The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.775
25(5), 775–783

Statistical Analysis of K-League Data using Poisson Model

Yang-Jin Kim[1]

[1]Department of Statistics, Sookmyung Women's University

(Received June 25, 2012; Revised August 28, 2012; Accepted September 24, 2012)

Abstract

Several statistical models for bivariate poisson data are suggested and used to analyze 2011 K-league data. Our interest is composed of two purposes: The first purpose is to exploit potential attacking and defensive abilities of each team. Particular, a bivariate poisson model with diagonal inflation is incorporated for the estimation of draws. A joint model is applied to estimate an association between poisson distribution and probability of draw. The second one is to investigate causes on scoring time of goals and a regression technique of recurrent event data is applied. Some related future works are suggested.

Keywords: Bivariate poisson data, diagonal inflation model, K-league, random effect, recurrent event data.

1. 서론

이변량 포아송 자료는 쌍으로 표현되는 서로 연관된 계수형 자료로 여러 분야에서 예를 찾아볼 수 있다. 의약학 분야에서 치료 전후의 종양 개수, 역학 분야의 예는 각 농장에서 구제역으로 죽은 돼지와 소의 수, 경제학에서는 각 회사에서 발생된 자발적 이직수와 비자발적 이직수 등 한 개체로부터 측정된 두 사건의 발생 건수 또는 서로 연관된 두 관측 개체로부터 측정된 계수형 자료 등이 그 예가 될 것이다. 이변량 포아송 자료에서 공변량의 효과를 추정하기 위해 여러가지 방법들이 적용되어 왔다. Kocherlakota와 Kocherlakota (2001)은 EM 알고리즘을 Ho와 Singer (2001)은 일반화 최소 제곱법을 사용하였다. 이러한 이변량 포아송 분포의 확장으로 Li 등 (1999)은 다변량 포아송 자료에 대한 통계적 모형을 제시하였다.

본 연구에서는 이변량 포아송 모형을 K-리그 축구 경기 결과 분석에 적용하고자 한다. 특히 각 팀의 득점은 상대방 팀의 공격성을 자극하고 수비력을 강화함으로 두 팀의 득점은 상관관계가 존재할 것으로 예상된다. 이러한 연관관계를 표현하기 위해 랜덤 효과를 포함한 이변량 포아송 분포를 적용한다. 이와 관련된 연구로 Karlis와 Ntzoufras (2003)는 포아송 확률 변수가 같이 공유하는 확률변수를 고려한 다음의 이변량 포아송 분포를 적용하였다.

$$\Pr(y_1, y_2) = \Pr(Y_1 = y_1, Y_2 = y_2) = \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\boldsymbol{\lambda_3}}{\boldsymbol{\lambda_1 \lambda_2}} \right)^k,$$

This work was supported by the National Research Fund(2011-0013221)

[1]Assistant Professor, Department of Statistics, Sookmyung Women's University, 52 Hyochangwon-gil Yongsan-gu, Seoul 140-742, Korea. E-mail: yjin@sookmyung.ac.kr

Table 2.1. Frequency of goals by two teams

$Y_1 \vert Y_2$	0	1	2	3	4	5	6
0	22	17	10	5			
1	25	25	16	11			
2	21	22	14	6	1		1
3	5	9	11	3	1		
4	1	3	1	2	1		
5		2	1				
6		1		1			
7	1	1					

여기서 두 팀의 득점은 확률 변수 Y_1과 Y_2로 표현되며 $E(Y_1) = \lambda_1 + \lambda_3$, $E(Y_2) = \lambda_2 + \lambda_3$, $\mathrm{Cov}(Y_1, Y_2) = \lambda_3$로 λ_3는 두 팀간의 연관 정도를 보여준다. 따라서 $\lambda_3 = 0$은 두 팀의 골점수가 독립임을 의미하며 이러한 경우 위 분포는 이중 포아송 분포(double poisson distribution)로 축소된다. 본 연구에서는 두 팀의 연관성을 모형화하기 위해 랜덤 효과를 고려한다. 본 논문의 2장에서는 K-리그 결과를 간략하게 설명하고 자료의 특징을 기술할 것이며 3장에서는 새로운 이변량 모형을 제안하고 그 추론 과정을 서술하고자 한다. 또한 4장에서는 득점 소요시간을 분석하기 위해 포아송 확률 과정에 근거한 재발 사건 자료 분석 기법을 적용할 것이다. 5장에서는 제안된 모형의 확장을 제시하고자 한다.

