

# 이변량 영과잉 포아송 및 대각확대 모형을 이용한 K-리그 골 득점 자료 분석<sup>†</sup>

허윤서<sup>1</sup> · 김경희<sup>2</sup>

<sup>1,2</sup>성신여자대학교 통계학과

접수 2018년 10월 29일, 수정 2018년 11월 16일, 게재 확정 2018년 11월 16일

## 요약

축구경기의 득점수 분석을 위한 연구는 꾸준히 이루어져 왔다 (Seong과 Chang, 2007; Lee, 2012). 본 연구는 2015~2018년의 케이-리그 (K-league) 경기 결과 자료에 이변량 영과잉 포아송 회귀모형을 포함한 8개의 회귀모형을 적합하였다. 반응변수는 홈팀과 원정팀의 전체 득점수 또는 후반전 득점수이며 설명변수는 각 팀의 전반전 득점수, 전반전 점유율이다. 모형판정기준에 의해 가장 적절한 모형은 전반전 득점수, 전반전 점유율을 설명변수로 한 이변량 포아송 회귀모형과 동점발생률에 대해 포아송 분포를 적용한 대각확대 회귀모형으로 나타났다. 홈팀의 전체 득점수에 대해 전반전 득점수가 점유율보다 높은 영향력을 미치는 반면 원정팀의 전체 득점수에 대해 전반전 득점수보다 전반전 점유율이 높은 영향력을 미침을 알 수 있었다.

주요용어: 대각확대 모형, 이변량 영과잉, 케이-리그, 포아송 모형.

## 1. 서론

가산 자료 (count data)란 사고 발생건수, 이직횟수와 같은 계수형 자료이다. 가산 자료에 대한 선형연구로는 Jeong과 Choi (2014)가 교통사고건수에 대해 포아송 회귀모형과 음이항 회귀모형을 적합하였으며 AIC 등의 모형 선택 기준에 의하여 음이항 회귀모형이 더 적절함을 보인 연구가 있고 Chun (2016)이 보험설계사들의 이직횟수 자료에 대해 포아송 회귀모형을 적합하여 분석한 연구 등이 있다.

가산 자료의 값에 0이 매우 많으면 이를 영과잉 (zero-inflated) 가산자료라고 한다. Lambert (1992)는 영과잉 가산자료에 대한 선형연구로는 영과잉 포아송 회귀모형을 이용하여 영의 발생률과 포아송 분포의 모수의 공변량 효과를 분석하였다. Li 등 (1999)은 일변량 결과를 다변량 영과잉 포아송 분포로 확장시켰으며 Kim과 Lee (2008)는 연체건수 자료에 대해 포아송 회귀모형과 영과잉 포아송 회귀모형을 적용하고 예측력 비교를 통해 신용평가 모형을 구축하였다. Kim과 Um (2010)은 영과잉 포아송 분포의 추정량인 적률추정량과 최우추정량을 비교하였고 Lee와 Nam (2012)는 영과잉 포아송 회귀모형을 활용하여 청소년 가출빈도 자료를 분석하였다. Park 등 (2015)은 영과잉 포아송 회귀모형, 영과잉 음이항 모형을 포함한 다양한 모형을 적용하여 중소기업 청년인턴의 이직횟수에 대한 자료를 분석하였다.

<sup>†</sup> 이 논문은 2018년도 성신여자대학교의 학술연구조성비 지원에 의하여 연구되었음.

<sup>1</sup> (02844) 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 자연과학대학 통계학과 석사과정.

<sup>2</sup> 고신저자: (02844) 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 자연과학대학 통계학과 조교수.

E-mail: arlenekim@sungshin.ac.kr

스포츠 분야에서의 대표적인 가산자료로서 경기에서 득점한 점수가 존재한다. 최근 스포츠 팬들은 스포츠 관람과 더불어 스포츠복권 베팅을 통해 간접적으로 경기에 참여하고 있는데, 이에 따라 각 팀의 득점수에 대한 정확한 분석과 예측이 무엇보다 중요해졌다. 특히 축구의 경우 다른 스포츠 경기에 비해 정규 경기 시간 내 득점수가 적으며, 양 팀의 득점이 하나도 없이 경기가 종료되는 결과를 많이 보이는 종목으로 축구경기의 득점수를 분석하기 위한 연구가 꾸준히 이루어져왔다 (Seong과 Chang, 2007; Lee, 2012).

축구 경기 자료에 대해 포아송 모형을 적용한 선행연구를 살펴보면 Dixon와 Coles (1997)가 두 개의 독립적인 포아송 분포를 가정한 팽창모형을 기반으로 축구게임 베팅 전략을 제안한 연구가 있다. Kalis와 Ntzoufras (2000)는 영국 프리미어리그 축구 자료를 포아송 모형 및 음이항 모형에 적용하여 모형의 타당성을 분석하였다. Kalis와 Ntzoufras (2003)는 Dixon과 Coles이 제안한 모형을 확장시켜 양 팀 득점 사이의 상관관계를 고려한 이변량 포아송 모형과 영과잉 대각확대 모형을 제안하고, 유럽축구 및 수구 데이터를 적용하여 공격력과 수비력을 분석하였다. 이들은 위의 모형에 대해 EM알고리즘을 이용한 최대 우도 추정 방법을 R프로그램으로 구현하고 호주의 건강 조사 자료와 이탈리아 축구 자료에 적용하였다.

