IE Interfaces
Vol. 20, No. 2, pp. 133-141, June 2007.

대한산업공학회 제2차 석사논문경진대회 수상논문

# 포아송 확률 모형을 이용한 축구 경기 결과 예측

성 현 · 장우진†

서울대학교 산업공학과

# Forecasting the Results of Soccer Matches Using Poisson Model

Hyun Seong · Woojin Chang

Department of Industrial Engineering, Seoul National University, Seoul 151-742

As the sales of the Sports Toto, the Korean lottery on sports games, have increased significantly in recent five years, interest in predicting the various results of sports matches has also been raised. Dixon and Coles (1997) proposed a bivariate Poisson model to predict the results of English soccer league matches. In this paper, we pay attention to the physical condition of players that may affect soccer match results and revise Dixon and Coles' model to consider probable fatigue due to the players' short rest followed by their frequent matches. We observed the fatigue effect in the match results, and found positive betting returns available when using our prediction model. Furthermore, the validity of probability-based odds in European and Korean betting markets is analyzed.

***Keyword:*** betting, soccer results, fatigue effect, maximum likelihood estimation, Poisson distribution

## 1. 서 론

2001년 10월부터 국내에 정식 발행되고 있는 스포츠토토의 매출액이 해마다 증가하면서, 스포츠 경기의 결과 예측에 대한 사람들의 관심이 커지고 있다. 스포츠토토는 축구, 농구 등의 스포츠 경기를 대상으로 경기가 시작되기 전에 결과를 예측하여 경기 결과에 따라 순위별로 환급금을 받는, (주)스포츠토토에서 발행하는 레저 게임을 말한다.

현재 발행되고 있는 스포츠토토의 방식에는 경기 결과를 맞춘 사람들이 총 매출액의 일부의 환급금을 나누어 가지는 패리뮤추얼(pari-mutuel) 방식과, 결과를 맞출 경우 각 가능한 결과에 미리 정해진 배당률(odds) 만큼의 배당을 받는 고정배당(fixed-odds) 방식이 있다. 패리뮤추얼 방식에서는 결과를 정확하게 예측했을 때에 받는 배당의 비율이 사람들의 베팅(betting) 분포에 따라 변하지만, 고정배당 방식에서는 이와 상관없이 배당에 따른 수익을 미리 계산해 볼 수 있기 있다. 따라서 고정배당 방식에서는 배당률이나 결과의 발생확률에 따라 포트폴리오를 구성하여 베팅하는 것이 가능하다. 이처럼 어느 한 경기에서 여러 가능한 결과가 일어날 각각의 확률을 구하는 것은 고정배당 방식에서 주요한 관심사가 되고, 이러한 확률을 정확하게 추정할 수 있다면 상대적으로 고평가된(over-valued) 배당률의 결과에 돈을 걸어 이익을 얻을 수 있는 기회가 생길 것이다.

스포츠토토에서 현재 발매 중인 스포츠 종류에는 축구, 야구, 농구, 배구, 골프, 씨름이 있지만, 이 논문에서는 국내 경기와 해외 경기를 포함하여 매주 발매가 가능하고 현재 가장 큰 비중을 차지하고 있는 축구 경기를 대상으로 한다.

축구 경기에 관한 모형과 분석에 대한 연구는 계속 되어오고 있는데, Dixon and Coles(1997)는 축구 경기에 대한 독립 포아송 분포(independent Poisson distribution)에 대한 가정을 바탕으로 한 각 팀의 득점에 대한 이변수 포아송 분포(bivariate Poisson distribution)를 통해 각 결과의 확률을 추정하고, 이를 이용하여 1990년대 중반의 잉글랜드 축구를 대상으로 한 고정배당 시장에서 수익을 낼 수 있는 기회가 존재한다는 것을 보여주었다. 또, Dixon and Robinson(1998)은 각 경기의 득점 시간을 기초로 한 Birth Process Model을 세우고, 어느 시점에서의 각 팀이 득점

†연락저자 : 장우진 교수, 151-742 서울시 관악구 신림동 산 56-1 서울대학교 산업공학과, Fax : 02-889-8560, E-mail : changw@snu.ac.kr
2007년 01월 접수, 1회 수정 후 2007년 05월 게재확정.

을 기록할 확률이 양 팀의 당시 득점 상황에 따라 다르다는 사실을 제시하였다. Greenhough *et al*.(2002)은 저득점 경기의 경우에는 일반적인 독립적인 포아송 분포로 축구의 득점 분포로 설명하는 것이 부적합하다는 근거를 보여주었고, Goddard *et al*.(2004)은 득점 데이터와 각 팀 간의 이동거리 등 여러 설명변수를 고려한 프로빗 회귀 모형으로 확률을 추정하였을 때에 양의 수익을 얻을 수 있음을 제시했다. 그리고 Dixon and Pope (2004)은 베팅 시장의 배당률이 실제의 확률에 비해 편향(biased) 되어있음을 보여주었다.

본 논문의 전개는 다음과 같다. 2장에서는 스포츠토토의 배당률과 확률, 그리고 여기에서 응용될 포아송 분포에 대하여 살펴보고, 3장에서는 짧은 휴식으로 인한 피로를 고려한 모형과 그 추정량을 살펴보고 기존의 Dixon and Coles(1997) 논문의 모형과 비교할 것이다. 4장에서는 이를 바탕으로 실제로 시장에 베팅했을 때의 결과에 대해 살펴보고, 마지막 5장에서는 결론과 향후 과제 및 앞으로의 연구방향에 대해 다룰 것이다.

## 2. 축구 베팅과 포아송 분포

### 2.1 스포츠토토와 베팅

스포츠토토는 정해진 스포츠 경기 결과를 예측하고, 그 예측이 맞았을 경우 배당률에 대한 배당금을 주는 일종의 복권이다. 패리뮤추얼 방식에서는 돈을 걸고 예측이 맞았을 때에 얻는 배당률이 시시각각 변화하여 베팅시점에 정확한 배당률을 알 수 없지만, 고정배당 방식에서는 돈을 걸기 전에 배당률이 정해지기 때문에 패리뮤추얼 방식에서와 같은 어려움이 없다.

