Journal of the Korean Data & Information Science Society
2014, **25**(6), 1221–1229

http://dx.doi.org/10.7465/jkdi.2014.25.6.1221
한국데이터정보과학회지

이변량 포아송분포를 이용한 K-리그 골 점수의 예측

이장택[1]

[1]단국대학교 응용통계학과

접수 2014년 7월 25일, 수정 2014년 8월 17일, 게재확정 2014년 9월 1일

요 약

30년 동안의 K-리그 자료를 득점과 실점이 서로 상관이 있다는 가정과 R 패키지를 사용하여 12개의 서로 다른 이변량 포아송모형에 적합시켰다. 그 결과 AIC와 BIC 판정기준 아래에서 공변량 효과가 상수인 이변량 포아송모형이 가장 타당하며, 영과잉 및 대각확대 모형은 필요하지 않은 것으로 나타났다. 제안된 모형은 홈경기의 효과, 팀별 공격능력과 수비능력 및 적합도를 구하는 데 사용될 수 있다.

주요용어: 대각확대 모형, 영과잉모형, 이변량 포아송모형, 케이-리그.

1. 머리말

축구는 11명이 한 팀을 이루어 두 팀이 겨루는 세계적으로 큰 인기를 누리는 스포츠이다. 축구 경기의 승패는 팬들에게 지대한 관심사이며, 경기 결과는 운에 따르기도 하지만 팀에 대한 통계를 통해 확률적으로 접근하여 많은 성과를 창출했는데, 2014 국제축구연맹 월드컵에서의 독일 우승은 실력도 탁월했지만 빅데이터 전략도 매우 큰 역할을 한 것으로 알려져 있다. 축구 경기에 있어서 고전적인 중요한 관심사는 골 득점의 확률분포에 관한 것인데, 선행연구들을 살펴보면 Moroney (1956)는 포아송분포를 이용하여 축구 경기 전체 및 각 팀별 골 득점의 분포를 모형화 하였으며, REEP 등 (1971)은 골 득점 분포를 음이항 분포로 설명하였다. 또한 Greenhough 등 (2002)은 169개국 장기리그 축구경기 골의 분포는 극단치분포를 필요로 한다는 사실을 지적하였으며, Germert (2010)은 축구 골은 영의 값이 과잉관측되는 경우가 많은 이유로 영과잉 포아송분포 (zero inflated Poisson distribution)를 이용하여 설명하였다. 이외에도 프리미어 리그나 분데스리가와 같은 해외유명 축구리그인 경우에는 선행연구들이 많이 있으나 비록 2002년 월드컵 4강까지 진출한 한국 축구지만 한국을 대표하는 K-리그에 대한 체계적인 연구들은 그리 많지 않다. 축구 경기 결과에 대한 국내연구로는 유럽 리그에서 득점과 실점을 이용한 승점 추정에 관한 연구 (Shin 등, 2009), 2010년 남아공 월드컵 축구의 최종 결과와 예측자료를 비교한 연구 (Hong 등, 2010), 축구 대표 팀의 선수들 간의 패스 정보를 이용한 사회네트워크분석 (Choi 등, 2011) 등이 있다.

한편 축구경기에서 득점과 실점은 서로 상관관계가 존재할 수도 있는데, K-리그와 같은 장기리그인 경우에는 전반에 득점 차이가 많이 나면 다음 경기에 대한 체력 안배 등의 이유로 구태여 후반전에 사력을 다할 필요가 없는 등의 이유가 발생하기 때문에 득점과 실점의 상관관계를 고려하지 않고 추정하는 경우에는 편의된 모수 추정치를 얻을 가능성이 있다. 하지만 Lee (1997)와 Karlis 등 (2000)의 경우를 보더라도 축구 경기의 골 득점은 강하지는 않지만 서로 종속이라는 사실은 모형화가 용이하지 않다

