Bayesian modelling of football outcomes

(using Skellam's Distribution)

Dimitris Karlis & Ioannis Ntzoufras
e-mails: {karlis, ntzoufras}@aueb.gr
Department of Statistics
Athens University of Economics and Business
Athens, Greece;

2007, 24-26th June, Manchester

Synopsis

1. Introduction.
2. Skellam's Distribution.
3. Bayesian infernece for the Skellam's distribution
4. The General Model.
5. Illustrative example and results.
6. Discussion and further work.

1 Introduction

- Poisson distribution → has been widely used as a simple modelling approach for predicting the number of goals in football (see, for example, Lee, 1997, Karlis and Ntzoufras, 2000).
- Empirical evidence has shown a (relatively low) correlation between the goals in a football game. This correlation must be incorporated in the model.
- It is reasonable to assume that the two outcome variables in football are correlated since the two teams interact during the game.

- We can assume → bivariate Poisson distribution (see Karlis and Ntzoufras, 2003).
- The marginal distributions are simple Poisson distributions, while the random variables are now dependent.
- Important issue → model extra variation of draws (mainly 0-0, 1-1, 2-2). This problem has been effectively handled by Dixon and Coles (1997) and Karlis and Ntzoufras (2003)

The Bivariate Poisson Model (Karlis and Ntzoufras, 2003)

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$$

where $BP(\lambda_1, \lambda_2, \lambda_3)$ is the bivariate Poisson distribution with parameters $\lambda_i,\ i = 1, 2, 3$ and probability function

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{k=0}^{min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k. \quad (1)$$

- Marginals: $X \sim Poisson(\lambda_1 + \lambda_3)$ and $Y \sim Poisson(\lambda_2 + \lambda_3)$
- Means and Variance: $E(X) = V(X) = \lambda_1 + \lambda_3$, $E(Y) = V(X) = \lambda_2 + \lambda_3$
- Covariance: $Cov(X, Y) = \lambda_3 > 0$
- Can be derived using latent variables $W_i \sim Poisson(\lambda_i)$, for $i = 1, 2, 3$ with $X = W_1 + W_3$ and with $Y = W_2 + W_3$.

- Covariates can be linked directly on the means of the latent variables as in Karlis and Ntzoufras (2003) or directly on the marginal means.
- A Simple Model (Karlis and Ntzoufras, 2003)
  - Response variables (X, Y) are the home and away goals in each game.
  - Consider the structure of Lee (1997) for λ_1 and λ_2

$$\log(\lambda_{1i}) = \mu + H + A_{HT_i} + D_{AT_i} \quad (2)$$
$$\log(\lambda_{2i}) = \mu + A_{AT_i} + D_{HT_i} \quad (3)$$

    μ: constant; H: home effect; A_k, D_k attacking and defensive parameters of team k; HT_i, AT_i home and away team in i game.
  - Constant covariance λ_3

Advantages

- Accounts for the covariance which can be modelled
- Has he latent variables decomposition which can be used for data augmentation
- Poisson Marginals
- Easy interpretation
- Easy extension to similar models.

Disadvantages

- How to model λ_3? Is constant λ_3 sufficient? Does λ_3 varies across teams or time?
- Underestimates the number of draws in certain cases (while it provides better results on this from the double Poisson model)
- Poisson Marginals while empirical evidence has shown slight over-dispersion: need to adjust
- The model allows only for positive correlation, if the data show negative correlation (even slow) the model cannot handle it

marginal means	λ_3	win	draw	loss	ratio
1.5	0.00	0.378	0.243	0.378	
	0.02	0.378	0.245	0.378	1.008
	0.05	0.376	0.248	0.376	1.012
	0.10	0.374	0.253	0.374	1.020
	0.20	0.368	0.264	0.368	1.044
2.0	0.00	0.396	0.207	0.396	
	0.02	0.396	0.208	0.396	1.006
	0.05	0.395	0.210	0.395	1.008
	0.10	0.394	0.213	0.394	1.014
	0.20	0.390	0.219	0.390	1.030

Table 1: The effect on the predicted probabilities when ignoring the covariance; the BP increases the probability of a draw.

Diagonal Inflated Bivariate Poisson Model

Under this approach a diagonal inflated model is specified by

$$P_D(x,y) = \begin{cases} (1-p)BP(x,y \mid \lambda_1,\lambda_2,\lambda_3), & x \neq y \\ (1-p)BP(x,y \mid \lambda_1,\lambda_2,\lambda_3) + pD(x,\boldsymbol{\theta}), & x = y, \end{cases} \quad (4)$$

where $D(x,\boldsymbol{\theta})$ is discrete distribution with parameter vector $\boldsymbol{\theta}$. Such models can be fitted using the EM algorithm.

