응용통계연구 (2011)
24(5), 915–930
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.915

한국 프로야구 경기결과에 관한 통계적 연구

최영근[1] · 김형문[2]

[1]건국대학교 응용통계학과, [2]건국대학교 응용통계학과

(2010년 10월 접수, 2011년 6월 채택)

요 약

경기의 결과를 모형 화하는 것은 다양한 방법을 통하여 이루어져 왔다. 특히 두 개의 팀만이 경기를 하는 경우에는 더욱 다양한 방법이 제안되었다. 그 중에서 Bradley-Terry 모형은 짝지어진 자료로부터 선호하는 크기의 특성을 얻을 수 있는 가장 넓게 사용되어지고 있는 모형이다. Bradley-Terry 모형은 스포츠 경기결과의 분석과 심리학에 관련된 분야들에서 다양하게 적용되어진다. 본 연구자는 한국 프로야구 자료에 Bradley-Terry 모형을 적용하였다. 그 결과 연속형 공변량의 경우 평균자책점과 세이브를 포함하는 모형이 최적으로 나타났고 관심의 대상이 되는 몇 가지 범주형 분석의 결과 동군과 서군, 골든글러브, 다승왕, 그리고 홈경기의 이점이 승부에 영향을 주는 것으로 나타났다. 하지만 이들의 조합은 단순히 연속형 변수만을 포함한 모형이 분석 결과 더 적절한 것으로 고려되어졌다. 제안된 모형은 경기의 승패를 예측하는 데도 유용하게 사용될 수 있다. 한 예로 한국시리즈에서 우승할 확률들의 순서가 2008년도의 결과와 정확히 일치하였다.

주요용어: Bradley-Terry 모형, Loglinear Bradley-Terry 모형, 한국프로야구.

1. 서론

1905년 우리나라에 야구가 도입된 이래 1982년에 프로야구가 창설되고 6개의 팀으로 한국 프로야구는 시작되었다. 최근에는 국제대회에서 뛰어난 성과를 내고 있고 또한 관중 수에 있어서도 국내의 프로리그 중에서는 가장 인기 있는 것이 프로야구이다. 하지만 국내프로야구에 대한 연구에 있어서는 프로야구 성적이나 경기력 향상에 대한 것에 비해서 마케팅관련 연구가 더 많은 현실이다. 그리고 한국 프로야구에 대한 기존의 연구를 보면 신상근 등 (2007)은 한 팀을 기준으로 분석을 하여 승리 요인에 대하여 연구를 하였고, 이장택과 조현식 (2009)은 로지스틱 회귀모형을 활용하여 한국프로야구에 대하여 연구를 하였다.

그래서 본 연구에서는 한국 프로야구에 대한 연구 방법을 기존에 사용하였던 모형과는 다른 모형인 Bradley-Terry 모형을 사용하여 프로야구에 대한 연구를 하려한다. 본 연구에서 사용하는 모형은 스포츠 경기나 심리학에서 많이 적용되는 모형으로 야구와 같이 단 두 개의 팀이 경기를 치러 승패를 가리는 경기에서 사용되어질 수 있는 모형이다.

Bradley-Terry(BT) 모형은 최초 Bradley와 Terry (1952)에 의해서 제안되었고 Davison (1970)에 의하여 결정하지 못하는 영역(무승부)을 포함하는 모형으로 확장되었다. 이후 많은 연구자들에 의해 기본적인 Bradley-Terry 모형이 확장되어졌다. 특히 Fienberg와 Larntz (1976)는 기본적인 BT모형을

이 논문은 2009학년도 건국대학교의 지원에 의하여 연구되었음.
[2]교신저자: (143-701) 서울시 광진구 화양동 1번지, 건국대학교 상경대학 응용통계학과, 교수.
E-mail: hmkim@konkuk.ac.kr

로그선형 모형으로 표현하였고 또한 모수의 추정에서 두 모형에서의 방법들이 동일함을 또한 보여주었다. Bradley-Terry 모형을 포함한 짝지어진 비교방법에 대한 포괄적인 개론은 Agresti (2002)나 David (1988)를 참조하면 된다.