[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

는 이유로 대부분 무시되어졌다. 이와 같은 문제점을 해결하기 위하여 학계에서는 일반적으로 두 팀 간의 경기에서 두 팀의 득점은 서로 상관관계가 존재할 것이라는 사실을 수용할 수 있는 이변량 포아송모형을 고려하기 시작하였으며, 이변량 포아송분포는 쌍으로 표현되는 서로 연관된 계수형 자료에 사용되는 분포로 주변확률분포는 각각 포아송분포지만 2개의 확률변수는 서로 종속이며 대각확대를 허용함으로써 축구에서 무승부의 가능성을 높일 수 있다. 이변량 포아송모형을 K-리그 경기 자료에 접목시킨 연구로는 2011년 K-리그 경기자료에 근거하여 각 팀의 수비력과 공격력을 측정하고 대각확대 모형을 이용하여 결합 모형의 적용을 통해 동점과 골 득점과의 연관성 여부를 추정한 Kim (2012)이 있는데, 본 연구는 2012년까지 K-리그 전 경기의 결과와 Karlis와 Ntzoufras (2003)이 다룬 12개의 이변량 포아송 모형 전부를 포함시켜 포괄적 비교를 시도하였다는 점이 다르다고 할 수 있다.

논문은 다음과 같이 구성된다. 2절에서는 이변량 포아송분포와 축구에 대한 포아송 로그 선형모형을 설명하며 3절에서는 사용된 데이터 및 여러 가지 포아송모형의 K-리그 적합결과를 제시하고 공격 및 수비능력을 이용한 승점의 회귀방정식을 제안하며 마지막 4절은 본 연구의 결론을 맺고자 한다.

2. 이변량 포아송 자료에 대한 모형

2.1. 이변량 포아송분포

모수가 각각 λ_i인 서로 독립인 포아송분포를 따르는 확률변수 X_i, $i = 1, 2, 3$에 대해 2개의 확률변수 $X = X_1 + X_3$와 $Y = X_2 + X_3$가 식 (2.1)과 같은 결합확률밀도함수를 가질 때, (X, Y)는 이변량 포아송분포 (bivariate Poisson distribution) $BP(\lambda_1, \lambda_2, \lambda_3)$를 따른다고 한다 (Karlis와 Ntzoufras, 2003).

$$f_{BP}(x, y|\lambda_1, \lambda_2, \lambda_3) = \exp[-(\lambda_1 + \lambda_2 + \lambda_3)]\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{k=0}^{\min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k \tag{2.1}$$

이변량 포아송 확률변수 (X, Y)의 주변확률분포는 각각 기댓값이 $E(X) = \lambda_1 + \lambda_3$와 $E(Y) = \lambda_2 + \lambda_3$인 포아송분포를 따른다. 또한 두 확률변수의 연관 정도를 나타내는 모수 λ_3가 0이면 두 확률변수가 독립임을 의미하며 일반적으로 이 경우 두 확률변수는 이중 포아송분포 (double Poisson distribution)를 따른다고 한다. 한편 축구경기에서 K-리그와 같은 장기리그인 경우에는 득점보다 승점이 중요한 이유로 승리할 가능성이 많아지면 최선을 다하지 않을 수도 있기 때문에 골 점수가 서로 종속이라는 가정은 일반적으로 타당하다고 할 수 있는데, 예를 들면 λ_1과 λ_2는 각각 홈팀과 원정팀의 순수한 능력에 의한 골 점수이며, λ_3는 관중의 수, 날씨, 운동장 조건과 같은 랜덤한 게임조건으로 간주할 수 있다.

