The Bivariate Poisson Distribution and its Applications to Football

Author: Gavin Whitaker. Supervisors: Dr. P. S. Ansell & Dr. D. Walshaw.

Newcastle University
School of Mathematics and Statistics

Introduction

Football: described by many as the worlds most unpredictable game. Here we look to see if we can use mathematics to predict football matches. We look at the bivariate Poisson distribution and how it simulates results. We specifically look at simulating the 2009/2010 Premier League.

The Bivariate Poisson Distribution

The bivariate Poisson distribution can be derived by taking the limit of the bivariate Binomial distribution. For two random variables X and Y it has joint probability function

$$G_{XY}(t_1,t_2) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{i=0}^{\infty} \frac{\lambda_1^i t_1^i}{i!} \sum_{j=0}^{\infty} \frac{\lambda_2^j t_2^j}{j!} \sum_{k=0}^{\infty} \frac{\lambda_3^k t_1^k t_2^k}{k!}.$$

It is clear through the summations that this is just a natural extension of the univariate Poisson distribution. The R package "Bivpois" was developed to analyse the bivariate Poisson distribution. The package has been used for general simulation, to model the demand for health care in Australia, to model water polo games and to model football matches; the last case was specifically used to model the 1991-1992 Italian Seria A season. We use the package to calculate probabilities from which we can simulate.

The Model

To predict the number of goals in a football match we use the bivariate Poisson distribution, where X is the number of goals scored by the home team and Y is the number of goals scored by the away team in a single match. Given a single match with team i playing at home and team j playing away we fit the model proposed by Karlis and Ntzoufras, (2003):

$$(X,Y) \sim BVP(\lambda_1,\lambda_2,\lambda_3)$$

where

$$\log(\lambda_1) = \mu + (\text{attack}_i) + (\text{defence}_j) + (\text{home effect})$$
$$\log(\lambda_2) = \mu + (\text{attack}_j) + (\text{defence}_i).$$

μ is the mean level of goals scored, attack and defence are the attack and defence parameters for a specific team and home effect is the advantage of playing at home. Note λ_3 is determined using "Bivpois."

Home Effect

We have included a home effect in our model but, "Is there a home effect in the Premier League?" Table (1) shows the points scored at home by every team over the 2009/2010 Premier League season, a team can obtain a maximum of 57 points.

Team	Points at home	Team	Points at home
Chelsea	51	Aston Villa	24
Man United	48	Birmingham	24
Arsenal	45	Burnley	21
Tottenham	42	Stoke	21
Liverpool	39	West Ham	21
Man City	36	Bolton	18
Everton	33	Hull	18
Fulham	33	Wigan	18
Blackburn	30	Portsmouth	15
Sunderland	27	Wolves	15

TABLE 1: Home points over the 2009/2010 season.

It is clear that the better teams, i.e. Chelsea obtain more points at home than the poorer teams, i.e. Wolves; this however gives no indication of a home effect, only that some teams are better than others. Consider Fulham, who got 33 of their 46 points at home, or Sunderland who got 27 of their 44 points at home. Both these teams got a large proportion of their points at home, and it was ultimately their home form that kept both these teams safely in the Premier league. On the evidence of these two teams it is clear that there is a home effect in the Premier League and it is needed in our model.

Predicting the Premier League

Table (2) shows the final table of results based on 100 simulations of the Premier League. We have taken the means so that the table reflects an average season. The simulated table is very close to the 2009/2010 Premier League table. It captures teams at the top and bottom of the league, however there are some discrepancies of order around the middle of the table but this is mainly due to how similar some teams are. Our simulations suggest that LIV dramatically underperformed over the season and a 7th place finish was a poor result. A team's average goals scored and conceded in the simulated table is reasonably accurate, so the model appears to have captured the rate at which teams attack and defend. The model seems to inflate the number of goals over a season slightly meaning that if we used it to predict scores then we would expect more goals than we would observe; this suggests that this model is a good predictor of results but perhaps not perfect scores.