2절에서는 연구에 사용된 모형에 대하여 소개 하고, 경기에서 승리를 예측하는데 있어서 승패이외에 추가적인 요인을 첨가한 모형에 대하여 설명하였다. 3절에서는 모형을 통한 분석으로 한국 프로야구의 경기결과에 추가적인 기록으로 승리에 대한 요인들과 모형에 대한 검정결과와 함께 각 팀별로 승리를 할 수 있는 확률이 어느 정도 차이가 있는지 표를 통하여 나타내었다. 연속형 공변량의 경우 평균자책점과 세이브를 포함한 모형이 최적의 모형으로 판단되어졌으며 관심의 대상이 되는 몇 가지 범주형 분류에 의한 분석의 결과 동군과 서군, 골든글러브, 다승왕, 그리고 홈경기의 이점이 승부에 영향을 미쳤다. 마지막으로 4절에서는 연구를 통한 한국 프로야구의 승리에 대한 요인을 확인하고, 연구를 함에 있어서 어려운 점과 앞으로 연구에 있어서 추가적으로 필요한 부분에 대하여 논의한다.

2. Bradley-Terry 모형

2.1. 기본적인 Bradley-Terry 모형

경기에서 두 팀에 대한 비교를 위해서 많은 방법들이 사용되어져왔다. 이러한 모형들 중 잘 알려져 있는 방법이 Bradley-Terry(BT) 모형이다. 기본적인 브래들리-테리 모형 (Bradley와 Terry, 1952)은 아래와 같다.

객체 O를 비교함에 있어 비교대상이 되는 j를 O_j로 표시하고, 또 다른 비교대상인 k를 O_k로 표시한다. 본 논문에서 각 객체는 프로야구팀을 의미한다. 그리고 먼저 비교 대상인 j (O_j)가 k (O_k)보다 더 선호되는 경우에 식 (2.1)과 같이 나타낼 수 있다. 여기에서 선호의 의미는 j팀과 k팀이 경기를 하였을 시 j팀이 승리하였음을 나타낸다.

$$p\left(O_j \succ O_k | \pi_j, \pi_k\right) = p_{(jk)j} = \frac{\pi_j}{\pi_j + \pi_k}, \quad \text{모두 } \ j \neq k. \tag{2.1}$$

식 (2.1)에서 π_j와 π_k는 음이 아닌 모수이고, 팀별 비교에 있어서는 해당 팀의 힘(strengths)이나 능력(abilities)을 나타내며, $p_{(jk)j}+p_{(jk)k} = 1$이다. J개의 팀이 있는 경우 모든 가능한 비교경우는 $\binom{J}{2} = J(J-1)/2$의 경우가 된다.

식 (2.1)의 모형은 Glickman (1993)의 방법으로 유도를 할 수 있다. 위 모형을 유도하기위해 아래를 가정한다. j팀이 경기를 할 때 그 팀의 점수 S_j가 생성된다. 이 점수는 관측이 불가능하며 상대팀의 점수 S_k와는 독립이다. 관측된 변수는 이 경기의 결과인데 이는 두 점수 중 큰 점수에 의해 결정되어진다. 점수 S_j는 위치모수 $\log \pi_j$를 갖는 극단치분포를 따르게 된다. 따라서 S_j의 누적분포함수인 $F_j(s)$는 식 (2.2)의 형태를 이루며,

$$F_j(s) = \exp\left(-e^{-(s-\log \pi_j)}\right) \tag{2.2}$$

점수 S_j와 S_k의 차이, 즉 $S_j - S_k$는 평균이 $\log \pi_j - \log \pi_k$인 로지스틱 분포를 따르며 이는 식 (2.3)과 같다.

$$S_j - S_k \sim F_j(s) = \frac{1}{1 + e^{-\left(s-\left(\log \pi_j - \log \pi_k\right)\right)}}. \tag{2.3}$$

따라서 누적분포의 성질을 이용하여 아래의 식 (2.4)를 유도할 수 있으며 이는 원래의 식 (2.1)과 동일하

다.

$$\Pr(S_j > S_k) = P(S_j - S_k > 0) = 1 - \frac{1}{1+e^{-\left(0-\left(\log\pi_j - \log\pi_k\right)\right)}}$$
$$= \frac{e^{\left(\log\pi_j - \log\pi_k\right)}}{1+e^{\left(\log\pi_j - \log\pi_k\right)}} = \frac{\pi_j/\pi_k}{1+\pi_j/\pi_k} = \frac{\pi_j}{\pi_j+\pi_k}. \quad (2.4)$$