2.2. 포아송 로그 선형모형

포아송 로그 선형모형 (Poisson log-linear model)은 로그연결함수를 사용하고 종속변수에 포아송분포를 가정하는 일반화선형모형이다. 본 연구에서 고려된 포아송 로그 선형모형은 이변량 포아송분포를 기초로 하는 R의 bivpois 패키지 (Karlis와 Ntzoufras, 2003)가 취급하는 식 (2.2)와 같은 모형이다. 사용된 기호의 정의는 n은 게임 수, h_i와 g_i는 i번째 게임에 해당되는 홈팀과 원정팀, X_i와 Y_i는 게임 i에서의 h_i와 g_i의 득점, λ_{1i}와 λ_{2i}는 h_i와 g_i의 득점의 기댓값, μ는 상수, h는 홈경기 효과모수, a_k와 d_k는 각각 팀 k의 공격모수와 수비모수로 약술할 수 있다.

$$\begin{aligned}(X_i, Y_i) &\sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),\ i = 1, 2, \cdots, n \\ \log(\lambda_{1i}) &= \mu + h + a_{h_i} + d_{g_i},\ \ \log(\lambda_{2i}) = \mu + a_{g_i} + d_{h_i}\end{aligned} \tag{2.2}$$

모수들의 추정가능성을 위해서 팀의 개수 p에 대하여 식 (2.3)과 같은 제약식을 사용하였는데 이 경우 공격모수와 수비모수는 전체 팀들의 공격능력과 수비능력의 차이를 의미함으로 해석을 수월하게 할 수 있다.

$$\sum_{i=1}^{p} a_i = \sum_{k=1}^{p} d_k = 0 \tag{2.3}$$

한편 당일 게임조건과 같은 두 팀에 공통적으로 적용되는 랜덤 효과로 해석될 수 있는 공분산 모수 λ_{3i}에 대해서는 식 (2.4)와 같은 구조로 정의하였다.

$$\log(\lambda_{3i}) = \gamma_0 + \gamma_1 \beta_{h_i}^{home} + \gamma_2 \beta_{g_i}^{away} \tag{2.4}$$

여기서 γ_0는 상수이며, $\beta_{h_i}^{home}$와 $\beta_{g_i}^{away}$는 각각 홈팀과 원정팀에 종속되는 모수이다. 그리고 γ_1과 γ_2는 0 또는 1의 값을 갖는 지시변수인데, 식 (2.4)에 대입하여 보면 $(\gamma_1, \gamma_2) = (0,0)$일 땐 상수인 공분산, $(\gamma_1, \gamma_2) = (1,0)$일 땐 공분산은 홈 팀에 종속되는 경우, $(\gamma_1, \gamma_2) = (0,1)$일 땐 공분산은 원정 팀에 각각 종속되는 경우를 각각 나타낸다.

본 연구에서 고려된 모형은 R의 bivpois 패키지에서 적합시킬 수 있는 모형-A부터 모형-L까지 모두 12개로 모형-A는 홈팀의 득점과 실점이 서로 독립인 이중 포아송분포를 따르는 경우를 의미하며, 모형-B부터 모형-F까지는 공변량의 효과가 있는 이변량 포아송분포를 따르는 경우인데, 모형-B는 λ_3가 $\gamma_1 = \gamma_2 = 0$인 상수인 경우, 모형-C는 λ_3가 $\gamma_1 = 1$, $\gamma_2 = 0$인 홈팀의 효과에만 종속되는 경우, 모형-D는 λ_3가 $\gamma_1 = 0$, $\gamma_2 = 1$인 원정팀의 효과에만 종속되는 경우, 모형-E는 λ_3가 $\gamma_1 = \gamma_2 = 1$인 홈팀과 원정팀 모두에 종속되는 경우, 모형-F는 λ_3가 상수인 영과잉 포아송분포를 따르는 경우를 각각 언급한다. 한편 축구경기와 같이 득점과 실점이 많지 않은 스포츠에서는 무승부 게임이 많기 때문에 동점발생 확률을 고려한 이변량 포아송 대각확대 모형을 고려할 필요가 있다 (Karlis와 Ntzoufras, 2003). 이와 같은 이유로 자주 발생하는 동점 (tie)을 고려하기 위한 모형-G부터 모형-K까지는 모두 식 (2.5)와 같은 확률분포를 갖는 대각확대 이변량 포아송모형 (diagonal inflated bivariate Poisson model)이다. 식 (2.5)의 혼합비율 (mixing proportion) p의 값이 클수록 동점이 될 가능성이 큼을 의미한다.

