The Bivariate Poisson Distribution and its Applications to Football

May 5, 2011

Author:
Gavin Whitaker

Supervisors:
Dr. P. S. Ansell
Dr. D. Walshaw

School of Mathematics and Statistics
Newcastle University

Abstract

We look at properties of univariate and bivariate distributions, specifically those involving generating functions. Using these properties we arrive at the bivariate Poisson distribution which we use to simulate football matches. We consider the home effect and the problems involved when estimating our parameters. We view two methods for estimating these parameters and use them to simulate football matches. We simulate the 2009/2010 Premier League, looking at the final results for the season, how the home effect influences the model and how teams compare and differ. We finish by considering how others simulate football matches, specifically the computer games industry.

Acknowledgements

Firstly Dr. Phil Ansell and Dr. Dave Walshaw for the knowledge they have passed on, and the valuable advice and encouragement they have given me. George Stagg for spending many nights on my sofa, Jack Sykes for his work ethic and Patrick Robertson for being a grammar tiger and tirelessly keeping me in check. Finally to those family, friends and strangers who have listened to my incessant ramblings, I thank you all.

Contents

Chapter 1

Introduction

In recent times there has been increasing development of the Internet, meaning nearly everyone now has web access. These developments have led to an increase in the betting market, with companies expanding onto the Internet. This new technology means spread betting companies can now offer a real time market which can update its odds quickly to match changing scenarios in sport. Football has arguably seen the greatest increase of interest, with punters able to bet on results, goal scorers, the times of goals etc. With all this money been bet on football the natural question that arises is, "Can we use mathematics to predict football matches?"

We begin by looking at the number of goals scored in a single match in the English Premier League.

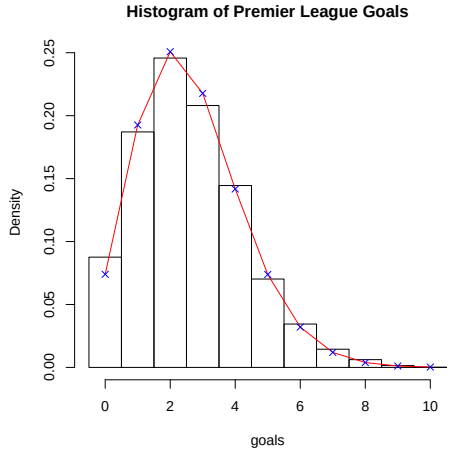

Figure 1.1: Number of goals in a Premier League match.

Figure (1.1) shows the number of goals scored by both teams during a game in the Premier League. The mean number of goals scored in a game is 2.604. The shape of the histogram suggests that a good way to start would be by modeling the goals using a Poisson distribution with mean 2.604, indicated by the red line on the histogram. Using this distribution we consider the probability of goals scored in the Sunderland, Liverpool game, (20$^{\text{th}}$ March 2011), and compare it with the odds offered by bet365.

Total Goals	0	1	2	3	4
Pois(2.604)	0.074	0.193	0.251	0.218	0.142
bet365 odds	0.12	0.24	0.282	0.221	0.137

If the probabilities from the Pois(2.604) are greater than the odds offered we would bet as we believe the event to be more likely than the bookmakers, and as such we should get good odds. Here the probability of 4 goals in the game is slightly greater than the odds offered so we may be tempted to bet. The probabilities are reasonably close to those offered by bet365; this suggests that this model, albeit basic, captures the distribution of goals and is a reasonable starting point for predicting football matches. However we are more interested in being able to predict the number of goals for an individual team. Let us now consider the average number of goals scored by the home team and the average number scored by the away team.

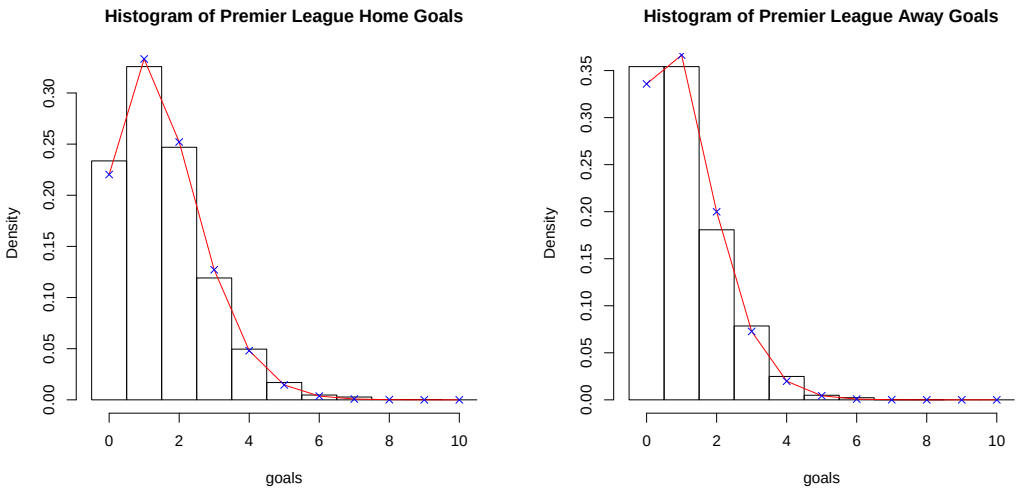

Figure 1.2: Home and away goals in the Premier League.

The shape of the histograms in Figure (1.2) again suggest a Poisson distribution. The mean of goals scored by the home team is 1.513 and the mean

of away goals is 1.091. The corresponding Poisson distributions are shown on the histograms by the red line. Thus, can a Poisson distribution for the home goals and a Poisson distribution for the away goals more accurately allow us to predict football matches?

Sunderland Goals (Home)	0	1	2	3
Pois(1.513)	0.220	0.333	0.252	0.127
bet365 odds	0.368	0.283	0.243	0.106

Liverpool Goals (Away)	0	1	2	3
Pois(1.091)	0.336	0.366	0.2	0.073
bet365 odds	0.312	0.365	0.222	0.101

These probabilities are again close to the odds offered. Using these distributions we may be tempted to bet on Sunderland scoring 1 goal as the probabilty is higher than the odds offered. However in this case the probabilities are slightly more erratic, i.e. the probability of Sunderland scoring 0 goals is not that accurate. This is because teams have different levels of attack and defence; we presume Chelsea (1^{st} in the 2009/2010 Premier League) score more goals than Portsmouth (20^{th}) for example. This leads to the question, "How do we assign these levels of attack and defence?"

Brian Clough once said "It only takes a second to score a goal," and when Ron Atkinson was asked for his feelings on an upcoming match he responded with, "Well, either side could win it, or it could be a draw." Over the following chapters we will try to predict a sport that has been described on many occasions as unpredictable. We will look at match results, how a team attacks and defends, a team's form and the question of the home effect. We will also consider how others have predicted football matches.

Chapter 2

Properties of Univariate and Bivariate Distributions

In this chapter we will look at some of the properties involved with univariate distributions, specifically those involving generating functions. We will then extend these to the bivariate case using examples from the bivariate Binomial distribution. We will use this distribution to derive the bivariate Poisson distribution, which we will be using to predict football matches.

2.1 The Univariate Case

2.1.1 Probability Generating Functions

For the univariate case, where X is a random variate taking values on a subset of the non-negative integers $0, 1...$, $p(x)$ is the probability mass function of X and the Probability Generating Function (PGF) is defined by:

$$G_X(t) = E[t^X] = \sum_x p(x)t^x. \tag{2.1}$$

If X and Y have identical PGFs, i.e. $G_X(t) = G_Y(t)$ then $p(x) = p(y)$. That is to say that identical PGFs imply that X and Y have identical distributions.

When considering PGFs there are some important properties to consider, for example:

- $G_X(1) = 1$.

- $E[X] = G_X'(1)$.
- $Var(X) = G_X''(1) + G_X'(1) - [G_X'(1)]^2$.
- If $A(t)$ is known to be a PGF of X then $P(X = k)$ can be obtained by differentiating $A(t)$ k times w.r.t t and setting $t = 0$.
- If the PGF of Y is $h(x)$ then

$$G_Y(t) = G_{h(x)}(t) = E[t^{h(x)}] = \sum_x p(x)t^{h(x)}.$$

Given the last property if $h(x)$ is relatively simple then it may be possible to express $G_Y(t)$ in terms of $G_X(t)$. For example if $Y = a + bx$ then

$$G_Y(t) = E[t^{a+bX}] = t^a E[(t^b)^X] = t^a G_X(t^b).$$

We will now consider examples using the univariate Poisson distribution and the univariate Binomial distribution.