식 (2.1)의 모수들이 식별가능하려면 $\pi_J = 1$과 같은 제약식이 필요하다. 모형은 $\binom{J}{2}$개의 확률들을 $J-1$개의 모수로서 나타내므로 자유도가 $\binom{J}{2} - (J-1)$이 된다. 로짓 모형에 적용할 때 확률변수 $Y_{(jk)j}$ (j를 선호하는 횟수)와 $Y_{(jk)k}$ (k를 선호하는 횟수)에 대한 가정은 모수 $n_{(jk)} = Y_{(jk)j} + Y_{(jk)k}$와 $p_{(jk)j}$ 또는 $p_{(jk)k}$를 가지는 독립인 이항변수이다. 각각의 $\ln(p_{(jk)j}/p_{(jk)k}) = \beta_j - \beta_k$에 대하여 β의 계수에 해당하는 J개의 설명변수를 만들어야 한다. 즉, 객체 J와 k를 비교할 때 β_j에 대한 변수는 1이고 β_k에 대한 변수는 -1이며 모든 다른 변수들은 0이다. 이러한 방법으로 계획행렬을 만들 수 있으며 이에 대한 분석은 로지스틱회귀에 관한 기본적인 소프트웨어를 사용하면 된다 (Agresti, 2002).

2.2. 로그선형 브래들리-테리 모형

2.2.1. 두 가지 반응 영역 BT모형은 로그선형 모형으로 표현될 수 있다 (Sinclair, 1982; Agresti, 2002; Dittrich 등, 1998). Fienberg와 Larntz (1976)는 기본적인 BT모형을(이항분포에 기초) 로그선형 모형으로(포아송분포에 기초) 표현하였고 로그선형 모형으로의 표현의 몇 가지 장점들을 제시하였다. 이는 기본적인 BT모형을 다변량으로 쉽게 확장할 수 있다는 것이다.

식 (2.1)의 확률은 아래와 같이 나타내어 질 수 있다 (Sinclair, 1982).

$$p_{(jk)j} = \frac{\pi_j}{\pi_j+\pi_k} = \frac{(\pi_j/\pi_k)^{\frac{1}{2}}}{(\pi_j/\pi_k)^{\frac{1}{2}} + (\pi_k/\pi_j)^{\frac{1}{2}}}. \quad (2.5)$$

비슷한 형태로 k팀이 j팀에 비해 선호되는 확률은 아래와 같다.

$$p_{(jk)k} = \frac{\pi_k}{\pi_j+\pi_k} = \frac{(\pi_k/\pi_j)^{\frac{1}{2}}}{(\pi_j/\pi_k)^{\frac{1}{2}} + (\pi_k/\pi_j)^{\frac{1}{2}}}.$$

식 (2.5)를 이용하여 기본적인 BT모형은 일반화선형모형, 즉 로그선형모형으로 나타내어 질 수 있다. 확률변수들 $Y_{(jk)j}$와 $Y_{(jk)k}$는 각각 독립인 포아송분포를 따른다고 가정한다. $n_{(jk)}$를 j와 k를 비교하는 횟수로 하고, $Y_{(jk)j}$는 j를 선호하는 횟수로 $Y_{(jk)k}$는 k를 선호하는 횟수로 정의한다. 총비교횟수 $n_{(jk)} = Y_{(jk)j} + Y_{(jk)k}$가 고정되고 각 팀의 경기결과가 독립이라면 확률변수 $Y_{(jk)j}$는 모수가 $n_{(jk)}$와 식 (2.5)의 확률을 가지는 이항분포를 따른다. 따라서 $Y_{(jk)j}$의 기댓값을 $m_{(jk)j}$로 표현하면 이는 $n_{(jk)}p_{(jk)j}$로 구할 수 있으며 이는

$$\ln m_{(jk)j} = \mu_{(jk)} + \lambda_j^O - \lambda_k^O,$$

여기에서 $\mu_{(jk)} = \ln n_{(jk)} - \ln[(\pi_j/\pi_k)^{1/2} + (\pi_k/\pi_j)^{1/2}]$, $\lambda_j^O = \ln\pi_j/2$ 그리고 $\lambda_k^O = \ln\pi_k/2$. 따라서 로그선형 브래들리-테리 모형(Loglinear Bradley-Terry Model; LLBT)은 아래의 식 (2.6)이 된다.

$$\ln m_{(jk)j} = \mu_{(jk)} + \lambda_j^O - \lambda_k^O,$$
$$\ln m_{(jk)k} = \mu_{(jk)} - \lambda_j^O + \lambda_k^O. \quad (2.6)$$

표 **2.1.** 단순 LLBT의 계획 구조

비교	결정	횟수	μ	λ_1^O	λ_2^O	λ_3^O
(12)	O_1	$y_{(12)1}$	1	1	−1	0
(12)	O_2	$y_{(12)2}$	1	−1	1	0
(13)	O_1	$y_{(13)1}$	2	1	0	−1
(13)	O_3	$y_{(13)3}$	2	−1	0	1
(23)	O_2	$y_{(23)2}$	3	0	1	−1
(23)	O_3	$y_{(23)3}$	3	0	−1	1