2. 2011 K-리그 득점 현황

K-리그는 2012년 현재 16개의 프로팀으로 구성되어 있으며 지난 2011년 리그에서는 전북 FC가 우승을 하였다. 전체 240번의 경기가 치루어졌으면 각 팀당 30번의 경기가 있었다. 경기는 홈경기와 원정경기로 두 팀이 번갈아 각자의 홈경기장에서 시합을 하게 된다. Table 2.1은 두 팀간의 골 점수의 빈도를 보여준다. 자료에 의하면 홈팀 승 : 무승부 : 원정팀 승의 비율은 107 : 65 : 68로 홈팀이 승리하는 빈도가 훨씬 더 높았다. 그리고 165시합(약 68.8%)이 골 득점차가 0 또는 1로($|Y_1 - Y_2| \leq 1$) 두 팀의 전력이 비슷함을 보였다. 본 연구에서는 각 경기당 홈팀과 원정팀간의 골점수의 연관 관계를 모형화하기 위해 랜덤 효과를 고려한다. 또한 무승부의 높은 발생 빈도를 위해, 동점 발생 확률을 포함한 결합 모형(joint model)이 적용될 것이다. 두 팀 간의 골 득점에 대한 상관 계수는 0.142 (p-value = 0.026)로 유의한 양의 상관관계가 존재함을 보여준다. 즉 상대팀의 득점은 서로에게 자극을 주어 득점을 이루어지게 하였음이 예상된다. 이에 대한 분석을 위해 재발 사건 방법론이 적용된다. 즉, 각 팀의 골 득점 소요시간에 대한 분포와 공변량의 효과를 분석한다. 재득점의 평균 소요 시간은 20.153분이고 가장 빠른 재득점 시간은 1분이었다.

3. 이변량 포아송 자료

본 연구에서는 한 경기당 두 팀의 골 득점에 대해 이변량 포아송 변수를 가정한다. 특히 랜덤 효과 (u_{ht_i}, u_{at_i})가 주어져 있을 때, 홈팀과 원정팀의 골점수, Y_{i1}와 Y_{i2}는 조건부적으로 서로 독립이며 그들의 평균 λ_{i1}와 λ_{i2}에 대해 다음의 모형이 각각 적용된다.

$$\log(\lambda_{i1} | x_{i1}, u_{ht_i}) = \beta_1' x_{i1} + u_{ht_i}, \tag{3.1}$$

$$\log(\lambda_{i2} | x_{i2}, u_{at_i}) = \beta_2' x_{i2} + u_{at_i}, \tag{3.2}$$

Table 3.1. Comparison of fitness of several models

	AIC	BIC
Bivariate Poisson $u_{ht} = u_{at}$	1434.9	1445.5
Bivariate Poisson $u_{ht,at} = u_{at,ht}$	1433.1	1442.6
동점을 고려한 모형: 결합 모형		
(i) logistic model (상수 가정)	1311.5	1424.6
(ii) logistic model (결합 모형)	1313.2	1429.2

여기서 $x_{i1} = (x_{i1,1}, \ldots, x_{i1,p_1})'$와 $x_{i2} = (x_{i2,1}, \ldots, x_{i2,p_2})'$는 각각 $p_1 \times 1$와 $p_2 \times 1$의 차원을 가진 공변량 벡터를 나타내며 β_1과 β_2는 연관된 회귀 계수 벡터들이다. u_{ht_i}와 u_{at_i}는 홈팀과 원정팀의 개별 효과를 위한 랜덤효과이다. 본 연구에서는 다음 두 가지 유형의 랜덤 효과를 고려한다.