우리나라의 축구리그인 케이-리그 (K-league)의 경기결과에 대한 연구로는 Shin 등 (2009)이 유럽 5대 리그 경기의 득점과 실점을 이용하여 회귀식을 추정하고 이를 K-리그 경기결과에 적용한 연구가 있고 Kim (2012)이 랜덤효과를 포함한 이변량 포아송 모형과 영과잉 모형을 적용하여 양 팀 득점 간의 상관관계를 분석한 연구가 있다. Lee (2012)는 각 팀별 승점과 실점이 서로 다른 포아송 분포를 따른다는 것을 제안하고 이러한 승률예측모형을 기존의 승률예측모형과 비교하였다. Lee (2014)는 K-리그의 공격력과 수비력 점수에 이변량 포아송 모형의 적합결과를 이용하여 골 득점을 예측했으며, 승점을 많이 취득하기 위해서는 공격력이 수비력보다 20%정도 더 중요하다는 결론을 내렸다.

본 연구는 2015년부터 2018년까지의 케이-리그 경기 결과 자료에서 양 팀의 득점수에 영향을 미치는 독립변수들을 고려하여 이변량 영과잉 포아송 회귀모형을 고려하였다. 축구 경기의 특성상 0 대 0 무승부 뿐 아니라 1 대 1 등의 무승부 상황이 빈번하므로 대각확대 모형도 고려하여 홈팀과 원정팀의 골 득점을 분석하였다. 본 연구의 순서는 다음과 같다. 2절에서 일변량 영과잉 포아송 모형과 이변량 영과잉 포아송 및 대각확대 모형 및 회귀모형을 설명한다. 3절은 분석을 위한 케이-리그 경기결과 자료 및 분석을 위한 사용한 모형에 대해 설명하고 분석결과를 제시한다. 마지막으로 4절에서 본 연구의 결론 및 연구방향을 제안한다.

## 2. 영과잉 포아송 모형

### 2.1. 일변량 영과잉 포아송 모형

영과잉 포아송 (zero-inflated Poisson) 모형이란 확률변수  $Y$ 가 0이 과도하게 많은 영과잉 자료를 설명하기 위해 포아송 모형을 변형한 것이다.  $p$ 의 확률로  $Y$ 가 0의 값을 가지고  $1 - p$ 의 확률로  $Y$ 가  $Poi(\lambda)$ 를 따를 경우  $Y \sim Poi(\lambda)$ 로 표기한다. 이때 0의 비율에 대한 확률  $p$ 는  $0 \leq p \leq 1$ 이고 포아송 분포의 평균  $\lambda$ 는  $\lambda > 0$ 이다. ZIP( $p, \lambda$ )의 확률질량함수는 다음과 같이 쓸 수 있다.

$$P(Y = y_i) = \begin{cases} p + (1-p)e^{-\lambda}, & y_i = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, & y_i = 1, 2, 3, \dots \end{cases}$$

## 2.2. 이변량 영과잉 포아송 및 대각확대 모형

이변량 영과잉 포아송 모형은 일변량 영과잉 포아송 모형을 2변량의 경우로 확장한 것이다. 먼저 이변량 포아송 (bivariate Poisson; BP) 분포는 3개의 모수  $(\lambda_{10}, \lambda_{20}, \lambda_{00})$ 로 정의되며  $U_1, U_2, Z$ 가 각각  $\lambda_{10}, \lambda_{20}, \lambda_{00}$ 을 평균으로 하는 포아송 분포를 따를 때,  $X_1 = U_1 + Z, X_2 = U_2 + Z$ 로 정의된  $(X_1, X_2)$ 는 BP를 따른다 (Marshall과 Olkin, 1985). 주변 분포는  $X_1 \sim Poi(\lambda_1), X_2 \sim Poi(\lambda_2)$ 로  $\lambda_1 = \lambda_{10} + \lambda_{00}, \lambda_2 = \lambda_{20} + \lambda_{00}$ 을 만족한다.

영이 과도하게 많은 이변량 영과잉 자료  $(Y_1, Y_2)$ 는 이변량 영과잉 포아송 (bivariate zero-inflated Poisson) 분포로 모형화할 수 있다.  $(Y_1, Y_2)$ 가  $p_0$ 의 확률로  $(0, 0)$ 을 갖고,  $p_1$ 의 확률로  $(Poi(\lambda_1), 0), p_2$ 의 확률로  $(0, Poi(\lambda_2)), p_{12}$ 의 확률로 이변량 포아송 분포를 따를 때,  $(Y_1, Y_2) \sim BZIP(p_0, p_1, p_2, \lambda_{10}, \lambda_{20}, \lambda_{00})$ 로 표기한다. 이때  $p_0 + p_1 + p_2 + p_{12} = 1$ 이고, 역시  $\lambda_1 = \lambda_{10} + \lambda_{00}, \lambda_2 = \lambda_{20} + \lambda_{00}$ 를 만족한다.