식 (2.6)에서 μ는 장애모수이고 객체들의 각각의 비교에서 포함된 객체들을 나타내는 상호작용모수로 해석되어질 수 있다. 이모형은 각각의 비교 (jk)에 두 가지 가능한 결과로 제한되어 있다. 객체 j를 선호하고 객체 k를 선호하지 않는 것 또는 객체 k를 선호하고 객체 j를 선호하지 않는 것이다. 따라서 이 모형은 $\binom{J}{2} \times J$의 불완비 이차원 (짝지어진 객체 × 반응영역) 분할표에 대한 로그선형모형으로 해석되어질 수 있다.

표 2.1은 식 (2.6)의 계획 구조를 보여주는 표로 비교대상이 3개가 있는 경우를 나타낸 것이다. 표의 열의 원소들은 횟수, 요인인 μ (3개의 수준을 가짐), λ_1, λ_2, 그리고 λ_3에 대한 변수들 O_1, O_2, 그리고 O_3가 있다. 표 2.1에서 비교라는 열에서 (12)는 객체 O_1과 객체 O_2를 비교하는 것이며 나머지는 비슷하게 해석되어진다.

이하 도출되는 모든 로그선형모형은 일반화선형모형의 특수한 경우이므로 모수추정, 통계적 추론은 몇 가지 표준적인 통계프로그램으로 쉽게 분석가능하다. 예를 들면 포아송 오차와 로그링크를 사용하여 GLIM (Francis 등, 1993)을 이용하면 되고 R(또는 Splus)에서는 glm 함수를 이용하며 SAS에서는 PROC GENMOD를 이용하면 된다.

2.2.2. 결정하지 못하는 영역이 포함된 세 가지 영역 비교를 하다보면 가끔 어느 한쪽을 선택할 수 없는 경우가 발생한다. Davidson과 Beaver (1977)는 이러한 경우에 LLBT를 아래와 같이 구하였다.

$$\begin{aligned} \ln m_{(jk)j} &= \mu_{(jk)} + \lambda_j^O - \lambda_k^O, \\ \ln m_{(jk)0} &= \mu_{(jk)} + \gamma, \\ \ln m_{(jk)k} &= \mu_{(jk)} - \lambda_j^O + \lambda_k^O. \end{aligned} \tag{2.7}$$

식 (2.7)에서 $m_{(jk)j}$는 객체 j를 선호하는 개수의 기댓값을 나타내고 $m_{(jk)k}$는 객체 k를 선호하는 개수의 기댓값을 나타내며, $m_{(jk)0}$는 (jk)의 비교에서 선택하지 않는 개수에 대한 기댓값을 표시한다. 그리고 γ는 결정되지 않는 영향이다. 영역이 지금과 같이 3가지의 경우에는 분석을 위한 자료를 만드는 경우에는 γ를 더미변수로 만들어서 분석을 하면 된다. 다음의 표 2.2는 표 2.1에서 γ를 더미변수로 만들어서 구축한 계획구조이다.

2.3. 확장된 모형의 특징과 설계

2.3.1. 로그선형 브래들리-테리 모형과 객체 공변량 LLBT의 확장된 모형은 식 (2.7)에서 선택되지 않는 영역이 포함된 비교 실험을 바탕으로 한다. 확장된 모형에서 고려해야 할 것은 선호하는 것에 대한 판단에 있어서 객체의 공변량의 영향을 포함한다는 것이다. 공변량은 비교가 되는 대상의 개별적인 특성을 말한다. 즉, 한 경기에서 각 팀만이 가지고 있는 특징이라고 할 수 있다.

표 **2.2.** 응답이 3개인 경우의 계획 구조

비교	결정	횟수	μ	γ	λ_1^O	λ_2^O	λ_3^O
(12)	O_1	$y_{(12)1}$	1	0	1	-1	0
(12)	no	$y_{(12)0}$	1	1	0	0	0
(12)	O_2	$y_{(12)2}$	1	0	-1	1	0
(13)	O_1	$y_{(13)1}$	2	0	1	0	-1
(13)	no	$y_{(13)0}$	2	1	0	0	0
(13)	O_3	$y_{(13)3}$	2	0	-1	0	1
(23)	O_2	$y_{(23)2}$	3	0	0	1	-1
(23)	no	$y_{(23)0}$	3	1	0	0	0
(23)	O_3	$y_{(23)3}$	3	0	0	-1	1

기본적인 아이디어는 객체들의 P개의 성질들을 나타내는 공변량 $X_1, \ldots, X_P$들의 선형결합으로 객체모수들을 나타내는 것이다 (Springall, 1973). 즉,

$$\lambda_j^O = \sum_{p=1}^{P} x_{jp}\beta_p^X. \tag{2.8}$$

식 (2.8)에서 x_{jp}는 객체 j의 p번째 특성을 기술하는 공변량을 나타낸 것이고 β_p^X는 x_{jp}에 대한 알려지지 않은 회귀 모수이다.

