Modelling the Scores of Premier League Football Matches

by: Daan van Gemert

The aim of this thesis is to develop a model for estimating the probabilities of premier league football outcomes, with the potential to form the basis for a profitable betting strategy against online bookmakers. Bets with a positive expected value can be made when the online bookies misprice their odds and when the model is accurate enough to detect these profitable betting opportunities. This thesis describes the development of the model and investigates the accuracy of the forecasts of this model compared to the predictions of online bookmakers. The model developed in this thesis estimates the final score of a football match, from which the predictions for the match result and the margin of victory can easily be calculated.

Data

The data described in this section contain information about all Premier League matches only. The championship games are not included for the analysis of the data, since the final purpose of this paper is to find an appropriate model which serves as a basis to place bets on matches played in the premier league competition. When estimating the model, information about all championship games from the past five seasons will also be included. This extra data gives more information about teams who played only one season in the premier league or are totally new to the premier league.

The dataset consists of 1900 matches played during the seasons 2004-2005 to 2008-2009. All games end after 90 minutes of play plus some injury time determined by the referees, so there is no extra time or penalty shootout involved. The result of a football game is expected to be influenced by a variety of explanatory variables. The model in this paper only makes use of the full time match results from the past. All other factors such as injuries and suspensions are ignored.

Over the past five seasons, an average number of 2,523 goals were scored in a premier league football match, with a variance equal to 2,640. The teams playing at home scored an average number of 1,468 goals, with a variance of 1,617. For the away teams, this average was 1,055 with a variance of 1,158. The observations are slightly overdispersed; the variance exceeds the mean for both the home and away goals.

Daan van Gemert

I obtained a BSc in Actuarial Sciences in june 2008 at the University of Amsterdam. This article is a summary of my master thesis written under the supervision of Dr. J.C.M. van Ophem in order to obtain the master degree in Econometrics at the same University. The subject of my thesis is mainly chosen because of my great interests in football and gambling.

Marginal Distributions for home and away goals

The number of goals scored by the home team and the number of goals scored by the away team are treated independently at first. The two simple count variables Y_1 and Y_2, the number of home and away goals respectively, with observations y_{1g} and y_{2g} (g=1,...,1900) are considered. This section compares the censored versions of the Poisson, zero inflated Poisson and Negative Binomial distributions to analyse these count data variables. Censoring takes place after five goals. The results of the fits of these distributions to the observed data are shown in tables 1 and 2.

The Negative Binomial distributions give a closer fit to the data then the Poisson distributions for both the home and away goals. This is not surprising, since the Poisson distribution is a special case of the Negative Binomial distribution (the Negative Binomial distribution converges to the Poisson distribution as the parameter r goes to infinity). The Negative Binomial distribution accounts for overdispersion by introducing the extra parameter r.

Looking at the Poisson models for the home and away goals, it seems that the Poisson distribution overestimates the number of matches where the home and away teams score one or two goals, at the cost of the games where the home and away teams do not score any goals. So the data contain more zeros than the Poisson distribution predicts. The zero inflated Poisson distribution includes a proportion (1- π) of extra zeros and therefore accounts for the problem of excess of zeros.

Table 1. Expected frequencies for the number of home goals, based on the censored Poisson, censored zero inflated Poisson and censored Negative Binomial distributions. Maximum likelihood parameter estimates are given at the bottom of the table, with their standard errors between parentheses.

Home Goals, censored distributions				
# Home Goals	Observed	Poisson	ZIP	Neg. Binomial
0	469	439.9	469.0	466.1
1	621	643.6	610.2	628.7
2	456	470.8	462.7	448.9
3	217	229.6	233.9	225.5
4	100	84.0	88.7	89.4
≥5	37	32.1	35.5	41.4
Total	1900	1900.0	1900.0	1900.0
	π_2 =		0.965 (0.016)	
Parameters	λ_2 =	1.463 (0.028)	1.516 (0.038)	1.465 (0.029)
	r_2 =			17.067 (7.27)

Table 2. Expected frequencies for the number of away goals, based on the censored Poisson, censored zero inflated Poisson and censored Negative Binomial distributions. Maximum likelihood parameter estimates are given at the bottom of the table, with their standard errors between parentheses.

