STATISTICAL MODELLING FOR SOCCER GAMES: THE GREEK LEAGUE

Dimitris Karlis and Ioannis Ntzoufras
Department of Statistics, Athens University of Economics and Business
[e-mail for correspodence: jbn@stat-athens.aueb.gr]

Abstract

Nowadays, the sports related industry increases dramatically and vast amounts of money are spent on such activities. Betting on the results of athletic competitions is very popular in all countries of the European Community and the electronic facilities allow participation in such bets from all over the world. In this paper we explore the possibility of developing statistical models for predicting the outcome of a soccer game. Some results based on careful examination of a large number of games are presented and used for modelling soccer data of the Greek league for season 1997-98. These models are used for interpretation of team performances and prediction of the results in future games.

Keywords: League Soccer Data; Poisson distribution; Home Effect; sport statistics; independence.

1 Introduction and Review

Sports is a blooming field for applying statistical methodologies as well as a platform for developing suitable methods for dealing with athletic data. Moreover, increased amounts of money are invested by sports related industries which have extremely grown over the last decades. Sports clubs have become well organized companies that invest large amounts of money every year. For example, a soccer team can earn a lot of money by simply participating in the 'European Champions League' competition. Thus statistics can help managers to make crucial decisions in certain circumstances and provide guidance that is not available without scientific search in large data sets.

Betting on the outcome of soccer matches has a long tradition in the UK and other European countries. Football pools typically involve the selection of matches whose the outcome is hard to predict. Such bets involve either the final score or other specific characteristics of the match, such as halftime result or the scorer of the first goal. In making bets, the challenge is to trace matches the probabilities of which have been determined inaccurately and thus the expected gain is high. Statistical models can be helpful tools for such purposes.

A lot of papers have been published presenting statistical methods for sports data, covering a great variety of sports, including association football (hereafter soccer), American football, basketball, baseball, ice hockey, skating and several other athletic competitions. There are many articles analysing soccer data and even more focused on other sports, but their methodology can be easily extended in order to be implemented in soccer data. Interesting results concerning soccer can be found in Maher (1982), Baxter and Stevenson (1988), Ridder *et al.* (1994), Jackson (1994), Fahrmeir and Tutz (1994), Clarke and Norman (1995), Dixon and Coles (1997), Lee(1997), Pollard and Reep (1997), Rue and Salvesen (1997), and Kuonen (1997a,b).

The remaining of the paper is organised as follows. Section 2 addresses some usual questions related to statistical modelling of soccer games. Section 3 describes the structure of soccer data used in this paper while section 4 proposes a loglinear formulation which enables model checking and model selection. These models are applied in the Greek League data in section 5. Concluding remarks can be found in section 6.

2 Problems Related to Statistical Modelling of Soccer Data

A fundamental question when modelling the number of goals scored by a team is whether the underlined distribution can be assumed to be the Poisson distribution. The Poisson distribution has a formal theoretical basis and is naturally used for events that occur randomly at a constant rate over the observed time period. In our case, we assume that the ability of a team to score a goal is constant throughout the championship. This assumption is restrictive since the ability as well as the composition and the physical conditions of each team vary with time. The assumption of varying scoring ability leads to mixed Poisson distributions with the negative Binomial as the first choice.

The Poisson distribution has the unique property that the mean is equal to the variance. We calculated the index of dispersion (that is the variance to the mean ratio) for 456 teams participated in 24 championships of different Eurpoean countries, including Germany, Spain, England, Italy, France and Nederlands. Of the 456 teams, 58.3% have an index of dispersion greater than one. This result is statistically significant using a sign test. If the Poisson assumption were valid we would expect that half of the teams would have an index of dispersion greater than one. The 95% confidence interval given by (0.538, 0.628) shows that the proportion of teams having a dispersion index higher than one is significanlty greater than 0.5. We can draw similar conclusions by examining the number of goals scored by the opponents of each team. Again, the 58.1% of teams show overdispersion, which is significantly different from 50%. This strongly implies that the distribution of the number of goals is overdispersed relative to the simple Poisson distribution.