3.1. 랜덤 효과를 이용한 이변량 포아송 자료 분석

A. 각 경기의 개별 효과가 존재: $u_{ht_i} = u_{at_i} = u_i$

공변량으로 측정할 수 없는 당일 기상 상태와 축구장의 **잔디** 조건 등 각 경기마다 두 팀에 똑같이 적용되는 요인이 존재한다. 이를 위해 두 팀에 공통적으로 적용되는 랜덤 효과를 가정한다. 랜덤 효과에 대한 분포로 평균이 영이고 분산이 σ_1^2인 정규 분포가 가정된다.

$$u_{ht_i} = u_{at_i} \sim N\left(0, \sigma_1^2\right). \tag{3.3}$$

B. 두 팀간의 상대 전력이 존재: $u_{ht_i,at_i} = u_{at_i,ht_i} = u_i$

두 팀의 골점수는 두 팀간의 개별 능력뿐만 아니라 상대 전력에 영향을 받을 수 있다. 예를 들어 서울과 수원은 라이벌이라는 특성 때문에 다른 팀들과의 경기보다 더 강한 정신력을 가지고 경기에 임하게 된다. 이러한 팀별 효과를 랜덤 효과로 모형화한다. 따라서 두 번의 경기(두 팀이 홈경기와 원정경기를 번갈아 시합)에서 두 팀은 동일한 팀별 효과(team pair-specific effect)를 가지게 된다. 이러한 경우 랜덤 효과는 다음의 분포를 가진다고 가정한다.

$$u_{ht_i,at_i} = u_{at_i,ht_i} \sim N(0, \sigma_2^2). \tag{3.4}$$

Table 3.1의 처음 두 행은 위의 두 모형을 적합한 후 계산된 AIC $(= -2l(\hat{\theta}) + 2p$, 여기서 $\hat{\theta}$는 추정된 모수값, l은 로그 우도이며 p는 모수의 개수)와 BIC $(= -2l(\hat{\theta}) + 2p\log(n), n$은 관측 개체수) 값을 보여준다. 이를 통해 두 번째 모형(팀별 랜덤효과)이 조금 더 나은 모형이라고 생각되어지며 다음 절에서도 팀별 랜덤효과가 적용될 것이다. 한편 랜덤 효과의 분산에 대한 추정치는 ($\hat{\sigma}_1^2 = 0.073$ (SE $= 0.039, p$-value $= 0.067$), $\hat{\sigma}_2^2 = 0.070$(SE $= 0.034, p$-value $= 0.043$))으로 비슷한 값을 보여주었다.

3.2. 동점 자료를 고려한 결합 모형

특히 스포츠 자료에서 자주 발생하는 동점(tie)을 고려하기 위해 결합 모형(joint model)을 제안한다. 이와 비슷한 주제에 대해 Walhin (2001)은 공변량을 고려하지 않은 이변량 영 확대 모형을 고려하였으며 Karlis와 Ntzoufras (2003)은 동점 발생 확률을 고려한 이변량 포아송 대각 확대 모형(Bivariate Poisson Diagonal Inflation Model; BPDIM)을 제안하였다. 또 다른 방법으로 골 득점차 $(Y_1 - Y_2)$를 이용한 영 확대 모형을 적용하였다. Wang 등 (2003)은 영확대 이변량 분포를 사고 건수 자료에 고려하였다. 본 논문에서는 랜덤 효과를 포함한 포아송 회귀 모형과 동점 모형의 결합 모형을 제안한다. i번째

경기에 대한 동점 확률을 $\phi_i = \Pr(Y_{i1} = Y_{i2})$로 정의하고 공변량과의 관계를 위해 다음의 로지스틱 회귀 모형들이 적용된다. 일반적으로 동점은 두 팀의 경쟁이 치열했음을 의미함으로 득점과 동점간의 연관관계(association)가 존재할 수 있다. 즉 팀들 간 상대전력에 대한 랜덤 효과($u_i = u_{ht_i,at_i}$)를 동점에 대한 확률에 적용함으로써 두 현상간의 관계를 추정하고자 한다. 이를 위해 다음의 모형이 고려된다.

$$\text{logit}(\phi_i(\gamma|u_i)) = \log \frac{\phi_i(\gamma|u_i)}{1-\phi_i(\gamma|u_i)} = \gamma_0 + \gamma_1 u_i, \tag{3.5}$$

γ_0는 상수항이며 γ_1는 골 득점과 동점 가능 여부와의 연관정도를 추정하기 위해 사용된다. 여기서 $\gamma_1 > 0$은 두 팀 간의 고득점에 대한 성향이 높을수록 동점이 될 가능성이 큼을 의미하며 $\gamma_1 = 0$은 득점의 크기와 동점 간에 연관정도가 없음을 의미한다. 동점 여부에 따라 $\Pr(Y_{i1} = y_1, Y_{i2} = y_2|u_i)$는 다음과 같이 두 경우로 나누어서 구할 수 있다.