Away Goals, censored distributions				
# Away Goals	Observed	Poisson	ZIP	Neg. Binomial
0	692	661.7	692.0	695.8
1	680	698.0	657.1	665.8
2	335	368.1	366.4	349.6
3	131	129.4	136.2	133.3
4	51	34.1	38.0	41.2
≥5	11	8.7	10.3	14.3
Total	1900	1900.0	1900.0	1900.0
	π_2 =		0.946 (0.024)	
Parameters	λ_2 =	1.055 (0.024)	1.115 (0.037)	1.055 (0.025)
	r_2 =			10.260 (3.79)

The validity of the different models for the marginal distributions of home and away goals are tested by performing chi-square goodness-of-fit tests. The test results are presented in table 3. According to these tests, the censored Negative Binomial distribution gives the best fit to both the home and away goals data.

Assuming that the marginal distributions for home and away goals are independent from each other is quite strong. Since teams compete each other in a football match, it is likely that there exists some correlation between the number of goals scored by the home team and the number of goals scored by the away team. In the next chapter, a model is presented which accounts for a dependence structure between the marginal distributions of home and away goals. Both the censored Negative Binomial as well as the censored zero inflated Poisson distributions are taken as the marginal distributions in the correlated model. The estimation results of the two different models are then compared.

The Model

The model presented in this section takes the following aspects into account (see also Dixon and Coles (1997)):

- The different abilities of teams in a match.
- Teams playing in their home stadium have a 'home ground advantage' over the away teams.
- A team's total ability is determined by its ability to attack and its ability to defend.
- The correlation of the scoring performances of two competing teams in a match.

The Goal Model: the marginal distributions

Consider a game between home team i and away team j. Let n be the number of different teams in the dataset. Let Y_{1ij} denote the number of goals scored by the home team and Y_{2ij} the number of goals scored by the away team. It is assumed that these two count variables are either drawn from a censored zero inflated Poisson distribution or from a censored Negative Binomial distribution. In both cases the parameters λ_{1ij} and λ_{2ij} are defined by

$$\lambda_{1ij} = \exp\{\gamma + \alpha_i - \beta_j\},$$
$$\lambda_{2ij} = \exp\{\alpha_j - \beta_i\}.$$

The parameter γ represents the advantage of the home team when playing at its home stadium. This home ground advantage parameter is assumed to be equal for every team[1]. The set of parameters α_i and β_i, $i = 1,\ldots,n$, measure team i's attack and defence strength. In words; in a match between home team i and away team j, the number goals scored by the home team depends on (home) team i's ability to attack, team j's ability to defend and the home ground advantage. The number of goals scored by the away team depends on team j's ability to attack and team

[1] In early stages of the development of the model, I have estimated different home effects for every team. This has led to insignificant parameter estimates.

i's ability to defend.

The Goal Model: the dependence structure

To model the relation between the two scoring processes, the method introduced by van Ophem (1999) is followed. Here, the bivariate cumulative normal distribution function is used to relate any two discrete random processes. If it is assumed that the two count variables Y_1 and Y_2 are marginally censored Negative Binomial distributed with unknown parameters (λ_{1ij}, r_1) and (λ_{2ij}, r_2), then the method to formalize the dependence structure between these two possibly correlated discrete stochastic variables is as follows:

- Set the numbers $\eta_1,...,\eta_5$ and $\gamma_1,...,\gamma_5$ equal to:

$$\eta_\kappa = \Phi^{-1}\left(\sum_{k=0}^{\kappa-1} p^{NB}\left(k;\lambda_{1ij},r_1\right)\right),$$
$$\gamma_\mu = \Phi^{-1}\left(\sum_{m=0}^{\mu-1} p^{NB}\left(m;\lambda_{2ij},r_2\right)\right)$$

  and define $\eta_6 = \gamma_6 = +\infty$. p^{NB} Denotes the density function of the censored NB distribution and Φ^{-1} is the standard normal quantile function.