예를 들어 한 객체 공변량의 효과를 포함하는 LLBT 모형은 아래의 식 (2.9)와 같다.

$$\begin{aligned} \ln m_{(jk)j} &= \mu_{(jk)} + \lambda_j^O - \lambda_k^O \\ &= \mu_{(jk)} + \beta_1^X x_{j1} - \beta_1^X x_{k1} \\ &= \mu_{(jk)} + \beta_1^X (x_{j1} - x_{k1}). \end{aligned} \tag{2.9}$$

다른 방정식들도 비슷하게 정의되어진다. 이모형은 포아송 오차와 로그링크를 사용한 일반화 선형 모형(Generalized Linear Model)으로 적합 시킬 수 있다. 이때의 계획구조는 표 2.2와 비슷하며 모수 λ대신 객체 공변량 $x_{jp} - x_{kp}$의 값을 가지는 P개의 열벡터를 포함한다.

2.3.2. 홈경기의 이점(Home Advantage) 홈경기의 이점이 존재할 경우 비교 (jk)와 비교 (kj)는 다른 의미를 가진다. 즉, (jk)의 경우 j팀이 홈이며 (kj)는 k팀이 홈을 의미한다. 따라서 홈경기의 이점을 나타내는 추가적인 모수가 하나 더 필요하며 이를 포함하는 확장된 모형은 식 (2.10)와 같이 나타내어질 수 있다 (David, 1988).

$$\begin{aligned} \ln m_{(jk)j\cdot j} &= \mu_{(jk)j} + \lambda_j^O - \lambda_k^O + \delta, \\ \ln m_{(jk)k\cdot j} &= \mu_{(jk)j} - \lambda_j^O + \lambda_k^O, \\ \ln m_{(jk)j\cdot k} &= \mu_{(jk)k} + \lambda_j^O - \lambda_k^O + \delta, \\ \ln m_{(jk)k\cdot k} &= \mu_{(jk)k} - \lambda_j^O + \lambda_k^O, \end{aligned} \tag{2.10}$$

여기에서 δ는 홈경기의 이점을 나타내며 $m_{(jk)j\cdot j}$는 (jk)의 비교에서 j팀이 홈일 경우 j팀이 선호되는, 즉 승리하는 기댓값을 나타낸다. 아래 첨자 $(jk)j\cdot j$의 의미는 (jk)팀이 경기를 하는 경우 $(\cdot)$ 뒤의 첨자는 홈팀을 나타내고 $(\cdot)$ 앞의 첨자는 승리하는 팀을 나타낸다. 따라서 이표기는 객체의 순서가 명확

표 **2.3.** 2가지 반응영역과 홈경기의 이점 모형의 계획 구조

비교	결정	횟수	μ	δ	λ_1^O	λ_2^O	λ_3^O
(12)	O_1	$y_{(12)1}$	1	1	1	-1	0
(12)	O_2	$y_{(12)2}$	1	0	-1	1	0
$\vdots$	$\vdots$	$\vdots$			$\vdots$		
(21)	O_2	$y_{(21)2}$	4	1	-1	1	0
(21)	O_1	$y_{(21)1}$	4	0	1	-1	0
$\vdots$	$\vdots$	$\vdots$			$\vdots$		

한 경우, 즉 (jk)의 경우 j팀이 홈이며 (kj)는 k팀이 홈을 의미, 간략히 나타내어 질 수 있다. 예를 들면 $m_{(21)1.2}$는 $m_{(21)1}$와 동일하다. 다음의 표 2.3은 두 가지 반응영역이 있고 홈경기의 이점이 있는 모형의 계획구조이다.

