Statistical Modelling in Sports

Martin Andla Supervisor George Foulds
Lancaster University

Introduction

Freely available football data will be used from:

- English Division 1,
- seasons 1993/1994 to 2012/2013.

Dixon and Coles (1997) proposed a Poisson model to estimate the probabilities of football results. In their model they included:

- attack and defence parameters for both the home and away teams;
- home advantage.

The Dixon and Coles model will be altered to see if attendance figures have any effect on home advantage.

What is home advantage?

It is the positive effect experienced by a player or a team playing at home.

Clarke (1996) proposed a non-parametric model for calculating home advantage of team i:

$$h_i = \frac{\text{team i's home goal diff} - \text{team i's away goal diff} - H}{N-2}, \qquad (1)$$

where

- $H = \frac{\text{total home goal difference of all teams}}{N-1}$,
- N = total number of teams in a league.

Using Equation 1 it can be shown that average home advantage in English League 1 for seasons 1993/1994 to 2012/2013 was **1.35** goals per match.

It can also be shown that home advantage has slightly decreased over the years (Figure 1), however the reasoning behind this remains unclear.

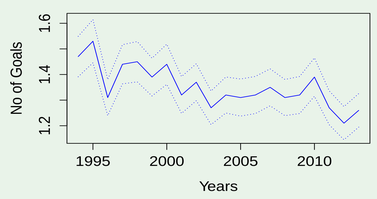

Figure 1: Mean home advantage (with standard error confidence interval)

Can we use a Poisson distribution?

Informally, frequencies for both home and away scores can be plotted and compared with data drawn from a Poisson distribution (Figure 2):

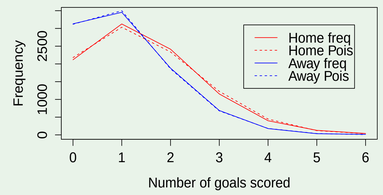

Figure 2: Verifying Poisson assumption

More formally, setting

H_0 : match scores follow Poisson distribution,
H_1 : match scores do not follow Poisson distribution,

and testing the hypothesis using the Pearson's χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where

- O_i = observed score by team i;
- E_i = expected score by team i;
- n = number of teams in a league

yields that p-value = 0.2642 and 0.2372 for home and away teams respectively. Thus the data can be assumed to follow a Poisson distribution.

Are the match scores independent?

In order to assess the validity of this assumption, Table 1 displays:

$$\frac{\tilde{p}(i,j)}{\tilde{p}_H(i)\tilde{p}_A(j)} \qquad (2)$$

for each home and away score (i,j), where $\tilde{p}$,$\tilde{p}_H$ and $\tilde{p}_A$ are the joint and marginal empirical probabilities for home and away scores respectively.

Home goals (i)	*Away goals (j)* 0	1	2	3	4
0	1.099	0.942	0.959	0.948	0.996
1	0.968	1.063	0.946	0.992	0.961
2	0.947	1.004	1.051	1.058	1.070
3	1.006	0.912	1.129	1.078	0.960
4	0.939	1.028	1.063	0.924	1.189

Table 1: Results of Equation 2

The values in Table 1 are close to 1, so the assumption of independence between scores is reasonable. Now a model can be formulated.

Introducing the model

The basic tool in finding out whether attendance figures have an effect on home advantage will be the likelihood function

$$L(\alpha_i, \beta_i, \gamma) = \prod_{k=1}^{N} e^{-\alpha_{i(k)}\beta_{j(k)}\gamma}(\alpha_{i(k)}\beta_{j(k)}\gamma)^{x_k} e^{-\alpha_{j(k)}\beta_{i(k)}}(\alpha_{j(k)}\beta_{i(k)})^{y_k},$$

where

- x_k is the number of goals scored by home team in match k;
- y_k is the number of goals scored by away team in match k;
- $\alpha_{i(k)} > 0$ are the measure of attack rates;
- $\beta_i(k) > 0$ are the measure of defence rates;
- $\gamma > 0$ measures home advantage in match k between team i and team j.

Also:

$$\gamma = \begin{cases} \gamma_1 \text{ if } a_k \geq A \\ \gamma_2 \text{ if } a_k < A \end{cases},$$

where a_k is crowd size in match k and $A \in \mathbb{N}$.

Conclusion

Plotting the ratio γ_1/γ_2 for A $\in (7100, 30000)$ (Figure 3) implies that:

- crowd sizes less than 16500 have little effect on home advantage, indicated by the random fluctuations of the ratio;
- crowd sizes larger than 16500 positively affect home advantage, indicated by the incline in the ratio.

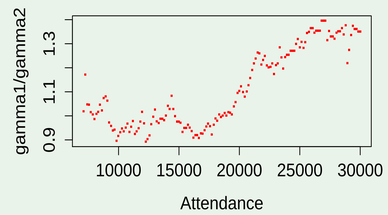

Figure 3: Effect of crowd size on home advantage

Thus, based on Figure 3, it is possible to claim that increases in the attendance figures have a positive effect on home advantage. However, this should be verified using less ambiguous statistical methods and also data from other divisions.

References

Dixon, M.J. and Coles, S.G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market.
In *Applied Statistics*, **46**, 2 (1997):265-280.

Clarke, S.R. (1996). Home advantages in balanced competitions: English soccer 1991-1996.
Proceedings of the 3rd Australian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland (1996): 111-116.

andla@exchange.lancs.ac.uk