If $X \sim \text{Pois}(\lambda)$, with

$$Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2...$$

then the PGF is

$$G_X(t) = \exp(\lambda(t-1)).$$

Thus

$$\begin{aligned} G_X'(t) &= \lambda \exp(\lambda(t-1)), \\ G_X''(t) &= \lambda^2 \exp(\lambda(t-1)). \end{aligned}$$

Hence

$$\begin{aligned} E[X] &= \lambda \exp(0) \\ &= \lambda, \\ Var(X) &= \lambda^2 \exp(0) + \lambda \exp(0) - (\lambda \exp(0))^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda. \end{aligned}$$

Thus if $X \sim \text{Pois}(\lambda)$, $E[X] = Var(X) = \lambda$, which is what we expect given our knowledge of the Poisson distribution.

For $X \sim \text{Bin}(n, p)$, we have

$$Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, ..., n,$$

with

$$G_X(t) = ((1-p) + pt)^n.$$

Thus

$$\begin{aligned} G_X^{'}(t) &= np((1-p) + pt)^{n-1}, \\ G_X^{''}(t) &= n(n-1)p^2((1-p) + pt)^{n-2}. \end{aligned}$$

Hence

$$\begin{aligned} E[X] &= np((1-p) + p)^{n-1} \\ &= np, \\ Var(X) &= n(n-1)p^2 + np - (np)^2 \\ &= np(1-p). \end{aligned}$$

Thus if $X \sim \text{Bin}(n, p)$, $E[X] = np$ and $Var(X) = np(1-p)$. This again matches our expectations.

2.1.2 Moment Generating Functions

For the random variable X, the Moment Generating Function (MGF) is defined as:

$$M_X(t) = E[e^{tX}]. \tag{2.2}$$

The MGF of a random variable is an alternative form of its probability distribution. Equation (2.2) allows us to find all the moments of the distribution. Recall that the series expansion of

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + ...$$

From this it follows that

$$M_X(t) = 1 + m_1 t + \frac{m_2 t^2}{2!} + \frac{m_3 t^3}{3!} + ...$$

where m_i is the i^{th} moment. We can calculate m_i, $i = 1, 2, ...$ by differentiating $M_X(t)$ i times and setting $t = 0$.

As in the case of the PGF there are a collection of simple results to ease calculations. For example:

- $M_X(0) = 1$.
- $E[X] = M_X^{'}(0)$.
- $Var(X) = M_X^{''}(0) - [M_X^{'}(0)]^2$.
- Given $X_1, X_2, ..., X_n$ are a sequence of independent random variables where $S_n = \sum_{i=1}^{n} c_i X_i$ (c_i's are constants) then

$$M_{S_n}(t) = M_{X_1}(c_1 t) M_{X_2}(c_2 t) ... M_{X_n}(c_n t).$$

- If X and Y are independent then

$$M_{X+Y}(t) = E[e^{(X+Y)t}] = E[e^{Xt} e^{Yt}] = M_X(t) M_Y(t).$$

To demonstrate these properties we again consider the univariate Poisson as described above. It has MGF

$$M_X(t) = \exp\big(\lambda\big(e^t - 1\big)\big).$$

Thus

$$\begin{aligned} M_X^{'}(t) &= \lambda e^t \exp\big(\lambda\big(e^t - 1\big)\big), \\ M_X^{''}(t) &= \lambda e^t \exp\big(\lambda\big(e^t - 1\big)\big) + \lambda^2 e^{2t} \exp\big(\lambda\big(e^0 - 1\big)\big). \end{aligned}$$

Hence

$$\begin{aligned} E[X] &= \lambda e^0 \exp\big(\lambda\big(e^0 - 1\big)\big) \\ &= \lambda, \\ Var(X) &= \lambda + \lambda^2 - (\lambda)^2 \\ &= \lambda. \end{aligned}$$

As in the case of the PGF these results are as expected. We now move to consider a continuous distribution, $X \sim N(\mu, \sigma^2)$. The MGF is given by

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Thus

$$\begin{aligned} M_X^{'}(t) &= (\mu + \sigma^2 t) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right), \\ M_X^{''}(t) &= \sigma^2 \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) + (\mu + \sigma^2 t)^2 \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right). \end{aligned}$$

Hence

$$\begin{aligned} E[X] &= \mu \exp(0) \\ &= \mu, \\ Var(X) &= \sigma^2 + \mu^2 - (\mu)^2 \\ &= \sigma^2. \end{aligned}$$

These results are what we expect given our knowledge of the Normal distribution.

2.1.3 Cumulant Generating Functions

The Cumulant Generating Function (CGF) is the log of the MGF. Note that if 2 distributions have identical moments then they will also have identical cumulants. The CGF is defined as:

$$K_X(t) = \log(M_X(t)) = \sum_x \frac{t^x}{x!}\kappa_x. \tag{2.3}$$

Here κ_x represent the cummulants of X and are:

- $\kappa_1 = E[X]$.
- $\kappa_2 = E[(X - E[X])^2] = Var(X)$.
- $\kappa_3 = E[(X - E[X])^3]$.
- $\kappa_4 = E[(X - E[X])^4] - 3[Var(X)]^2$.

For the Poisson distribution

$$\begin{aligned} K_X(t) &= \log\left\{ \exp\left(\lambda\left(e^t - 1\right)\right)\right\} \\ &= \lambda\left(e^t - 1\right) \\ &= \lambda\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + ...\right). \end{aligned}$$

Therefore all the cummulants of the Poisson distribution are given by λ.

For the Normal distribution

$$\begin{aligned} K_X(t) &= \log\left\{ \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)\right\} \\ &= \mu t + \frac{1}{2}\sigma^2 t^2. \end{aligned}$$

Hence $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$, the mean and variance respectively. Again for the Poisson distribution and the Normal distribution these results are as expected.

2.2 The Bivariate Case

Taking the functions defined above it is now a natural progression to expand the PGF, MGF and CGF to encompass bivariate data.

2.2.1 Bivariate PGF

Given a pair of discrete random variables, X and Y, with probability function $p(x, y)$ we can define the PGF as $E[t_1^X t_2^Y]$. Hence:

$$G_{XY}(t_1, t_2) = \sum_{x,y} t_1^x t_2^y p(x, y). \tag{2.4}$$

From equation(2.4) we can determine the probability function. This is achieved by using the fact that the PGF can be differentiated continuously w.r.t t_1 and t_2, and then evaluated at $(0, 0)$:

$$p(x, y) = \frac{1}{x!}\frac{1}{y!}\frac{\partial^{x+y}}{\partial t_1^x \partial t_2^y} G_{XY}(t_1, t_2)|_{t_1=0, t_2=0}.$$

2.2.2 Bivariate MGF

As in the case of the bivariate PGF, the bivariate MGF is an extension of the univariate case. If we take X and Y as defined above in section 2.2.1, the MGF is $E[e^{t_1 X + t_2 Y}]$. Thus,

$$M_{XY}(t_1, t_2) = \sum_{x,y} e^{t_1 x + t_2 y} p(x, y). \tag{2.5}$$

It is worth noting that this definition of the MGF assumes that all the moments do exist. We define the joint moments $\mu'_{r,s}$ as $E[X^r Y^s]$, by expanding the exponentials equation (2.5) becomes

$$M_{XY}(t_1, t_2) = \sum_{r,s} \frac{t_1^r}{r!}\frac{t_2^s}{s!}\mu'_{r,s}.$$

Here $\mu'_{r,s}$ can also be defined as the mixed partial derivative

$$\frac{\partial^{r+s}}{\partial t_1^r \partial t_2^s} M_{XY}(t_1, t_2)|_{t_1=0, t_2=0}.$$

From the definitions of the PGF and MGF (2.4, 2.5) the following relationship can be assumed:

$$M_{XY}(t_1, t_2) = G_{XY}(e^{t_1}, e^{t_2}).$$

2.2.3 Bivariate CGF

Given the definition above for the CGF in the univariate case we can define the bivariate CGF as:

$$K_{XY}(t_1, t_2) = \log\left(M(t_1, t_2)\right) = \sum_r \sum_s \frac{t_1^r}{r!} \frac{t_2^s}{s!} \kappa_{r,s}.$$

Where $\kappa_{r,s}$ is the cumulant of order (r,s).