스포츠토토에서의 배당률이란, 1단위의 금액을 걸었을 때 얻을 수 있는 수입의 비율을 뜻한다. 예를 들어 배당률이 3일 때 1단위를 걸었다면, 예측이 맞을 경우 걸었던 1단위를 포함하여 총 3단위의 수입을 얻게 되고, 예측이 틀렸을 경우 걸었던 1단위를 잃게 된다. 따라서 $a$라는 배당률을 정한 회사 측에서 생각하는 특정 결과가 일어날 확률, 즉, 배당률에 내제된 (implied) 확률 $P_{maker}$는 식 (1)과 같다.

$$P_{maker} = \frac{1}{a} \tag{1}$$

실제로 한 축구 경기의 경기 결과(홈팀(home team)의 승, 무, 혹은 패)를 맞추는 게임의 배당률을 통해 회사(bookmaker) 측의 내재된 확률을 알아보면 확률들의 전체 합이 1보다 크다는 것을 알 수 있는데, 이러한 배당률의 설정을 통하여 회사 측에서는 수수료(take)와 같은 이익을 남길 수 있다(Dixon and Coles, 1997).

회사 측에서 사건이 일어날 확률 $P_{actual}$을 정확하게 추정하여 배당률을 설정한다면, 회사는 확실한 이득을 복권 구입자들로부터 얻어낼 수 있다. 반면, 베팅을 하는 사람의 입장에서는 $P_{actual}$과 $P_{maker}$의 비율이 1이 넘는 대상을 찾아서 베팅을 한다면 양의 수익을 올릴 수 있을 것이다. 하지만 $P_{actual}$을 정확하게 알 수 없기 때문에, 모형을 통해 $P_{actual}$을 대체할 수 있는 확률을 추정함으로써 이와 같은 문제를 해결할 수 있을 것이다. 따라서 이러한 확률을 구하기 위해 다음과 같은 모형을 세워 추정하도록 한다.

### 2.2 포아송 분포

포아송 분포는 이산적(discrete)이고 사건의 발생 비율(arrival rate)이라는 매개 변수에 의해 분포 형태가 좌우된다는 점에서 매우 다양하게 응용되고 있다. 축구 경기의 득점에서도 양 팀의 전력을 고려하여 제대로 된 득점 발생 매개 변수를 추정할 수 있다면 포아송 분포를 통해 축구 경기의 승부를 정확하게 예측할 수 있는 유용한 모형을 만들어 낼 수 있다. 이러한 포아송 분포의 성질은 이미 여러 축구 관련 논문(Dixon and Coles, 1997; Greenhough *et al*., 2002)에서 검증된 바 있고, 본 연구에서도 3.2절의 카이제곱 적합도 검정(Chi-square goodness-of-fit test)을 통해 입증하였다.

Dixon and Coles(1997)는 홈팀 i와 원정팀(away team) j의 각각의 득점을 각각 $X_{ij}$, $Y_{ij}$라 하고, 이는 식 (2)와 같은 포아송 분포를 따른다고 하였다.

$$\begin{aligned} X_{i,j} &\sim Poisson(\alpha_i \beta_j \gamma) \\ Y_{i,j} &\sim Poisson(\alpha_j \beta_i) \end{aligned} \tag{2}$$

여기에서 $\alpha_i$와 $\beta_i$는 각각 팀 $i$의 공격력과 수비력을 나타내는 양(positive)의 매개변수(parameter)이고, $\gamma$는 홈팀의 유리한 정도를 나타내는 양의 매개변수이다. 따라서 $\gamma$가 1보다 큰 추정량을 가진다면, 이는 홈팀이 더 유리하다는 것을 뜻한다.

위의 이변수 포아송 분포를 이용하여, 어느 한 경기에서 홈팀 득점이 $x$, 원정팀 득점이 $y$일 확률을 다음 식 (3)과 같이 나타냈다.

$$\begin{aligned} &\Pr(X_{i,j} = x, Y_{i,j} = y) \\ &= \tau_{\mu,\lambda} \frac{\lambda^x \exp(-\lambda)}{x!} \frac{\mu^y \exp(-\mu)}{y!} \end{aligned} \tag{3}$$

$$\text{단, } \lambda = \alpha_i \beta_j \gamma, \quad \mu = \alpha_j \beta_i$$

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho & \text{if} \quad x = y = 0, \\ 1 + \lambda\rho & \text{if} \quad x = 0, y = 1, \\ 1 + \mu\rho & \text{if} \quad x = 1, y = 0, , \\ 1 - \rho & \text{if} \quad x = y = 1, \\ 1 & otherwise. \end{cases}$$

$$\max(-1/\lambda, -1/\mu) \le \rho \le \max(1/\lambda\mu, 1)$$

여기에서 $\lambda$는 홈팀 $i$가 원정팀 $j$를 상대할 때의 홈팀 $i$의 득점의 평균이라고 볼 수 있고, 마찬가지로 $\mu$는 원정팀 $j$의 득점의 평균이라고 볼 수 있다. $\rho$는 홈팀과 원정팀 득점의 분포가

서로 독립적(independent)인지를 나타내주는 매개변수로서, $\rho$ = 0이면 홈팀과 원정팀의 득점 분포가 서로 독립임을 말해준다.

그리하여 총 N개의 경기에서 k번째(단, 1≤k≤N) 경기의 홈팀과 원정팀의 득점을 각각 $x_k$, $y_k$라 하면, 위의 매개변수를 추정하기 위한 우도함수(likelihood function)는 식 (4)와 같이 나타낼 수 있다.

$$L_t(\alpha_i, \beta_i, \rho, \gamma;\ i=1, \cdots, n) = \prod_{k=1}^{N} \tau_{\lambda,\mu}(x_k, y_k)\exp(-\lambda_k)\lambda_k^{x_k}\exp(-\mu_k)\mu_k^{y_k} \quad (4)$$

위의 우도함수를 최대화(maximization)시킴으로써 각 매개변수의 추정량을 구할 수 있는데, 이러한 방법을 최대우도추정(Maximum Likelihood Estimation, MLE)라고 한다. 그리고 이러한 추정 방법에서 $\alpha_i$가 지나치게 커지는 것을 막기 위해, 식 (5)와 같은 제약식을 설정한다.

$$n^{-1}\sum_{i=1}^{n}\alpha_i = 1 \quad (5)$$

하지만 식 (4)에서 구한 각 팀의 매개변수는 그 팀이 과거에 어느 팀과 어떠한 내용의 경기를 펼쳤느냐에 따라 다르게 추정되고, 추정하는 시점에 따라 매개변수의 값은 계속 변화하게 된다. 또한, 최근의 경기 결과가 현재 능력치에 더 많은 영향을 끼치도록 식 (4)에서 시간에 대한 가중치(weight)를 주면, 더 정확한 매개변수의 추정량을 얻을 수 있을 것이다. 따라서 시간에 대한 가중치를 준, 시점 t에서의 우도함수는 식 (6)과 같이 나타낼 수 있다.