본 연구에서는 분석의 편의상  $p_1 = p_2 = 0$ 을 가정하여 Kalis와 Ntzoufras가 제안한  $BZIP(p_0, \lambda_{10}, \lambda_{20}, \lambda_{00})$  모형을 고려하였다.  $\lambda = \lambda_{10} + \lambda_{20} + \lambda_{00}$ 이라 놓으면,

$$f_{BZIP}(y_1, y_2) = \begin{cases} p_0 + p_{12} \exp(-\lambda) & \text{if } y_1 = y_2 = 0 \\ (1 - p_0)f_{BP}(y_1, y_2 | \lambda_{10}, \lambda_{20}, \lambda_{00}) & \text{if } y_1 \neq 0, y_2 \neq 0, \end{cases}$$

여기에서

$$f_{BP}(y_1, y_2 | \lambda_{10}, \lambda_{20}, \lambda_{00}) = \sum_{j=0}^{\min(y_1, y_2)} \frac{\lambda_{10}^{y_1-j} \lambda_{20}^{y_2-j} \lambda_{00}^j}{(y_1-j)!(y_2-j)!j!} \times \exp(-\lambda), \quad y_1, y_2 = 1, 2, \dots$$

즉 0-0 결과를 나타낸 경우를 제외한 나머지 경우는  $BP(\lambda_{10}, \lambda_{20}, \lambda_{00})$ 을 따른다고 가정한 것이다. 축구경기 결과의 경우 0-0뿐 아니라 1-1, 2-2 와 같은 동점이 발생하는 상황이 빈번하므로 0-0을 포함한 동점발생 확률  $p_D$ 를 고려하여 대각확대 (diagonal-inflated) 이변량 포아송 모형도 적합하였다. 모수벡터  $\boldsymbol{\theta}$ 로 정의되는 이산형 분포를  $f_D(y_1 | \boldsymbol{\theta})$ 라 하며  $f_D(y_1 | \boldsymbol{\theta})$ 는 주로 포아송 분포, 기하분포, 단순한 이산형 분포 ( $discrete(J), J = 0, 1, 2, \dots, j$ )이다.  $J = 0$ 일 때 0-0 상황만 고려하는 것이고  $J = 1$ 일 때 0-0과 1-1 상황을,  $J = j$ 일 때 0-0, 1-1,  $\dots, j-j$  상황을 고려하는 것을 뜻한다. 즉  $J = 0$ 이면 이변량 영과잉 포아송 모형이고, 대각확대 이변량 포아송 모형에서  $P_D = 0$ 이면 이변량 포아송 모형이다. 대각확대모형의 확률질량함수는 다음과 같다.

$$f_{DIBP}(y_1, y_2) = \begin{cases} (1 - p_D)f_{BP}(y_1, y_2 | \lambda_{10}, \lambda_{20}, \lambda_{00}) & \text{if } y_1 \neq y_2, \\ (1 - p_D)f_{BP}(y_1, y_2 | \lambda_{10}, \lambda_{20}, \lambda_{00}) + p_D f_D(y_1 | \boldsymbol{\theta}) & \text{if } y_1 = y_2, \end{cases}$$

여기에서  $f_D(y_1 | \boldsymbol{\theta})$ 로 다음과 같은 세 가지 분포를 고려한다.

1)  $f_D(y_1 | \boldsymbol{\theta}) \sim Poi(\theta)$

$$P(Y = y_1) = \frac{e^{-\theta} \theta^{y_1}}{y_1!}, \quad y_1 = 0, 1, 2, \dots$$

2)  $f_D(y_1 | \boldsymbol{\theta}) \sim Geometric(\theta)$

$$P(Y = y_1) = (1 - \theta)^{y_1 - 1} \theta, \quad y_1 = 1, 2, 3, \dots.$$

3)  $f_D(y_1 | \boldsymbol{\theta}) \sim \text{discrete}(J)$

$$P(Y_1 = y_1) = \theta_{y_1}, \quad \sum_{y_1=0}^J \theta_{y_1} = 1.$$

### 2.3. 이변량 영과잉 포아송 및 대각확대 회귀모형

$\lambda_{10}, \lambda_{20}, \lambda_{00}$ 에 대해 로그연결함수를 가정하고  $p_0$ 에 대해 로짓연결함수를 가정한 이변량 영과잉 포아송 회귀모형의 형태는 다음과 같다.

$$BZIP(p_0, \lambda_{10}, \lambda_{20}, \lambda_{00}),$$

여기에서

$$\log \lambda_{10} = A_1^T \boldsymbol{\alpha}_1, \quad \log \lambda_{20} = A_2^T \boldsymbol{\alpha}_2, \quad \log \lambda_{00} = A_3^T \boldsymbol{\alpha}_3, \quad \log \frac{p_0}{1 - p_0} = B^T \boldsymbol{\beta}.$$

$p_0, \lambda_{10}, \lambda_{20}, \lambda_{00}$ 은 서로 다른 변수에 의해 설명되므로  $A_1^T, A_2^T, A_3^T, B$ 는 각각 절편을 포함한  $1 \times k_1, 1 \times k_2, 1 \times k_3, 1 \times k_4$  설명변수 벡터이고  $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}$ 는 설명변수에 대한  $k_1 \times 1, k_2 \times 1, k_3 \times 1, k_4 \times 1$  회귀계수 벡터이다.  $(Y_1, Y_2)$ 에 대각확대 모형을 고려할 경우 2.2절에서 정의한 모형을 연결함수를 이용하여 적절히 변형한다.