$$f_{IBP}(x,y) = \begin{cases} (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3), & x \neq y \\ (1-p) f_{BP}(x,y|\lambda_1,\lambda_2,\lambda_3) + p f_D(x|\boldsymbol{\theta}), & x = y \end{cases} \tag{2.5}$$

식 (2.5)에서 $f_D(x|\boldsymbol{\theta})$는 모수벡터 $\boldsymbol{\theta}$를 갖는 이산분포인데, $p = 0$이면 단순이변량 포아송분포가 된다. 분포 $f_D(x|\boldsymbol{\theta})$는 보통 기하분포, 간단한 이산분포 또는 포아송분포를 많이 이용하는데, 기하분포는 축구 득점의 패턴과 유사하며 간단한 이산분포로는 $x = 0, 1, 2, \cdots, S$에 대하여 $P(X = x) = \theta_x$, $\sum_{x=0}^{S} \theta_x = 1$을 고려하였는데, $S = 0$인 경우가 영과잉 포아송분포이다. 모형-G부터 모형-K까지는 대각확대 이변량 포아송분포를 이용한 모형으로 모형-G, 모형-H, 모형-I, 모형-J, 모형-K는 추가분포로 각각 기하분포, $S = 1$, $S = 2$, $S = 3$인 이산분포, 포아송분포를 고려한 모형을 나타내며, 모형-L은 추가분포로 포아송분포를 고려한 대각확대 이중 포아송모형을 나타낸다.

3. K-리그 분석 결과

3.1. 사용된 데이터

본 연구에서 사용된 데이터는 한국 프로축구리그인 K-리그 결과만을 포함하였으며, 챔피온스 리그는 제외하였다. 데이터는 1983년부터 2012년까지의 총 4218 경기의 결과로 모든 경기는 연장전을 제외한 정규시간만의 결과이며, 모든 자료의 출처는 K-리그 공식사이트 (http://www.kleague.com)이다.

K-리그 경기결과를 간단하게 약술하면, 1983년부터 2012년까지 30년간의 모든 경기에 대하여 홈팀이 승리한 경우가 1669번 (39.6%), 무승부가 1249번 (29.6%), 원정팀이 승리한 경우가 1300번 (30.8%)로 8.8% 정도 홈팀이 승리를 많이 하였다. 이 사실로부터 홈팀의 이점이 실제적으로 약간 있음을 알 수 있으며, 또한 골 득점차가 0인 경우가 29.6%, 1인 경우가 43.5%, 2인 경우가 18.2%로 세 가지 경우가 전체의 91.3%로 나타나 팀들 간의 전력이 크지 않음을 알 수 있다. 그리고 0-0 게임이 10.6%로 게임 결과 중 가장 빈도수가 많았다. 따라서 무승부 경기가 많은 관계로 영과잉 모형이나 대각확대 모형을 고려해 보는 것도 필요하다고 생각되어진다.

3.2. 모형의 적합도

고려된 12개 모형의 모수들은 RStudio (Ver. 0.97.320)와 bivpois 패키지를 사용하여 추정하였다. 바람직한 모형의 선택을 위하여 사용한 판정기준은 값이 클수록 적절한 로그우도함수 (log-likelihood function), 적은 것을 선호하는 모형에 포함된 모수의 개수, 값이 작은 것을 선호하는 Akaike의 정보기준 AIC와 베이지안 정보기준 BIC와 같은 모두 4가지이다.