무승부와 홈경기의 이점이 동시에 포함된 모형은 아래와 같이 표시되어질 수 있으며 계획구조는 표 2.2와 비슷하게 더미변수 γ을 추가하여 만들어주면 된다.

$$\begin{aligned}
\ln m_{(jk)j\cdot j} &= \mu_{(jk)j} + \lambda_j^O - \lambda_k^O + \delta,\\
\ln m_{(jk)o\cdot j} &= \mu_{(jk)j} + \gamma,\\
\ln m_{(jk)k\cdot j} &= \mu_{(jk)j} - \lambda_j^O + \lambda_k^O,\\
\ln m_{(jk)j\cdot k} &= \mu_{(jk)k} + \lambda_j^O - \lambda_k^O + \delta,\\
\ln m_{(jk)o\cdot k} &= \mu_{(jk)k} + \gamma,\\
\ln m_{(jk)k\cdot k} &= \mu_{(jk)k} - \lambda_j^O + \lambda_k^O.
\end{aligned} \tag{2.11}$$

3. 실증자료분석

3.1. 자료수집

본 연구를 수행하기 위해서 한국야구위원회에서 발간되는 2007, 2008, 2009년 프로야구연감에서 자료를 수집하였다. 최근 3년 페넌트레이스 자료를 통하여 각 팀별 경기 결과와 각 팀별 세부적인 자료를 수집하여 분석에 사용하였다. 한 팀당 한 해의 총 경기 수는 126경기이고, 한해의 전체 팀의 경기 수는 504경기였다. 그런데 2008년의 경우 경기 규칙에서 무승부 제도가 없이 오직 승패만 인정되는 경기 규칙이 있었기 때문에 2008년의 자료의 경우 무승부는 전혀 없다. 그리고 2008년에 창단된 히어로즈는 2006년과 2007년의 자료가 존재하지 않기 때문에 그 전신인 현대 유니콘스의 2006년과 2007년 자료를 히어로즈의 해당연도의 자료라고 가정하였다.

3.2. 기초자료 분석

각 팀별 페넌트레이스의 경기 결과는 표 3.1에서 나와 있듯이 3년의 경기 결과로는 단순히 승패와 승률을 비교할 수 있는 정도이다. 어느 팀이 어떠한 환경에서, 어느 조건을 가지고 있을 때 승률에 차이가 있는지 현재의 자료만을 가지고는 알 수 가 없는 것이다. 또한 어떤 경기 규칙을 가지고 있느냐에 따라 승률에 영향이 있는지 바로 알기는 어렵다. 팀들 중에서 '두산'만을 보더라도 알 수가 있다.

'두산'의 2007년과 2008년의 경기결과에서 이긴 경기의 수는 같지만 무승부가 2007년에는 있지만 2008년에 없기 때문에 승률에 차이가 있다. 이것은 앞에서 언급한 것과 같이 2008년의 경기 규칙에서

표 **3.1.** 2006~2008년 각 팀별 페넌트레이스 성적

팀	연도	승	무	패	승률
SK	2006	60	1	65	0.480
	2007	73	5	48	0.603
	2008	83	0	43	0.659
두산	2006	63	3	60	0.512
	2007	70	2	54	0.565
	2008	70	0	56	0.556
롯데	2006	50	3	73	0.407
	2007	55	3	68	0.447
	2008	69	0	57	0.548
삼성	2006	73	3	50	0.593
	2007	62	4	60	0.508
	2008	65	0	61	0.516
한화	2006	67	2	57	0.540
	2007	67	2	57	0.540
	2008	64	0	62	0.508
기아	2006	64	3	59	0.520
	2007	51	1	74	0.408
	2008	57	0	69	0.452
히어로즈	2006	70	1	55	0.560
	2007	56	1	69	0.448
	2008	50	0	76	0.397
LG	2006	47	4	75	0.385
	2007	58	6	62	0.483
	2008	46	0	80	0.365

무승부가 없는 경기규칙을 시행하여서 이러한 승률이 나왔다고 할 수 있다. 표 3.1의 팀 열거 순서는 2008년의 페넌트레이스 성적을 기준으로 하였다.

3.3. 자료 분석

분석에 있어서 분석방법은 LLBT방법을 사용하여 자료를 분석하였다. BT모형이 아닌 LLBT방법을 사용하여 자료를 분석하는 것은 앞에서도 언급한 것과 같이 로그선형모형을 사용하면 다변량으로의 확장이 쉬울 뿐만 아니라 로그선형모형이 GLM에 포함되므로 이를 분석하기위한 일반적인 통계프로그램을 사용할 수 있다는 장점이 있다.

3.3.1. 승패만 있는 모형 식 (2.6)의 모형을 기준으로 오직 승패만을 가지고 분석을 실시하였다. 분석에 있어서 3년간의 자료 중에서 2008년은 무승부가 없기 때문에 분석하는데 있어서 상관이 없었지만 2006년과 2007년의 경우 무승부가 있어서 승패만을 고려한 모형에서는 무승부 경기의 결과는 분석에서 제외하고 분석을 실시하였다. 그렇기 때문에 분석에 있어서는 오직 승패의 결과만이 분석에서 사용되었다.