## 3. 자료 설명 및 분석 결과

### 3.1. 자료 설명

본 연구에서 사용한 자료는 2015년 3월 7일부터 2018년 9월 9일까지 실시된 케이-리그 경기 결과 자료이며 R의 bivpois패키지 (Karlis, 2005)를 이용하여 분석하였다. 케이-리그 경기 결과 자료에는 각 경기별 날짜, 홈팀 및 원정팀 이름, 정규시간 내 골 득점 수, 전반/후반 골 득점수, 전반/후반 점유율에 대해 기록되어 있다 (한국프로축구연맹). 2015년과 2016년에 각각 452번의 경기, 2017년에 412번의 경기, 2018년에 297번의 경기로 총 1613번의 경기가 이루어졌다. 경기는 주로 4~10월에 이루어졌으며 2015년과 2016년에는 23개의 팀이 참여하여 경기를 진행하였고 2017년과 2018년에는 고양, 충주의 K리그 탈퇴 및 안산 팀의 창단으로 22개의 팀이 경기를 진행하였다. 경기결과 중 무승부는 약 28%이고 홈팀이 승리한 경우는 약 38%, 원정팀이 승리한 경우는 약 33%이다. 무승부 결과 중 약 35%가 0-0이고 약 41%가 1-1, 약 19%가 2-2의 결과를 보였다 (Figure 3.1).

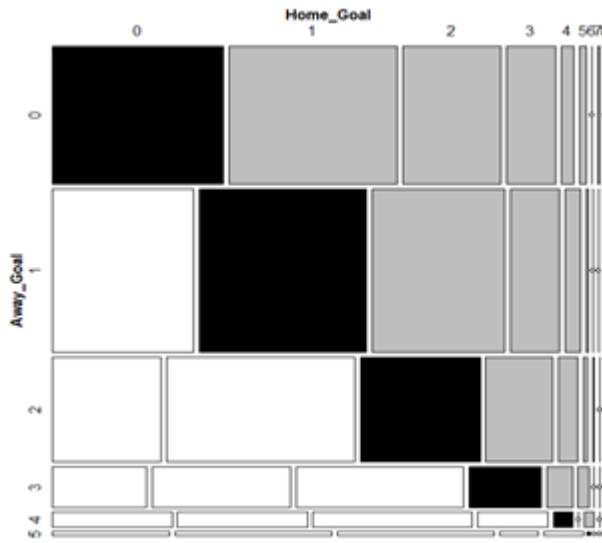


Figure 3.1 No. of goals at home and away

### 3.2. 변수 및 모형 설명

사전 분석을 실시한 결과 전체적인 경기 상황으로 경기시간 내 전체 득점수를 설명하는 것보다 전반전의 경기 상황으로 전체 득점수 또는 후반전 득점수를 설명하는 것이 더욱 적절하다고 판단하였다. 2.3절의 회귀모형에서 반응변수로  $Y_1$  = 홈팀의 전체 득점수,  $Y_2$  = 원정팀의 전체 득점수를 고려하거나  $Y_1$  = 홈팀의 후반전 득점수,  $Y_2$  = 원정팀의 후반전 득점수를 고려하였다. 특정 팀이 홈팀인지 원정팀인지에 따라 응원 관객 수, 경기장 적응정도 등으로 인한 경기력이 다를 수 있으므로 설명변수  $A_1^T, A_2^T$ 로 양 팀의 홈팀 및 원정팀의 팀 자체 효과, 홈팀 및 원정팀의 전반전 득점수, 전반전 점유율과 같은 전반전의 경기 상황을 고려하였다. 전반전 득점수는 first half goals, 전반전 점유율은 first half possession rate, 전체 득점수는 total goals, 후반전 득점수는 second half goals로 표기하고 홈경기의 경우 home, 원정경기의 경우 away를 괄호 안에 표기하였다. Table 3.1은 변수들의 요약통계량을 보여준다.

Table 3.1 Summary statistics

Variables	Mean	Min	Q1	Median	Q3	Max
first half goals (home)	0.55	0.00	0.00	0.00	1.00	5.00
first half goals (away)	0.49	0.00	0.00	0.00	1.00	5.00
first half possession rate (home)	0.51	0.05	0.47	0.51	0.56	0.73
first half possession rate (away)	0.49	0.27	0.44	0.49	0.53	0.95
total goals (home)	1.35	0.00	0.00	1.00	2.00	8.00
total goals (away)	1.23	0.00	0.00	1.00	2.00	5.00
second half goals (home)	0.79	0.00	0.00	1.00	1.00	4.00
second half goals (away)	0.74	0.00	0.00	1.00	1.00	4.00

본 연구는 각 팀에 대해 홈팀 및 원정팀의 팀 자체 효과와 함께 홈팀의 전반전 경기상황으로 홈팀의 전체 득점수/후반전 득점수를 설명하고 원정팀의 전반전 경기상황으로 원정팀의 전체 득점수/후반전

득점수를 설명하였다. 홈팀의 전체 득점수 또는 후반전 득점수 기댓값을  $\lambda_{10}$ , 원정팀의 전체 득점수 또는 후반전 득점수 기댓값을  $\lambda_{20}$ 로 가정하고 공통적으로 영향을 미치는 모수  $\lambda_{00}$ 는 상수로 가정하였다.

