Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Odds Modelling and Testing Inefficiency of Sports Bookmakers Rmodel

®yo Lian Hu, ENG
Coursera / Scibrokes®

Abstract

In this paper I am applied a diagonal inflated biviriate poisson as well as a simple staking model whereby evaluate the efficiency of odds price offered by 29 sports bookmakers.

Keywords: keywords, Bivariate poisson, Multivariate discrete model, Betting strategy, Soccer, English Premier League, Expected return, Maximum likelihood, Statistical forecast, Bookmakers, R, Excel.

1. Introduction

The odds modelling in Europe and United States are very popular since decades. However statistical odds modelling and algorithmic staking has not yet popular in Far East Asia.

By refer to *Dixon & Coles 1996*[1], *Karlis & Ntzoufras 2005*[2] and also *Dixon & Pope 2004*[3] I tried to collect soccer data from year 2006 to 2011. The purpose of the research is testing the inefficiency of soccer odds offered by 29 bookmakers as well as making profit from bookmakers.

The paper *Dixon & Coles 1996* inspired by the *Maher 1982*[4] to identify the offence and defence index of every single team where *Karlis & Ntzoufras 2005* enahanced to be more complicated model. *Moya 2012*[5] taken 40000 customers' data from bwin to analayse and profit and lose and applied diversified staking strategies to make profit from bookmakers. *Goddard 2004*[6] model an ordered probit regression and placed stkes on English soccer leagues

[1]Refer to reference paper 02
[2]Refer to reference paper 08
[3]Refer to reference paper 05
[4]Refer to reference paper 01
[5]Refer to reference paper 10
[6]Refer to reference paper 09

from 1998 to 2002 and finally yeild `1998/99` = 0.116, `1999/00` = 0.008, `2000/01` = -0.008, `2001/02` = 0.160.

Well, *Dixon & Robinson 1997*[7] has built a rebirth model on 90 minutes In-Play soccer gaming. *Crowder, Dixon, Anthony & Robinson 2001*[8] applied MCMC[9] model for soccer result prediction and do a comparison with previous *Dixon & Coles 1996* model where concludes that previous model forecast more precisely.

Similar with *Dixon & Coles 1996*, *Karlis & Ntzuofras 1998*[10] has encountered an issue which is a number of nil-nil tied games. while *Dixon & Coles 1996* applied an inflation on low scores games while *Karlis & Ntzoufras 2005* built an extra distribution parameter to settle it.

The latest research paper wrote by Dixon is that *Dixon & Pope 2004* which have reviewed the previous model and testing the efficiency on correct score of 3 major firms in UK. *Karlis & Ntzuofras 2007* make a summary of evolution on his research which is apply Skellam's distribution on bivariate poisson model to resolve the obstacle of draw games.

Section 2 discribe a statiscal model applicable to soccer odds modelling. Section 3 talk about the dataset while section 4 model focus on staking model. Section 5 present the result and last section conclude.

1.1. Code formatting

Don't use markdown, instead use the more precise latex commands:

- `Java`
- **plyr**
- `print("1+2")`

2. Modelling

2.1. Basic Model

As mentioned in *Karlis & Ntzuofras 2005*, bivariate Poisson models are appropriate for modeling paired count data exhibiting correlation. Paired count data arise in a wide context including:

- marketing (number of purchases of different products)
- epidemiology (incidents of different diseases in a series of districts)
- accident analysis (number of accidents in a site before and after infrastructure changes)
- medical research (the number of seizures before and after treatment)

[7] Refer to reference paper 03
[8] Refer to reference paper 04
[9] Markov Chain Monte Carlo model
[10] Refer to reference paper 06

- sports (the number of goals scored by each one of the two opponent teams in soccer)
- econometrics (number of voluntary and involuntary job changes)

Where I just to name a few among the use.

Bivariate Poisson regression models

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

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 $Y = W_2 + W_3$.

Figure 1: Bivariate Poisson regression models

From above formula, bivariate poisson basically measure the correlationship between X and Y compare to double poisson models. However, as I mentioned which is *Dixon & Coles 1996* modified a little on the score 0-0, 1-0, 1-1 and vice versa.

- 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

Figure 2: Double Poisson regression models

A Double poisson model can be easily applied by generalized linear model. The covariates is a constant parameter across all soccer matches or teams as we know from figure 2.

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

Figure 3: Diagonal Inflated Bivariate Poisson regression models

Diagonal Inflated Bivariate Poisson regression models

Since the bivariate is not accurate enough and applicable to predict the real life soccer result. *Karlis & Ntzuofras 2005* introduced a more complicated model which able to inflated the probabilities of the occurrance on draw games.

