basketball, Baseball, American Football, Ice Hockey and Tennis
- Wide range of interesting problems to work on
- Actively recruiting!

Building a Football model

- Suppose we decide to build a football model for the English football leagues
- Here we model the divisions Premier League, Championship, League 1 and League 2
- There are 92 teams in total to model
- We want to predict the probability of team A winning against team B where team A and team B could be from any of the 4 leagues




Literature review

- Maher (1982) assumed independent Poisson distributions for home and away goals
  - Means based on each teams' past performance
- Dixon and Coles (1997) took this idea further by accounting for fluctuations in performance of individual teams and estimation between leagues
- Dixon and Robinson (1998) modelled the scores during a game as a two-dimensional birth process

Model formulation

- Assume that home and away goals follow a Poisson distribution

$$Pr(\text{x goals}) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$Pr(\text{y goals}) = \frac{\mu^y e^{-\mu}}{y!}$$

- To estimate the probabilities of x and y goals we need λ and μ




Model 1: Mean goals




- Assume that home and away teams are expected to score the same number of goals
- Take average goals scored in a game in England as 2.56 and divide by two

$$\lambda = 1.28$$

$$\mu = 1.28$$

- However we may believe that there is some advantage associated with playing at home

Model 2: Home Advantage

- Include a term to take account of home advantage

$$\lambda = \gamma \times \tau$$

$$\mu = \gamma$$

- γ is the common mean and τ represents the home advantage

Model 2: Home Advantage (Cont)

- Mean goals scored by the away team in the four leagues we model English Leagues is 1.10 giving

$$\gamma = 1.10$$

- This implies mean goals scored by the home team are $2.56 - 1.10 = 1.46$
- Using the above we can estimate τ as

$$\tau = 1.46/1.10 = 1.33$$




Model 3: Team Strengths

- Previous attempts assumed all teams of equal strength
- Can add team strength parameters for each team
- Better teams score more goals. Give each team an attack parameter denoted α
- Better teams concede fewer goals. Give each team a defence parameter denoted β




Model 3: Team Strengths (Cont)

- Write λ and μ in terms of the attack and defence parameters of the home and away teams, which we denote by i and j, giving

$$\lambda = \gamma \times \tau \times \alpha_i \times \beta_j$$

$$\mu = \gamma \times \alpha_j \times \beta_i$$

- The model is overparameterised, so we apply the constraints

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 1, \; \frac{1}{n}\sum_{i=1}^{n} \beta_i = 1.$$




Model 3: Pseudolikelihood

- The pseudolikelihood for this model is:

$$L(\gamma, \tau, \alpha_i, \beta_i; i = 1, \ldots, n) =$$

$$\prod_k \{\exp(-\lambda_k)\lambda_k^{x_k} \exp(-\mu_k)\mu_k^{y_k}\}^{\phi(t-t_k)}$$

- $\phi(\cdot)$ is an exponential downweighting function, which allows us to place less weight on older games
- Other downweighting functions could be used




Estimation techniques

- Obtaining the parameter estimates is not straightforward
- In this example we have 186 parameters to estimate
- Various optimisation techniques could be used to obtain parameter estimates (numerical maximisation of the likelihood function, MCMC)
- High dimensional problems may also require more sophisticated computing solutions (MPI)




Parameter estimates

- These are Smartodds' current estimates of the attack and defence parameters of the top 6 teams in the Premier League

Team	Attack Parameter	Defence Parameter
Chelsea	3.15	0.34
Man Utd	3.08	0.35
Arsenal	2.84	0.37
Man City	2.44	0.42
Tottenham	2.22	0.44
Liverpool	2.12	0.39




Predicting outcomes

- Suppose Man Utd are playing at home to Man City
- Using the parameter estimates we get

$$\lambda = 1.10 \times 1.33 \times 3.08 \times 0.42 = 1.89$$

$$\mu = 1.10 \times 2.44 \times 0.35 = 0.94$$

- We can use λ and μ to obtain the probability of Man Utd winning the match




Predicting outcomes (Cont)

- The probability of a specific score is given as follows

$$Pr(x,y) = \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^y e^{-\mu}}{y!}$$

- So the probability of the score, Man Utd 2 Man City 1, is

$$Pr(2,1) = \frac{1.89^2 e^{-1.89}}{2!} \frac{0.94^1 e^{-0.94}}{1!} = 0.099$$

Predicting outcomes (Cont)

- Obtain the probability matrix of all possible scores

	0	1	2	3	4	...
0	0.059	0.112	0.105	0.066	0.031	...
1	0.055	0.105	0.099	0.062	0.029	...
2	0.026	0.049	0.047	0.029	0.014	...
3	0.008	0.015	0.015	0.009	0.004	...
4	0.002	0.004	0.003	0.002	0.001	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Predicting outcomes (Cont)

- Sum over all events where home goals are greater than away goals

	0	1	2	3	4	...
0	0.059	0.112	0.105	0.066	0.031	...
1	0.055	0.105	0.099	0.062	0.029	...
2	0.026	0.049	0.047	0.029	0.014	...
3	0.008	0.015	0.015	0.009	0.004	...
4	0.002	0.004	0.003	0.002	0.001	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$




Predicting outcomes (Cont)

- Giving the probability that Man Utd win at home to Man City as 59.6%

	0	1	2	3	4	...
0	0.059	0.112	0.105	0.066	0.031	...
1	0.055	0.105	0.099	0.062	0.029	...
2	0.026	0.049	0.047	0.029	0.014	...
3	0.008	0.015	0.015	0.009	0.004	...
4	0.002	0.004	0.003	0.002	0.001	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Practical issues