고려한 모형은 총 8개로 모형 1~4는 설명변수로 각 팀의 팀 자체 효과와 전반전 득점수를 고려하였으며, 모형 5~8는 각 팀의 팀 자체 효과와 전반전 득점수와 전반전 점유율을 설명변수로 고려하였다. 모형 1과 모형 5는 이변량 포아송 회귀모형이며 모형 2와 모형 6은 동점발생 확률에 대해 단순한 이산형 분포와  $J = 0$ 을 고려한 대각확대모형, 즉 이변량 영과잉 포아송 회귀모형이다. 모형 3, 모형 7은 동점발생 확률에 대해 단순한 이산형 분포를 고려하고  $J = 1$ 을 고려한 대각확대모형이다. 모형 4와 모형 8은 동점발생 확률에 대해 포아송 분포를 고려한 대각확대모형이다. 즉,

1) 모형1 / 모형5

$$(Y_1, Y_2) \sim BP(\lambda_{10}, \lambda_{20}, \lambda_{00}) \\ \log \lambda_{10} = A_1^T \boldsymbol{\alpha}_1, \quad \log \lambda_{20} = A_2^T \boldsymbol{\alpha}_2.$$

2) 모형2 / 모형6

$$(Y_1, Y_2) \sim BZIP(p_0, \lambda_{10}, \lambda_{20}, \lambda_{00}) \\ \log \lambda_{10} = A_1^T \boldsymbol{\alpha}_1, \quad \log \lambda_{20} = A_2^T \boldsymbol{\alpha}_2.$$

3) 모형3 / 모형4 / 모형7 / 모형8

$$(Y_1, Y_2) \sim DIBP(p_D, \lambda_{10}, \lambda_{20}, \lambda_{00}) \\ \log \lambda_{10} = A_1^T \boldsymbol{\alpha}_1, \quad \log \lambda_{20} = A_2^T \boldsymbol{\alpha}_2.$$

이때  $A_1^T$ 는 홈팀의 득점수를 설명하기 위한 홈팀 전반전 상황 설명벡터이고  $A_2^T$ 는 원정팀의 득점수를 설명하기 위한 원정팀 전반전 상황 설명벡터이다.

### 3.3. 분석 결과

Table 3.2는 홈팀 및 원정팀의 전체 득점수를 반응변수  $Y_1, Y_2$ 로 할 경우 회귀모형을 적합한 결과이다. 모든 판정기준에 의해 이변량 포아송 회귀모형이 가장 좋은 결과를 보였는데, loglikelihood와 AIC판정기준으로는 팀 자체 효과, 각 팀의 전반전 득점수, 각 팀의 전반전 점유율을 설명변수로 고려한 모형 5가 가장 적절하였으며 BIC판정기준으로는 팀 자체 효과와 각 팀의 전반전 득점수를 설명변수로 고려한 모형 1이 가장 적절하였다.

Table 3.3은 홈팀 및 원정팀의 후반전 득점수를 반응변수  $Y_1, Y_2$ 로 할 경우 회귀모형을 적합한 결과이다. Log likelihood 판정기준으로는 팀 자체 효과와 각 팀의 전반전 득점수, 각 팀의 전반전 점유율을 설명변수로 하고 동점발생확률에 대해 포아송 분포를 고려한 대각확대 회귀모형인 모형 8이 가장 적합하였다. AIC와 BIC 판정기준으로는 이변량 포아송 회귀모형이 가장 적절한 결과를 보였는데, AIC 판정기준의 경우 팀 자체 효과, 각 팀의 전반전 득점수, 각 팀의 전반전 점유율을 설명변수로 고려한 모형 5를 적합하였을 때 더 적절했으며 BIC 판정기준의 경우 팀 자체 효과, 각 팀의 전반전 득점수를 설명변수로 고려한 모형 1을 적합하였을 때 더 적절하였다.

**Table 3.2** Model fit ( $Y_1$  =total goals (home),  $Y_2$  =total goals (away))

Models	Explanatory variables	No. of parameters	Log likelihood	AIC	BIC
1. BP	team effect	51	-4137.71	8376.06	8686.09*
2. BZIP( $J=0$ )	first half goals	52	-4140.70	8378.05	8694.16
3. DIBP( $\text{discrete}, J=1$ )		53	-4140.30	8380.07	8702.26
4. DIBP( $\text{dist}=\text{poi}$ )		53	-4140.14	8380.04	8702.23
5. BP	team effect	53	-4135.56*	8375.74*	8697.93
6. BZIP( $J=0$ )	first half goals	54	-4138.55	8377.74	8706.00
7. DIBP( $\text{discrete}, J=1$ )	first half possession rate	55	-4138.15	8379.75	8714.10
8. DIBP( $\text{dist}=\text{poi}$ )		55	-4138.00	8379.73	8714.07

즉 2015~2018년 경기 자료에 각각 다른 8개의 모형을 적합하였을 때 모형 5 > 모형 1 > 모형 8 순으로 우수한 결과를 보였다. 또한 전반전 경기상황으로 전체 득점수를 설명한 것 보다 후반전 득점수를 설명한 경우에 Log likelihood, AIC, BIC값이 더 작았다.

이 결과를 고려하여 이번량 포아송 회귀모형 중 우수한 모형5와 대각확대 회귀모형 중 우수한 모형8을 적합하고 회귀 계수를 비교하였으며 2018년 리그에서 1~6위를 차지한 팀의 팀별 효과를 Table 3.4, Table 3.5에 기록하였다.