$$L_t(\alpha_i, \beta_i, \rho, \gamma;\ i=1, \cdots, n) = \prod_{k\in A_t} \tau_{\lambda,\mu}(x_k, y_k)\exp(-\lambda_k)\lambda_k^{x_k}\exp(-\mu_k)\mu_k^{\phi(t-t_k)} \quad (6)$$

$$\text{단, } A_t = k : t_k < t,\quad \phi(t) = \exp(-\xi t)$$

여기에서 $t_k$는 $k$번째 경기가 치러진 시점이고, $\phi(t)$는 모든 $\xi > 0$인 범위에서 $t$에 관한 단조감소함수이다. 그리고 식 (7)의 값이 최대가 되는 지점에서의 $\xi$의 값을 찾는다.

$$S(\xi) = \sum_{k=1}^{N}(\delta_k^W \log p_k^W + \delta_k^D \log p_k^D + \delta_k^L \log p_k^L) \quad (7)$$

위의 식에서 $p_k^W$는 $k$번째 경기 시점에서 식 (7)을 최대화시키는 모수를 이용하여 구한 홈팀이 이길 확률이고, 마찬가지로 $p_k^D$와 $p_k^L$는 각각 홈팀이 비길 확률과 질 확률이다. 또한, k번째 경기에서 홈팀이 이기면 $\delta_k^H = 1$이고 그렇지 않으면 $\delta_k^H = 0$이며, $\delta_k^D$과 $\delta_k^L$도 각각 비길 때와 홈팀이 질 때 위와 같은 값을 가진다.

### 2.3 선수들의 피로도

전 세계의 축구리그 시즌의 시작 시점은 각각 저마다 차이가 있지만, 대게 한 시즌은 약 8~9개월 동안 계속된다. 또한, 리그 경기 사이에는 자국 리그가 속해 있는 대륙별 클럽대회나 자국 내의 컵 대회가 있는데, 이러한 대회에 참가하는 팀은 리그 경기 때보다 더 먼 거리를 이동하고 참가하지 않는 팀보다 더 빡빡한 경기일정을 소화하게 된다. 따라서 이러한 일정을 소화하는 팀은 그렇지 않는 팀보다 더 많은 피로를 느낄 수 있고, 이는 경기력에 영향을 끼칠 수 있다.

위의 식 (2)에서는 이러한 짧은 휴식에서 오는 피로(fatigue) 측면을 고려하지 않는다. 이러한 경우, 어느 한 팀이 이전 경기를 크게 이기고 짧은 휴식 후에 다음 경기를 치른다면, 피로가 고려되지 않고 오히려 식 (6)에서 큰 시간의 가중치로 인해 더 큰 $\alpha$를 가지게 되는 문제점이 생길 것이다.

다음 장에서는 이와 같은 선수들의 피로를 고려한 모형을 세우고, 그 결과를 식 (2)의 결과와 비교해보았다.

## 3. 모형 및 추정 결과

### 3.1 모형

2.3절에서 언급한 것처럼, 짧은 휴식 기간으로 인한 피로를 고려해주기 위해 이전의 식 (2)를 식 (8)과 같이 나타낸다.

$$X'_{i,j} \sim Poisson(\alpha_i\beta_j\gamma\eta^{c_i})$$
$$Y'_{i,j} \sim Poisson(\alpha_j\beta_i\eta^{c_j}) \quad (8)$$

$$\text{단, } c_i = \begin{cases} 1 & \text{if } \ t - t_{i,last} \le 4, \\ 0 & elsewhere \end{cases},$$

$t_{i,last}$ : 팀 $i$의 가장 최근의 경기 시점

여기에서 $\eta$는 휴식 기간이 4일 이내인 팀의 피로도를 나타내고, 이는 0보다 큰 값을 가진다. 만약, $\eta$이 1보다 작게 추정된다면, 이는 짧은 휴식기간이 팀의 공격력에 영향을 끼쳤다고 볼 수 있다. 여기에서 짧은 휴식을 4일로 설정한 이유는 대부분의 리그 경기는 주말에 벌어지고, 그 밖의 추가적인 리그 외의 경기나 리그 일정을 고려한 추가적인 리그 경기는 주로 주중에 한 경기씩 있기 때문이다.

그 밖의 우도함수는 식 (6)과 동일하게 사용하여 매개변수를 추정하였다. 지금부터 표현상의 편의를 위해, Dixon and Coles (1997)의 모형을 Model 1, 피로도를 고려한 모형을 Model 2로 나타내도록 한다.

### 3.2 자 료

이 논문에서는 2003~2006시즌 K리그, 02/03 ~05/06 시즌의

잉글랜드 프리미어 리그, 스페인 프리메라 리그, 이탈리아 세리에 A의 경기일정과 경기결과 자료와 같은 시기에 있었던 대륙 간 클럽대회, 각 국가의 컵 대회의 자료를 이용하였고, 자료의 출처는 RSSSF(http://www.rsssf.com), Football -data(http://www.football-data.co.uk), Naver 스포츠 축구(http://news.naver.com/sports)이다.

그리고 해외리그의 경우, 02/03～04/05 시즌의 잉글랜드, 스페인, 이탈리아 리그 자료를 각 리그 별로 누적하여 05/06 시즌의 각 리그의 경기 결과에 대한 확률을 추정하였고, 05/06 시즌 동안 9개의 해외 베팅 업체에서 제시한 배당률과 (주)스포츠토토에서 제시한 배당률을 가지고 위 모형들의 타당성을 판단하였다.