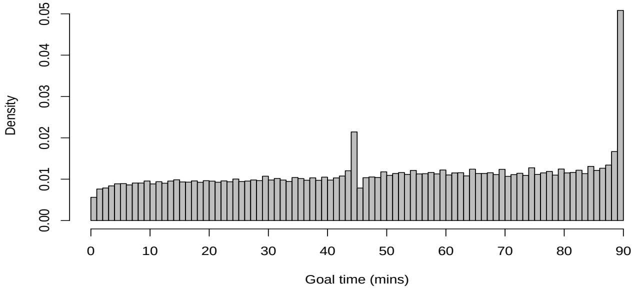

- Betfair's odds imply Man Utd has a 63% chance of winning the game, potentially leaving value for a bet on Man City. However, should we bet?
- These models take into account no external information about match circumstances
  - Injuries
  - Motivation
  - Fatigue
  - Newly signed players
- So betting off a mathematical model would be dangerous!

Shortcomings of the model

- If we compare the expected full-time scores under the model with the observed scores, we find our modelling assumptions don't hold
  - Goals don't have a Poisson distribution
  - Goals scored by the home and away teams aren't independent
- Dixon and Coles corrected for this by modifying the predicted distribution to increase probability of draws and 0-1 and 1-0 scores
- However this isn't entirely satisfactory — would be better to model what is happening directly




Goal time distribution



- Goals in injury time at the end of each half are recorded as 45 / 90 min goals
- Goal rate steadily increases over the course of the game
- Notice the spikes every 5 minutes in the second half - due to rounding?

Dixon and Robinsons' model

- If we assume that the goal scoring processes for the home and away teams are independent homogeneous Poisson processes then our model reduces to the full time model discussed previously.
- For match k between teams i and j

$$\lambda_k(t) = \lambda_k = \gamma \times \tau \times \alpha_i \times \beta_j$$

$$\mu_k(t) = \mu_k = \gamma \times \alpha_j \times \beta_i$$




Dixon and Robinsons' model (continued)

- Three changes:

1. Goal-scoring rate dependent on the current score
2. Modelling of injury time
3. Increasing goal-scoring intensity through the game (due to tiredness of players)

(1) Goal-scoring rate dependent on current score

- Assume that home and away scoring processes are independent Poisson processes
- Scoring rates are piecewise constant
  - Home and away intensities are constant until a goal is scored and only change at these times
- Denote λ_{xy} and μ_{xy} as parameters determining the scoring rates when the score is (x,y)
- Scoring rates are now

$$\lambda_k(t) = \lambda_{xy}\lambda_k$$

- and

$$\mu_k(t) = \mu_{xy}\mu_k$$

Estimates of $\lambda(x, y)$ and $\mu(x, y)$

- $\hat{\lambda}(0,0) = 1$
  $\hat{\mu}(0,0) = 1$
- $\hat{\lambda}(1,0) = 0.88$
  $\hat{\mu}(1,0) = 1.35$
- $\hat{\lambda}(0,1) = 1.10$
  $\hat{\mu}(0,1) = 1.07$

(2) Increase the scoring rate during injury time

- Goals scored during injury time are recorded as having occurred at either 45 or 90 minutes.
- Define two new parameters ρ_1 and ρ_2 to model injury time.
- The adjusted scoring rates are

$$\lambda_k(t) = \begin{cases} \rho_1 \lambda_{xy} \lambda_k & t \in (44, 45]\text{mins}, \\ \rho_2 \lambda_{xy} \lambda_k & t \in (89, 90]\text{mins}, \\ \lambda_{xy} \lambda_k & \text{otherwise} \end{cases}$$

- and similarly for $\mu_k(t)$




(3) Increasing goal-scoring intensity

- Allow the scoring intensities to increase over time
- Model scoring rates as time inhomogeneous Poisson processes with a linear rate of increase
- Replace $\lambda_k(t)$ and $\mu_k(t)$ with

$$\lambda^*_k(t) = \lambda_k(t) + \xi_1 t,$$

$$\mu^*_k(t) = \mu_k(t) + \xi_2 t$$

- ξ_1 and ξ_2 could be constrained to be positive to ensure that the hazard functions above are constrained to always be positive, but in practice this is not neccessary
- Scoring rates are estimated to be about 75% higher at the end of the game then at the start of the game.




Model usage

- This ‘in-running’ model can be useful in its own right (for deriving in-running prices)
- Also explains the home/away dependencies and non-Poisson pdfs observed in the data




Summary

- The Dixon-Coles model is a simple and robust full-time score model, but not all of its assumptions are met
- A continuous time model such as the Dixon-Robinson model can model dependencies between home and away scoring rates
- Mathematical models cannot model team news (unless this is incorporated into the model somehow)
- These models can be extended to other sports by changing the distributions, eg
  - Normal distribution for American Football
  - Negative binomial for baseball

References

- M.J. Maher, 1982, Modelling association football scores, *Statist. Neerland.*, 36, 109-1188
- M. Dixon and S.G. Coles, 1997. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Applied Statistics*, 46(2), 265-280
- M. Dixon and M. Robinson, 1998. A birth process model for association football matches. *JRSS D*, 47(3), 523-538

Interested?

- If you are interested in sports modelling and possess the following skills:
  - Post graduate qualification (at least MMath / MSc, PhD. preferred) in mathematics, statistics or another subject with considerable mathematical content
  - Experience in developing and implementing mathematical / statistical models
  - Experience of computer programming (preferably in C++, C, R or Python)
  - Enthusiasm, self-motivation and the ability to work under pressure to strict deadlines
- Then email us at careers@smartodds.co.uk
- For more information see our website: http://www.smartodds.co.uk