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

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

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

Figure 4: Skellam's Distribution for Football Scores

Well, when we talk about the parameter to measure the correlationship. How can we know what models might fit into it? *Karlis & Ntzuofras 2005* has compare few models which are :

- Discrete distribution (with an adjustable paramters)
- Poisson distribution
- Geometric distribution

They built 12 statistical models to compare and get the best fit model. For more details

kindly refer to the paper.

2.2. Model Enhancement

There has a popular quote in sportsbook betting industry which is term as `FORM`. There is a flutuation of the ability and aggresiveness on sports competition as time goes by. Lets review the *Dixon & Coles 1996* model and fit the decay parameter into our basic model.

Choice of Weighting Function ϕ

There are various possible choices for the weighting function ϕ in equation One possibility would be

$$\phi(t) = \begin{cases} 1 & t \leqslant t_0, \\ 0 & t > t_0, \end{cases}$$

in which case, at time t, all results within the last t_0 time units would be given equal weight in the inference. Instead, we work with the model

$$\phi(t) = \exp(-\xi t),$$

in which all previous results, downweighted exponentially according to a parameter $\xi > 0$, are included in the inference at time t. The static model arises as the special case $\xi = 0$, whereas taking increasingly large values of ξ gives relatively more weight to the most recent results.

Optimizing the choice of ξ is problematic, since equation defines a sequence of non-independent 'likelihoods', whereas we require ξ such that the overall predictive capability of the model is maximized. In fact, in subsequent sections, we restrict attention to the prediction of match outcomes rather than match scores. Therefore it is pragmatic to choose ξ to optimize the prediction of outcomes. First note that the probability of a home win in match k is estimated as

$$p_k^{\mathrm{H}} = \sum_{l,m \in B_{\mathrm{H}}} \Pr(X_k = l,\ Y_k = m)$$

where $B_{\mathrm{H}} = \{(l,\ m):\ l > m\}$, and the score probabilities are determined from the maximization of model (4.5) at $t(k)$, the time of match k. Similar expressions hold for p_k^{A} and p_k^{D}, the probabilities of an away win and a draw respectively. Now define

$$S(\xi) = \sum_{k=1}^{N} (\delta_k^{\mathrm{H}} \log p_k^{\mathrm{H}} + \delta_k^{\mathrm{A}} \log p_k^{\mathrm{A}} + \delta_k^{\mathrm{D}} \log p_k^{\mathrm{D}})$$

where, for example, $\delta_k^{\mathrm{H}} = 1$ if match k is a home win and $\delta_k^{\mathrm{H}} = 0$ otherwise, and p_k^{H}, p_k^{A} and p_k^{D} are the maximum likelihood estimates from model , with weighting parameter set at ξ. Considering only the outcomes, and not the scores, equation is the analogue of a predictive profile log-likelihood. A plot of $S(\xi)$ against ξ, with time units taken to be half-weeks, is given in Fig. 1. The function is maximized at $\xi = 0.0065$, and all subsequent results are given with respect to this choice of ξ, though in fact the results are robust across a range of ξ-values.

Figure 5: decay rates

After simulation, I get the optimal decay rate which is almost `0.0065` and similar with *Dixon & Coles 1996*. However, due to I consider the soccer matches has come out result once the whistle is blew. Therefore I've tried to build another model which is similar with Weibull model to make the decay rate flexible compare to constantly annum. few models, which are:

- Count in the soccer result once a soccer match is finished to get a dynamic decay rates.

- Follow *Dixon & Coles 1996* which taken a constant decay rates for a soccer session.

I got a vector of decay rates around `0.0045` with the standard deviation not more than 1~10%. which is similar with the model at MatchOdds.org.

3. Data

3.1. Soccer Sports Dataset

3.2. Odds Price Dataset

I manually copy and paste the odds price from 500Wan.com. You are feel free to browse over the dataset via 200611 EngAllOdds[11]

4. Staking Model

4.1. Betting Strategies

As I mentioned in section Model Enhancement on the decay rates. In order to test the efficiency and the return of investment, I've taken both models in algorithmic simulations.

4.2. Preview of Returns.

5. Conclusion

5.1. Conclusion

5.2. Future Works

6. Appendices

6.1. Documenting File Creation

It's useful to record some information about how your file was created.