Team	Points	Games Won	Games Drawn	Goals	Conc	Goaldif
Chelsea	90.5	28.44	5.18	111.54	36.86	74.68
Man United	88.4	27.37	6.29	95.82	31.34	64.48
Arsenal	78.35	23.97	6.44	91.29	44.72	46.57
Liverpool	71.23	20.89	8.56	67.32	39.29	28.03
Tottenham	71.1	20.99	8.13	75.63	45.96	29.67
Man City	70.14	20.78	7.8	77.08	48.16	28.92
Aston Villa	63.78	718.11	9.45	58.4	42.59	15.81
Everton	61.78	17.9	8.08	66.24	52.19	14.05
Sunderland	48.78	13.27	8.97	52.12	60.72	-8.6
Fulham	48.07	12.62	10.21	44.46	50.48	-6.02
Birmingham	46.46	12.05	10.31	41.58	51.14	-9.56
Blackburn	42.86	11.37	8.75	45.46	60.97	-15.51
West Ham	42.8	11.57	8.09	53.12	71.32	-18.2
Stoke	41.31	10.42	10.05	36.95	54.48	-17.53
Bolton	38.3	9.92	8.54	46.85	73.6	-26.75
Wolves	35.06	8.43	9.77	33.72	60.83	-27.11
Burnley	31.81	8.2	7.03	46.48	89.14	-42.66
Portsmouth	30.91	7.48	8.47	36.99	73.89	-36.9
Wigan	27.92	6.82	7.46	40.95	87.65	-46.7
Hull	27.65	6.55	8	36.28	82.95	-46.67

TABLE 2: The Premier League based on 100 simulations.

Estimating Parameters

Table (2) was simulated using attack and defence parameters estimated over a season, however we are interested in including a team's form in the model. A team's form describes how a team are playing at that point in the season, and it can go up and down depending on how well the team is doing. Adding a team's form into the model will allow us to accurately model how a team's attack and defence changes over the season. This should provide us with more accurate results as we have included how a team performance changes over the season. Figures (1 & 2) show moving averages for the parameters over the season obtained using 100 games and a time step of 20 games.

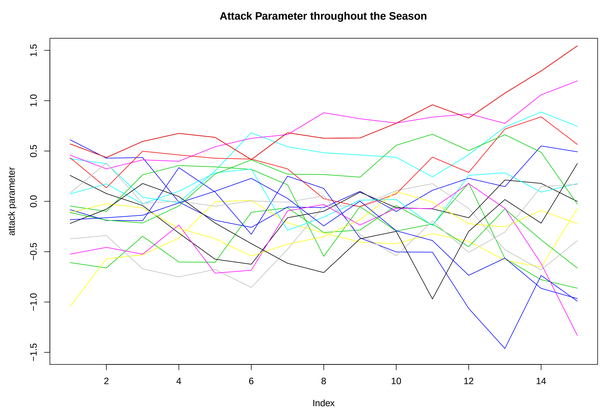

FIGURE 1: Moving average of the attack parameters.

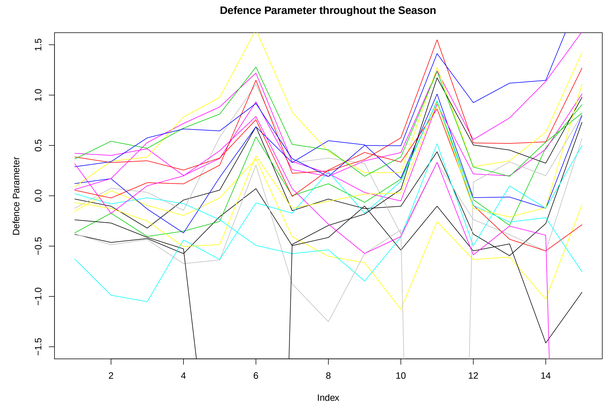

FIGURE 2: Moving average of the defence parameters.

Looking at Figure (1) we see that the right half has more variation than the left. This is due to the January transfer window; this window is regarded as expensive and so only the better clubs with more money can buy new players, meaning that the better teams get better whilst some of the poorer teams loose their key players. In Figure (2) there is not as much variation and the transfer window does not have the same effect. This is most likely due to the fact that defending is a team aspect whereas goals are scored by individuals; hence it is easier to replace a good defender in a team than it is a good attacker. There are increases around 6 and 11 for nearly all teams, this is because the parameters sum to 0.

By including these changing estimates and simulating the fixtures in the order they occurred during the season we can gain more accurate results. We should be able to capture how a team's attack and defence changes throughout a season and our results should reflect these changes.

Conclusion

Considering the results we have observed it appears we are able to simulate an entire season reasonably accurately; we can capture its final standings and model its general trends, such as a team's points or goals scored. The problems arise when we consider a specific match. We can predict the match result but it is more difficult to obtain an accurate scoreline, including the moving average estimates would hopefully rectify this.

Reference: Karlis, D. and Ntzoufras, I. (2003), Analysis of Sports Data Using Bivariate Poisson Models, Journal of the Royal Statistical Association, D (Statistician), 52, 381-393