Important: diagonal inflation improves in several aspects: better draw prediction, overdipsersed marginals, introduce correlation

Skellam's distribution (1)

For any pair of variables (X,Y) that can be written as

$$\begin{array}{llll} X = & W_1 + W_3, & Y = & W_2 + W_3 \text{ with} \\ W_1 \sim & Poisson(\lambda_1), & W_2 \sim & Poisson(\lambda_2) \text{ and} \\ W_3 \sim & \text{any distribution with parameters } \boldsymbol{\theta}_3 & & \end{array}$$

then

$$Z = X - Y \sim PD(\lambda_1,\lambda_2)$$

(Poisson difference or Skellam's distribution with parameters λ_1 and λ_2).

Skellam's distribution (2)

$$Z = X - Y \sim PD(\lambda_1,\lambda_2)$$

Poisson difference or Skellam's distribution with $\begin{cases} \text{Mean} & E(Z) = \lambda_1 - \lambda_2 \\ \text{Variance} & Var(Z) = \lambda_1 + \lambda_2 \end{cases}$

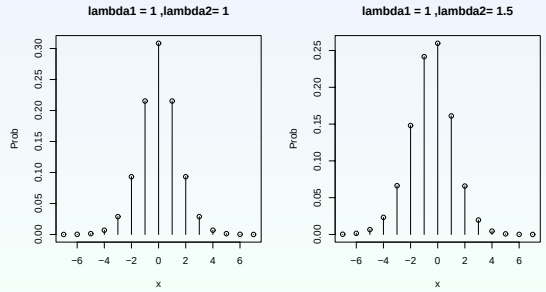
and density function

$$f_{PD}(z|\lambda,\lambda_2) = P(Z=z|\lambda_1,\lambda_2) = e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}\left(2\sqrt{\lambda_1\lambda_2}\right). \quad (5)$$

for all $z \in \mathcal{Z}$, where $I_r(x)$ is the modified Bessel function of order r

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{k=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^k}{k!\Gamma(r+k+1)}.$$

(see Abramowitz and Stegun, 1974, pp. 375).



Advantages

- Removes additive covariance (do not need to model it)
- Has an Poisson latent variable interpretation
- Does not assumes Poisson marginals (i.e. one may assume an overdispersed marginal distribution)
- Easy interpretation
- Simpler Model set-up than the corresponding Biv. Poisson model

Disadvantages

- Discards part of the information
- Cannot model covariance and final full score (only differences).

Bayesian approach

(see, Karlis and Ntzoufras,2006, SIM)

Used Discrete differences of dental data to

- test for differences of means between two repeated measurements
- test for zero inflated components

using the Bayes Factor approach (and RJMCMC to estimate it).

The same approach can be used to quantify (using the Bayes factor) Used Discrete differences of dental data to

- the importance of home effect (by testing for the equality of expected home and away goals) and
- the excess of draws (by testing the importance of the zero inflated component).

Skellam's Distribution for Football Scores

- Response variable: $Z = X - Y$ the goal difference in each game.
- Same structure for parameters λ_1 and λ_2 as in Bivariate Poisson:

$$\log(\lambda_{1i}) = \mu + H + A_{HT_i} + D_{AT_i} \quad (6)$$
$$\log(\lambda_{2i}) = \mu + A_{AT_i} + D_{HT_i} \quad (7)$$

  μ: constant; H: home effect; A_k, D_k attacking and defensive parameters of team k; HT_i, AT_i home and away team in i game.

- Use the zero inflated variation of Skellam's distribution to model the excess of draws. Hence we define the zero inflated Poisson Difference (ZPD) distribution as

$$f_{ZPD}(0|\, p, \lambda_1, \lambda_2) = p + (1-p) f_{PD}(0|\, \lambda_1, \lambda_2) \quad \text{and}$$
$$f_{ZPD}(z|\, p, \lambda_1, \lambda_2) = (1-p) f_{PD}(z|\, \lambda_1, \lambda_2), \quad (8)$$

  for $z \in \mathcal{Z} \setminus \{0\}$; where $p \in (0,1)$ and $f_{PD}(z|\, \lambda_1, \lambda_2)$ is given by (5).