Table 3.1 Results of the fitted models for the K-league data

Year	Model	number of parameters	loglikelihood	AIC	BIC
total	Model-A	56*	-11890.04	23892.09	24286.34
	Model-B	57	-11826.68	23767.36*	24168.66*
	Model-C	84	-11815.81	23799.63	24391.01
	Model-D	84	-11816.56	23801.12	24392.50
	Model-E	111	-11800.44*	23822.89	24604.36
	Model-F	58	-11826.68	23769.36	24177.70
	Model-G	59	-11826.68	23771.36	24186.74
	Model-H	59	-11826.67	23771.34	24186.72
	Model-I	60	-11826.67	23773.35	24195.76
	Model-J	61	-11826.67	23775.35	24204.80
	Model-K	59	-11826.68	23771.36	24186.74
	Model-L	58	-11872.00	23860.00	24268.34
2012	Model-A	32*	-985.0999	2034.200	2180.017
	Model-B	33	-981.4139	2028.828*	2179.201*
	Model-C	48	-974.3435*	2044.687	2263.412
	Model-D	48	-974.8681	2045.736	2264.462
	Model-E	63	-963.8021	2053.604	2340.681
	Model-F	34	-981.4140	2030.828	2185.758
	Model-G	35	-981.4140	2032.828	2192.315
	Model-H	35	-981.4140	2032.828	2192.315
	Model-I	36	-981.3520	2034.704	2198.748
	Model-J	37	-981.1883	2036.377	2204.977
	Model-K	35	-981.4140	2032.828	2192.315
	Model-L	34	-984.5685	2037.137	2192.067
2011	Model-A	32*	-687.1875	1438.375	1572.726
	Model-B	33	-680.3109	1426.622*	1565.172*
	Model-C	48	-672.1555	1440.311	1641.838
	Model-D	48	-674.8345	1445.669	1647.196
	Model-E	63	-663.2657*	1452.531	1717.036
	Model-F	34	-680.1118	1428.224	1570.972
	Model-G	35	-680.1458	1430.292	1577.238
	Model-H	35	-680.1119	1430.224	1577.170
	Model-I	36	-680.0012	1432.002	1583.148
	Model-J	37	-680.0012	1434.002	1589.346
	Model-K	35	-680.1119	1430.224	1577.170
	Model-L	34	-684.8492	1437.698	1580.447

Table 3.1 은 적합된 12가지 모형에 대한 모형판정기준 값들을 보여준다. 고려된 경우는 3가지로 전체, 2011년, 2012년 데이터의 결과를 각각 보여준다. 각 판정기준의 경우에 최상의 모형을 표시한 *를 살펴보면 4개의 판정기준이 모두 동일한 모형은 없지만 AIC와 BIC 판정기준은 모두 모형-B가 최고임을 보여주며, 모수의 계수도 거의 최소 수준임을 알 수 있는데, 이 사실은 공분산은 있지만 홈팀과 원정팀의 영향을 받지 않는 상수인 경우를 의미하며 영과잉 모형이나 대각확대 모형은 생각보다 모형의 적합도를 높이지 못했다. 대각확대 이변량 포아송모형의 경우인 모형-G부터 모형-K까지는 AIC와 BIC 판정기준에 의하면 근소한 차이지만 식 (2.5)에서 추가분포로 $S = 1$인 이산분포를 고려한 경우가 바람직하다. Table 3.2는 2012년 K-리그 자료와 모형-B를 이용하여 추정한 모수 추정값을 보여주는데, 식 (2.2)과 비교설명하면 μ의 추정값이 -0.618, h의 추정값이 0.205이 된다. 또한 식 (2.4)에서의 γ_0의 추정값은 -1.665이다. Table 3.3은 모형-H의 모수 추정값으로 Table 3.2와 비교하면 공통모수들은 거의 비슷한 값으로 추정되며, 혼합비율 p와 θ_1는 2.2절에서 정의된 모수들이다.