**Table 1.** Number of matches which ended in following home and away scores for three soccer leagues (EPL, LFP, and Serie A indicate English Premier League, Spanish Primera League, and Italian Serie A respectively.)

| # of goals scored | EPL (1520 matches) | | LFP (1520 matches) | | Serie A (1372 matches) | |
|---|---|---|---|---|---|---|
| | Home | Away | Home | Away | Home | Away |
| 0 | 358 | 535 | 339 | 504 | 306 | 457 |
| 1 | 493 | 531 | 532 | 554 | 462 | 490 |
| 2 | 383 | 292 | 369 | 308 | 344 | 285 |
| 3 | 176 | 112 | 182 | 104 | 177 | 108 |
| 4 | 79 | 40 | 64 | 38 | 60 | 30 |
| 5 | 24 | 7 | 28 | 10 | 18 | 1 |
| 6 | 5 | 3 | 5 | 2 | 5 | 1 |
| 7 | 2 | 0 | 1 | 0 | 0 | 0 |
| Mean | 1.4914 | 1.0947 | 1.4796 | 1.1158 | 1.4876 | 1.1042 |
| p-value | 0.5852 | 0.1539 | 0.4844 | 0.4502 | 0.9720 | 0.2980 |

또한, 국내리그의 경우는 마찬가지로 2003～2005 시즌의 자료를 누적하여 2006년 3월부터 2006년 8월까지의 경기 결과에 대한 확률을 추정하여, (주)스포츠토토의 고정배당률 게임의 배당률을 가지고 위 모형들의 타당성을 판단하였다.

<Table 1>은 잉글랜드 프리미어 리그와 스페인 프리메라 리그, 이탈리아 세리에 A리그의 4년간의 경기 결과를 각각 홈팀과 원정팀을 기준으로 하여 나타낸 것이다. 한 팀의 경기 득점 분포가 포아송 분포를 따른다는 가설을 카이제곱 적합도 검정(Chi-square goodness-of-fit test)을 통해 p-value를 구했을 때, 모두 90% 유의수준에서 위의 가설을 기각할 수 없음을 알 수 있다. <Table 1>에 포함하지 않은 K리그의 경우에도 90% 수준에서 위의 가설을 기각할 수 없었다. 이는 포아송 분포로 팀의 득점 분포를 나타내는 것이 문제가 되지 않음을 뒷받침해 준다.

### 3.3 매개 변수의 추정

<Table 2>와 <Table 3>은 각각 2006년 5월 7일을 기준으로 한 05/06 잉글랜드 프리미어 리그를 각각 Model 1과 Model 2로 매개변수를 추정한 결과와, 2006년 5월 20일을 기준으로 한 05/06 스페인 프리메라 리그의 추정 결과이다. 전체적으로 Model 1과 Model 2의 추정량이 서로 비슷함을 알 수 있다. 추정시의 $\xi$는 0.002를 사용하였다.

<Table 2>와 <Table 3>에서 각각 대륙별 클럽 대회에 진출한 팀의 $\Delta\hat{\alpha}$의 평균은 전체 팀의 $\Delta\hat{\alpha}$의 평균과 차이를 보이고, 각각의 p-value는 0.0244, 0.0284로, 95%의 신뢰수준에서 모두 유의하다.

<Table 4>에서 잉글랜드의 경우 $\hat{\eta}$이 1보다 크고, 스페인의 경우 $\hat{\eta}$이 1보다 작음을 알 수 있는데, 이는 Table 1에서 잉글랜드 내의 대륙별 클럽 대회에 진출한 팀의 $\Delta\hat{\alpha}$이 리그 평균의 $\Delta\hat{\alpha}$보다 작고, 스페인의 경우에는 진출한 팀의 $\Delta\hat{\alpha}$가 리그 평균 $\Delta\hat{\alpha}$보다 크다는 것과 연관시켜 생각해볼 수 있다.

**Table 2.** Maximum Likelihood Estimates for $\hat{\alpha}$ and $\hat{\beta}$, on May 7th, 2006, for English premium league teams (* indicates the team who advanced to European club competition in 2005/2006.; $\Delta\hat{\alpha}$ = ($\hat{\alpha}$ of Model 1) - ($\hat{\alpha}$ of Model 2))

| Team | Model 1 | | Model 2 | | $\Delta\hat{\alpha}$ |
|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\alpha}$ | $\hat{\beta}$ | |
| Arsenal* | 1.556 | 0.688 | 1.546 | 0.681 | 0.010 |
| Aston Villa | 0.934 | 1.124 | 0.937 | 1.111 | -0.003 |
| Birmingham | 0.734 | 1.025 | 0.736 | 1.014 | -0.002 |
| Blackburn | 0.973 | 0.934 | 0.975 | 0.923 | -0.001 |
| Bolton* | 1.032 | 0.956 | 1.031 | 0.945 | 0.001 |
| Charlton | 0.904 | 1.161 | 0.906 | 1.149 | -0.002 |
| Chelsea* | 1.472 | 0.488 | 1.464 | 0.483 | 0.008 |
| Everton* | 0.842 | 1.040 | 0.844 | 1.026 | -0.001 |
| Fulham | 1.075 | 1.224 | 1.078 | 1.211 | -0.003 |
| Liverpool* | 1.158 | 0.682 | 1.149 | 0.673 | 0.009 |
| Man City | 0.972 | 0.999 | 0.974 | 0.987 | -0.002 |
| Man United* | 1.388 | 0.711 | 1.382 | 0.703 | 0.006 |
| Middlesboro* | 1.053 | 1.138 | 1.047 | 1.125 | 0.007 |
| Newcastle | 1.045 | 0.998 | 1.042 | 0.987 | 0.003 |
| Portsmouth | 0.885 | 1.251 | 0.886 | 1.235 | -0.002 |
| Sunderland | 0.544 | 1.438 | 0.544 | 1.426 | -0.001 |
| Tottenham | 1.082 | 0.924 | 1.087 | 0.914 | -0.005 |
| West Brom | 0.673 | 1.228 | 0.675 | 1.217 | -0.002 |
| West Ham | 1.084 | 1.218 | 1.083 | 1.208 | 0.001 |
| Wigan | 0.942 | 1.111 | 0.945 | 1.096 | -0.003 |