$$\Pr(Y_{i1} = y_1, Y_{i2} = y_2|u_i) = \begin{cases} \phi_i(\gamma|u_i) + \{1-\phi_i(\gamma|u_i)\} f(y_1, y_2|\beta_1, \beta_2, u_i), & \text{if } y_1 = y_2, \\ \{1-\phi_i(\gamma|u_i)\} f(y_1, y_2|\beta_1, \beta_2, u_i), & \text{if } y_1 \neq y_2, \end{cases}$$

여기서 u_i가 주어진 경우, 조건부 독립에 의해 $f_i(y_1, y_2|\beta_1, \beta_2, u_i) = f_{1i}(y_1|\beta_1, u_i) f_{2i}(y_2|\beta_2, u_i)$이 유도되며 식 (3.1)과 (3.2)의 평균을 이용하여 다음의 혼합 포아송 분포의 질량함수를 적용한다.

$$f_{1i}(y_1|\beta_1, u_i) = \frac{e^{-m_{i1}} m_{i1}^{y_1}}{y_1!}, \qquad f_{2i}(y_2|\beta_2, u_i) = \frac{e^{-m_{i2}} m_{i2}^{y_2}}{y_2!},$$

여기서 $m_{i1} = \beta_1^{'} x_{i1} + u_i$, $m_{i2} = \beta_2^{'} x_{i2} + u_i$이며 $u_i = u_{ht_i,at_i} = u_{at_i,ht_i}$가 적용된다. 우도 함수를 유도하기 위해 동점여부를 나타내는 지시함수, $\delta_i = I(Y_{i1} = Y_{i2})$가 정의된다. 미지의 모수들의 집합을 $\theta = (\beta_1, \beta_2, \gamma_0, \gamma_1, \sigma_2^2)$로 표시한다. 랜덤 효과 $u = (u_1, \ldots, u_n)$가 주어져 있을 때, 다음의 조건부 우도 함수(conditional likelihood function)가 유도된다.

$$L(\theta|u) = \prod_{i=1}^{n} \Pr(Y_{i1} = y_1, Y_{i2} = y_2|u_i).$$

랜덤 효과에 대한 분포 가정을 이용하여 주변 우도 함수(marginal likelihood function)는 다음과 같다.

$$L(\theta) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \Pr(Y_{i1} = y_1, Y_{i2} = y_2|u_i) g\left(u_i|\sigma_2^2\right) du_i \tag{3.6}$$

위의 식 (3.6)을 최대화하는 추정량을 구하기 위해 랜덤 효과에 대한 적절한 처리가 필요하다. 이를 위해 u_i를 결측치로 간주한 후 EM 알고리즘을 적용한다. 이를 위해 u_i가 알려져 있다고 간주하여 다음의 완전 자료 로그 우도 함수(complete data log likelihood function)를 유도한다.

$$\begin{aligned} l(\theta) &= l_1 + l_2, \\ l_1(\beta, \gamma) &= \sum_{i=1}^{n} (1-\delta_i) \log\left[(1-\phi_i(\gamma)) f(y_{i1}|\beta_1, u_i) f(y_{i2}|\beta_2, u_i)\right] \\ &\quad + \delta_i \log\left[\phi_i(\gamma) + (1-\phi_i(\gamma)) f(y_{i1}|\beta_1, u_i) f(y_{i2}|\beta_2, u_i)\right], \\ l_2(\sigma_2^2) &= -\frac{1}{2}\left\{ n \log\left(2\pi\sigma_2^{-2}\right) + \sum_{i=1}^{n} \frac{u_i^2}{\sigma_2^2} \right\}, \end{aligned}$$

여기서 l_1은 랜덤 효과가 조건부로 주어졌다고 가정할 때, 동점을 고려한 이변량 자료의 조건부 로그 우도함수이며, l_2는 랜덤 효과의 분포로 조건부 로그 우도 함수에 대한 변화 정도를 반영하는 것이다. 모수추정 과정은 다음의 EM algorithm을 통해 적용된다.