- Maximize the following log likelihood function with respect to α_i ,β_i, γ, r_1, r_2 and ρ:

$$\log L\left(\alpha_i,\beta_i,\gamma,r_1,r_2,\rho;i=1,...,n\right) =$$
$$\sum_{g=1}^{N}\log\left[B\left(\eta_{y_{1g}+1},\gamma_{y_{2g}+1};\rho\right)-B\left(\eta_{y_{1g}},\gamma_{y_{2g}+1};\rho\right)\right. \quad (1)$$
$$\left.-B\left(\eta_{y_{1g}+1},\gamma_{y_{2g}};\rho\right)+B\left(\eta_{y_{1g}},\gamma_{y_{2g}};\rho\right)\right]$$

  where $B(\bullet,\bullet;\rho)$ is the bivariate normal cumulative distribution function with expectations 0, variances 1 and correlation ρ.

  The log likelihood function from the second step is obtained after setting $\Pr(Y_{1g} \le y_{1g},\ Y_{2g} \le y_{2g}) = B\left(\eta_{y_{1g}+1},\gamma_{y_{2g}+1};\rho\right)$. So the multivariate joint distribution function is constructed by specifying the dependence structure as a bivariate normal distribution on the transformed marginal distributions of Y_1 and Y_2, i.e. the Gaussian copula is used.

Results

Estimation of the model yields an estimated dependence parameter of $\hat{\rho}$ = 0.074 (0.018) for the model with Negative Binomial marginal distributions and a dependence parameter estimate of $\hat{\rho}$ = 0.070 (0.017) for the zero inflated Poisson model. So the number of goals scored by the home team and the number of goals scored by the away team are positively correlated. For both models the dependence parameter ρ is highly significant.

The estimation results of the ability parameters for the NB model, including their standard errors, can be found in table 4. The ZIP model gives comparable estimates. The teams are ranked according to the total ability; the sum of the estimated attacking and defending strength parameters. The standard errors of the estimated ability parameters are typically about 0.05 – 0.07.

The home ground advantage parameter γ is estimated to be equal to 0.30 (0.018) for the NB model and 0.31 (0.026) for the ZIP model. Thus; home teams have a highly significant home ground advantage over the away teams. The parameter estimates for π_1 and π_2 in the ZIP model are equal to $\hat{\pi}_1$ = 1.040 (0.010) and $\hat{\pi}_2$ =1.049 (0.016). These estimates are significantly bigger than one with p-values of 0.007% and 0.160%. This implies a reduction in the probability of zero goals, which is not what we expected since we observed an excess of zeros in the data for home and away goals. This result may be caused by the fact that in the correlated model, the dependence parameter already accounts for an extra proportion of the score 0-0. Estimation of the NB model yields dispersion parameter estimates of and $\hat{r}_1$ = 2,675,100 and $\hat{r}_2$ = 1,138,100. These extreme high estimates suggest that using marginal censored Poisson distributions for home and away goals might be equally appropriate as using the

Table 3. Results of chi square goodness of fit tests for the number of home goals and the number of away goals. The degrees of freedom for these tests are 4, 3 and 3 respectively.

Null-hypothesis		*p*-value	Conclusion
H_0:	*Number of home goals follows a…*		
-	censored Poisson distribution	10.39%	do not reject H0 at 10% level
-	censored Zero - Inflated Poisson distribution	38.89%	do not reject H0 at 10% level
-	censored Negative Binomial distribution	52.09%	do not reject H0 at 10% level
H_0:	*Number of away goals follows a…*		
-	censored Poisson distribution	0.80%	reject H0 at 1% level
-	censored Zero - Inflated Poisson distribution	4.21%	reject H0 at 5% level
-	censored Negative Binomial distribution	25.64%	do not reject H0 at 10% level

Table 4. Attacking and defending parameter estimates and standard errors from the NB model for all teams included in the dataset. Teams are ranked according to their total strength. The estimates are obtained with maximum likelihood and each match has equal weight in determining the parameter estimates.