However, the overdispersion is relatively small. The upper 95% percentile of the distribution of the index of dispersion is 1.55. This shows that the deviation from the Poisson distribution is not of much concern. Differences in winning probabilities between Poisson and negative Binomial distributions are minimal for this range of overdispersion.

To conlcude with, the distribution of goals is slightly overdispersed relative to the Poisson assumption. Hence, using the Poisson distribution will lead to similar results while the computational gain is high. Moreover, there is not any overdispersion pattern supporting a particular mixed Poisson distribution.

Another question which naturally arises in a same context is whether the number of goals scored by the two opponents in the same game are independent. For each championship, a χ^2 test was performed to examine possible dependencies. In 15 out of 24 cases the independence assumption was not rejected. If we combine the results of all the championships in one overall χ^2 test, the null hypothesis is rejected. However, the rejection of the independence hypothesis was rather an artifact due to the large sample size. Spearman's correlation coefficient was highly significant even at a 0.001 level of significance, but its value was considerably small, namely 0.03, revealing that there is no strong dependence between the two variables.

Moreover, the following result was derived. Suppose that the random variates X and Y represent the number of goals scored by each team, and that (X, Y) jointly follow a bivariate Poisson distribution (Kocherlakota and Kocherlakota, 1992). Then, the probability of the outcome (win,draw or loss) does not depend on the amount of correlation between X and Y. A proof of this result can be seen by calculating the probability of a win of the first team. This probability is equal to $P(X > Y)$. After tedious algebraic manipulations the resulting expression does not contain the covariance parameter of the bivariate Poisson distribution. The above result implies that dependence is of no concern for the outcome of a game and therefore can be ignored in modelling. The above results are used for modelling purposes in the sequel.

3 Greek League Soccer Data

Data referring to the season 1997-98 of the Greek First Division are considered. In general, soccer data form a kind of a three-way 'contingency table' with counts of the goals scored by team A, against team B, playing in ground C. The factors used in the model are the scoring team A (determining the offensive parameters), the team B against which these goals are scored (determining the defensive parameters) and the home effect (C). This 'contingency table' must be cautiously utilized since it involves zero counts and structural zeros (in the diagonal of scoring and defending teams).

Greek First Division league has 18 teams playing with each opponent twice, once at home and once in away football grounds. Each team plays 34 games, 17 in home and 17 away. The final league consists of 306 soccer games. Every win gives three points to the winner and every draw one point to each opponent. The team which collects the highest number of points is the winner of the league. Positions 2-4 are of crucial interest since they give the right of playing in European cups such as 'Champions league' and UEFA cup. Finally, the three teams which collect the lowest number of points are relegated in the lower division and are substituted in the next season by the three best teams of the second division.

Greek league data were taken by International Soccer Server web-site available in the URL address `http://sunsite.tut.fi/rec/riku/soccer2.html`. Many interesting soccer data and statistics are also available in `http://www.risc.uni-linz.ac.at/non-official/rsssf/`.

4 Poisson Model Formulation

The simplest type of model that we may use is the Poisson log-linear model. Due to the large number of zeroes, the chi-square approximation of the deviance is not suitable. The full Poisson model has the following form

$$n_{ijk} \sim Poisson(\lambda_{ijk}),$$

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik} + a.d_{jk} + h.a.d_{ijk},$$

where n_{ijk} and λ_{ijk} are the observed number and the mean, respectively, of the goals scored by team j, playing against team k, in football ground i (away/home); μ is a constant parameter, h_i is the home effect parameter, a_j encapsulates the parameter for the offensive performance of j team and d_k the defensive performance of k team. The dots indicate the interaction term, e.g. $a.d$ denotes the interaction between offensive and defensive parameters. The rest of the parameters can be interpreted accordingly. The full model implies that the offensive and defensive abilities vary in each game depending on the playing ground, the scoring and defending ability of the competing teams. Such a model is not useful for prediction since we need full league data (which are not available) to estimate these parameters. Additionally, data of previous years may not reflect performances in present time and estimation of these parameters is problematic.