분석결과는 표 3.2로 나타내었다. 표 3.2는 'LG'를 기준으로 분석을 실시하였다. 그렇기 때문에 'LG'의 추정치와 표준편차는 0이 되면서 기준이 된다. 표 3.2에서 '추정치'가 분석에 기준이 되는 'LG'보다 큰 값 즉, 양수가 나오면 'LG'보다 승리할 확률이 더 높아지게 되는 것이고, 음수가 나오면

표 **3.2.** 승패만을 고려한 결과

팀	추정치	표준편차	π_j
SK	0.3096	0.0706	0.16386
두산	0.2439	0.0700	0.14367
롯데	0.1142	0.0700	0.11086
삼성	0.2340	0.0700	0.14086
한화	0.2174	0.0699	0.13626
기아	0.0937	0.0696	0.10639
히어로즈	0.1098	0.0697	0.10987
LG	0	0	0.08823

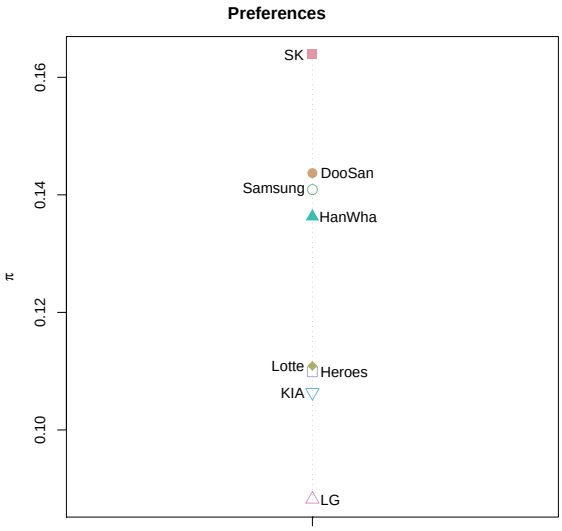

그림 **3.1.** 승패만을 고려한 모형의 팀별 특성(π_j)

'LG'보다 승리할 확률이 작아진다고 할 수 있다. 그리고 각 팀의 특성 또는 능력을 나타내는 가치모수 'π_j'의 합은 1이다. 결과를 살펴보면 경기를 하였을 경우 가장 승리를 할 수 있는 확률이 높은 팀은 'SK'이고, 가장 낮은 팀은 'LG'이다. 표 3.2에서 추정치의 값이 음수로 나오는 추정치가 없고 또한 π_j의 값이 'LG'의 경우 가장 작은 값이 나오기 때문이다.

모형의 적합도를 검정하기 위한 검정통계량은 자유도가 21인 카이제곱 분포를 따르고 유의확률이 0.7074로 나왔다. 따라서 승패만을 가지고 한 분석의 모형은 현재의 자료를 바탕으로 타당하다고 할 수 있다. 이 연구에서 사용한 유의수준은 0.05를 기준으로 하였다. 모형 선택의 기준중 하나인 AIC(Akaike Information Criterion)의 값은 373.1로 나왔다.

그림 3.1은 표 3.2의 π_j를 한 번에 파악할 수 있는 그림으로 나타낸 것이다. 그림 3.1의 가로 축은 표 3.2의 팀에 해당하고, 세로축은 표 3.2의 π_j에 해당된다.

3.3.2. 무승부를 고려한 모형 무승부를 고려한 모형은 식 (2.7)의 모형을 사용하였다. 이것은 앞에서 분석한 승패를 고려한 모형에 무승부 결과를 추가하여 분석을 하였기 때문이다. 이번에는 2006년과 2007년의 경기 결과는 모두 사용하였고, 2008년의 경기 결과의 경우 각 팀의 자료에서 무승부는 모두

표 **3.3.** 각각의 변수조합에서 최적의 모형들

변수의 수	모형에 포함된 변수들	AIC
7	타율, 도루, 실책, 장타율, 평균자책점, 세이브, 홀드	373.10
6	타율, 도루, 실책, 장타율, 세이브, 홀드	371.10
5	도루, 장타율, 평균자책점, 세이브, 홀드	369.16
4	도루, 장타율, 평균자책점, 세이브	367.20
3	실책, 평균자책점, 세이브	365.30
2	평균자책점, 세이브	363.53
1	평균자책점	364.44

0으로 놓고 분석을 하였다. 이렇게 되는 것은 2008년에 경기 규칙에 무승부가 없기 때문이다.