	Team	$\hat{\alpha}$	se($\hat{\alpha}$)	$\hat{\beta}$	se($\hat{\beta}$)		Team	$\hat{\alpha}$	se($\hat{\alpha}$)	$\hat{\beta}$	se($\hat{\beta}$)
1	Chelsea	1.60	0.06	1.95	0.10	29	Watford	1.01	0.06	0.85	0.06
2	Man United	1.66	0.06	1.76	0.09	30	Cardiff	0.91	0.06	0.91	0.06
3	Liverpool	1.51	0.06	1.66	0.08	31	Bristol City	0.87	0.10	0.94	0.10
4	Arsenal	1.66	0.06	1.52	0.08	32	Hull	0.90	0.07	0.88	0.07
5	Everton	1.26	0.07	1.38	0.07	33	Burnley	0.85	0.06	0.92	0.06
6	Tottenham	1.38	0.06	1.19	0.07	34	Southampton	0.95	0.06	0.81	0.06
7	Aston Villa	1.33	0.06	1.16	0.07	35	Sheffield Weds	0.84	0.07	0.85	0.07
8	Man City	1.19	0.07	1.24	0.07	36	Colchester	1.06	0.09	0.63	0.09
9	Blackburn	1.21	0.07	1.17	0.07	37	Plymouth	0.81	0.06	0.88	0.06
10	Bolton	1.19	0.07	1.18	0.07	38	Leeds	0.77	0.08	0.91	0.08
11	Reading	1.19	0.06	1.15	0.07	39	Derby	0.92	0.06	0.76	0.06
12	Middlesbrough	1.16	0.07	1.11	0.07	40	Blackpool	0.85	0.10	0.80	0.09
13	Wigan	1.10	0.07	1.16	0.07	41	Leicester	0.72	0.07	0.93	0.07
14	Portsmouth	1.14	0.07	1.12	0.07	42	QPR	0.82	0.06	0.81	0.06
15	Newcastle	1.17	0.07	1.08	0.06	43	Norwich	0.91	0.06	0.72	0.06
16	Fulham	1.16	0.07	1.07	0.06	44	Coventry	0.85	0.06	0.74	0.06
17	West Ham	1.14	0.07	1.07	0.06	45	Doncaster	0.63	0.15	0.95	0.14
18	Birmingham	1.00	0.07	1.15	0.07	46	Millwall	0.62	0.11	0.92	0.10
19	Sunderland	1.06	0.06	1.02	0.06	47	Luton	0.95	0.09	0.59	0.08
20	West Brom	1.15	0.06	0.92	0.06	48	Barnsley	0.79	0.08	0.67	0.07
21	Sheffield United	0.99	0.06	1.04	0.06	49	Nott Forest	0.72	0.10	0.73	0.09
22	Swansea	1.03	0.12	0.99	0.14	50	Crewe	1.00	0.09	0.43	0.08
23	Wolves	1.00	0.06	0.96	0.06	51	Gillingham	0.68	0.15	0.71	0.12
24	Preston	0.98	0.06	0.98	0.06	52	Scunthorpe	0.70	0.15	0.68	0.12
25	Crystal Palace	0.96	0.06	0.99	0.06	53	Southend	0.72	0.14	0.53	0.11
26	Stoke	0.89	0.06	1.02	0.06	54	Brighton	0.55	0.11	0.68	0.09
27	Charlton	1.00	0.07	0.90	0.06	55	Rotherham	0.43	0.17	0.67	0.12
28	Ipswich	1.05	0.06	0.83	0.06						

marginal censored Negative Binomial. Apparently, the censored Negative Binomial distribution is not a very suitable choice for the marginal distributions in correlated model specified in the previous chapter. Taking censored Poisson distributions as the marginal distributions in the correlated model produces the same estimation results for all the parameters and standard errors (at least up to two decimals).