(step i) E-step:

$$Q\left(\theta;\theta^{k-1}\right)=Q_1\left(\beta,\gamma;\theta^{k-1}\right)+Q_2\left(\sigma_2^2;\theta^{k-1}\right),$$

여기서

$$Q_1\left(\beta,\gamma;\theta^{k-1}\right)=E\left(l_1(\beta,\gamma)|x_i,y_i,\delta_i\right),\qquad Q_2\left(\sigma_2^2;\theta^{k-1}\right)=E\left(l_2(\sigma_2^2)|x_i,y_i,\delta_i\right)$$

위의 기대값을 계산하기 위해 다음의 랜덤 효과의 조건부 분포를 이용한다.

$$h(u_i|x_i,y_i;\theta)=\frac{f(y_1,y_2|u_i,\theta)g(u_i|\sigma_2^2)}{\int_{-\infty}^{\infty}f(y_1,y_2|u_i,\theta)g(u_i|\sigma_2^2)du_i}. \tag{3.7}$$

따라서 랜덤 효과 함수의 조건부 기대값(conditional expectation)은

$$E(u_i|x_i,y_i;\theta)=\int_{-\infty}^{\infty}u_ih(u_i|x_i,y_i;\theta),\qquad E[\log u_i|x_i,y_i;\theta]=\int_{-\infty}^{\infty}\log(u_i)h(u_i|x_i,y_i;\theta)$$

로 표현된다. 그러나 식 (3.7) 분모의 적분은 닫힌 형태(closed form)를 가지지 않으며 따라서 수치 적분이나 MCMC 방법을 고려해 볼 수 있다. 본 연구에서는 Gauss-Hermite 알고리즘을 통해 적분 계산이 구해졌다.

(step ii) M-step: E-step에서 구한 랜덤 효과의 조건부 기대값이 미지의 랜덤 효과를 대체한다. Newton-Raphson 방법을 이용하여 Q_1과 Q_2에 포함된 모수 $(\beta_1,\beta_2,\gamma_1,\gamma_2,\sigma_2^2)$의 최대값을 구한다.

(step iii) (step i)과 (step ii)를 수렴할 때까지 반복 시행한다.

수렴이 이루어진 후 각 추정량의 표준 오차는 관측된 정보 행렬을 이용하여 구한다. Table 3.2는 각 팀의 공격력과 수비력 및 동점 확률을 추정하기 위해 세 가지 모형을 적합한 결과로 모수의 추정치와 표준오차를 보여준다.

모형1 (동점을 고려하지 않음):

$$\log\lambda_{i1}=\mu+\text{home}_i+\text{att}_{ht_i}+\text{def}_{at_i}+u_{ht_i,at_i},\qquad \log\lambda_{i2}=\mu+\text{att}_{at_i}+\text{def}_{dt_i}+u_{ht_i,at_i}.$$

모형2 (모형1 + 동점 (상수)):

$$\text{logit}(\phi_i(\gamma))=\log\frac{\phi_i}{1-\phi_i}=\gamma.$$

모형3 (모형1 + 결합모형):

$$\text{logit}(\phi_i(\gamma))=\log\frac{\phi_i}{1-\phi_i}=\gamma_0+\gamma_1u_{ht_i,at_i}.$$

Table 3.1의 마지막 두 행의 결과를 통해, 두 동점 모형 중에서 상수 가정 모형이 결합 모형보다 더 나은 모형임을 알 수 있다. 또한 결합 모형에서 추정된 $\hat{\gamma}_1=6.470\,(\text{se}=11.04)$은 비록 통계적으로 유의하지는 않았지만 양의 부호를 통해 골 득점이 많을수록 동점 가능 여부가 커지는 경향이 있음을 알 수 있었다. 또한 Table 3.2의 결과에 의하면 공격력이 강한 팀(양의 값이 클수록 강한 팀을 의미)은 전북, 포항, 서울 순이며, 수비력이 강한 팀(음의 값이 클수록 강한 팀을 의미)은 전남, 울산, 포항 순으로 추정되었다. 또한 동점이 발생할 확률들은 각각 $1/(1+\exp(1.330))=0.209$ (상수 모형)와 $1/(1+\exp(1.475))=0.186$ (결합모형)이었다. 또한 세 모형에서 모두 홈경기장의 효과는 유의적으로 추정되었다.