**Table 3.** Maximum Likelihood Estimates for $\hat{\alpha}$ and $\hat{\beta}$, on May 20th, 2006, for Spanish primera league teams (* indicates the team who advanced to European club competition in 2005/2006.; $\Delta\hat{\alpha}$ = ( of Model 1) - ( of Model 2))

| Team | Model 1 | | Model 2 | | $\Delta\hat{\alpha}$ |
|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\alpha}$ | $\hat{\beta}$ | |
| Alaves | 0.765 | 1.197 | 0.765 | 1.210 | 0.000 |
| Ath Bilbao | 1.040 | 1.096 | 1.039 | 1.105 | 0.001 |
| Ath Madrid | 0.974 | 0.904 | 0.973 | 0.916 | 0.001 |
| Barcelona* | 1.621 | 0.787 | 1.642 | 0.792 | -0.021 |
| Betis* | 0.980 | 1.135 | 0.988 | 1.147 | -0.008 |
| Cadiz | 0.823 | 1.156 | 0.830 | 1.175 | -0.007 |
| Celta | 1.007 | 0.873 | 1.006 | 0.881 | 0.000 |
| Espanol* | 0.956 | 1.209 | 0.965 | 1.220 | -0.009 |
| Getafe | 1.055 | 1.059 | 1.058 | 1.072 | -0.003 |
| La Coruna | 1.070 | 1.037 | 1.078 | 1.047 | -0.007 |
| Malaga | 0.874 | 1.363 | 0.871 | 1.373 | 0.003 |
| Mallorca | 0.937 | 1.255 | 0.934 | 1.266 | 0.003 |
| Osasuna* | 0.976 | 1.077 | 0.982 | 1.090 | -0.005 |
| Real Madrid* | 1.538 | 0.922 | 1.558 | 0.933 | -0.020 |
| Santander | 0.909 | 1.256 | 0.905 | 1.268 | 0.004 |
| Sevilla* | 1.136 | 0.938 | 1.156 | 0.947 | -0.020 |
| Sociedad | 1.094 | 1.341 | 1.090 | 1.351 | 0.004 |
| Valencia | 1.268 | 0.784 | 1.276 | 0.792 | -0.008 |
| Villarreal* | 1.168 | 0.921 | 1.186 | 0.928 | -0.017 |
| Zaragoza | 1.037 | 1.195 | 1.049 | 1.216 | -0.012 |

**Table 4.** Maximum Likelihood Estimates for $\hat{\gamma}$, $\hat{\rho}$, and $\hat{\eta}$ using Model 2, on May 31st, 2006, for four soccer leagues (KL indicate K-league.)

| | $\hat{\gamma}$ | $\hat{\rho}$ | $\hat{\eta}$ |
|---|---|---|---|
| KL | 1.1224 | -0.12174 | 1.0395 |
| EPL | 1.4138 | -0.00438 | 1.0233 |
| LFP | 1.2584 | -0.05006 | 0.93703 |
| Serie A | 1.3201 | -0.15152 | 0.91234 |

<Table 4>는 Model 2를 이용하여 2006년 5월 31일을 기준으로 한 각 리그의 $\hat{\gamma}$, $\hat{\rho}$, $\hat{\eta}$의 최대우도추정량을 나타낸 표이다. 각 리그마다 홈팀의 유리한 정도($\hat{\gamma}$)와 홈팀과 원정팀의 득점 분포의 의존성(dependence) 정도($\hat{\rho}$), 피로도가 미치는 영향($\hat{\eta}$)이 서로 다르다는 것을 알 수 있다. 그리고 한국과 잉글랜드의 경우는 짧은 휴식이 공격력에 미치는 영향이 1보다 크게 추정됐는데, 이는 피로도가 오히려 공격력에 긍정적인 영향을 끼친 것으로 해석할 수 있다. 이러한 결과는 우리가 의도한 바와는 거리가 먼 것이라고 할 수 있다. 반면, 1보다 작은 추정량은 피로도가 공격력에 부정적인 영향을 끼쳤음을 뜻하고, 스페인과 이탈리아의 경우에 이러한 추정량을 가졌다.

Model 1과 Model 2의 추정량을 통해 구한 확률의 적합성을 비교하기 위해서, 다음과 같은 식 (9)를 이용하였고 <Table 5>는 이를 이용하여 구한 결과이다.

$$V = \frac{1}{n}\sum_{k\in B}(\delta_k^W p_k^W + \delta_k^D p_k^D + \delta_k^L p_k^L) \tag{9}$$

단, B = {$k$ | $k$는 K리그의 경우, 2006 시즌, 나머지는 05/06 시즌의 경기},
n = 집합 B의 원소의 개수

**Table 5.** Comparison of the validity between two models in each soccer league

| | KL | EPL | LFP | Serie A |
|---|---|---|---|---|
| Model 1 | 0.34542 | 0.42588 | 0.37372 | 0.41661 |
| Model 2 | 0.34355 | 0.42057 | 0.37687 | 0.41732 |

식 (9)에서의 $V$는 경기 이전에 각 모형으로 추정했던 확률이 실현된 결과의 평균값으로서, 더 큰 V를 가지는 모형이 주어진 자료에서 더 좋은 결과를 얻었다고 할 수 있다. 두 모형 사이의 $V$는 서로 큰 차이를 보이지는 않지만, K리그와 잉글랜드 프리미어 리그의 경우에는 Model 1이 더 크고, 스페인 프리메라 리그와 이탈리아 세리에 A의 경우에는 Model 2가 더 큰 것을 알 수 있다. 이는 <Table 4>에서 추정된 $\hat{\eta}$이 한국과 잉글랜드의 경우는 1보다 크다는 것과 연관시켜 생각해볼 수 있고, 1보다 크게 추정된 $\hat{\eta}$이 Model 2의 적합성에 부정적인 영향을 주었을 것이라 추측해 볼 수 있다.

## 4. 베팅 결과

### 4.1 베팅 방법

3장에서의 추정치를 이용하면 각 축구 경기 결과에 대한 확률을 구할 수 있는데, 위의 모형이 적합하다면 실제 고정배당률 시장에서 양의 수익을 얻을 수 있는 베팅을 할 수 있을 것이다. 여기에서 여러 방법을 통하여 사건의 확률과 배당률을 이용하여 베팅할 수 있겠지만, 이 논문에서는 식 (10)과 같은 간단한 방법으로 베팅하도록 한다.

$$\frac{P_{estimated}}{P_{maker}} = P_{estimated} \times Odds \geq r \tag{10}$$

여기에서 $P_{estimated}$는 위의 모형을 이용해 추정한 확률이고, $P_{maker}$와 Odds는 각각 식 (1)에서와 같이 발행 회사 측에서 생각하는 확률과 배당률이다. 어느 한 사건에 대한 식 (10)의 좌변이 r보

다 크다면, 그 결과에 1단위를 거는 방식이다. 따라서 위에서 추정한 확률 $P_{estimated}$가 사건의 실제 발생 확률 $P_{actual}$에 가깝다면, 위의 전략은 $r \geq 1$인 범위에서 양의 수익을 얻을 수 있을 것이다.