Posterior Distributions

The key element for building an MCMC algorithm for the above models is to use augmented data

- Sample latent data w_{1i} and w_{2i} from

$$f(w_{1i}, w_{2i}|z_i = w_{1i} - w_{2i}, \lambda_{1i}, \lambda_{2i}) \propto \frac{\lambda_{1i}^{w_{1i}}}{w_{1i}!} \frac{\lambda_{2i}^{w_{2i}}}{w_{2i}!} I(z_i = w_{1i} - w_{2i})$$

  where $I(A) = 1$ if A is true and zero otherwise.

- Sample $[\delta_i|z_i, \lambda_{1i}, \lambda_{2i}] \sim Bernoulli(\tilde{p}_i)$ from

$$\tilde{p}_i = \frac{p}{p + (1-p) f_{PD}(z_i|\ \lambda_{1i}, \lambda_{2i})}$$

  [δ_i indicates the mixing (zero inflated or PD) component]

Then we MCMC algorithm for model parameters is similar to the one used in Poisson regression models.

We can additionally model p as in logistic regression models.

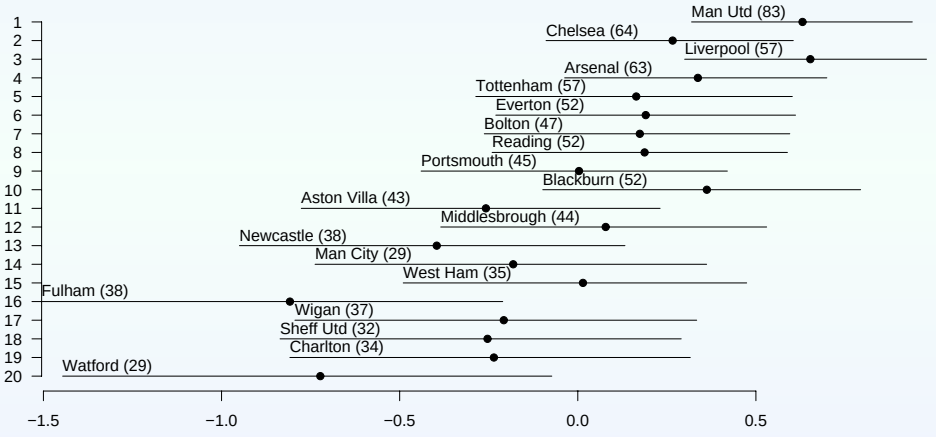
Example: Premiership season 2006-7 data

Data from the premiership for the 2006-7 season are used to fit the PD and ZPD models.

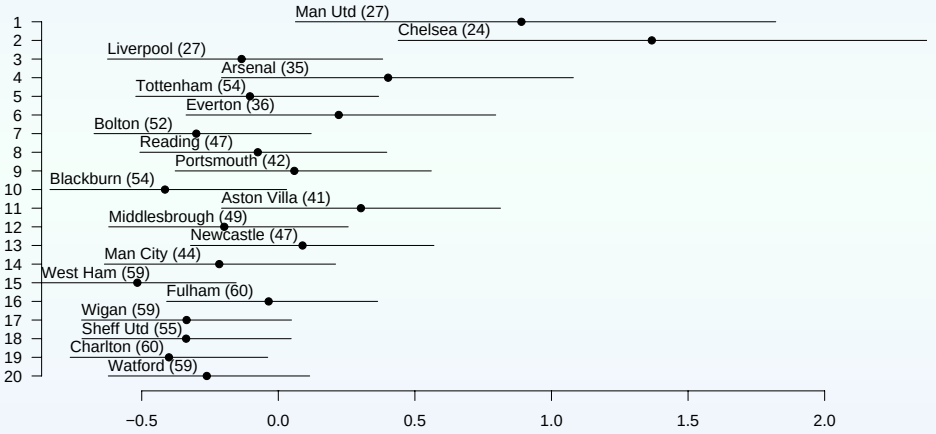
Results using PD

Posterior summaries for μ and home effect

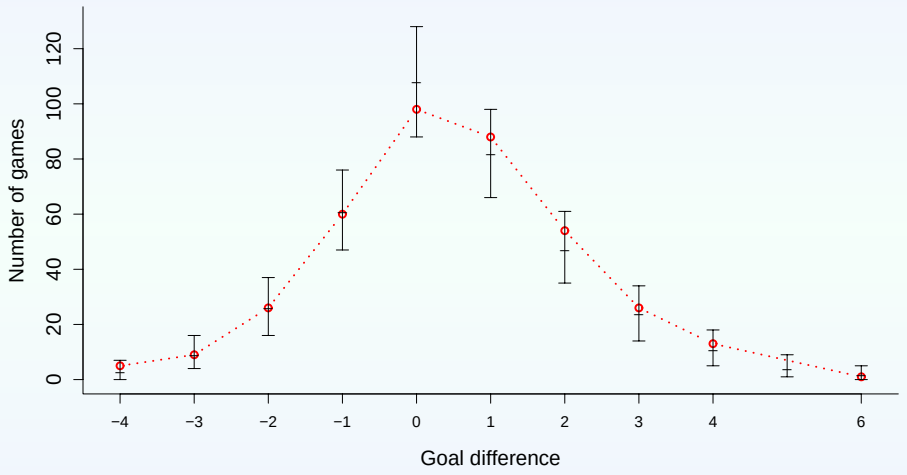
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	s.d.
mu	-0.793	-0.457	-0.390	-0.386	-0.316	-0.024	0.109
home	0.165	0.384	0.432	0.436	0.486	0.720	0.074