Table 3.2. Estimates(standard errors) of attacking and defensive ability by team under several models

	$u_{ht_i,at_i} = u_{at_i,ht_i}$		동점 로지스틱(상수)		동점 로지스틱(회귀모형)	
	공격력	수비력	공격력	수비력	공격력	수비력
홈경기장(home)	0.241(0.078)		0.310(0.089)		0.309(0.089)	
동점			−1.330(0.198)		−1.475(0.374)	
부산	0.232	0.121	0.222	−0.039	0.224	−0.044
	(0.142)	(0.145)	(0.165)	(0.175)	(0.167)	(0.176)
대구	−0.096	0.365	−0.161	0.181	−0.160	0.182
	(0.165)	(0.130)	(0.208)	(0.177)	(0.210)	(0.178)
대전	−0.197	0.049	−0.305	0.495	−0.305	0.492
	(0.175)	(0.150)	(0.223)	(0.153)	(0.224)	(0.159)
광주	−0.190	−0.029	−0.262	0.036	−0.259	0.038
	(0.172)	(0.156)	(0.206)	(0.173)	(0.207)	(0.175)
인천	−0.229	−0.187	−0.398	−0.079	−0.413	−0.097
	(0.175)	(0.173)	(0.250)	(0.214)	(0.250)	(0.214)
전북	0.533	0.117	0.724	−0.142	0.729	−0.139
	(0.125)	(0.148)	(0.139)	(0.204)	(0.141)	(0.205)
제주	0.129	−0.340	0.143	0.070	0.143	0.070
	(0.149)	(0.181)	(0.189)	(0.183)	(0.189)	(0.184)
전남	−0.182	−0.007	−0.337	−0.577	−0.340	−0.584
	(0.170)	(0.156)	(0.213)	(0.235)	(0.213)	(0.231)
경남	0.051	0.065	0.061	−0.051	0.066	−0.041
	(0.154)	(0.148)	(0.173)	(0.172)	(0.174)	(0.175)
강원	−1.019	−0.176	−1.038	0.161	−1.040	0.159
	(0.255)	(0.171)	(0.278)	(0.157)	(0.279)	(0.158)
포항	0.405	−0.033	0.551	−0.237	0.539	−0.244
	(0.132)	(0.160)	(0.149)	(0.197)	(0.153)	(0.199)
서울	(0.357)	0.256	0.414	−0.013	0.417	−0.009
	(0.135)	(0.139)	(0.151)	(0.183)	(0.152)	(0.185)
상주	−0.061	0.157	0.020	0.331	0.027	0.314
	(0.164)	(0.145)	(0.190)	(0.153)	(0.191)	(0.156)
성남	0.107	−0.180	0.055	0.128	0.058	0.133
	(0.151)	(0.171)	(0.178)	(0.170)	(0.179)	(0.171)
수원	0.262	−0.343	0.280	−0.132	0.294	−0.114
	(0.140)	(0.181)	(0.151)	(0.188)	(0.153)	(0.191)
울산	−0.187	0.013	−0.104	−0.264	−0.108	−0.270
	(0.170)	(0.028)	(0.185)	(0.197)	(0.186)	(0.199)

4. 득점 소요 시간에 대한 재발 사건 자료 분석의 적용

매 경기에서 발생하는 골 득점은 주어진 시간에서 발생되는 재발 사건으로 간주할 수 있을 것이다. 재발 사건 자료 분석은 각 관측 개체가 동일한 종류의 사건을 반복적으로 경험하는 것으로 그 발생 시점의 분포에 대한 통계적 모형을 구축하는 것이 주 목적이다. 재발 자료의 적용 예는 간질병 환자의 빈번한 발작, 암환자의 종양 재발, 성범죄자의 재범, 구직자의 빈번한 이직, 기계 부품의 지속적인 고장 등 여러 분야에서 그 예를 찾아볼 수 있을 것이다. 재발 사건 자료 분석은 생존 분석과 마찬가지로 연속형인 사건 발생 시간의 분포 추정과 공변량과의 연관성을 추정하기 위해 모수적, 비모수적, 그리고 준모수적 방