2.2.4 Marginal Distributions

It may be of interest to observe the behaviour of the variables independently of each other. For this we use the marginal distributions. Taking the probability function of X and Y as $p(x, y)$, the marginal probability function for x is

$$g(x) = \sum_y p(x, y);$$

and the marginal probability function for y is

$$h(y) = \sum_x p(x, y).$$

This gives the marginal PGF for x as:

$$\begin{aligned} G_X(t) &= \sum_x g(x) t^x \\ &= \sum_x t^x \sum_y p(x, y) \\ &= \sum_x \sum_y p(x, y) t^x = G_X(t, 1). \end{aligned} \tag{2.6}$$

Similarly the marginal PGF for y is given by $G_Y(1,t)$.

We can work out the marginal MGFs using a similar method. These are seen to be:

$$M_X(t) = M(t,0),$$

and

$$M_Y(t) = M(0,t). \tag{2.7}$$

2.2.5 Convolutions

Bivariate distributions can also be generated using convolutions of random variables. Take

$$X = X_1 + X_3$$

and

$$Y = X_2 + X_3$$

with X_1, X_2, X_3 independently distributed. Thus X and Y are jointly distributed. Now taking PGFs as defined above, the joint PGF of (X,Y) is given by

$$G_{XY}(t_1,t_2) = G_{X1}(t_1)G_{X2}(t_2)G_{X3}(t_1t_2). \tag{2.8}$$

The joint MGF of (X,Y) is similarly given by

$$M_{XY}(t_1,t_2) = M_{X1}(t_1)M_{X2}(t_2)M_{X3}(t_1+t_2). \tag{2.9}$$

2.3 The Bivariate Binomial distribution

As in the univariate case, the bivariate Binomial distribution is a continuation of the Bernoulli distribution. One bivariate Bernoulli trial measures two random variables, both with outcomes 0 and 1. Each trial therefore has four possible outcomes, $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$, where the probabilities of the outcomes remain constant, and the trials are independent of each other.

Consider a sequence of n bivariate Bernoulli trails, measuring the two random variables I_1 and I_2 which can take the values 0 or 1. We define the probability of $I_1 = a$ and $I_2 = b$ for $a = 0,1$ and $b = 0,1$ as

$$P\{I_1 = a, I_2 = b\} = p_{ab}.$$

Taking

$$X = \sum_{i=1}^{n} I_{1i},$$

and

$$Y = \sum_{i=1}^{n} I_{2i}.$$

The PGF of X,Y is:

$$\begin{aligned} & G_{XY}(t_1, t_2) \\ &= E[t_1^X t_2^Y] \\ &= \{E[t_1^{I_1} t_2^{I_2}]\}^n \\ &= (t_1 t_2 p_{11} + t_1 p_{10} + t_2 p_{01} + p_{00})^n \\ &= (1 + p_{1+}(t_1 - 1) + p_{+1}(t_2 - 1) + p_{11}(t_1 - 1)(t_2 - 1))^n. \end{aligned} \quad (2.10)$$

Where p_{1+} is the sum over b when $a = 1$, p_{+1} is the sum over a when $b = 1$, and p_{11} is the probability that $a = 1$ and $b = 1$.

Using equation (2.6) the marginal PGF for X is

$$\begin{aligned} G_X(t) &= (1 + p_{1+}(t - 1) + p_{+1}(1 - 1) + p_{11}(t_1 - 1)(1 - 1))^n \\ &= (p_{0+} + p_{1+}t)^n. \end{aligned}$$

Similarly the marginal PGF for Y is

$$G_Y(t) = (p_{+0} + p_{+1}t)^n.$$

Since $G_X(t)$ has the same form as the PGF of a $\text{Bin}(n, p)$ we can state that

$$X \sim \text{Bin}(n, p_{1+});$$

similarly

$$Y \sim \text{Bin}(n, p_{+1}).$$

We can see that the bivariate Binomial distribution is just a continuation of the Binomial distribution to higher dimensions. In the univariate case we are considering the number of successes against the number of failures, whereas in the bivariate case we are interested in how many times the events X and Y have occurred.

2.4 The Bivariate Poisson distribution

Just as in the univariate case the bivariate Poisson distribution can be derived by taking the limit of the bivariate Binomial distribution which has PGF

(2.10). We take λ_1, λ_2 and λ_3 to be positive constants independent of n and that

$$p_{1+} = \frac{\lambda_1}{n},$$

$$p_{+1} = \frac{\lambda_2}{n}$$

and

$$p_{11} = \frac{\lambda_3}{n}.$$

Substituting into equation (2.10) gives the PGF as:

$$G_n(t_1, t_2) = \left(1 + \frac{\lambda_1(t_1 - 1)}{n} + \frac{\lambda_2(t_2 - 1)}{n} + \frac{\lambda_3(t_1 - 1)(t_2 - 1)}{n}\right)^n. \quad (2.11)$$

Taking the limit of equation (2.11) as $n \to \infty$ and using the result

$$\lim_{n\to\infty} \left(1 + \frac{\lambda}{n}\right)^n \sim \exp(\lambda),$$

we get

$$G_{XY}(t_1, t_2) = \exp(\lambda_1(t_1 - 1) + \lambda_2(t_2 - 1) + \lambda_3(t_1 - 1)(t_2 - 1)). \quad (2.12)$$

Rearranging equation (2.12) and reparameterising gives:

$$G_{XY}(t_1, t_2) = \exp(\lambda_1(t_1 - 1) + \lambda_2(t_2 - 1) + \lambda_3(t_1 t_2 - 1)). \quad (2.13)$$

This is the PGF of the bivariate Poisson distribution with parameters λ_1, λ_2 and λ_3 for two random variables X and Y. Comparing this with the PGF for the univariate Poisson distribution, which is given by

$$G_X(t) = \exp(\lambda(t - 1));$$

we see that equation (2.13) is an extension of the univariate Poisson distribution, just as in the case of the Binomial distribution. Here the univariate case considers λ and the bivariate case considers λ_1, λ_2 and λ_3.

Using equation (2.6) on (2.13) gives the marginal PGF for X as

$$\begin{aligned} G_X(t) &= G_X(t, 1) \\ &= \exp(\lambda_1(t - 1) + \lambda_2(1 - 1) + \lambda_3(t - 1)) \\ &= \exp((\lambda_1 + \lambda_3)(t - 1)). \end{aligned} \quad (2.14)$$

Similarly the marginal PGF for Y is given by

$$G_Y(t) = \exp\left((\lambda_2 + \lambda_3)(t-1)\right).$$

Hence the marginal distribution for X is

$$X \sim Po(\lambda_1 + \lambda_3), \tag{2.15}$$

and the marginal distribution for Y is

$$Y \sim Po(\lambda_2 + \lambda_3). \tag{2.16}$$

Expanding equation (2.13) in powers of t_1 and t_2 gives the joint probability function

$$G_{XY}(t_1, t_2) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{i=0}^{\infty} \frac{\lambda_1^i t_1^i}{i!} \sum_{j=0}^{\infty} \frac{\lambda_2^j t_2^j}{j!} \sum_{k=0}^{\infty} \frac{\lambda_3^k t_1^k t_2^k}{k!}. \tag{2.17}$$

Equation (2.17) indicates (through the summations) why the marginal distributions of X and Y are Poisson distributions.

We will conclude this chapter by considering some examples involving the bivariate Poisson distribution. Firstly consider a case when $\lambda_1 = 2$, $\lambda_2 = 1$ and $\lambda_3 = 1$. From the above we can see that this gives a PGF:

$$\begin{aligned} G_{XY}(t_1, t_2) &= \exp\left(2(t_1 - 1) + (t_2 - 1) + (t_1 t_2 - 1)\right) \\ &= \exp\left(2t_1 + t_2 + t_1 t_2 - 6\right). \end{aligned}$$

The marginal distributions are given by

$$X \sim Po(3)$$

and

$$Y \sim Po(2).$$

Using the R package "Bivpois" which was developed to analyze the bivariate Poisson distribution, we are able to calculate probabilities for the distribution using the commands `bivpois.table` and `pbivpois`. 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. Here we will use the command `bivpois.table` to analyze the distribution.

```
> bivpois.table(4,4,lambda=c(2,1,1))
```

0.018	0.018	0.009	0.003	0.001
0.037	0.055	0.037	0.015	0.005
0.037	0.073	0.064	0.034	0.012
0.024	0.061	0.067	0.044	0.019
0.012	0.037	0.049	0.039	0.021

Element (i, j) of the above matrix represents the $Pr(X = i - 1, Y = j - 1)$. From this we can see that $X = 2$, $Y = 1$ (element (3,2)) has the highest probability which is what we expect; there is also more probability for higher values of X than there are for higher values of Y and the probability is mainly concentrated around the top left corner of the matrix.