95% Posterior intervals for attacking coefficients (A_i)



95% Posterior intervals for defensive coefficients $[(-1) \times D_i]$



95% Posterior intervals for predicted differences

(in red=observed differences; '-' indicates the posterior median)

Posterior Predictive Table				Observed Final Table			
Pred.(Obs.)		Post. Expectations		Obs.		Obs. values	
Rank	Team	Pts	G.Dif.	Rank	Team	Pts	G.Dif.
1	(1) Man Utd	86.7	56.0	1	Man Utd	89	56
2	(2) Chelsea	81.0	40.0	2	Chelsea	83	40
3	(3) Arsenal	70.5	28.0	3	Liverpool	68	30
4	(3) Liverpool	69.4	30.2	4	Arsenal	68	28
5	(6) Everton	62.5	16.0	5	Tottenham	60	3
6	(8) Reading	55.5	5.4	6	Everton	58	16
7	(5) Tottenham	54.0	3.2	7	Bolton	56	-5
8	(9) Portsmouth	53.3	2.8	8	Reading	55	5
9	(10) Blackburn	51.8	-2.1	9	Portsmouth	54	3
10	(11) Aston Villa	51.5	1.6	10	Blackburn	52	-2
11	(12) Middlesbrough	49.0	-4.7	11	Aston Villa	50	2
12	(7) Bolton	49.0	-5.7	12	Middlesbrough	46	-5
13	(13) Newcastle	43.8	-8.8	13	Newcastle	43	-9
14	(14) Man City	41.8	-14.8	14	Man City	42	-15
15	(15) West Ham	38.6	-24.3	15	West Ham	41	-24
16	(17) Wigan	38.1	-21.6	16	Fulham	39	-22
17	(18) Sheff Utd	37.0	-23.0	17	Wigan	38	-22
18	(19) Charlton	35.7	-25.9	18	Sheff Utd	38	-23
19	(16) Fulham	33.1	-22.0	19	Charlton	34	-26
20	(20) Watford	29.7	-30.3	20	Watford	28	-30

Posterior Predictive Table				Observed Final Table			
Pred.(Obs.)		Post. Expectations		Obs.		Obs. values	
Rank	Team	Pts	G.Dif.	Rank	Team	Pts	G.Dif.
1	(1) Man Utd	86.7	56.0	1	Man Utd	89	56
2	(2) Chelsea	81.0	40.0	2	Chelsea	83	40
3	(3) Arsenal	70.5	28.0	3	Liverpool	68	30
4	(3) Liverpool	69.4	30.2	4	Arsenal	68	28
5	(6) Everton	62.5	16.0	5	Tottenham	60	3
6	(8) Reading	55.5	5.4	6	Everton	58	16
7	(5) Tottenham	54.0	3.2	7	Bolton	56	-5
8	(9) Portsmouth	53.3	2.8	8	Reading	55	5
9	(10) Blackburn	51.8	-2.1	9	Portsmouth	54	3
10	(11) Aston Villa	51.5	1.6	10	Blackburn	52	-2
11	(12) Middlesbrough	49.0	-4.7	11	Aston Villa	50	2
12	(7) Bolton	49.0	-5.7	12	Middlesbrough	46	-5
13	(13) Newcastle	43.8	-8.8	13	Newcastle	43	-9
14	(14) Man City	41.8	-14.8	14	Man City	42	-15
15	(15) West Ham	38.6	-24.3	15	West Ham	41	-24
16	(17) Wigan	38.1	-21.6	16	Fulham	39	-22
17	(18) Sheff Utd	37.0	-23.0	17	Wigan	38	-22
18	(19) Charlton	35.7	-25.9	18	Sheff Utd	38	-23
19	(16) Fulham	33.1	-22.0	19	Charlton	34	-26
20	(20) Watford	29.7	-30.3	20	Watford	28	-30

Deviations between observed and predictive measures

Comparison	Deviation
1. Relative Frequencies (counts/games)	1.06%
2. Frequencies (counts)	4.04
3. Relative Frequencies of win/draw/lose	2.20%
4. Frequencies of win/draw/lose	8.30
5. Expected points	3.02
6. Expected goal difference	0.28

$$Deviation = \frac{1}{K}\sqrt{\sum_{i=1}^{K}\left(E(Q_i^{Pred}|\boldsymbol{y}) - Q_i^{obs}\right)^2}$$

where K is the lengths of vector $\boldsymbol{Q}$ ($K = 13$ for comparisons 1-4, $K = 20$ for comparisons 5-6).