Two simpler candidate models are of great interest. The first model is given by

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k. \tag{1}$$

This simple model assumes that the offensive and defensive performances are the same for home and away games and that the home effect is the same for all teams. In this model we use sum-to-zero constraints on offensive and defensive parameters and corner constraints on the playing ground with baseline level the away grounds. Therefore, we have $\sum a_j = \sum d_k = h_1 = 0$; where h_1 indicates goals scored by away teams. This parametrization implies a straightforward interpretation of the model parameters: μ is the average of log-mean of goals scored in away games, h_2 is the constant home effect while a_j and d_k are the offensive and defensive performances of j and k teams respectively, expressed in deviations from μ. It is obvious that the greater the offensive parameter the better the offensive performance of the corresponding team, while the lower the defensive parameter the better the defensive performance of the corresponding team.

This model was also used by Lee (1997) and has the great advantage that can be used for prediction in games played on a neutral ground. The second model is given by

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik}. \tag{2}$$

The motivation for using this more complicated model is the plausible assumption that the offensive and defensive abilities of each team differ between home and away games. In (2) we use the same constraints as in Model 1 and thus we have

$$\sum_{j=1}^{p} a_j = \sum_{k=1}^{p} d_k = 0, \quad h_1 = h.a_{1j} = h.d_{1k} = 0, \quad \sum_{j=1}^{p} h.a_{2j} = \sum_{k=1}^{p} h.d_{2k} = 0.$$

This parametrization facilitates an easy to use interpretation of model parameters similar to Model 1 implying that μ is the average of the log-mean of goals for away games, h_2 is the difference of the average of the log-mean of goals of home games from away games, a_j is the away offensive ability of team j expressed in deviations from μ, d_k is the away defensive ability of k team expressed in deviations from μ and $h.a_{2j}$ is the home-away difference in the offensive abilities of team j, and $h.d_{2k}$ is home-away difference in the defensive abilities of k.

5 Application to Greek Data

5.1 Model Selection: Are the Offensive and Defensive Performances of Each Team Different in Home and Away Games?

Initially, some tests are performed to check whether the goals scored by the two opponents in the same game can be assumed independent. A crosstabulation of home and away games truncating at 4 goals resulted in an approximate p-value of 0.34 for Peason's χ^2 test of independence not rejecting the independence assumption. The Spearman correlation coefficient was -0.056, a rather small value which was not significant.

Our main interest lies in selecting the best model from the models described in section 4. The model selection procedure including deviance and AIC statistics are presented in Table 1. Using the Akaike criterion (Akaike, 1973) we select model (1) as the best. Moreover, the approximate χ^2 likelihood ratio test rejects the hypothesis that $h.a$ and $h.d$ interactions are significant terms for the model and therefore all results fully support model (1). Therefore, more weight is given to Model (1). However, some results of Model (2) are also presented but differences from Model (1) are minimal.

	Model	Removed Term	Deviance	D.F.	P.value for fit	P.value for change	AIC
1	`Saturated`		000.00	000			1224.0
2	`H*A+H*D+A.D`	`-H.A.D`	340.57	271	0.0026	0.0026	1022.6
3	`H*A+H*D`	`-A.D`	652.41	542	0.0008	0.0444	792.4
4	`H*A+D`	`-H.D`	673.49	559	0.0006	0.1920	779.5
5	`H+A+D`	`-H.A`	699.22	576	0.0003	0.0802	771.2
[a]	`A+D`	[5]`-H`	747.61	577	<0.0001	<0.0001	817.6
[b]	`H+A`	[5]`-A`	822.43	593	<0.0001	<0.0001	860.4
[c]	`H+D`	[5]`-D`	744.03	593	0.0002	0.0003	782.4

Table 1: Model Selection Details

5.2 'Who is the Best' Analysis

In this section we examine which team was the best in terms of performance according to the model of constant home effect. The estimated parameters are used to generate replications of leagues.