We now vary the value of λ_3 which represents the covariance between X and Y; keeping λ_1 and λ_2 as before and setting $\lambda_3 = 3$ we see:

```
> bivpois.table(6,6,lambda=c(2,1,3))
```

0.002	0.002	0.001	0.000	0.000	0.000	0.000
0.005	0.012	0.010	0.005	0.001	0.000	0.000
0.005	0.020	0.029	0.019	0.008	0.003	0.001
0.003	0.018	0.039	0.041	0.025	0.010	0.003
0.002	0.012	0.033	0.050	0.044	0.024	0.009
0.001	0.006	0.020	0.040	0.047	0.036	0.018
0.000	0.002	0.010	0.023	0.036	0.036	0.024

Here we see that the probabilities have been severely reduced from the case above, this is because as we increase λ_3 we expect to see higher values of X and Y. However we still keep most of the traits described above with more probability for higher values of X than higher values of Y. We now see that most of the probability is concentrated in the bottom right corner and again this is down to the expectation of higher values.

Finally if we set $\lambda_3 = 0$ we get:

```
> bivpois.table(4,4,lambda=c(2,1,0))
```

0.050	0.050	0.025	0.008	0.002
0.100	0.100	0.050	0.017	0.004
0.100	0.100	0.050	0.017	0.004
0.066	0.066	0.033	0.011	0.003
0.033	0.033	0.017	0.006	0.001

It is notable here that the probabilities are greater than in the other cases although most of the density is still found in the top left corner of the matrix. Setting $\lambda_3 = 0$ means that there is no covariance between X and Y; this model is often referred to as the double Poisson distribution, which is just the sum of independent Poisson distributions. From our knowledge of independent Poisson distributions we expect an outcome of 3 to occur here; this can happen in any combination, e.g. $(X = 3, Y = 0), (X = 2, Y = 1)$, etc. Looking at the matrix we see that the higher probabilities are around these combinations, with the outcomes involving a higher X having a slightly greater probability than those with a larger Y.

Chapter 3

Simulating Football Matches

We now look at how the theory of chapter 2 can be used to predict football matches and address some of the problems we face when simulating results. We will be using regression to simulate results and we begin by looking at some of the aspects of regression.

3.1 Regression

Regression is used to model relationships between random variables. In its simplest form the relationship is a straight line and we model it using linear regression. By using regression we aim to get estimates for these variables using data and then use the estimates to predict future events. We will be estimating a teams attack and defence to predict their goals and the goals of their opponents, from which we can determine who would win the game.

Most estimates in regression are obtained by the method of least squares. To estimate the parameters for a simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1...n$$

where $\epsilon_i \sim N(0, \sigma^2)$ we look to minimise the residual sum of squares

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3.1}$$

All the computations involved use a few summary statistics from the data meaning that for simple regression we can easily calculate estimates of the variables.

We can extend this simple case to include several predictors, this is known as multiple regression.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon_i, i = 1...n$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. This equation can be written in matrix form, suppose that we have n observations, then

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

where

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ & & & & \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} = \begin{pmatrix} \vec{x_1}^T \\ \\ \\ \vec{x_n}^T \end{pmatrix}$$

$$\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$$

$$\vec{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T.$$

For multiple regression the residual sum of squares becomes

$$\sum_{i=1}^{n} (Y_i - \vec{x_i}^T \vec{\beta})^2 = (\vec{Y} - X\vec{\beta})^T (\vec{Y} - X\vec{\beta}). \tag{3.2}$$

However the relationship between our variables is not a straight line so we must consider alternatives to linear regression. We will be using the Expectation-Maximisation (EM) algorithm. The EM algorithm is used when we can assume that the data comes from a multivariate distribution; in our case this is the bivariate Poisson distribution. It allows us to compute maximum likelihoods given missing data; in our case this is results from other seasons. The algorithm is an iterative process consisting of two steps:

1. The E-step, expectation.

2. The M-step, maximisation.

In the first step missing data are estimated given the observed data and the parameters are estimated for the current state using the conditional expectation. The second step maximises the likelihood function assuming that the missing data are known using estimates from step 1. The algorithm guarantees convergence as the likelihood increases with each iteration. If we were estimating θ then at the i^{th} iteration the current estimate would be θ_i.

3.2 The Model

To predict the number of goals in a football match we will be using the bivariate Poisson distribution as described in section 2.4, 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

$$\begin{aligned} \log(\lambda_1) &= \mu + (\text{attack}_i) + (\text{defence}_j) + (\text{home effect}) \\ \log(\lambda_2) &= \mu + (\text{attack}_j) + (\text{defence}_i) \end{aligned} \tag{3.3}$$

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. We will address the problems of obtaining these attack and defence parameters later. Note λ_3 is determined using "Bivpois."

3.3 The Premier League

The Premier League was established in 1992 and is the most watched football league in the world. To date only four teams have won the Premier League, Manchester United (11 times), Arsenal (3 times), Chelsea (3 times) and Blackburn Rovers (once). The league itself consists of 20 teams who over a season play every other team both home and away; therefore every team plays 38 games, meaning a season compromises of 380 fixtures. Every year the bottom three teams are relegated to be replaced by three teams from the lower division, and the top teams (or cup winners) are entered into European competition.

If a team wins a game, meaning they score more goals than their opponents, they are awarded 3 points; if the game is drawn, i.e. both teams score the same number of goals, then both teams are awarded 1 point. At the end of the season the team with the most points wins. If the points are the same between two teams then goal difference is used to separate them. Goal difference is the number of goals scored by a team minus the number of goals conceded. The team with the greater goal difference occupies the higher position in the league. If the goal difference is the same then the number

of goals scored is used to separate them. If the number of goals scored is also the same then a playoff at a neutral venue is organised between the two teams to see who will occupy the higher position, if the position matters, i.e. to determine who wins the league, who gets relegated or who plays in Europe; this has never happened in the history of the Premier League.

3.4 The Home Effect

We now consider the question of whether there is a home effect in the Premier League. Table (3.1) 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 3.1: 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. The home effect is calculated directly using "Bivpois."

3.5 Attack and Defence Parameters

We now look at how we obtain the estimates for the attack and defence parameters in our model. We will obtain these estimates using regression techniques described above, but we must consider how long a time period we use. It is excessive to use all the results from the Premier League's history as teams are relegated and promoted over time, and squads change, meaning that a team which was very good 10 years ago may not necessarily be good today, and thus our estimates may be distorted.

We believe it is a reasonable starting place to use a season's worth of results when trying to predict the final table. Doing this gives a reasonable estimate of how a team performs over the entire season; any periods of increased or decreased attack and defence are averaged out and as such the final estimates give the average rate for a team over that season. This method has advantages in the fact that it is reasonably quick to carry out so we can easily perform a lot of simulations. However it doesn't really capture a team's form. To elucidate, Hull in 2008/2009 got off to a great start and found themselves high in the table; however they barely won a game from November onwards and only avoided relegation by a single point. The question is, "How do we get our model to replicate this behaviour?"

One method is to have a dynamic model and use smaller time periods, thus creating a moving average estimate for the parameters. We observe the first 100 games (roughly 10 weeks) and assume over this period teams are trying to integrate players into their squads and are generally finding their feet in the league, and as such may be slightly erratic in their performances. When simulating we will use the previous seasons estimates for these games, making sensible judgments to replace those relegated by those promoted. We then obtain estimates using these 100 games, after which we will move forward one time step and again use 100 games to gain new estimates. There are several options for the length of the time step; 10 games (roughly a weekly time step), 20 games (roughly a 2 week time step) or 40 games (roughly a monthly time step). We have calculated the estimates using all 3 time steps and plotted the results. Figure (3.1) shows the plots of the attack parameters. You can see that all the time steps keep the general shape with a time step of 10 having the most interference and a time step of 40 being the most smooth. A time step of 20 models the changes in the parameters well without being too smooth, and as it will be quicker to simulate than a time step of 10 we will use this when simulating using a moving average. The results of using a whole seasons results or the alternative of a time step can be seen in chapter 4.