Results using ZPD

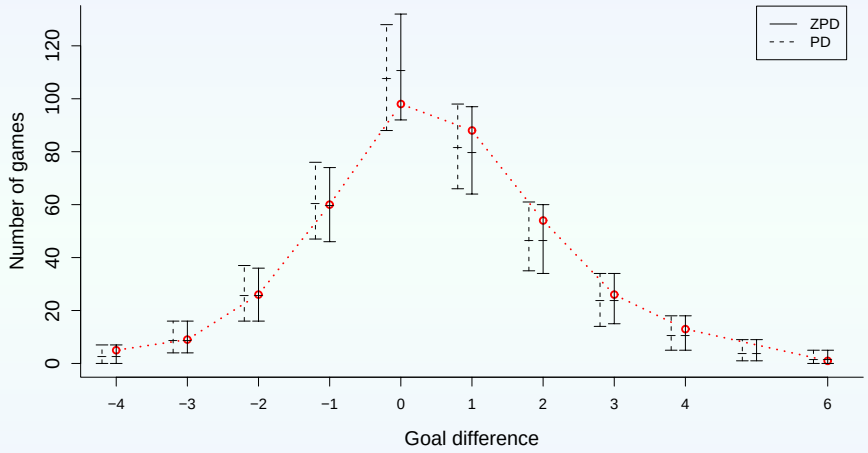
Posterior summaries for μ and home effect

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	s.d.
mu	-0.785	-0.432	-0.355	-0.360	-0.284	-0.019	0.115
home	0.184	0.378	0.432	0.434	0.486	0.705	0.082
p	0.000	0.006	0.014	0.018	0.025	0.114	0.015

- The posterior distribution of p is close to zero
- Small deviations between μ for PD and ZPD
- Home effect is similar in both models

Results from PD

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	s.d.
mu	-0.793	-0.457	-0.390	-0.386	-0.316	-0.024	0.109
home	0.165	0.384	0.432	0.436	0.486	0.720	0.074



95% Posterior intervals for predicted differences

(in red=observed differences; '-' indicates the posterior median)

Comment: No major differences between the two models are observed.

Deviations between observed and predictive measures

	Deviation	
Comparison	PD	ZPD
1. Relative Frequencies (counts/games)	1.06%	1.32%
2. Frequencies (counts)	4.04	5.04
3. Relative Frequencies of win/draw/lose	2.20%	2.80%
4. Frequencies of win/draw/lose	8.30	10.65%
5. Expected points	3.02	3.07
6. Expected goal difference	0.28	0.40

Comments

- Zero inflation component does not improves the performance (as expected)
- Prediction of goal differences for ZPD are worse than the corresponding ones for PD
- Nevertheless, differences in the final rankings are minimal

Discussion and further research

- Although in the simple and bivariate Poisson model we underestimate draws here using PD we have overestimated draws.
- No zero inflation is needed. We might need to add a component which will reduce the predicted draws.
- Can covariates on p improve the model?
- Apply Bayesian variable selection and Bayesian model averaging techniques.
- Apply other distributions defined on the same range.

Our Related Publications

1. Karlis D. and Ntzoufras, I. (2006). Bayesian analysis of the differences of count data. *Statistics in Medicine*, **25**, 1885-1905.
2. Karlis, D. and Ntzoufras, I. (2005). Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R. *Journal of Statistical Software*, Volume **10**, Issue 10.
3. Karlis, D. and Ntzoufras, I. (2003). Analysis of Sports Data Using Bivariate Poisson Models. *Journal of the Royal Statistical Society*, D, (Statistician), **52**, 381 – 393.
4. Karlis, D. and Ntzoufras, I. (2000). On Modelling Soccer Data. *Student*, **3**, 229–244.

Additional References

- Lee, A.J. (1997). Modeling scores in the Premier League: Is Manchester United really the best? *Chance*, **10**, 15-19.
- Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in football betting market. *Applied Statistics*, **46**, 265-280.