For each replication, the total points gained by each team were calculated, as well as the rank of each team in this replicated league. This analysis will account for corrections of games that were surprisingly unfair or won by luck.

According to the estimated model parameters, Panathinaikos had a slightly better attack and defence than Olympiakos which won the league. On the other hand, Athinaikos, Kalamata and Ethnikos had the worst attack. The three teams with the worst defence were Panahaiki, Kavala and Proodeutiki.

We generated 10,000 leagues using the estimates of Model 1 and 2. Both models support that

		Observed	Av.Points		Observed	Model 1 Parameters	
	Team	Points	Model 1	Model 2	Goals	Offensive	Defensive
1	Olympiakos	88	78.9	79.4	88-27	0.683	-0.465
2	Panathinaikos	85	81.5	82.1	90-24	0.702	-0.581
3	AEK	74	66.2	67.0	61-30	0.319	-0.394
4	PAOK	70	65.6	66.4	74-41	0.527	-0.064
5	Ionikos	62	57.2	57.7	46-30	0.038	-0.380
6	Iraklis	51	50.5	49.2	49-45	0.118	-0.003
7	OFI	49	44.0	44.3	45-33	0.043	0.157
8	Skoda-Xanthi	45	48.6	48.3	52-52	0.187	0.146
9	Veria	42	41.8	41.2	38-48	-0.132	0.048
10	Paniliakos	36	40.9	40.9	41-54	-0.049	0.170
11	Panionios	36	40.8	39.5	41-54	-0.049	0.170
12	Apollon	36	39.6	39.0	37-51	-0.156	0.108
13	Kavala	35	38.5	38.6	40-58	-0.069	0.241
14	Proodeutiki	34	35.4	35.3	35-57	-0.204	0.217
15	Ethnikos	33	32.3	33.1	27-51	-0.472	0.095
16	Panahaiki	32	29.1	29.9	29-62	-0.386	0.294
17	Kalamata	29	32.8	32.7	27-56	-0.473	0.075
18	Athinakos	27	27.4	28.0	23-55	-0.627	0.166

Table 2: Data and Model Details (Home Constant=0.428, Away Constant=-0.058 and Home Effect=0.486).

	Team	1	1.5	2	2.5	3-4	4.5	5-10.5	10-15.5	16-18
1	Olympiakos	33.8	4.4	47.8	2.4	11.2	0.2	0.3		
2	Panathinaikos	58.2	4.7	29.6	1.5	5.9	0.1	0.1		
3	AEK	1.6	0.5	7.3	1.9	69.3	2.6	16.8		
4	PAOK	1.2	0.3	5.9	1.8	69.5	3.2	18.1		
5	Ionikos	0.1	0.1	0.6	0.2	22.8	3.6	70.9	1.4	
6	Iraklis			0.1		5.5	1.4	84.1	8.5	0.3
7	OFI					0.7	0.2	65.6	30.7	2.8
8	Skoda-Xanthi					3.2	0.9	81.8	13.3	0.7
9	Veria					0.4	0.3	54.0	40.0	5.4
10	Paniliakos					0.2	0.1	48.3	43.7	7.6
11	Panionios					0.3	0.1	47.6	44.6	7.4
12	Apollon					0.2		41.1	48.6	10.1
13	Kavala					0.1	0.1	34.6	52.8	12.4
14	Proodeutiki							19.5	56.0	24.3
15	Ethnikos							9.7	48.7	41.4
16	Panahaiki							3.7	34.3	62.0
17	Kalamata							10.8	50.5	38.7
18	Athinakos							1.9	25.4	72.7