**Table 3.3** Model fit ( $Y_1$  =2nd half goals (home),  $Y_2$  =2nd half goals (away))

Model	Explanatory variables	No. of parameters	Log likelihood	AIC	BIC
1. BP	team effect	51	-3688.23	7477.56	7787.59*
2. BZIP( $J=0$ )	first half goals	52	-3689.46	7478.28	7794.39
3. DIBP( $\text{discrete}, J=1$ )		53	-3688.76	7480.27	7802.46
4. DIBP( $\text{dist}=\text{poi}$ )		53	-3688.05	7478.57	7800.75
5. BP	team effect	53	-3683.00	7471.10*	7793.29
6. BZIP( $J=0$ )	first half goals	54	-3683.77	7472.06	7800.32
7. DIBP( $\text{discrete}, J=1$ )	first half possession rate	55	-3683.63	7474.04	7808.39
8. DIBP( $\text{dist}=\text{poi}$ )		55	-3682.88*	7471.98	7806.32

Table 3.4와 Table 3.5는 각각 모형5와 모형8의 회귀모형 적합 결과를 보여준다. 모형 5와 모형 8 모두 홈팀의 전체 득점수에 대해 홈팀의 전반전 점유율보다 전반전 득점수가 더 영향이 있고, 원정팀의 전체 득점수에 대해 원정팀의 전반전 득점수보다 전반전 점유율이 약간 더 영향이 있음을 보여준다. 또한 홈팀과 원정팀의 후반전 득점수에 대해 전반전의 득점수의 영향력보다 전반전 점유율의 영향력이 큰 편으로 볼 수 있다.

특히 홈팀에 비해 원정팀의 전반전 점유율이 높을수록 후반전 득점에 대한 가능성성이 높아진다. 각 팀의 전반전 득점수는 후반전 득점수와 음의 상관관계가 있는 반면 전체 득점수와 양의 상관관계가 있고, 각 팀의 전반전 점유율은 전체 득점수보다 후반전 득점수에 큰 영향을 준다. 모형 8의 경우 전체득점수를 반응변수로 이용한 모형에서는 영과잉 확률의 추정값이 거의 0으로 모형 5의 적합값들과 거의 같은 값을 보여주고 있다.

Table 3.6는 모형 5와 모형 8을 적용한 결과 해당 팀이 홈팀일 경우를 가정하고 적합한 득점수 ( $\hat{Y}_1$ )의 평균과 원정팀일 경우를 가정하고 적합한 득점수 ( $\hat{Y}_2$ )의 평균을 실제 홈팀일 경우의 평균득점수 및 원정팀일 경우의 평균득점수와 비교한 결과이다. Table 3.6에서 제시한 RMSE (root mean square error)는 각 경기에 대한 실제 득점수 와 적합 값의 차이를 제곱해서 평균한 값에 제곱근을 취한 값이다.

반응변수로 홈팀 및 원정팀의 전체 득점수를 고려하였을 때 모형 5와 8은 홈팀일 경우의 득점수 적합

**Table 3.4** Estimated regression parameters (Model 5)

$(Y_1, Y_2)$	team name	home effect	away effect	$\hat{\alpha}_1$			$\hat{\alpha}_2$			$\hat{\alpha}_3$	
				interce pt	goals (first half)	possessi on (first half)	interce pt	goals (first half)	possessi on (first half)	interc ept	
$Y_1 = \text{total goals (home)}$	Jeonbuk	0.35	-0.46	-0.17	0.54	0.29	-0.44	0.58	0.70	-3.41	
	Gyeongnam	0.16	-0.31								
$Y_2 = \text{total goals (away)}$	Ulsan	0.15	-0.28								
	Pohang	0.24	-0.29								
	Suwon	0.27	-0.25								
	Jeju	0.25	-0.27								
$Y_1 = \text{2nd half goals (home)}$	Jeonbuk	0.64	-0.66	0.54	-0.01	0.79	-0.85	-0.03	1.31	-3.14	
	Gyeongnam	0.31	-0.46								
$Y_2 = \text{2nd half goals (away)}$	Ulsan	0.22	-0.46								
	Pohang	0.33	-0.45								
	Suwon	0.40	-0.38								
	Jeju	0.42	-0.41								

**Table 3.5** Estimated regression parameters (Model 8)

$(Y_1, Y_2)$	team name	home effect	away effect	$\hat{\alpha}_1$			$\hat{\alpha}_2$			$\hat{\alpha}_3$		$p_D$
				interce pt	goals (first half)	possessi on (first half)	interce pt	goals (first half)	possessi on (first half)	interc ept		
$Y_1 = \text{total goals (home)}$	Jeonbuk	0.35	-0.46	-0.17	0.54	0.29	-0.44	0.58	0.70	-3.46	0.00	
	Gyeongnam	0.16	-0.31									
$Y_2 = \text{total goals (away)}$	Ulsan	0.15	-0.28									
	Pohang	0.24	-0.29									
	Suwon	0.27	-0.252									
	Jeju	0.25	-0.270									
$Y_1 = \text{2nd half goals (home)}$	Jeonbuk	0.67	-0.640	-0.54	-0.01	0.77	-0.86	-0.03	1.33	-4.00	0.03	
	Gyeongnam	0.33	-0.45									
$Y_2 = \text{2nd half goals (away)}$	Ulsan	0.24	-0.45									
	Pohang	0.37	-0.42									
	Suwon	0.43	-0.37									
	Jeju	0.43	-0.40									