### 4.2 해외 고정배당 시장의 경우

<Figure 1>～<Figure 3>은 각각 잉글랜드 프리미어 리그와 스페인 프리메라 리그, 그리고 이탈리아 세리에 A의 05/06 시즌 해외 9개 고정배당 시장의 배당률 자료를 이용하여, 각 모형으로 확률을 추정하여 베팅했을 때의 $r$의 변화에 따른 평균 수익률의 변화를 나타낸 것이다.

<Figure 1>을 보면, 05/06 잉글랜드 프리미어 리그 자료에서는 Model 1이 Model 2보다 더 높은 수익률을 얻었다. 그리고 <Figure 3>의 05/06 이탈리아 세리에 A 자료에서는 Model 2가 Model 1보다 r이 더 커짐에 따라 더 높은 수익률을 얻었다. <Figure 2>의 경우에는 r에 따라 점점 증가하다가 갑자기 감소한 것을 볼 수 있는데, 이는 아래의 <Table 6>에서 볼 수 있는 것과 같이 r이 커질수록 이 범주에 해당하는 자료의 상대적인 부족으로 인한 이상치의 결과일 것이라고 추측해 볼 수 있다.

위의 <Figure 1>～<Figure 3>, 그리고 <Table 6>에서 보는 것과 같이, r이 커짐에 따라 베팅할 수 있는 대상(target match)은 줄어들지만 그 평균수익률이 대체로 증가한다는 것을 알 수 있다. 이는 식 (10)을 고려했을 때, $P_{estimated}$가 $P_{actual}$에 가깝게 추정된 것이라고 볼 수 있다. 그리고 해외 베팅 시장의 배당률이 평균적으로 회사에게 10% 정도 유리하게 책정되는 것을 감안하면, 위의 3개 리그의 경우에서 $r \geq 1.15$일 때 전체 경기의 약 15～25% 정도에 해당하는 경기에 단위 베팅을 하는 위의 전략으로 양의 수익률을 얻을 수 있었음을 알 수 있다.

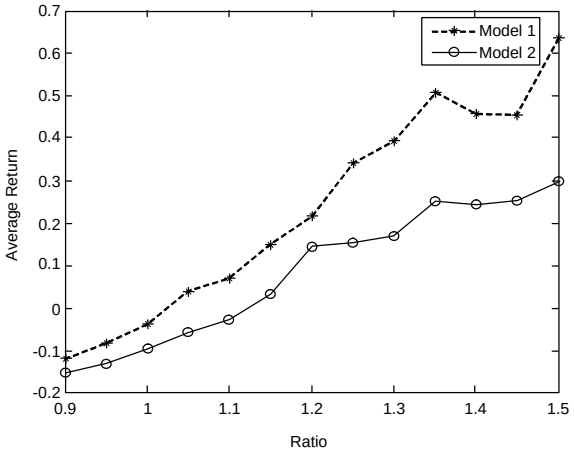


**Figure 1.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in English Premier League 05/06 season

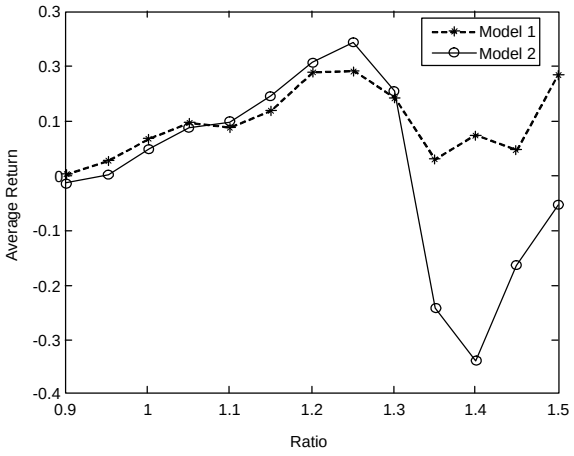


**Figure 2.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in Spanish Primera League 05/06 season

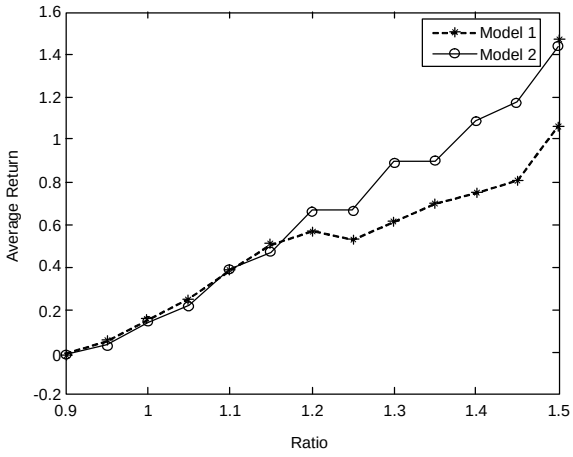


**Figure 3.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in Italian Serie A 05/06 season

**Table 6.** The percentages of target matches for betting in three 05/06 leagues using two models

| r | EPL | | LFP | | Serie A | |
|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| 0.90 | 70.29% | 77.25% | 95.56% | 99.62% | 81.29% | 81.90% |
| 0.95 | 62.16% | 68.54% | 84.88% | 86.37% | 70.96% | 71.49% |
| 1.00 | 44.18% | 48.60% | 65.06% | 65.99% | 54.50% | 54.12% |
| 1.05 | 30.91% | 34.27% | 45.85% | 45.96% | 41.40% | 40.76% |
| 1.10 | 21.20% | 24.42% | 32.31% | 32.75% | 31.02% | 29.82% |
| 1.15 | 14.82% | 17.34% | 22.08% | 23.25% | 23.60% | 21.75% |
| 1.20 | 10.53% | 12.49% | 15.35% | 16.78% | 17.92% | 16.05% |
| 1.25 | 7.66% | 8.54% | 11.02% | 11.29% | 14.21% | 12.51% |
| 1.30 | 5.67% | 6.35% | 7.92% | 8.42% | 12.16% | 9.94% |
| 1.35 | 4.12% | 4.53% | 5.85% | 5.82% | 9.56% | 8.04% |
| 1.40 | 3.10% | 3.45% | 3.74% | 4.06% | 7.51% | 6.35% |
| 1.45 | 2.25% | 2.60% | 2.84% | 2.92% | 6.23% | 5.26% |
| 1.50 | 1.73% | 2.08% | 2.19% | 2.25% | 4.97% | 4.04% |

### 4.3 국내 고정배당 시장의 경우

(주)스포츠토토에서 판매하는 고정배당 상품인 프로토(Proto)는 국내 유일의 고정배당 방식의 상품이고, 배당률이 회사에게 15% 정도 유리하게 책정되며, 2~10 경기 범위를 선택하여 모두 맞추면 선택된 경기의 배당의 곱만큼 환급해준다. 이처럼 상대적으로 배당률이 작고 동시에 여러 경기를 맞춰야 하기 때문에, 국내의 고정배당 방식은 해외 고정배당 시장에서보다 상대적으로 불리한 조건 하에서 이루어진다고 볼 수 있다.