Table 3: Probability of Ranks of Model 1

Panathinaikos (2nd in rank) had a higher expected value of points than Olympiakos which won the league. This can easily be interpreted since Panathinaikos had a better defence and attack but lost important games with Olympiakos and AEK (3rd in rank). Differences between the two models are minimal. Both models indicate that Panathinaikos is the best team with average difference of 2.5 points. Both models also indicate that Xanthi was better than OFI although OFI collected four points more than Xanthi. Model 1 indicates that Ethnikos, Panahaiki and Athinaikos are the three worst teams while Model 2 indicates Kalamata, Panahaiki and Athinaikos (the 3 relegated teams). In both models the differences between Ethnikos and Kalamata are very small. Note that the average goals of each team from the simulated results of Model (1) are very close to the actual results. The results of Model (2) are ommited since they are almost identical to those of Model (1).

Table 3 gives the distribution of ranks for Model 1. The results for model 2 are similar. Half ranks imply that there was a tie between two teams and thus both teams were assigned this half rank. According to this table, Panathinaikos had a higher probability to win the league than Olympiakos. The rest of the resulted probabilities are consistent with the observed data. Four teams had a probability of relegation higher than 30%: Athinaikos, Panahaiki, Ethnikos and Kalamata. Three of them have finally been relegated.

6 Discussion

In this article, some models for soccer games were discussed. Adopting a log-linear formulation one can examine several models at the same time, allowing for interactions between the different parts of the model. The results of section 2 reveal that assuming independent Poisson distributions suffices for describing the game. However, improvements of the model can be considered by truncating the number of goals or adding some other covariates that account for psychological factors.

Another relevant problem, that of using such models for betting purposes, has not been attempted in this paper. The derived models can be useful devices for this purpose. However, due to the incosistency of football bets (outcomes with small probability give small returns) refined models must be developed.

References

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, 267–281, Budapest: Akademia Kiado.

[2] Baxter, M. and Stevenson, R. (1988). Discriminating Between the Poisson and negative Binomial Distributions: An Application to Goal Scoring in Association Football. *Journal of Applied Statistics*, **15**, 347-438.

[3] Clarke, S.R. and Norman, J.M. (1995). Home Ground Advantage of Individual Clubs in English League. *The Statistician*, **44**, 509-521.

[4] Dixon, M.J. and Coles, S.G. (1997). Modelling Association Football Scored and Inefficiencies in Football Betting Market. *Applied Statistics*, **46**, 265-280.

[5] Fahrmeir, L. and Tutz, G. (1994). Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison System. *Journal of the American Statistical Association*, **89**, 1438-1449.

[6] Jackson, D.A. (1994). Index Betting on Sports. *The Statistician*, **43**, 309-315.

[7] Kocherlakota S. and Kocherlakota K. (1992). *Bivariate Discrete Distributions.* Marcel and Decker, NY.

[8] Kuonen, D. (1997a). Statistical Models for Knock-out Soccer Tournaments. *Technical Report*, Department of Mathematics, Chair of Applied Statistics, Ecole Polytechnique Federale De Lausanne.

[9] Kuonen, D. (1997b). Modelling the Success of Football Teams in the European Championships (in French). *Technical Report*, Department of Mathematics, Chair of Applied Statistics, Ecole Polytechnique Federale De Lausanne.

[10] Lee, A.J. (1997). Modeling Scores in the Premier league: Is Manchester United Really the Best? *Chance*, **10(1)**, 15-19.

[11] Maher, M.J. (1982). Modelling Association Football Scores. *Statistica Nederlandica* **36**, 109-118.

[12] Pollard, R. and Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *The Statistician*, **46**, 541-550.

[13] Ridder, G., Cramer, J.S. and Hopstaken P. (1994). Down to Ten: Estimating the Effect of a Red Card. *Journal of the American Statistical Association*, **89**, 1124-1127.

[14] Rue, H. and Salvesen, O. (1997). Predicting Soccer Matches in a League. *Technical Report*, Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway.