Modeling and simulating association football scores for the Greek Superleague

Thomas Uriot

June 28, 2021

Abstract

The static Poisson model developed by Dixon and Coles (1996) is used in to predict the outcome of the 2015-2016 season of the Greek Superleage. The aim is then to compare the results to the actual rankings at the end of the season. Scores of the past 2 seasons were recorded and a Monte Carlo approach was used to simulate the outcome of the league a thousand times. The probabilities are then computed. It is shown that the very bottom and top of the league are almost known with certainty while the mid-table is very homogeneous. The static model fails to account for the dynamics of the sport and a dynamic model may be more appropriate.

Keywords: Greek Superleague; Football(soccer); Poisson distribution; Static model

1 Introduction

Nowadays, with the use of technology, betting has never been easier. As a result, the betting industry has been steadily growing and is now worth around £435bn to £625bn a year, out of which 70% comes from betting on football [1]. Being able to come up with reliable and accurate odds is very important for the betting companies, as it is the core of their business. They need to build statistical models to predict the betting outcomes such that it is not possible to find a positive expected return. The number of sports represented and type of bets available is enormous. For instance, one can bet on whether the total points scored by both teams in a basketball match is odd or even. This relies purely on luck and one does not need any particular knowledge or betting talent to try his luck. In other words, it can not be statistically modeled. In this article, the focus is on predicting football scores using a static Poisson model proposed by Dixon and Coles (1997). The number of goals scored by a team is modeled by a Poisson with mean depending on the attack rate of the team, the defense rate of the opposition and a home effect if the team plays at home. There are generally two classes of statistical model: static and dynamic. A static model does not account for in games events such as red cards or events happening before the game such as injuries or current form.

The model used from Dixon and Coles is based on a paper by Maher (1982). This paper by Maher is one of the earliest attempt to model football scores in a match between specific teams, accounting for the quality of the teams involved. Maher assumed independent Poisson distributions for the number of goals scored by each team in a game. However, some authors have rejected the Poisson model in favor of the Negative Binomial distribution. Reep, Pollard and Benjamin (1971) confirmed what has been first demonstrated by Moroney (1951) that the Negative Binomial gives a better fit than the Poisson

[1] http://www.bbc.co.uk/sport/0/football/24354124

for the number of goals scored by a team in a match. One of the advantage of the Negative Binomial is that is allows for different mean and variance. Furthermore, Reep and Benjamin (1968) observed that chance can be more important than skills for a single game, while over the season, the most skilled team will be likely to finish on top. This is particularly true for football since there are few goals scored per game and random events such as red cards or referring mistakes can affect the outcome of a game.

As mentioned earlier, this article focuses on predicting football scores using the static Poisson model formulated by Dixon and Coles. While the static model is not very good to predict individual football scores, due to random in games events and many changing factors that may affect the result, it is a solid model in order to predict the long run, such as the ranking of the teams after a season. While it is expected to find the same teams at the very bottom and top of the league after a large number of simulations, the mid-table ranks may not be very well predicted due to homogeneous attack and defense rate. In this paper, the ability of the Poisson static model to predict the long run football ranking will be tested. The test will be performed on the Greek Superleague, the highest professional association football league in Greece and spans the period from 17 August 2013 to 4 January 2016. In total, 762 games are used to estimate the attack and defense rates of the 22 teams playing during the period. The probability of each team winning the 2015-2016 championship will be calculated, including its standard error.

2 Data structure and modeling

The Poisson model arises as an approximation to the binomial when the number of trials is very large and the probability of success p is sufficiently small. Indeed, in a football game, each team has a very large number of ball possession and with each possession there is a very small probability of scoring. And so the number of goals scored for each team in a game is modeled by two Poissons with different means. Suppose that the mean of the Poisson for the home team is made of three variables: its attack rate, the opponent's defense rate and something called the 'home effect'. The 'home effect' is simply due to the observation that teams tend to perform better at home due to various factors. Figure 1 shows the distribution of away and home goals over the period the data was recorded. It clearly show that, indeed, there seem to be an 'home effect' as the number of goals scored by the home team is usually larger than the number of goals scored by the away team.

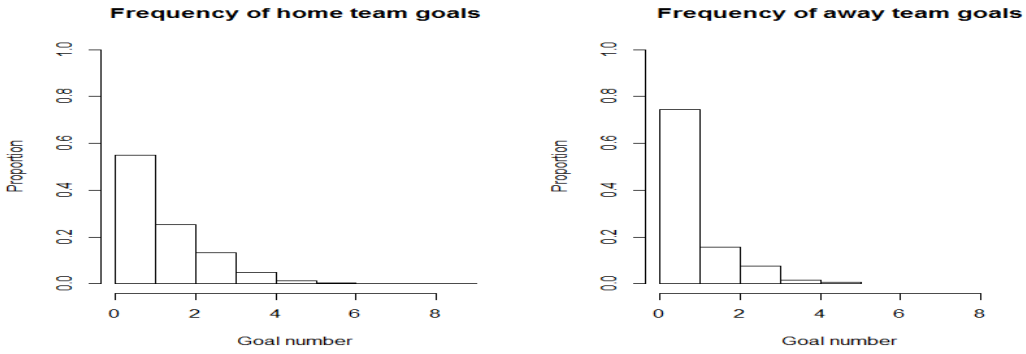

Figure 1: Plots of the residuals vs the fitted values for both the Exponential and Hill models for each type of detectors (Diamond, EFD, PFD).

In this model, the home effect is the same for all the teams. In a similar fashion, suppose that the mean of the Poisson for the away team is made of 2 variables: its attack rate and the opponent's defense rate. So that in a game with teams labeled by i and j, and let $X_{i,j}$ and $Y_{i,j}$ be the number of goals scored by the home team i and the away team j, respectively. Then suppose that the number of goals for both team is given by

$$X_{i,j} \sim \text{Pois}(\alpha_i \beta_j \gamma) \quad Y_{i,j} \sim \text{Pois}(\alpha_j \beta_i), \tag{1}$$

where $X_{i,j}$ and $Y_{i,j}$ are assumed to be independent and α_i, $\beta_i > 0$ for all i. The α_i measures the attack rate of the teams and β_i the defense rate. The 'home-effect' is measured by γ and is the same for all teams. Furthermore to avoid overparametrisation, the following constraint is imposed

$$\sum_{i=1}^{N_b} \frac{\alpha_i}{N_b} = 1, \tag{2}$$

where N_b is the number of teams. So for N independent games, the likelihood function for the number of goals in a game labeled k factorizes in N products of two Poisson densities. Let $f(x_k, y_k)$ be the joint distribution of the scoreline (intersection of goals scored by both away and home teams) in game k, then the likelihood function is

$$L(\alpha, \beta, \gamma | \mathbf{x}, \mathbf{y}) = \prod_{k=1}^{N} f(x_k, y_k) \tag{3}$$

$$\propto L(\alpha, \beta, \gamma | \mathbf{x}, \mathbf{y}) = \prod_{k=1}^{N} \exp(-\lambda_k) \lambda_k^{x_k} \cdot \exp(-\mu_k) \mu_k^{y_k},$$

,

where $\lambda_k = \alpha_{i(k)} \beta_{j(k)\gamma}$ and $\mu_k = \alpha_{j(k)} \beta_{i(k)\gamma}$. For instance, $\alpha_{2(5)}$ is the attack rate of team 2, playing in game number 5.

In order to maximize the likelihood and get estimates for the parameters, an historic of games is needed. Here, as mentioned in the introduction, the data collected concern the Greek Superleague and the score of 762 games was recorded. A number of 22 teams were playing over the period the data were collected, however, only 16 of these 22 teams are currently playing in the 2015-2016 season. The aim is thus to use the maximum likelihood estimates in order to predict the outcome of the 2015-2016 season. One may argue rightfully that using scores from more than a season ago does not improve the accuracy of our estimates, it is even likely to make them worse. This is particularly true in football, where players change team very often and the level of a team a couple of years back is often not reflecting the current level. This is an issue that could be resolved by giving more weight in the likelihood to the more recent games.

In order to come up with a probability of any team finishing at a certain position in the league at the end of the 2015-2016 season, one has to simulate each remaining game of the season. The software R was used in order to simulate 1000 times the outcome of the 2015-2016 season. Note, that only 1000 simulations were carried out as it took too long for the computer to deal with 10,000 simulation. Then, the number of times a team finishes in a certain position is recorded and divided by N in order to obtain the probability.

3 Results

After running 1000 simulations for all the games left in the league, it was possible to come up with an estimate of the probability of each team to finis at a certain rank. The standard error estimates associated to the probabilities are given in brackets. These estimates are shown in Table 1 and Table 2, for practical reasons, rounded up to 2 decimals. Table 1 shows from the $1st$ to the $8th$ position in the league. Table 2 shows from the $9th$ to the $16th$. The probabilities have been calculated using the R code given in Appendix A. I have also written a code given in Appendix A, for the bootstrap procedure, but after 3 hours, the computer was still processing the actions and no output was produced. The estimated standard errors were computing assuming normality and for a probability p and a sample size of N, we get

$$ese(\hat{p}) = \sqrt{p(1-p)/N},$$

where ese is the estimated standard error.

	Position in the league							
Teams	1st	2nd	3rd	4th	5th	6th	7th	8th
Olympiacos	1	0	0	0	0	0	0	0
AEK	0	0.54(0.12)	0.33(0.12)	0.11(0.08)	0.02(0.04)	0	0	0
Panathinaikos	0	0.39(0.12)	0.44(0.12)	0.13(0.08)	0.04(0.05)	0	0	0
Paok	0	0.06(0.06)	0.18(0.10)	0.49(0.12)	0.18(0.10)	0.05(0.05)	0.03(0.04)	0.01(0.02)
Tripolis	0	0.01(0.02)	0.04(0.05)	0.17(0.09)	0.37(0.12)	0.18(0.10)	0.10(0.08)	0.06(0.06)
Panionios	0	0	0	0.039(0.12)	0.12(0.08)	0.229(0.10)	0.15(0.09)	0.16(0.09)
Levadiakos	0	0	0	0.019(0.02)	0.03(0.04)	0.08(0.07)	0.10(0.08)	0.12(0.08)
Pas Giannina	0	0	0	0.03(0.03)	0.12(0.08)	0.15(0.09)	0.17(0.09)	0.15(0.09)
Iraklis	0	0	0	0.01(0.02)	0.03(0.04)	0.08(0.07)	0.10(0.08)	0.13(0.08)
Platanias	0	0	0	0	0.02(0.02)	0.04(0.05)	0.07(0.07)	0.08(0.07)
Veria	0	0	0	0	0	0.01(0.02)	0.02(0.02)	0.03(0.03)
Skoda	0	0	0	0	0.01(0.02)	0.05(0.06)	0.08(0.07)	0.08(0.07)
Panaitolikos	0	0	0	0.01(0.02)	0.01(0.02)	0.02(0.02)	0.05(0.06)	0.05(0.06)
Atromitos	0	0	0.01(0.02)	0	0.06(0.06)	0.12(0.08)	0.14(0.08)	0.14(0.08)
Panthrakikos	0	0	0	0	0	0	0.0	0
Ael Kalloni	0	0	0	0	0	0	0	0

Table 1: Estimates of the probabilities and the associated estimated standard errors for a team to finish at any rank between the 1st and 8th position in the 2015-2016 season.

Teams	Position in the league							
	9th	10th	11th	12th	13th	14th	15th	16th
Olympiacos	0	0	0	0	0	0	0	0
AEK	0	0	0	0	0	0	0	0
Panathinaikos	0	0	0	0	0	0	0	0
Paok	0.01(0.02)	0	0	0	0	0	0	0
Tripolis	0.04(0.05)	0.02(0.02)	0.01(0.02)	0	0	0	0.10(0.08)	0
Panionios	0.10(0.08)	0.08(0.07)	0.06(0.06)	0.05(0.06)	0.02(0.02)	0.01(0.02)	0	0
Levadiakos	0.12(0.08)	0.13(0.08)	0.11(0.08)	0.12(0.08)	0.11(0.08)	0.07(0.06)	0.02(0.02)	0
Pas Giannina	0.14(0.08)	0.10(0.08)	0.10(0.08)	0.04(0.05)	0.02(0.02)	0.01(0.02)	0	0
Iraklis	0.15(0.08)	0.15(0.09)	0.12(0.08)	0.11(0.08)	0.08(0.07)	0.04(0.05)	0.01(0.02)	0
Platanias	0.11(0.08)	0.12(0.08)	0.13(0.08)	0.12(0.08)	0.15(0.09)	0.13(0.08)	0.03(0.03)	0
Veria	0.05(0.06)	0.08(0.07)	0.11(0.08)	0.15(0.09)	0.23(0.11)	0.25(0.11)	0.09(0.08)	0
Skoda	0.10(0.08)	0.12(0.08)	0.14(0.08)	0.15(0.09)	0.12(0.08)	0.10(0.08)	0.04(0.05)	0
Panaitolikos	0.08	0.10	0.12	0.15	0.17	0.21	0.04	0
Atromitos	0.13	0.11	0.11	0.08	0.07	0.05	0.01	0
Panthrakikos	0	0.01(0.02)	0.01(0.02)	0.03(0.03)	0.04(0.05)	0.13(0.08)	0.64(0.12)	0.14(0.08)
Ael Kalloni	0	0	0	0	0	0.01(0.02)	0.13(0.08)	0.86(0.09)

Table 2: Estimates of the probabilities and the associated estimated standard errors for a team to finish at any rank between the 8th and 16th position in the 2015-2016 season.

Perhaps the most striking is that out of 1,000 simulations, Olympiacos won the league every single time. Furthermore, we can see that only 4 teams are realistically competing for the $2nd$ and $3rd$ positions if one consider that Atromitos finishing 3rd is a fluke. As one may have expected, for the mid-table, a large number of teams are competing to finish from the $5th$ to the $11th$ position. The very bottom of the league is similar to the very top in the sense that only a couple of teams (Panthrakikos and Ael Kalloni) are competing for the $16th$ position. If the model used to predict the score were to be true, Olympiacos fans could already be celebrating the victory of their team. The most likely teams to qualify for the play-off round in order to compete for a spot in the Champions League would be AEK, Panathinaikos, Paok and Tripolis. While Ael Kalloni would get relegated.

4 Discussion

My findings above show that is it harder to predict the ranking in the mid-table than to predict the very top or the very bottom of the league. With the Greek Superleague only being in the $3rd$ or $4th$ game into the second part of the season, which was the point where the results were not known, it makes very little sense to compare the predicted ranking to the current ranking. To be able to compare our prediction though, one may look at the odds for the final ranking, which I was not able to find. No other research was done on predicting the 2015-2016 Greek Superleague ranking and I cannot compare it to other models. However, Dixon and Coles themselves pointed out that a static model is not very adequate as "in reality, a team's performance tends to be dynamic, varying from one time period to another". They then enhanced the static model that I used in order to take into account the time factor. It would be of interest to enhance this simple static model to a dynamic one and do this analysis again. I would then be able to come up with a similar table with probabilities and I would then compare my results with the actual Greek Superleague at the end of the 2015-2016 season. While the champions and the relegate team are very likely to remain the same, we may get a better prediction for the teams in the middle of the table. However, as I mentioned earlier, with sufficient data, it is likely that the simple static model and a more elaborated model give very similar predictions for the end of the season rankings. However, a dynamic model is better in order to predict the outcome of a single game as it can incorporate recent factors that may affect the outcome of the game.

Appendix A

Simulating scores for a season

```
matches <- which(is.na(sl))
season <- sl[sl[,6]>=71,1:4]
sim_season <- function(season,alpha,beta,gam,rho)
{
complete_season <- season
first <- min(which(is.na(season[,3])))
last <- max(which(is.na(season[,3])))
for(i in first:last)
{
alpha_H <- alpha[season[i,1]]
alpha_A <- alpha[season[i,2]]
beta_H  <- beta[season[i,1]]
beta_A  <- beta[season[i,2]]
lambda  <- alpha_H*beta_A*gam
mu      <- alpha_A*beta_H
complete_season[i,3:4] <- sim(1,lambda,mu,rho)
}
return(complete_season)
}
```