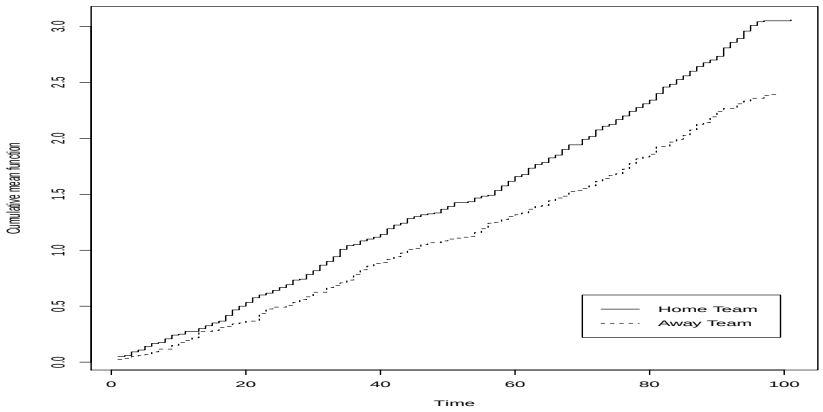

Figure 4.1. Cumulative mean function of home team and away team

법들이 적용되어 왔다. 특히 반복적으로 발생되는 사건들은 한 개체내에서 서로 연관성을 가지기 때문에 추론 시 종속 관계가 고려되어야 한다. Cook과 Lawless (2007)는 재발 사건 자료 분석에 대해 여러 가지 방법을 자세하게 소개하고 있다. 재발 사건 분석을 위해서는 분석한 시간 변수에 따라 두 가지 유형으로 나눌 수 있다. 즉 재발 사건 발생 시점을 그대로 사용하는 경우 $(0 = t_0 < t_1 < t_2 < \cdots < t_m)$와 재발 사건 사이 소요된 시간(gap time: $s_1 = t_1 - t_0, \ldots, s_m = t_m - t_{m-1}$)을 사용하는 경우이다. 두 가지 중 선택은 연구의 목적과 자료의 특성에 따라 정해질 것이다. 본 연구에서는 골 재득점까지 소요되는 시간의 분포와 그 분포에 영향을 주는 요인들을 구하고자 한다.

4.1. 누적 평균 함수를 이용한 재발 사건 분포의 추정

t시점까지 발생한 사건의 총 건수를 $N(t)$이라고 할 때, 누적 평균 함수(cumulative mean function), $M(t) = E(N(t))$를 추정하고자 한다. $i\,(= 1, \ldots, n)$번째 경기에서 t시점까지 발생되는 사건의 총 수를 $N_i(t)$라고 하자. 만약 모든 n개의 경기가 무사히 경기를 마친다고 할 때, $\hat{M}(t)$는 아래 같이 정의된다.

$$\hat{M}(t) = \frac{N(t)}{n}, \quad \text{여기서 } N(t) = \sum_{i=1}^{n} N_i(t).$$

Figure 4.1은 홈팀과 원정팀에 대해 따로 추정된 누적 평균 함수이다. 특히 이 함수가 보여주는 평원(plateau) 현상을 통해 사건 발생이 어떤 시점에 덜 빈번한지를 알 수 있으며 기울기를 통해 어느 시점이 골 득점 확률이 높을 지를 추론할 수 있다. Figure 4.1의 그림을 통해 전반 30–40분간 골 득점의 가능성이 가장 높으면 전반 종료 후반 시작 10–20분(전체 경기 50–60분)까지 골 득점의 빈도가 잠깐 주춤한 경향을 보인다. 또한 두 그룹의 득점 차이는 전반 15분까지 크게 차이가 없다가 이후 전 시점에 걸쳐 홈팀이 원정팀에 비해 골 득점이 높음을 알 수 있다. 그 차이는 거의 일정하며 이는 다음 절에 적용될 비례 위험(proportional hazard) 모형 가정의 적합성을 제시한다.

4.2. 준모수방법을 이용한 골 득점 소요시간에 대한 회귀 모형

이전 골 득점에서 재득점까지 소요 시간을 $s_{ij} = t_{ij} - t_{ij-1}$로 정의하였을 때, 공변량의 효과를 추정하기 위해 다음의 강도 함수가 적용된다.