Betting Results

The data used in this section contain all matches played from the beginning of season 2009/2010 up to the 7th of February 2010 (a total of 242 games), including information about the odds from more than 30 bookmakers. Match odds, total goals odds and Asian handicap odds are all available in the data[2]. Similar to Kuypers (2000), the following betting strategy is applied: place €1 on outcome i of a particular match j if

[2] Match odds are the odds for betting on the match result (home win, draw or away win). Total goal odds are the odds for betting on the total number of goals scored in the match. Only two outcomes are separated; over or under than 2.5 goals. Asian handicap is a form of betting on the match result, where one of the two teams receives a virtual head start. The team who scores the most with the handicap applied is deemed the winner.

$$1 + r_1 < \text{Probability Ratio} = \frac{\text{Predicted Probability from Model}}{1/O_{ij}} < 1 + r_2$$

where O_{ij} are the best available odds on outcome i for match j.

Overall, the results of the Poisson model seem to be better than the results of the ZIP model. For betting on match results, the Poisson model would have made a maximum profit of €9.71. This profit was generated for $r_1 = 6\%$ and $r_2 = 13\%$, placing a total number of 62 bets, 24 of which were winning. So a total return of 15.66% was made. For betting on the total score of the match, most betting strategies would have resulted in a loss, and both models do not seem to be capable of generating a profit against the bookmakers. For Asian handicap betting however, profits are made for almost every strategy, and both models seem to perform better against the bookmaker on this form of betting. For the betting strategy based on the Poisson model with $r_1 = 2\%$ and $r_2 = 15\%$ a profit of €12.09 was made, generated by placing 76 bets (out of which 37 bets were winning and 17 bets were voided). This equals a total return of 15.91%.

The maximum returns of 15.66% and 15.91% for betting on match results and Asian handicap betting can be compared to the returns realized by using a random betting strategy: placing a €1 bet on one outcome at random in 62/ 76 randomly chosen matches. For betting on 62 randomly chosen match results, a simulation of 100,000 returns resulted in an average profit of - €1.98 (-3.20%) with an estimated standard deviation of 13.34. For Asian Handicap betting, the estimated expected profit of 76 random bets equals - €3.17 (-4.17%) with an estimated standard deviation of 9.22. So a random betting strategy obviously results in negative returns, but the total profit after 62 or 76 random bets show high standard deviations, and the positive returns of €9.71 and €12.09 are only 0.9 and 1.7 standard deviations above the returns obtained from random betting. Although some of the betting strategies seem to be profitable, this still does not say much about the capability of the model to make profit over the bookmakers. Because of the high standard deviations it is also not clear which model gives the better results.

Concluding remarks

This paper developed a model for estimating the scores of Premier League matches. Previous literature was extended by allowing for correlation between the scoring performances of the home and away team. A correlation coefficient of about 7% was found, suggesting that the number of goals scored by the home and away team are positively correlated. A structural shortcoming of the model presented in this paper is that every match has an equal weight in determining the ability parameters α_i and β_i. So it is assumed that teams have a constant attacking and defending strength over the whole period. This is very unrealistic, since a team's strength varies over time and tends to be dynamic. The stochastic development of the ability parameters should be specified to improve the model on this topic. It was found that profitable betting opportunities based on the model estimations do exist, comparing the odds from more than 30 bookmakers. However, due to the high variance of the realized returns the betting results have to be interpreted with caution.

References

Dixon, M. J. and S. G. Coles. "Modelling association football scores and inefficiencies in the football betting market." *Applied Statistics* 46.2 (1997):265 – 280.

Kuypers, T. "Information and efficiency: an empirical study of a fixed odds betting market." *Applied Economics* 32 (2000):1353 – 1363.

Ophem, H.v. "A general method to estimate correlated discrete random variables." *Econometric Theory* 15 (1999):228 – 237.