값의 평균인  $\bar{Y}_1$ 과 원정팀일 경우 득점수 적합값의 평균  $\bar{Y}_2$ 에 대해 전북 > 수원 > 제주 순으로 큰 값을 나타냈다. 전북의 경우 홈팀, 원정팀 여부에 상관없이 팀 자체 효과가 높다고 할 수 있다. 실제로 전북팀은 2017리그 2위를 제외하고 2015, 2016, 2018 리그에서 1위를 차지한 팀이다. 포항팀의 경우 홈팀 및 원정팀 일 때 실제 평균 득점수는 각각 1.535, 1.157로 홈팀인 경우 더 높은 평균 득점수를 나타냈지만 모형 적합 결과 두 모형 모두 각각 1.380, 1.318로 실제 득점수에 비해 적은 차이를 보였다. 반면 수원팀은 홈팀 및 원정팀 일 때 실제 평균 득점수는 각각 1.571, 1.563으로 홈팀 또는 원정팀의 여부와 큰 상관없이 비슷한 득점수를 보이지만 모형 적합 결과 두 모형에서 모두 홈팀일 때 1.686, 원정팀일 때 1.446으로 홈팀일 때 더 높은 득점수 적합값을 나타냈다. 두 모형의 RMSE는 홈팀의 경우보다 원정팀의 경우가 근소하게 낮게 나타났다.

반응변수로 홈팀 및 원정팀의 후반전 득점수를 고려하였을 때 모형 5와 모형 8은 홈팀일 경우의 후반전 득점수의 평균인  $\bar{Y}_1$ 에 대해 전북 > 경남 > 수원 순으로 큰 값을 나타냈고, 원정팀일 경우의 후반전 득점수의 평균인  $\bar{Y}_2$ 에 대해 전북 > 제주 > 수원 순으로 큰 값을 나타냈다. 울산팀은 실제 후반전 평균 득점수가 홈팀일 때 0.686이지만 원정팀일 때는 0.761로 더 높은 후반전 평균 득점수를 보인다. 그러나

**Table 3.6** Average goals and average of fitted goals using model 5 and 8

$(Y_1, Y_2)$	team name	average goals		model 5		model 8	
		$\bar{Y}_1$	$\bar{Y}_2$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_1$	$\hat{Y}_2$
$Y_1 = \text{total goals}$ (home)	Jeonbuk	1.855	1.778	1.831	1.799	1.831	1.799
	Gyeongnam	1.472	1.366	1.511	1.326	1.511	1.326
$Y_2 = \text{total goals}$ (away)	Ulsan	1.300	1.239	1.359	1.178	1.359	1.178
	Pohang	1.535	1.157	1.380	1.318	1.380	1.318
	Suwon	1.571	1.563	1.686	1.446	1.686	1.446
	Jeju	1.714	1.352	1.656	1.407	1.656	1.407
RMSE				0.941	0.923	0.940	0.923

$(Y_1, Y_2)$	team name	average goals		model 5		model 8	
		$\bar{Y}_1$	$\bar{Y}_2$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_1$	$\hat{Y}_2$
$Y_1 = \text{2nd half goals}$ (home)	Jeonbuk	1.145	1.028	1.137	1.037	1.154	1.052
	Gyeongnam	0.931	0.789	0.899	0.828	0.905	0.831
$Y_2 = \text{2nd half goals}$ (away)	Ulsan	0.686	0.761	0.750	0.700	0.746	0.694
	Pohang	0.986	0.629	0.835	0.776	0.849	0.787
	Suwon	0.800	0.930	0.887	0.830	0.894	0.836
	Jeju	0.929	0.817	0.882	0.845	0.882	0.844
RMSE				0.877	0.852	0.881	0.855

모형 5 적합 결과, 홈팀일 때 0.75로 원정팀일 때 적합값 0.70보다 높은 후반전 평균 득점수를 나타냈다. 수원팀도 울산팀과 마찬가지로 실제로는 홈팀일 때 후반전 평균 득점수가 원정팀일 때 후반전 평균 득점수에 비해 높지만 모형 5, 모형 8 모두 원정팀일 때 후반전 평균 득점수 적합값이 홈팀일 때 후반전 평균 득점수 적합값보다 높게 나타났다. 두 모형의 전반적인 해석은 비슷하나 모형 8의 RMSE보다 모형 5의 RMSE이 더 작으므로 각 팀의 자체 효과와 전반전 득점수, 전반전 점유율을 고려한 이변량 포아송 회귀 모형이 홈팀 및 원정팀의 후반전 득점수를 더 잘 설명하는 모형이라고 판단하였다.

#### 4. 결론 및 논의

본 연구에서는 반응변수로 홈팀 및 원정팀의 전체 득점수 또는 후반전 득점수를 고려하고 설명변수로 팀 자체효과, 전반전 득점수, 전반전 점유율을 고려한 이변량 포아송 회귀모형을 포함한 8개의 회귀모형을 이용하여 케이-리그 경기 결과를 분석하였다. 이변량 포아송 회귀모형과 동점발생확률에 대해 포아송 분포를 고려한 대각화 회귀모형이 우수한 결과를 보여 두 모형을 이용한 회귀계수와 적합값을 비교하였다. 비교 결과, 전북은 홈팀, 원정팀의 여부와 관계없이 팀 자체 효과가 높은 것으로 나타났다. 홈팀은 전반전 득점수가 전반전 점유율보다 홈팀 전체 득점수에 더 큰 영향이 있고 원정팀은 전반전 득점수보다 전반전 점유율이 원정팀 전체 득점수에 더 큰 영향이 있는 것으로 나타났다.