Table 3.4는 모형-B, 식 (2.2)와 식 (2.3)을 이용하여 추정한 팀별 모수 추정값인데 공격능력 (attack)은 큰 값이 우수하며, 수비능력 (defense)은 작은 값이 우수한 것을 의미한다. 또한 공격능력과 수비능력의 합은 각각 0이며, 상대적으로 공격능력이 가장 강한 팀은 전북, 수비능력이 가장 뛰어난 팀은 인천으로 나타났다. 승점 (point)이 높은 팀들은 모두 상대적으로 공격 및 수비능력이 우수한 사실을 확인할 수 있다.

Table 3.2 Estimates of parameters of model-B for 2012 K-league data

Parameter	estimates
Intercept for $\log(\lambda_1)$	-0.413
Intercept for $\log(\lambda_2)$	-0.618
Intercept for $\log(\lambda_3)$	-1.665
Home effect	0.205

Table 3.3 Estimates of parameters of model-H for 2012 K-league data

Parameter	estimates
Intercept for $\log(\lambda_1)$	-0.414
Intercept for $\log(\lambda_2)$	-0.619
Intercept for $\log(\lambda_3)$	-1.661
Home effect	0.205
Mixing proportion p	6.758781e-06
θ_1	3.150834e-07

Table 3.5는 2012년 K-리그 전체 경기의 득점 빈도수와 괄호 안에 표기된 추정된 모형-B를 이용하여 구한 기대빈도수를 나타내는데, 전반적으로 기대빈도와 실제빈도는 비슷하나 홈팀이 3-0으로 이긴 경우는 7회, 원정팀이 3-0으로 이긴 경우는 15회가 되는 등과 같은 비상식적 결과는 오차가 크게 나타났다. 실제로 2012년 홈팀과 원정팀의 경기결과는 홈팀 승:무승부:원정팀 승의 비율이 148:93:111로 홈팀이 승리하는 빈도가 훨씬 더 많았다.

3.3. 승점의 추정

K-리그의 각 팀 성적들은 승률이 아니라 승점이다. 현재 승점은 승리에 3점, 무승부에 1점을 주는 승점제도를 사용하고 있는 데, 승점을 많이 획득하기 위하여 Table 3.4의 공격능력과 수비능력의 균형이 상대적으로 어떻게 나타나는 지는 매우 큰 관심사가 될 수 있다. 따라서 팀의 공격능력 (X_1), 수비능력 (X_2)을 독립변수로 사용하여 게임당 승점 (Y)을 회귀분석을 통하여 추정하였는데, 데이터는 최근 경기결과인 2011년 16팀의 결과 246개를 이용하여 추정한 X_1과 X_2각 16개, 2012년 16팀의 결과 352개를

Table 3.4 Estimated team parameters of model-B for 2012 K-league data

Number	Team Name	Attack	Defense	Point
01	BUSAN	-0.329	-0.192	53
02	CHUNNAM	-0.209	0.160	53
03	DAEGU	-0.016	0.107	61
04	DAEJEON	-0.197	0.303	50
05	GANGWON	0.039	0.326	49
06	GWANGJU	0.011	0.287	45
07	GYEONGNAM	-0.013	0.038	50
08	INCHEON	-0.317	-0.383	67
09	JEJU	0.353	-0.044	63
10	JEONBUK	0.493	-0.209	79
11	POHANG	0.373	-0.225	77
12	SANGJU	-0.751	0.386	27
13	SEONGNAM	-0.202	0.092	52
14	SEOUL	0.431	-0.344	96
15	SUWON	0.219	-0.111	73
16	ULSAN	0.116	-0.191	68

Table 3.5 Expected frequencies using bivariate Poisson distribution for 2012 K-league data