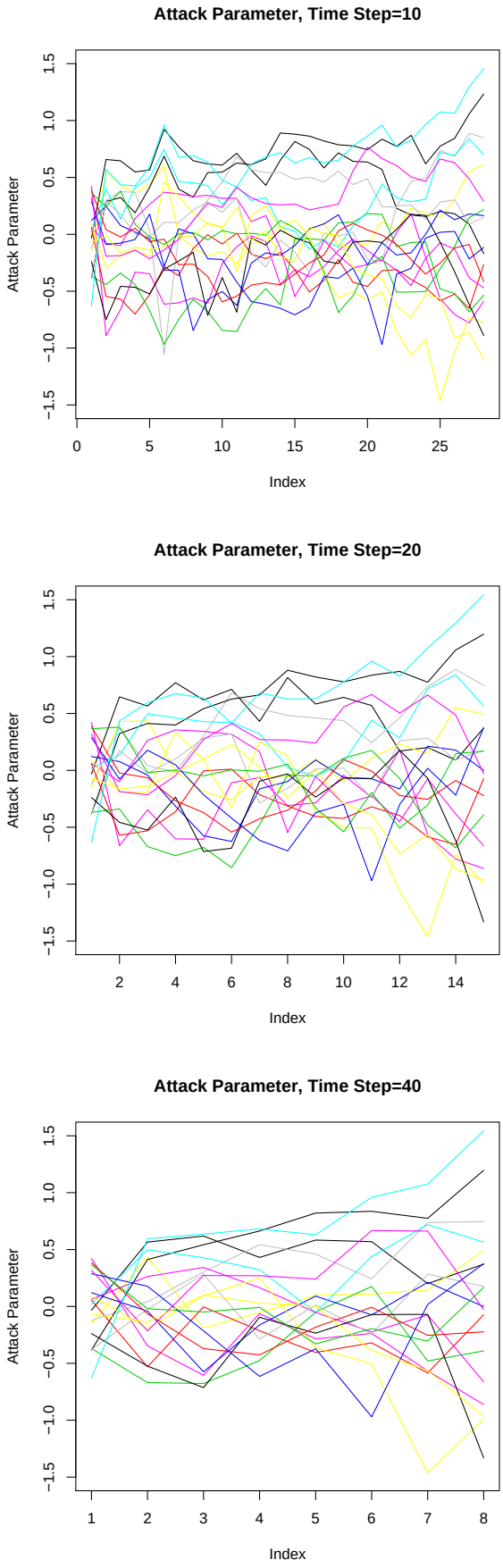

Figure 3.1: Attack parameters using different time steps.

3.6 Simulating Results

We now have 2 methods from which we can simulate; however we must consider what further conditions to use. We must first determine the range of results we should simulate from. We use "Bivpois" and the command `bivpois.table` to determine this. The 2009/2010 Premier League season had a home goals mean of 1.697 and an away goals mean of 1.074; we set λ_1 and λ_2 to be these values respectively and $\lambda_3 = 0.137$. Looking at the probabilities up to 8-8 under these conditions we see that they total 0.999; it therefore seems reasonable to simulate from results up to 8-8. It is also reasonable to assume that an 8-0 is the same as a 10-0 in the long run, and as such the higher scores can be discounted.

We must also consider the order of games we simulate from. For the method using estimates from the whole season this is not important as the estimate never change and games can take place in an order most convenient for simulation. The order is a problem though when using the method based on the moving average estimates. As the estimates change over time we require the games to be simulated in the same order that they occurred in the real season; this is so that any changes in form will be replicated in our results.In theory this should give more accurate results and therefore a more accurate representation of the Premier league.

We simulate results by generating probabilities using `pbivpois,` with the λ's being determined by equation (3.3) and our attack and defence estimates. We number the match results 1-81, and by sampling from 1-81 with each number having the `pbivpois` probabilities we gain an accurate result for each match. Converting the sampled numbers to scores and including the involved teams we have a full set of results for a season, from which we can determine the final league table, as well as look at results between specific teams.

Chapter 4

Results

We now consider the two methods described in chapter 3 and discuss issues with the home effect. During this chapter I will use the following to represent the teams in the Premier League:

Team	Code	Team	Code	Team	Code
Arsenal	ARS	Everton	EVE	Stoke	STO
Aston Villa	AST	Fulham	FUL	Sunderland	SUN
Birmingham	BIR	Hull	HUL	Tottenham	TOT
Blackburn	BLR	Liverpool	LIV	West Ham	WES
Bolton	BOL	Man City	MNC	Wigan	WIG
Burnley	BUR	Man United	MNU	Wolves	WOL
Chelsea	CHE	Portsmouth	POR		

4.1 Method 1: Season Estimates

First we shall consider the method which uses the whole season to obtain its estimates. We begin by plotting the attack parameters against the defence parameters for each team.

Figure (4.1) shows attack plotted against minus defence so that the higher the value, the better a team's attack and defence is. Thus we see that CHE (17) and MNU (5), who were the title contenders, occupy the top right corner with the higher values; it can be argued that the greater attack parameter of CHE is what won them the league despite having a worse defence than MNU. Conversely the teams involved in the relegation battle occupy the bottom left corner, that is HUL (2), POR (6), WIG (11) and BUR (16); while a greater attack won CHE the league it is possible that the slightly greater defence of

WIG in relation to BUR is what kept them in the league. FUL (19), BIR (13) and STO (7) are of particular interest as they have very poor attack parameters but relatively good defensive parameters. All these teams who arguably started the season with the view of staying in the league finished safely mid-table which suggests that a greater defence is better than a greater attack when trying to stay in the Premier League.

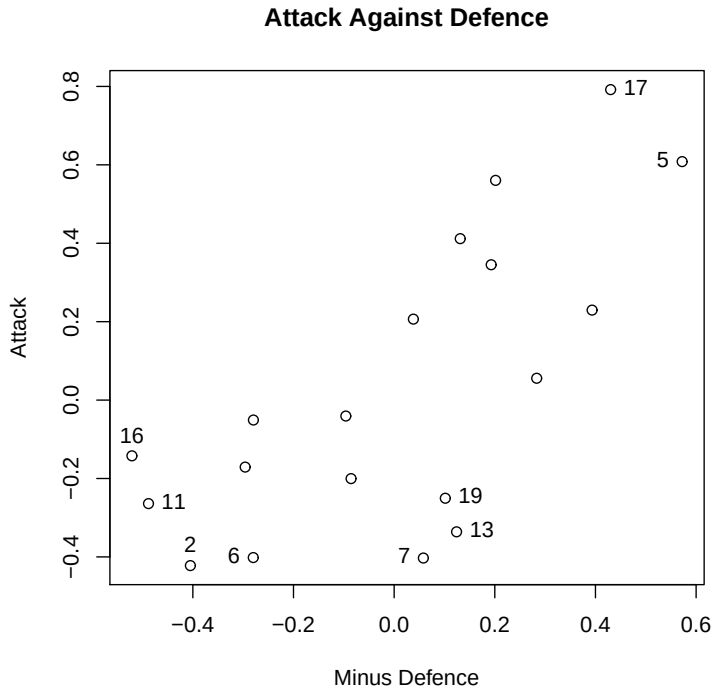

Figure 4.1: Attack against defence.

We are able to simulate this model in R. Table (4.1) shows the final table of results based on 100 simulations of the 2009/2010 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; teams at the top using our simulations are the teams at the top of the Premier League, and similarly for teams found at the bottom of the league. 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 7^{th} place finish was a poor result. Looking at the number of goals scored and conceded we see that this method appears to have captured the rate at which teams attack

and defend. A team's average goals scored and conceded in the simulated table is reasonably accurate when compared with the actual Premier League. 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 4.1: The Premier League based on 100 simulations.

Figure (4.2) shows box plots representing the final points for each team over 100 simulated seasons. They have been ordered to represent the final standing of the league table based on 100 simulated seasons obtained above. CHE and MNU have very similar box plots and this mirrors how close they were in the Premier League, (they were separated by 1 point). CHE and MNU both occupy the top numbers and there is a difference to ARS in 3^{rd}. LIV, TOT and MNC all have very similar boxplots representing the fight for 4^{th}; any one of these teams could have occupied this position and TOT perhaps did because of the slightly longer tails, although LIV had a poorer season and

finished 7^{th}. AST and EVE are also very similar and this shows the struggle for the last European place. AST eventually got it but EVE only missed out by 1 point showing how similar these teams are. All the box plots from SUN to HUL overlap significantly and it was considered at some point during the season that any of these teams could be relegated. The teams to the left of this group pulled away from those to the right as the season went on and this is most likely due to their higher means. Looking at BUR, POR, WIG and HUL we see that there is very little between them. BUR may count themselves slightly unlucky to be relegated given they have a higher mean than WIG, but it is too close to really call, showing how unpredictable the Premier League really is. POR have a much higher mean and finishing position using the simulations compared to the actual Premier League but this is due to them being docked 9 points for entering administration which our model does not account for.

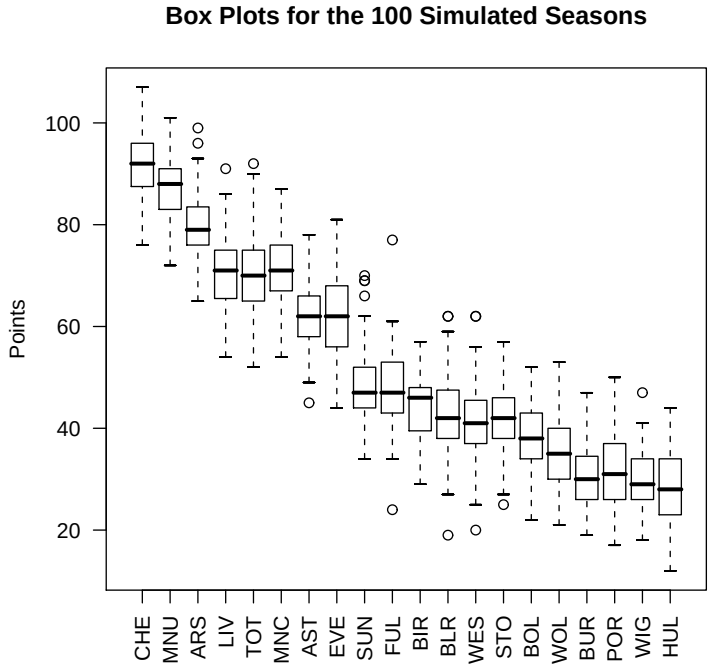

Figure 4.2: Box plots of 100 simulated seasons.

From the box plots it appears that some teams are very similar in nature. W1‘e will now look at these teams and investigate the home effect between them. For the 2009/2010 Premier League the home effect over the entire

season was 0.787, the mean number of home goals was 1.697 and the mean away goals was 1.074; the probability of a home win was 0.508, an away win 0.240, with a draw 0.253. Looking at the box plots and the final points from the simulations we begin by considering the teams EVE to WOL. The parameter estimates change slightly as they need to sum to zero, but the rough ordering stays the same with the better teams having the better parameters; this holds whichever teams are used. Using only the results between these 9 teams we see that the home effect is 0.384, a big decrease from the entire season; however the mean home goals has decreased to 1.458 and the mean away goals has decreased to 0.917. From this we conclude that even though the home effect is less it has a bigger effect overall as the mean number of goals has decreased. The probability of a draw has increased to 0.333 reflecting how similar these teams are, and the probability of an away win is now 0.181, adding further evidence that the home effect has more importance between these 9 teams than it does over the entire league.

Removing WOL and considering the teams EVE to BOL we find that the home effect is 0.415, the mean home goals is 1.482, the mean away goals is 0.875, the probability of a home win is 0.5 and the probability of an away win is 0.179. Removing WOL has increased the home effect between the 8 teams; the decreased mean away goals also implies that the home effect has more impetus here than it did before. We still have a high proportion of draws but the increase in home wins clearly shows that these teams tend to win at home and struggle more with the away games, thus there is a greater home effect between them.

Finally if we remove EVE and consider SUN to BOL the home effect is 0.402, the mean home goals is 1.405 and the mean away goals has decreased to 0.786; the probability of a home win is 0.524, a draw 0.286 and an away win 0.190. The proportion of results between these 7 teams most accurately reflects the Premier League but there is still a reduced number of away wins; there is a significant increase in the number of home wins adding to the evidence for an increased home effect between these teams. With the mean number of away goals decreasing again this lower home effect does in fact have the bigger impact on results. From these investigations we can conclude that although the home effect decreases in the middle of the Premier League there is in fact a greater home effect between the teams involved.

4.2 Method 2: Moving Average Estimates

Figure (4.3) shows a moving average of the attack parameters obtained using increments of 20. The left graph shows this for all teams and the right shows specifically CHE, MNU and BUR. When looking at the left graph 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. From the right graph we can clearly see that the better teams are higher up and have the higher averages over the season; a team's average is denoted by the line and the order of these averages is very similar to the order of the final table of the Premier League. We see that CHE and MNU are fairly similar over the season; CHE dip slightly around 8, which was during January when they lost their leading striker Didier Drogba to the African Cup of Nations and they did not score as many goals. They increase dramatically at the end when they won games 7-0, 7-1 and 8-0. This was because the league was extremely close and CHE needed to win every game to ensure they won the league; they did this by ensuring they scored goals. BUR at the other end of the league started well and had a famous win over MNU; however results started to slip and their manager left to join another club. It took the new manager a while to turn things around and they never really succeeded. They increase at the end when they were facing relegation; they had to win all the games they could and as such had to score more goals.

Figure (4.4) shows the defence parameters for the same conditions as above; note that there is not as much variation here compared to the attack parameters 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. Considering the right graph we see that the averages roughly follow the ordering of the Premier League just as in the case of the attack parameters. CHE and MNU are again roughly similar apart from around 11; this was due to CHE loosing their first choice keeper to injury and their reserve keeper struggled to keep clean sheets. BUR have a worse defence but they are reasonably consistent until the end where they start to concede more goals; this is due to the fact that they needed to score goals and in committing men to attack they ultimately left themselves open at the back.

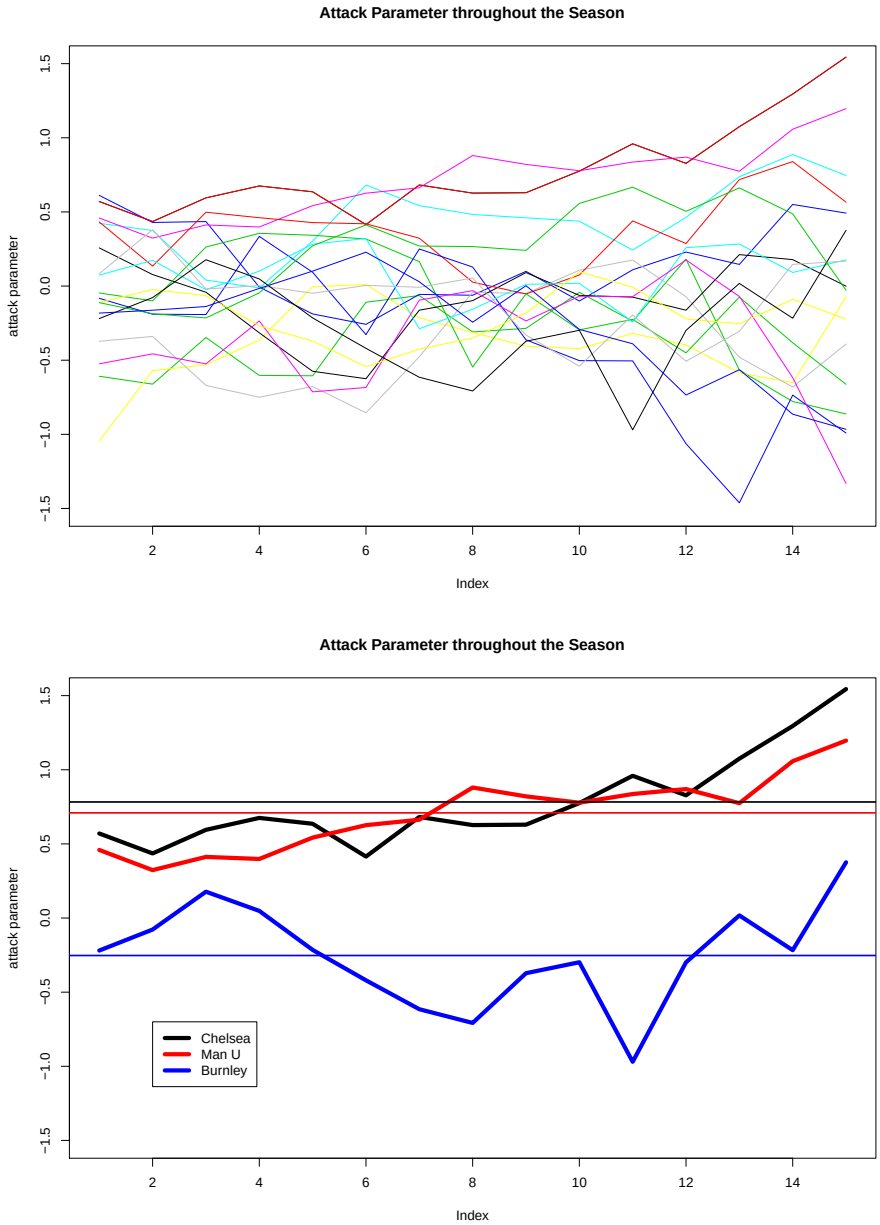

Figure 4.3: Moving average of the attack parameters.

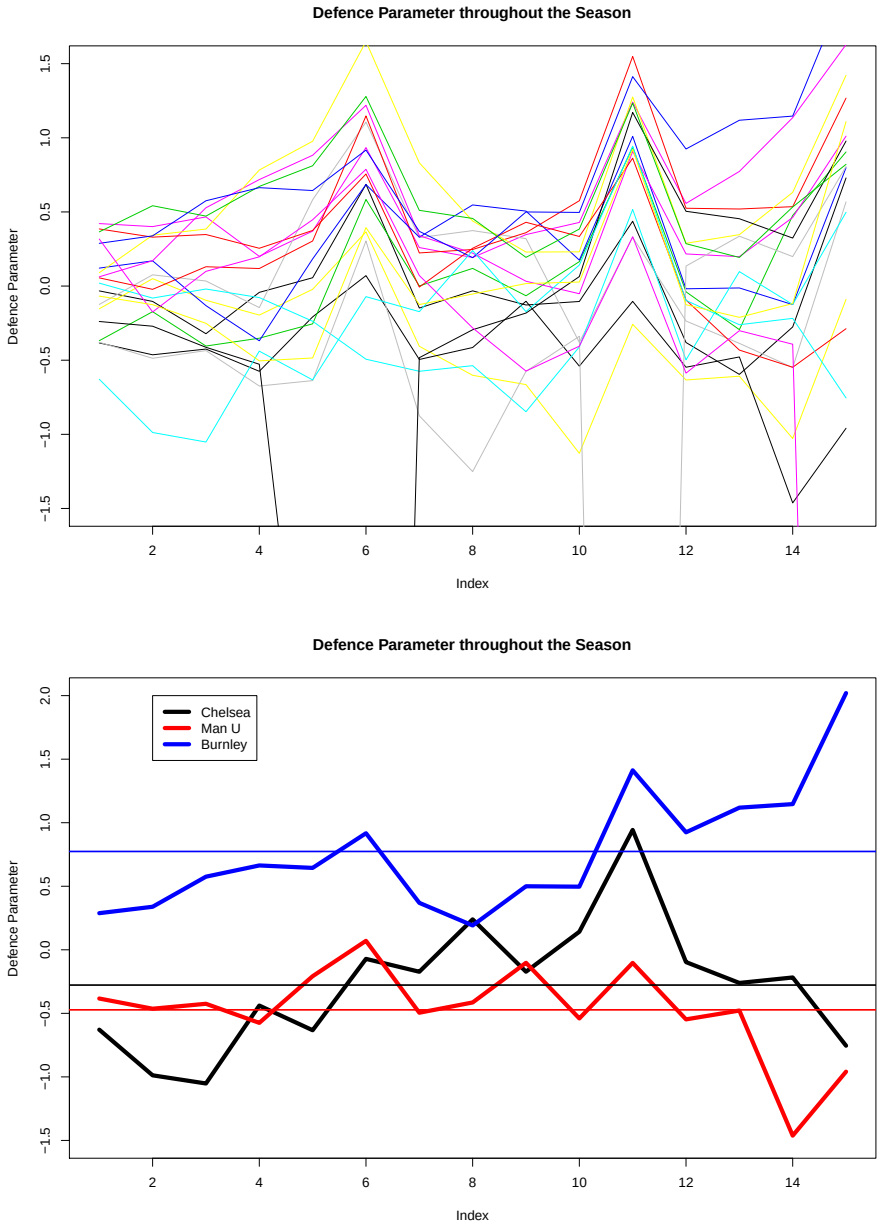

Figure 4.4: Moving average of the defence parameters.

Using these changing estimates we can again simulate the 2009/2010 Premier League. Table (4.2) shows the final table of results based on 100 simulations, again we have taken the means. We see that this method does not capture the final order of the Premier League very well, but it does retain the impression of the teams involved at the top and bottom. The table shows an increased number of goals over the Premier League, specifically for the teams in the bottom half. We find that the probability of a home win is 0.426, an away win 0.396 and a draw 0.178. Thus we see a higher than average number of away wins, suggesting that the home effect is underestimated. This discrepancy with the home effect is most likely the reason why the ordering is inaccurate, and why the teams towards the bottom the Premier League appear to score too many goals.

Team	Points	Goals	Conc	Goaldif
Man United	70.96	105.82	53.71	52.11
Tottenham	69.13	97.84	57.77	40.07
Aston Villa	67.38	80.01	52.57	27.44
Chelsea	65.52	105.42	64.55	40.87
Arsenal	65.25	106.40	68.36	38.04
Man City	62.84	99.48	78.67	20.81
Everton	61.89	87.26	65.29	21.97
Liverpool	60.38	82.75	61.85	20.90
Blackburn	55.63	74.14	75.38	-1.24
Birmingham	54.24	63.10	68.24	-5.14
Stoke	49.95	58.55	66.68	-8.13
Bolton	49.81	78.56	90.26	-11.70
Sunderland	45.75	67.37	85.41	-18.04
Fulham	45.60	53.36	64.22	-10.86
West Ham	44.20	68.38	92.36	-23.98
Wolves	42.64	55.73	84.46	-28.73
Hull	41.24	62.06	104.31	-42.25
Portsmouth	40.54	58.44	90.95	-32.51
Burnley	40.34	65.06	105.95	-40.89
Wigan	39.12	54.94	93.68	-38.74

Table 4.2: The Premier League based on 100 simulations using moving average estimates.

Figure (4.5) shows a moving average for the home effect over the season using a time step of 20. The red line shows the home effect we obtained when

using the whole season to gain our estimates. We see that the moving average changes greatly throughout the season, often with big changes between 2 consecutive time steps. We see that for most of the time the moving average is below the home effect obtained using the whole season. This is why we have a reduced home effect for this model and why we have a high number of away wins. This suggests that although this method takes into account a team's form throughout the season, the lack of accuracy when working out the home effect leads to a decreased home effect within the model. This means we see a high proportion of away wins over a season and as such our final results are not that accurate.

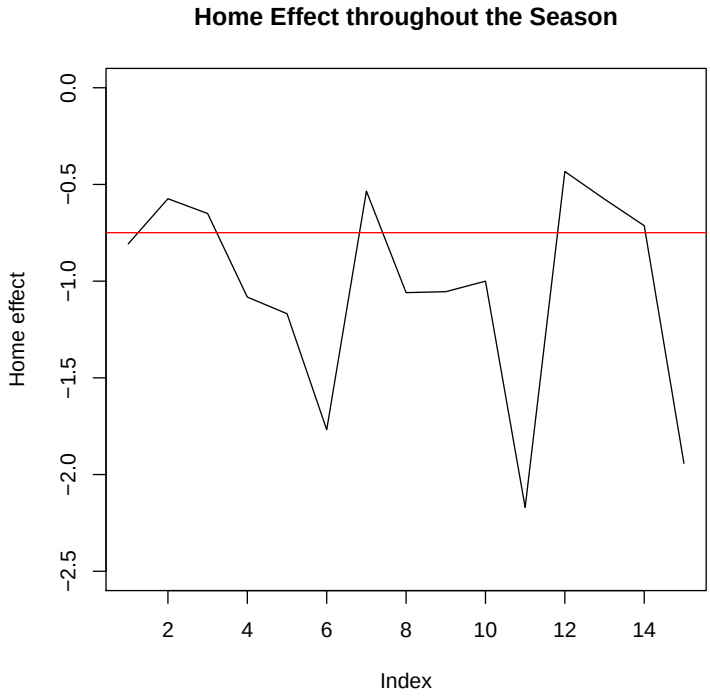

Figure 4.5: Moving average of the home effect.

4.3 A "Real" Life Example

The simulation of football matches does not just take place in a betting context; many video games use the principle, for example Fifa, Pro Evolution Soccer and Football Manager. Here we will take a look at Football Manager 2011 and examine how it simulates matches. Football Manager is a PC game developed by Sega and Sports Interactive. You manage a team controlling the transfers, tactics, finances etc. with the aim of winning matches. Most aspects are controlled by the player and can be changed at any time to effect a result. We note that Football Manager simulates depending on players, not teams; this means that players have a greater effect on the result and if we transfered the MNU players to WOL we would find that WOL started winning the Premier League - not a very realistic outcome.

To simulate a season in the game you need to manage a team. We therefore chose to manage Torquay United (League 2) and resigned on the first day. We chose Torquay as we believed it didn't matter who the manager of the League 2 side was, and it would have no effect on the final outcome of the Premier League. Note the simulations include the transfer windows and teams have budgets reflecting reality, with MNC having unlimited funds and WOL running on a shoestring. By going on holiday we were able to simulate the entire season without any influence. Table (4.2) shows the average points for the Premier League based on 20 Football Manager simulations. The simulations include the teams Newcastle (NCL), West Brom (WBR) and Blackpool (BPL) instead of POR, BUR and HUL who were relegated in the 2009/2010 season. We believe the new teams are similar to the relegated ones allowing us to compare the results. We first note that the ordering is roughly similar with MNU and CHE at the top with very little between them. These simulations however seem to place MNC lower down the table; this is perhaps because the makers of the game did not account for their continued rise and spending power. Football Manager also appears to capture how close the teams in the middle of the Premier league are with very little between their average points. The top teams have lower average points; this may be down to the player element in the simulation with certain players able to win games for lower teams which they should perhaps not win; it could also be because Football Manager doesn't appear to account for a home effect.

Team	Points
Man Utd	73.95
Chelsea	70.25
Arsenal	69.05
Tottenham	63.4
Liverpool	61.55
Everton	54.25
Man City	54.2
Aston Villa	52.35
Fulham	50.45
Stoke	48.6
Sunderland	45.8
Bolton	44.25
Birmingham	44.1
Wolves	43.6
Blackburn	42.9
West Ham	40.85
Newcastle	39.4
Wigan	34.3
West Brom	33.3
Blackpool	30

Table 4.3: The Premier League based on Football Manager 2011 simulations

If we consider one average simulated season we see that the probability of a home win is 0.413, an away win 0.35 and a draw 0.237. The probability of an away win is very high, with most of this increase taken from the home wins; this suggests that Football Manager does not account for a home effect, or if it does, it does not give a large enough advantage to the home team. The probability of a draw is reasonably accurate and it appears that Football Manager simulates these well. If we also consider the goals scored over the season we find that scores tend to be lower than we would expect, with fewer extreme results; this could also be down to the lack of a home effect with the home teams not scoring as many goals as they should do.

Football Manager therefore simulates with reasonable accuracy when looking at the end results but perhaps there are too many factors when considering the accuracy of individual results, with lower scores overall and no obvious home effect to help model results accurately.

Chapter 5

Conclusion

By considering properties of univariate and bivariate distributions, specifically those involving generating functions, we arrived at the bivariate Poisson distribution and looked at some of its properties. We used this distribution and the model proposed by Karlis and Ntzoufras, (2003) to predict football matches. We looked at the home effect in the Premier League and why it is needed in our model. We also considered how we obtain our attack and defence parameter estimates and the problems involved when choosing them.

We considered two methods for predicting matches; one where we took estimates over the whole season, and the other where we established a moving average. Under the first method we observed the slight differences in attack or defence which can win a team the league or cause them to get relegated. We also gained an understanding of the similarities and differences between teams from their attack and defence parameters.

We simulated the first method in R and produced an average final league table for the 2009/2010 Premier League. From the table we gained an understanding of how accurate this method was. It appears that this method is reasonably accurate when considering the final ordering of teams; it captures the teams at the top and bottom well and reflects how similar the teams in the middle are. This method slightly inflates the number of goals scored by a team and because of this we see an increase in higher scores over a season. This means that this method is good for predicting results in a general sense; that is, it can predict which team will win a game; we expect this generality as we are using estimates based on an entire season, and as such they are the most general estimates. Over a long period this method gives an accurate feel for what is occurring but it is not so good for specific match scores; this is because the result will have an increased number of goals in it.

By considering box plots of 100 simulated seasons we were able to see the differences between certain teams and we learned how similar some teams are, specifically those in the middle of the Premier League. We decided to take these teams and investigate the home effect between them. We found that amongst these teams, even though the numbers went down, there was in fact a greater home effect between them. This was most notable between the teams of Sunderland to Bolton where over half the games were won by the home team, with only 19% of games won by the away team. This means that the matches between these teams over a simulated season are not accurately represented as there are likely to be too many away wins. We could alter this method slightly by incorporating different rates of the home effect between certain teams to help rectify this problem.

Using the second method we looked at estimates for the attack and defence parameters which changed over time. We saw that the attack parameters had more variation than the defence parameters, leading us to believe that attacking is affected by the transfer window, whereas the team elements of defending mean that it is less effected. From this we concluded that it is easier to replace good defenders than good attackers. We saw that these changing estimates tell the story of the season reasonably well; the way they changed over time gave a good indication of a team's form throughout the season.

We again simulated this method in R, we found that this method is not as accurate and this appeared mainly down to the home effect. We saw that it had a changing home effect over the season which was generally too small; as such we saw an increased number of away wins and a high number of goals for the teams at the bottom of the Premier League, with teams scoring too many goals away from home.

We have also considered how Football Manager 2011 simulates football matches. We found that the game captured the final standings of the Premier League reasonably well; like the method based on season estimates we saw that it captured the general ordering of the league. There were some discrepancies with the ordering, with Manchester City for example being rather low in the league; this could be down to how the game was programmed or because the game appears to simulate using players rather than teams. Teams at the top of the league appeared to have a slightly lower points total than we expected and when we examined a season's results we found that the probability of an away win was 0.35. This is high compared with the 2009/2010 Premier League and shows that Football Manager does not include a home effect when simulating, or if it does, it is not as great as it should be. This decreased home effect means that we observed too many away wins and some teams

won away games which they possibly should not have won. The number of goals over a season was a little low and we did not observe as many extreme results as we possibly should have. This lack of goals is possibly down to the lack of a home effect, with home teams not scoring as many goals as they should do. Therefore Football Manager simulates the final results reasonably well but is not so good when considering specific match results.

If we had more time then we could consider how to change the home effect in the second method to make the results more accurate. This would incorporate a more accurate home effect and the form of each team into the model, which should in theory give us more accurate results. We could also consider different or more complicated models to see if we can improve the way in which we simulate football matches. Finally we could consider the home effect between the teams in the middle of the Premier League and look to incorporate that into our model. This would mean that we would have more accurate results between these specific teams; this idea could also be extended to include an individual home effect for every team.

Throughout this project we have looked at various ways to simulate football matches. 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. The methods we have looked at tend to overestimate the number of goals, whereas other people's methods (Football Manager) tend to underestimate due to the lack of a home effect. Also there is a lot of uncertainty in the outcome of a match as there are many variations of scorelines. All this uncertainty shows that although we may be able to get a picture of the general outcome, the specifics remain beyond us due to how unpredictable football can be with its many outcomes and unexpected results.

Appendix A

Bibliography

I/ Kocherlakota, S. and Kocherlakota, K. (1992), Bivariate Discrete Distributions, M. Dekker

II/ 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

III/ Dixon, M.J. and Coles, S.G. (1997), Modeling Association Football Scores and Inefficiencies in the Football Betting Market, Applied Statistics, 46, 265-280

IV/ Johnson, N., Kotz, S. and Kemp, A. (1992), Univariate Discrete Distributions, John Wiley & Sons

V/ Weisberg, S. (1985), Applied Linear Regression, John Wiley & Sons

VI/ Karlis, D. and Ntzoufras, I. (2007), "Bivariate Poisson Models Using The EM Algorithm",
http://www.stat-athens.aueb.gr/ jbn/papers/paper14.htm
(accessed 14/10/2010)

VII/ "The Premier League",
http://en.wikipedia.org/wiki/Premier_ League
(accessed 14/10/2010)

VIII/ "Official Site of The Premier league",
http://www.premierleague.com/page/Home/0,,12306,00.html
(accessed 14/10/2010)

IX/ "Football-Data.co.uk",
http://www.football-data.co.uk/englandm.php
(accessed 14/10/2010)

X/ "Probability Generating Function",
http://en.wikipedia.org/wiki/Probability-generating_function
(accessed 20/11/2010)

XI/ "Moment Generating Function",
http://en.wikipedia.org/wiki/Moment-generating_function
(accessed 20/11/2010)

XII/ "Cumulant",
http://en.wikipedia.org/wiki/Cumulant
(accessed 20/11/2010)

XIII/ "bet365 - Online Sports Betting",
http://www.bet365.com/home
(accessed 13/03/2011)

XIV/ Borman, S. (2004), "The Expectation Maximization Algorithm A short tutorial",
http://www.isi.edu/natural-language/teaching/cs562/2009/readings/B06.pdf
(accessed 02/04/11)

"They think it's all over! It is now!"
Kenneth Wolstenholme, 1966