케이-리그의 경기 결과는 매해 형태가 일정하지 않고 본 연구에서 설명변수로 사용한 전반전 득점수, 전반전 점유율 이외에 경기 결과에 영향을 주는 다양한 요인이 존재할 것이라 예상할 수 있다. 각 팀의 득점수를 보다 면밀하게 분석하기 위해 날씨, 관람객 수 등 선수들의 경기력에 영향을 주는 다양한 변수를 고려한 모형을 이용하면 더욱 정확한 분석과 예측이 가능할 것으로 사료된다.

## References

- Chun, H. J. (2016). The factors of insurance solicitor's turnovers of life insurance using Poisson regression. *Journal of the Korean Data & Information Science Society*, **27**, 1337-1347.
- Dixon, M. J. and Coles, S. C. (1997) Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, **46**, 265-280.
- Jeong, J. P. and Choi, J. H. (2014). Poisson regression and negative binomial regression model fit for traffic accidents. *Journal of the Korean Data Analysis Society*, **16**, 165-172.
- Karlis, D. and Ntzoufras, I. (2000) On modelling soccer data. *Student*, **3**, 229-245. Karlis, D. and Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *The Statistician*, **52**, 281-393.
- Karlis, D. and Ntzoufras, I. (2005). Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *Journal of Statistical Software*, **14**, 1-36.
- Kim, J. Y. and Lee, S. K. (2008). A case study on the credit scoring model with zero-inflated Poisson regression. *Journal of the Korean Data Analysis Society*, **10**, 3255-3265.
- Kim, K. M. and Um, H. J. (2010). Comparisons of the estimators for the zero-inflated Poisson distribution. *Journal of The Korean Data Analysis Society*, **12**, 1113-1124.
- Kim, Y. J. (2012). Statistical analysis of K-league data using Poisson model. *The Korean Journal of Applied Statistics*, **25**, 775-783.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.
- Lee, H. Y. (2012). A study on forecasting the winning rate of soccer games using the Poisson distribution. *Journal of the Korean Data Analysis Society*, **14**, 499-507.
- Lee, J. T. (2014). Prediction of K-league soccer scores using bivariate Poisson distributions. *Journal of the Korean Data & Information Science Society*, **25**, 1221-1229.
- Lee, S. M. and Nam, S. H. (2012). The factors affecting the increase of youths runaway episode -Analysis using zero-inflated Poisson regression model-. *Korean Journal of Youth Studies*, **19**, 85-108.
- Li, C. S., Lu, J. C., Park, J., Kim, K. and Peterson, J. (1999) Multivariate zero-inflated Poisson models and their applications. *Technometrics*, **41**, 29-38.
- Park, S. I., Ryu, J. S. and Kim, J. H. (2015). The study on the determinants of the number of job changes. *Journal of the Korean Data & Information Science Society*, **26**, 387-397.
- Seong, H. and Chang, W. J. (2007). Forecasting the results of soccer matches using Poisson model. *IE Interfaces*, **20**, 133-141.
- Shin, S. K. Cho, Y. J. and Cho, Y. S. (2009). A study on points per game using scored goal per game and lossed goal per game in the union of European football professional league. *Journal of the Korean Data & Information Science Society*, **20**, 837-844.

## Analysis of K-league data using bivariate Poisson and diagonal inflated model<sup>†</sup>

Yun Seo Heo<sup>1</sup> · Kyoung Hee Kim<sup>2</sup>

<sup>1,2</sup>Department of Statistics, Sungshin Women's University

Received 29 October 2018, revised 16 November 2018, accepted 16 November 2018

### Abstract

There has been a steady research for analyzing number of goals of soccer game (Seong and Chang, 2007; Lee, 2012). In this study, eight regression models including the bivariate zero inflated Poisson regression model were fitted to K-league data for 2015–2018. The response variable is the number of total goals or the second half goals for home and away teams. Explanatory variables are the number of goals for the first half and the first half ball possession rate of each team. We chose bivariate Poisson regression model and the diagonal inflated regression model with Poisson tie probability distribution following several model selection criteria such as log likelihood, AIC and BIC. We found that the first half goals of home teams have a higher influence on total goals of home teams than the first half ball possession rate, but vice versa for away teams.

*Keywords:* Bivariate zero-inflated poisson, diagonal-inflated, K-league.

---

<sup>†</sup> This work was supported by the Sungshin University Research Grant of 2018.

<sup>1</sup> Graduate student, Department of Statistics, College of Natural Sciences, Sungshin Women's University, Bomun-ro 34 Da Gil, Seongbuk-Gu, Seoul 02844, Korea.

<sup>2</sup> Corresponding author: Assistant professor, Department of Statistics, College of Natural Sciences, Sungshin Women's University, Bomun-ro 34 Da Gil, Seongbuk-Gu, Seoul 02844, Korea. E-mail: arlenekim@sungshin.ac.kr