		Away Goals						
		0	1	2	3	4	5	total
HomeGoals	0	36 (29.34)	31 (27.03)	20 (14.08)	15 (5.47)	1 (1.76)	0 (0.49)	103 (78.17)
	1	35 (36.89)	35 (38.02)	27 (21.28)	2 (8.66)	1 (2.87)	1 (0.82)	101 (108.54)
	2	28 (26.35)	34 (29.08)	17 (17.15)	9 (7.24)	0 (2.46)	0 (0.72)	88 (83.00)
	3	7 (14.12)	16 (16.26)	11 (9.92)	5 (4.30)	4 (1.49)	0 (0.44)	43 (46.53)
	4	0 (6.33)	4 (7.47)	3 (4.65)	0 (2.05)	0 (0.71)	0 (0.21)	7 (21.42)
	5	2 (2.50)	1 (2.99)	2 (1.88)	2 (0.83)	0 (0.29)	0 (0.08)	7 (8.57)
	6	3 (0.90)	0 (1.08)	0 (0.68)	0 (0.30)	0 (0.10)	0 (0.03)	3 (3.09)
total		111 (116.43)	121 (121.93)	80 (69.64)	33 (28.85)	6 (9.68)	1 (2.79)	352 (349.32)

이용하여 추정한 X_1과 X_2각 16개, 합계 32개의 자료를 이용하였으며 수비능력의 값이 크면 수비를 더 잘 하는 것으로 나타내기 위해서 X_2는 원래 값에 -1을 곱한 값을 이용하였다.

Table 3.6 Model summary for regression analysis

R Square	Adjusted R Square	Durbin-Watson
0.907	0.901	1.934

Table 3.7 Estimated regression model coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
	B	Std. Error	Beta			
(Constant)	1.355	.021		63.161	.000	
attack	.586	.056	.632	10.497	.000	1.133
defense	.803	.091	.528	8.781	.000	1.133

회귀분석 결과를 설명하면 분산분석의 p값은 p <0.001로 회귀직선은 유의수준 1%에서 매우 유의한 것으로 나타났다. 또한 Table 3.6에서 알 수 있듯이 결정계수가 90.7%로 높게 나타났으며 더빈-왓슨 통계량의 값은 1.934로 오차는 자기상관의 영향이 거의 없음을 알 수 있다. Table 3.7은 추정된 회귀식, 표준화 회귀계수 및 VIF를 보여주는데, VIF의 값이 1.133으로 다중공선성의 문제는 없는 것으로 나타났다. 따라서 추정된 회귀식은 다음과 같이 기술할 수 있으며 표준화 회귀계수를 이용하면 승점을 취득

하는데 약 1.2배 정도 공격능력이 수비능력보다 기여하는 바가 큰 것으로 나타났다.

$$\hat{Y} = 1.355 + 0.586X_1 + 0.803X_2$$

Figure 3.1은 게임당 승점과 추정된 게임당 승점의 관계를 보여주는 산점도이다. 대체적으로 실제 게임당 승점을 추정된 게임당 승점이 잘 예측하고 있으나 게임당 승점의 최대 및 최솟값에서 잔차가 커짐을 알 수 있다.