- File creation date: 2016-05-06
- R version 3.2.3 (2015-12-10)
- R version (short form): 3.2.3

[11] The spreadsheet file locate inside my previous project which is **Odds Modelling and Testing Inefficiency of Sports-Bookmakers** *2008-2010 by ®yo Eng Lian Hu*.

- **rticles** package version: 0.2
- File version: 1.0.0
- File latest updated date: 2016-05-05
- Author Profile: ®yo, Eng Lian Hu
- GitHub: Source Code
- Additional session information

[1] "2016-05-05 13:46:30 EDT" setting value
version R version 3.2.3 (2015-12-10) system x86_64, linux-gnu
ui X11
language (EN)
collate en_US.UTF-8
tz America/New_York
date 2016-05-05
sysname release "Linux" "3.10.0-229.20.1.el7.x86_64" version nodename "#1 SMP Tue Nov 3 19:10:07 UTC 2015" "scibrokes" machine login "x86_64" "unknown" user effective_user "ryoeng" "ryoeng"

6.2. Speech and Blooper

Firstly I do appreciate those who shade me a light on my research. Meanwhile I do happy and learn from the research. I do appreciated to take some spared time to write this thesis where the research has start from 2008 and finish in 2012. Infact I've finished my research on 2010 before I wrote a proposal to acquire the Ladbrokes[12] trading and hedge fund project in Scicom (MSC) Bhd and extended dataset soccer matches until 2012. Unfortunately the project has closed but I keep up learning journey to run my own company Scibrokes[13] some other days. I'll started work as customer service executive but in somewhere else next week, I am currently studying distance course data science at Coursera.org. You are feel free to browse over my CV at ®yo Eng Lian Hu.

I started my research journey when I decided to resign from Caspo Inc. to be an customer service operator in Scicom (MSC) Bhd. I've search, collected and read through thousands of research papers to get the applicable model in our real life investment. Fortunately I found and know a person Boffins -vs- Bookies (The Man Who Broke the World Leading Bookmakers) and start my learning from an outsider which don't know any statistical tools for modelling until successfully completed the research in year 2012. Kindly refer to My personal WordPress blog for more experience and bloopers.

Now I would like to share some bloopers during process this thesis.

- **Remarks** : Due to the mathematical LaTeX formula and greek letters unable use in **rticles** package. Here I forced to use some image for substitution.

[12]Ladbrokes is a world leader in the betting and gaming industry with over 2,700 betting outlets in the UK, Ireland, Belgium and Spain and over 800,000 active online customers. British public listed company which in the Fortune 500 and over hundred years business group.
[13]A registered company but not yet in operation. A prospective statistical hedge fund company.

- Due to the Microsoft Excel file inside my previous project **Odds Modelling and Testing Inefficiency of Sports-Bookmakers** *2008-2010 by ®yo Eng Lian Hu* is very huge. I tried to convert it to pdf format and attached as appendices but system keep endless processing there but no outcome. Secondly, huge dataset make it trouble to read into ®Studio and summarise and plotting.

6.3. Reference

1. **Modelling association football scores** *1982 by M.J Maher*
2. **Modelling Association Football Scores and Inefficiencies in the Football Betting Market.** *1996 by Mark Dixon and Stuart Coles*
3. **A Birth Process Model for Association Football Matches.** *1997 by Mark Dixon and Michael Robinson*
4. **Dynamic Modelling and Prediction of English Football League Matches for Betting.** *2002 by Martin Crowder, Mark Dixon, Anthony Ledford and Mike Robinson*
5. **The value of statistical forecasts in the UK association football betting market.** *2004 by Mark Dixon and Peter Pope*
6. **Statistical Modelling for Soccer Games: The Greek League.** *1998 by Dimitris Karlis and Ioannis Ntzoufras*
7. **Bayesian modelling of football outcomes (using Skellam's Distribution).** *2007 by Dimitris Karlis and Ioannis Ntzoufras*
8. **Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R.** *2005 by Dimitris Karlis and Ioannis Ntzoufras*
9. **John Goddard and Ioannis Asimakopoulos** *2004 by John Goddard and Ioannis Asimakopoulos*
10. **Statistical Methodology for Profitable Sports Gambling** *2012 by Fabián Enrique Moya*

Affiliation:

®yo Lian Hu, ENG
Coursera / Scibrokes®
09-11-02, Block Chengal, Taman Desaminium, 43300 Seri Kembangan, Selangor, Malaysia.
E-mail: englianhu@gmail.com / englianhu@scibrokes.com, +6017-2100905
URL: https://github.com/scibrokes/owner

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:*
MMMMMM YYYY *Accepted:* yyyy-mm-dd