이 논문에서는 국내의 고정배당 방식이 해외에서와 마찬가지로 한 경기만 선택할 수 있다는 전제 하에 위의 해외 고정배당 시장과 같은 방법으로 분석하였다. 프로토는 2006년 3월부터 발매가 시작한 관계로, 2006년 3월부터 2006년 8월 31일까지 발매 대상이었던 축구 경기 자료를 이용하였다. <Table 7>과 <Figure 4>는 한국 K리그, 잉글랜드 프리미어 리그, 스페인 프리메라 리그, 그리고 이탈리아 세리에 A의 국내 고정배당 시장의 배당률 자료를 이용하여, r의 변화에 따른 평균 수익률의 변화를 각각 표와 그림으로 나타낸 것이다. <Figure 4>의 경우, 상대적으로 적은 수의 표본(sample)을 감안하여 $0.9 \leq r \leq 1.25$의 범위에서 나타내었다.

<Figure 4>에서 국내의 고정배당 시장 역시 해외 시장과 마찬가지로 r의 증가와 함께 평균 수익률이 같이 증가하는 것을 볼 수 있고, Model 1과 Model 2의 평균수익률은 서로간의 뚜렷한 차이를 확인할 수 없었다. 상대적으로 적은 배당률의 표본 때문일 수도 있겠지만, 앞에서의 유럽 고정배당 시장에서의 그래프와 비교했을 때 기울기가 더 크다는 것을 알 수 있다. 이는 상대적으로 대상 경기의 수는 적지만, 위의 모형과 베팅 방법을 이용했다면 해외 시장에서보다 더 큰 수익률을 국내 고정배당 시장에서 얻을 수 있었음을 보여준다.

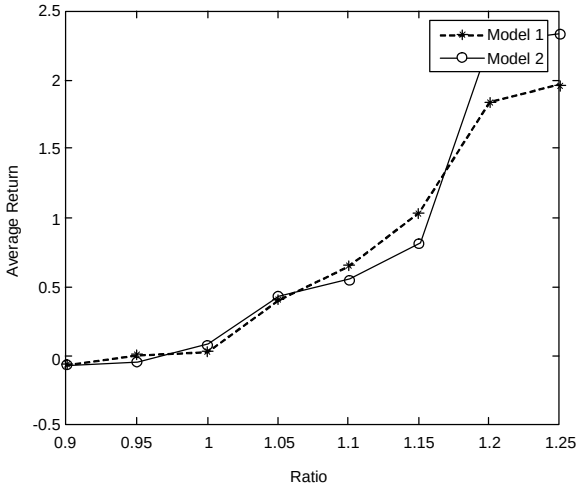


**Figure 4.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in Korean fixed-odds market

**Table 7.** Average return versus the ratio and the percentages of target matches for betting at Korean fixed-odds market in four 05/06 leagues using two models(in 316 matches)

| r | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | Average return | Target matches % | Average return | Target matches% |
| 0.90 | -0.10 | 56.60% | -0.10 | 54.40% |
| 0.95 | 0.00 | 34.91% | 0.01 | 33.33% |
| 1.00 | 0.02 | 20.44% | 0.19 | 18.87% |
| 1.05 | 0.40 | 14.15% | 0.46 | 12.58% |
| 1.10 | 0.65 | 9.12% | 0.69 | 8.81% |
| 1.15 | 1.03 | 6.29% | 1.00 | 6.29% |
| 1.20 | 1.84 | 4.09% | 2.02 | 3.14% |
| 1.25 | 1.96 | 2.83% | 3.34 | 2.52% |
| 1.30 | 3.43 | 1.89% | 4.32 | 1.57% |
| 1.35 | 5.65 | 1.26% | 5.65 | 1.26% |

### 4.4 모형의 재고

위의 결과에서, $\hat{\eta}$이 1보다 작은 스페인 프리메라 리그와 이탈리아 세리에 A에서는 Model 2가 Model 1보다 비교적 더 좋은 적합성과 수익률을 얻었던 반면, $\hat{\eta}$이 1보다 큰 K리그나 잉글랜드 프리미어 리그의 경우는 Model 1이 Model 2보다 더 좋은 적합성과 수익률을 얻었다. $\hat{\eta}$이 1보다 클 경우는 4일 이내의 휴식 기간을 취한 팀의 다음 경기에 더 높은 득점의 평균치를 부여하게 되므로, 이는 우리가 처음에 가정했던 $\eta$의 역할과는 다르게 작용한 것이다. 또 <Table 5>와 <Figure 1>에서 $\hat{\eta}$이 1보다 클 경우에는 $\eta = 1$로 설정한 Model 2라고 할 수 있는 Model 1이 더 좋은 예측 능력을 보여주었다는 점에서, 1보다 크게 추정된 $\hat{\eta}$이 Model 2의 예측 능력을 떨어뜨리는 것이라고 생각해 볼 수 있다.

따라서, 여기에서 위의 문제점들을 해결하기 위해, Model 2로 추정하여 $\hat{\eta}$이 1보다 큰 값으로 추정될 경우에는 다시 Model 1을 이용하여 그 추정량으로 확률을 구하는 모형을 정의하고 이를 Model 3이라 나타낸다. <Figure 5>와 <Figure 6>은 위와 같은 세 개의 모형을 이용했을 때의 해외 베팅 시장과 국내 베팅 시장에서의 성과를 각각 비교해 본 것이다.