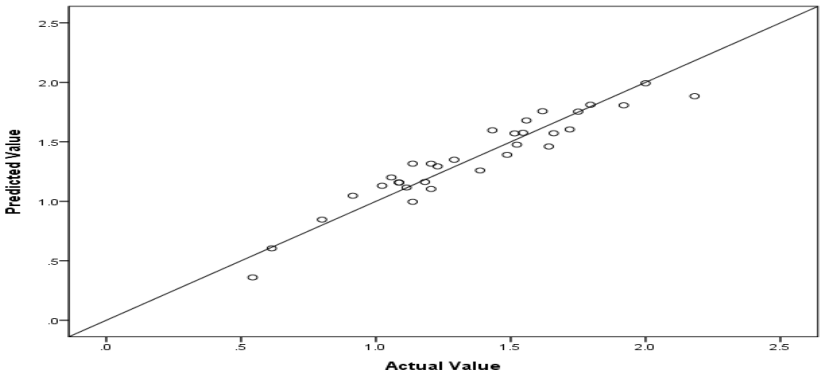

Figure 3.1 Scatterplot of predicted vs. actual values

4. 결론

우리가 일상생활에서 접하는 통계수치 중의 하나인 축구경기 결과를 통계 모형화 하는 것은 매우 의미있는 일이다. 본 연구에서는 포아송분포를 축구경기에 적용시킬 수 있다는 고전적인 상식의 연장선에서 2012년까지의 K-리그 전 경기 데이터를 각 팀의 득점은 상대방 팀의 공격을 유도하고 수비를 강화할 가능성이 크기 때문에 현실적으로 존재하는 득점과 실점의 상관관계를 허용하며 축구에서 빈번하게 발생되는 동점까지 처리할 수 있는 이변량 포아송분포에 적합시켰다. 그 결과 K-리그에서는 공분산이 일정한 이변량 포아송분포가 가장 적당하다고 나타났으며 외국의 프로축구 장기리그인 경우에 흔히 필요성이 대두되는 대각확대 모형 등은 특별하게 필요하지 않은 것으로 나타났다.

추정된 모형은 K-리그에서 두 팀 간의 승부를 예측할 수 있으며, 공격능력과 수비능력의 효과도 확인할 수 있는 데, 승점을 많이 취득하는 데는 공격능력이 수비능력보다 20% 정도 더 중요하다고 나타났다. 향후 연구관제로 본 연구에서 다루어지지 못한 여러 가지 다양한 축구모형들을 K-리그에 적용하여 보고, 또한 챔피온스 리그와 같은 단기리그인 경우를 본 연구의 결과와 비교해보는 일은 매우 흥미로운 일이라고 간주된다.

References

Choi, S. B., Kang, C. W., Cho, H. J. and Kang, B. Y. (2011). Social network analysis for a soccer game. *Journal of the Korean Data & Information Science Society*, **22**, 1053-1063.

Gemert, D. (2010). *Modelling the scores of premier league football matches*, Master's Thesis, University of Amsterdam, Amsterdam, Netherlands.

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

Hong, C. S., Jung, M. S. and Lee, J. H. (2010). Prediction model analysis of 2010 South Africa World Cup. *Journal of the Korean Data & Information Science Society*, **21**, 1137-1146.

Karlis, D. and Ntzoufras, I. (2000). On modelling soccer data. *Student*, **3**, 2 29-245.

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

Kim, Y. J. (2012). Statistical analysis of K-league data using Poisson model. *The Korean Journal of Applied Statistics*, **25**, 775-783.

Lee, A. J. (1997). Modeling scores in the premier league: Is Manchester United really the best? *Chance*, **10**, 15-19.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109-118.

Moroney M. J. (1956). *Facts from figures*, 3rd edition, Penguin, London.

Reep C., Pollard R. and Benjamin B. (1971). Skill and chance in ball games. *Journal of the Royal Statistical Society A*, **134**, 623-629.

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.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1221–1229

http://dx.doi.org/10.7465/jkdi.2014.25.6.1221
한국데이터정보과학회지

Prediction of K-league soccer scores using bivariate Poisson distributions

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 25 July 2014, revised 17 August 2014, accepted 1 September 2014

Abstract

In this paper we choose the best model among several bivariate Poisson models on Korean soccer data. The models considered allow for correlation between the number of goals of two competing teams. We use an R package called bivpois for bivariate Poisson regression models and the data of K-league for season 1983-2012. Finally we conclude that the best fitted model supported by the AIC and BIC is the bivariate Poisson model with constant covariance. The zero and diagonal inflated models did not improve the model fit. The model can be used to examine home-away effect, goodness of fit, attack and defense parameters.

Keywords: Bivariate Poisson distribution, K-league, zero and diagonal inflated models.

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 448-701, Korea.
E-mail: jtlee@dankook.ac.kr