$$\lambda(s_{ij}) = \lambda_0(s_{ij})\exp(\eta' \tilde{x}_i),$$

여기서 $\lambda_0(\cdot)$는 기저 강도 함수(baseline intensity function)이며 η는 공변량 $\tilde{x}_i$의 회귀 계수 벡터이다. 재발 사건간의 연관 관계에 대한 규정에 따라 여러 가지 방법이 적용될 수 있다. 조건부 방법(conditional approach), 주변 방법(marginal approach), 계수 과정(counting process) 방법 중 하나의 방법이 자료의 특성과 연구자의 관심 내용에 따라 결정된다. 이에 대한 자세한 내용은 Kelly와 Lim (2000)을 참고하길 바란다. 본 연구에서는 주변 방법을 적용할 것이다. 즉 각 개체내 발생되는 재발 사건들은 모두 독립이라고 가정한다. 실제 이 가정은 옳지 않으며 따라서 이 모형에 근거하여 구한 추정량은 이러한 모형의 오지정성(mis-specification)을 보상하기 위해 로버스트 분산을 유도하게 된다. 본 연구에서는 공변량으로 총 관중수($\tilde{x}_1$), 홈팀 여부($\tilde{x}_2$)와 이전 득점이 상대방 팀에 위한 것($\tilde{x}_3$)인지 여부가 고려되었으며 이들과 골 득점 소요 시간간의 관계를 조사하기 위해 모형을 적용하였다. 세 공변량의 추정치(표준 오차)는 각각 $(\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3) = 0.048(0.084), -0.002(0.004)$ 그리고 $0.278(0.085)$이었다. 즉, 홈팀과 관중수는 골 재득점 소요시간에 유의한 영향을 주지 못하는 반면에 이전 팀이 상대 팀인 경우 골 획득 시간이 짧아짐을 알 수 있다.

5. 결론 및 토의

본 논문에서는 K-리그 경기자료에 근거하여 각 팀의 수비력과 공격력을 측정하고 영확대 모형(zero-inflation model)의 확장으로 동점에 대해서는 대각 확대 모형이 적용되었다. 특히 결합 모형의 적용을 통해 동점과 골 득점과의 연관성 여부를 추정해 보았다. 이러한 연구의 확장으로 동점에 대한 다범주 모형을 적용해 볼 수 있을 것이다. 즉, 본 논문에서는 동점 여부에 따라 반응변수를 두 그룹으로 나누어 로지스틱 회귀 모형을 고려하였다. 하지만 동점내에도 (0 : 0, 1 : 1, 2 : 2, 3 : 3, 4 : 4)처럼 골 득점 차이가 존재할 경우 모든 동점을 동일 그룹으로 간주하는 대신에 골 득점을 고려한 순서형 다범주 자료(ordinal categorical data)가 적용될 수 있을 것이다.

또한 골 득점 소요 시간에 대한 모형의 확장으로 다단계 모형(Multi-state model) (예를 들어 A팀 → B팀, A팀 → A팀, B팀 → B팀, B팀 → A팀)을 적용함으로써 좀 더 세분한 관계를 유도할 수 있을 것이다.

또 다른 연구 문제로 골 소요 시간에 대한 기저 함수의 가정에 관한 것이다. 본 논문에서는 기저함수를 명시하지 않는 준모수 모형이 적용되었다. 만약 process의 시간대 별 변화정도에 대해 통계적 유의성을 구하고자 한다면 기저 강도 함수에 대해 모수적 가정 또는 조각 상수(piecewise constant) 모형이 고려될 수 있을 것이다.

더욱 더 흥미로운 연구의 확장은 더 많은 팀별 정보를 통해 팀들 간의 승부와 득점에 대한 예측을 하는 것이다. 이를 위해 스포츠 전문가의 전문적 지식을 필요로 할 것이다.

References

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*, Springer, New York.

Ho, L. and Singer, J. (2001). Generalized least squares methods for bivariate poisson regression, *Communications in Statistics - Theory and Methods*, **30**, 263–277.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models, *Journal of the Royal Statistical Society D (The Statistician)*, **52**, 381–393.

Kelly, P. J. and Lim, L. Y. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases, *Statistics in Medicine*, **19**, 12–33.

Kocherlakota, S. and Kocherlakota, K. (2001). Regression in the bivariate Poisson distribution, *Communications in Statistics - Theory and Methods*, **30**, 815–827.

Li, C., Lu, J., Park, J., Kim, K. and Peterson, J. (1999). Multivariate zero-inflated Poisson models and their applications, *Technometrics*, **41**, 29–38.

Walhin, J. (2001). Bivariate ZIP models, *Biometrical Journal*, **43**, 147–160.

Wang, K., Lee, A., Yau, K. and Carrivick, P. (2003). A bivariate zero inflated Poisson regression model to analyze occupational injuries, *Accident Analysis and Prevention*, **35**, 625–629.