<Figure 5>와 <Figure 6>에서 보는 것과 같이, Model 3은 Model 1과 Model 2보다 전체적으로 더 높은 평균수익률을 얻을 수 있음을 알 수 있다. 이는 Model 2에서의 $\hat{\eta}$이 1보다 더 높게 추정될 경우에는 기존의 모형보다 실제의 확률과 더 큰 차이를 가지는 확률을 추정하게 됨을 뜻한다. 또 상대적으로 r이 클수록 Model 3의 성과가 더 좋게 나타나는 것을 살펴볼 수 있다. 어느 한 경기가 큰 r을 가진다는 것은 상대적으로 실제 배당률에 내재된 확률에 비해 추정한 확률의 비가 더 크다는 것을 뜻하는데, 이러한 경우에서의 평균수익률이 높다는 것은 Model 3이

Model 1과 Model 2보다 큰 이익을 얻을 수 있는 실제 확률과 추정 확률 사이의 큰 편향(bias)을 더 잘 찾아낸다고 할 수 있다.

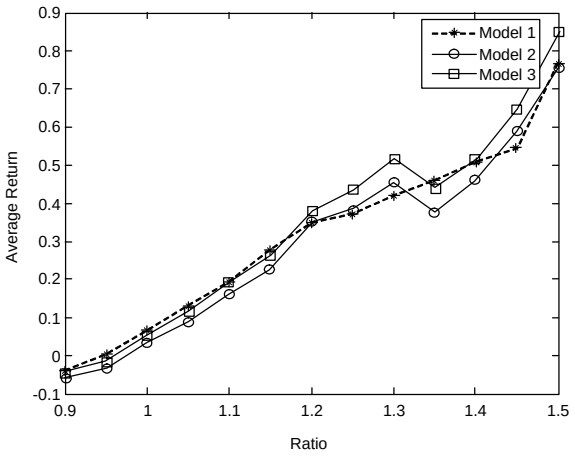


**Figure 5.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in foreign fixed-odds market

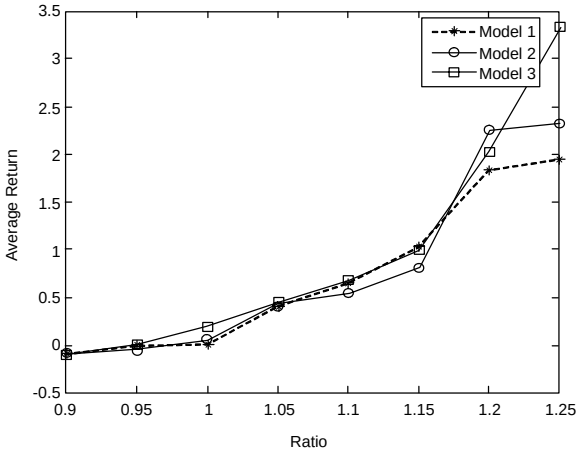


**Figure 6.** Average return plotted against the ratio, r, of estimated probabilities to bookmakers' implied probabilities in Korean fixed-odds market

## 5. 결론 및 향후 과제

지금까지 우리는 이 논문을 통해 Dixon and Coles(1997)의 기존의 모형(Model 1)과 짧은 휴식이라는 요소가 경기에 미치는 영향을 고려한 모형(Model 2)을 비교하면서 해외 고정배당 시장과 국내 고정배당 시장에서 양의 수익률을 얻을 수 있음을 보았고, 기존의 모형에 피로도를 고려해줌으로써 모형의 적합성과 실제 수익률에서의 향상을 얻을 수 있었다. 그리고 이 두 모형을 효율적으로 조합한 Model 3을 통해 앞의 두 모형에 비해 더 나은 수익률을 얻을 수 있었다. 위의 모형은 기존의 모형(Model 1)과 같이 확률을 추정하고자 하는 팀의 과거의 경기 일정과 각 경기의 득점, 상대팀의 득점만 알 수 있다면, 확률을 추정할 수 있다는 점에서 많은 자료를 추가적으로 필요로 하지 않는다는 장점이 있다. 그리고 매개변수만 추정할 수 있다면, 간단한 포아송 분포를 이용하여 원하는 사건의 확률을 쉽게 계산할 수 있다는 장점을 가진다.

하지만, 앞으로 이루어져야 할 과제들이 존재한다. 우선 많은 양의 자료를 통하여 위에서 발견한 사실들을 확고히 입증할 필요가 있다. 현재 이용한 배당률 자료는 최근 1년으로 국한된 것이므로, 넓은 구간 내에서 좀 더 검증할 필요가 있을 것이다. 그리고 현재 국내에서 이용할 수 있는 고정배당 상품의 경우는 두 경기 이상을 베팅해야 한다는 제약 때문에, 위와 같은 수익을 얻기 위해서는 이러한 상황을 고려한 베팅 방법 또한 연구되어야 할 것이다. 그리고 피로도와 같은 경기 결과에 지장을 줄 수 있는 다른 요인들을 기존의 모형에 새로운 변수를 추가하여 추정함으로써 모형을 더욱 정교하게 만들 수 있을 것이다.

## 참고문헌

Clarke, S. R. and Norman, J. M. (1995), Ground Advantage of Individual Clubs in English Soccer, *The Statistician*, 44(4), 509-521.

Dixon, M. J. and Coles, S. C. (1997) Modelling association football scores and inefficiencies in the football betting market, *Applied Statistics*, 46, 265-280.

Dixon, M. J. and Robinson, M. E. (1998), A Birth Process Model for Association Football Matches, *The Statistician*, 47(3), 523-538.

Dixon, M. J. and Pope, P. F. (2004), The value of statistical forecasts in the UK association football betting market, *International Journal of Forecasting*, 20, 697-711.

Forrest, D. and Simmons, R. (2000), Forecasting sports: the behaviour and performance of football tipsters, *International Journal of Forecasting*, 16, 317-331.

Goddard, J. and Asimakopoulus, I. (2004), Forecasting Football Results and the Efficiency of Fixed-odds Betting, *Journal of Forecasting*, 23, 51-66.

Greenhough, J., Birch, P. C., Chapman, S. C., and Rowlands, G. (2002), Football goal distributions and extremal statistics, *Physica A*, 316, 615-624.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer-Verlag New York, Inc.

Rue, H. and Salvesen, O. (2000), Prediction and retrospective analysis of soccer matches in a league, *The Statistician*, 49(3), 399-418.

**성 현**
한국과학기술원 산업공학과 학사
현재: 서울대학교 산업공학과 석사과정
관심분야: 확률모형, 금융공학, Data Mining

**장우진**
서울대학교 섬유고분자공학과 학사
Georgia Tech OR 석사
Georgia Tech 산업 공학 박사
Rensselaer Polytechnic Institute 조교수
현재: 서울대학교 산업공학과 조교수
관심분야: 금융공학, 계량 마케팅, 유비쿼터스 센서 네트워크