Points function

```
score <- function(complete_season)
{
teams <- sort(unique(season[,1]))
no_teams <- length(teams)
team_score <- matrix(no_teams,1)
for(i in 1:no_teams)
{
games_home <- which(complete_season[,1]==i)
games_away <- which(complete_season[,2]==i)
no_wins_home <- sum(apply(complete_season[games_home,c(3,4)],
1,diff)< 0)
no_draws_home<- sum(apply(complete_season[games_home,c(3,4)],
1,diff)==0)
no_wins_away <- sum(apply(complete_season[games_away,c(3,4)],
1,diff)> 0)
no_draws_away<- sum(apply(complete_season[games_away,c(3,4)],
1,diff)==0)

team_score[i] <- 3 * (no_wins_home+no_wins_away) +
no_draws_home + no_draws_away
}
order_results <- order(team_score,decreasing=TRUE)
return(cbind(teams[order_results],team_score[order_results]))
}
```

Computing probabilities

```
no_of_teams1516 <- length(unique(sl[sl[,6]>=71,1]))
N <- 1000
rankings <- array(dim=c(no_of_teams1516,2,N))
for(j in 1:N)
{
X <- score(sim_season(season,alpha,beta,gam,rho))
rankings[,,j] <- X
}


C <- 16
Champions <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
two <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
three <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
four <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
five <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
six <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
seven <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
eight <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
```

```
nine <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
ten <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
eleven <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
twelve <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
thirteen <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
forteen <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
fifteen <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
sixteen <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

for(j in 1:N)

{

for(i in 1:C)
{
        if(rankings[1,1,j]==i)
            {
                Champions[i] <- Champions[i]+1
            }

if(rankings[2,1,j]==i)
            {
                two[i] <- two[i]+1
            }
if(rankings[3,1,j]==i)
            {
                three[i] <- three[i]+1
            }
if(rankings[4,1,j]==i)
            {
                four[i] <- four[i]+1
            }
if(rankings[5,1,j]==i)
            {
                five[i] <- five[i]+1
            }
if(rankings[6,1,j]==i)
            {
                six[i] <- six[i]+1
            }
if(rankings[7,1,j]==i)
            {
                seven[i] <- seven[i]+1
            }
if(rankings[8,1,j]==i)
            {
                eight[i] <- eight[i]+1
            }
```

```
if(rankings[9,1,j]==i)
            {
                nine[i] <- nine[i]+1
            }
if(rankings[10,1,j]==i)
            {
                ten[i] <- ten[i]+1
            }
if(rankings[11,1,j]==i)
            {
                eleven[i] <- eleven[i]+1
            }
if(rankings[12,1,j]==i)
            {
                twelve[i] <- twelve[i]+1
            }
if(rankings[13,1,j]==i)
            {
                thirteen[i] <- thirteen[i]+1
            }
if(rankings[14,1,j]==i)
            {
                forteen[i] <- forteen[i]+1
            }
if(rankings[15,1,j]==i)
            {
                fifteen[i] <- fifteen[i]+1
            }
if(rankings[16,1,j]==i)
            {
                sixteen[i] <- sixteen[i]+1
            }
}
}

Champions/1000
two/1000
three/1000
four/1000
five/1000
six/1000
seven/1000
eight/1000
nine/1000
ten/1000
eleven/1000
twelve/1000
thirteen/1000
```

```
forteen/1000
fifteen/1000
sixteen/1000
```

4.1 Bootstrap attempt

```
R <- 100
for (h in 1:R){
X <- NULL
no_of_teams1516 <- length(unique(sl[sl[,6]>=71,1]))
N <- 1000
rankings <- array(dim=c(no_of_teams1516,2,N,h))
for(j in 1:N)
{
X <-sim_season(season,alpha,beta,gam,rho)
rankings[,,j,h] <- score(X)
}

first <- array(0,c(16,16,R))
for (j in 1:N){
for(i in 1:16){
for (k in 1:16){
if ( rankings[k,1,j,h] == i)
{
first[17-k,i,h] <- first[17-k,i,h] + 1
}
else {
first[17-k,i,h] <- first[17-k,i,h]
}}}}}
```

References

Reep, C.and Benjamin, B. (1968) Skill and chance in ball games.*J R. Statist. Soc.* A, **131**, 581-585.

Maher, M.J. (1982) Modelling association football scores. Statist. *Neerland.*, **36**, 109-118.

Mark J. Dixon, Stuart G. Coles. (1997) Modelling Association Football scores and inefficiencies in the football betting market. *Applied Statistics*, **46**, Issue 2, 265-280.