분석 결과를 보면 표 3.2에서 나타나는 수치와 비슷하게 추정치, 표준편차 그리고 π_j가 도출되었다. 따라서 그림 3.1에서 나온 것과 순서가 동일한 비슷한 그래프를 구할 수 있었다. 모형에 대한 적합도 검정 결과 이모형은 자유도가 48인 카이제곱 분포를 따르는데 모형에 대한 유의확률이 0.1987로 나와서 무승부를 고려한 모형이 유의하다고 할 수 있다.

그러나 모형 선택의 한 기준인 AIC의 값이 445.11로 나와 승패만 고려한 모형의 AIC값인 373.1과는 많은 차이를 보였다. 참고로 이후 나오는 모든 모형에서 무승부를 고려한 모형이 그렇지 않은 모형에 비해 항상 AIC값이 크게 나왔다. 표 3.1을 살펴보면 무승부의 경기 숫자가 워낙 작게 나타났기 때문에 이러한 결과를 도출한 것 같다. 여기에서 주의할 점은 단순히 AIC의 값만을 비교해서는 안 된다는 것이다. 왜냐하면 무승부를 고려한 모형에서는 지난 3년간의 페넌트레이스의 모든 자료를 이용하였으나 승패를 고려한 모형에서는 무승부를 제외하고 분석한 AIC이기 때문이다. 비슷한 이유로 이후 나오는 모형들에서 AIC의 단순비교(승패, 무승부 포함 모형)는 주의하여야 한다. 이러한 이유로 이후 나타나는 모형에서는 무승부를 제외한 자료를 가지고 주로 분석하고자 한다.

3.3.3. 공변량을 고려한 모형 객체에 대한 공변량을 고려한 경우 이모형은 식 (2.8)을 이용하여 분석할 수 있다. 각 팀의 특성들을 공변량을 통하여 분석하였을 때 경기에서 승리할 확률이 어느 쪽이 더 높아지는지 알아볼 수 있는 분석이다.

(1) 각 팀의 특성을 나타내는 연속형 변수를 0이용한 모형들

각 팀의 특성을 대표하기위해 지난 3년간의 팀 야수 성적 중에서 타율, 도루의 갯수, 장타율, 실책의 갯수를 평균하여 이를 사용하고 팀 투수 성적을 대표하는 평균자책점, 세이브의 갯수, 그리고 홀드의 갯수를 마찬가지로 평균하여 총 7개의 연속형 변수를 가지고 각 팀의 특성을 표현하여 기존의 승패만 있는 모형에서 보다 더 작은 개수의 모수를 가지고 승패를 잘 설명할 수 있는 모형을 찾고자 한다. 여기에서 기준은 AIC를 사용하고자 한다.

최적의 모형을 찾기 위해 모든 가능한 경우의 모형을 전부 고려하였다. 아래의 표에 각각의 변수의 개수별로 나타나는 최적의 모형조합을 나타내었다. 즉, 변수의 수가 하나인 모든 일곱 가지 가능한 모형 중에서 평균자책점을 변수로 하는 모형이 최소의 AIC를 갖는 모형으로 분석되어졌다. 편차를 이용한 근사검정 결과 아래 표 3.3에 나타난 모든 경우에 있어 기존의 7개의 변수를 사용하는 모형에 비해 자료를 더 잘 설명하는 것으로 나타났다. AIC값은 변수의 개수가 적을수록 감소하다가 변수의 개수가 하나일 때 증가하였다.

이중 최소의 AIC를 갖는 두 가지 모형에 대해 좀 더 고찰해 보기로 하자. 먼저 최소의 AIC값을 갖는 평균자책점과 세이브를 변수로 하는 모형의 결과는 표 3.4와 같다. 평균자책점에 대한 유의확률은 0에

표 **3.4.** 평균자책점과 세이브를 변수로 하는 모형의 결과

변수	추정치	표준오차
평균자책점	−0.2541	0.0640
세이브	0.0060	0.0035

표 **3.5.** 평균자책점을 변수로 하는 모형의 결과

변수	추정치	표준오차
평균자책점	−0.2986	0.0585

표 **3.6.** 승패만 있는 경우에 동군과 서군의 결과

분류	추정치	표준편차
동군	0.1182	0.0345

가까운 값이 나왔으나 세이브에 대한 유의확률은 0.0888이 나왔다. 이모형에 대한 AIC의 값은 모든 가능한 전체모형 중에서 최소의 값인 363.53이 나왔다. 모형의 검정을 하기위해 편차를 사용하였다. 전체변수를 포함한 모형에서 이모형의 편차(Deviance)의 차이값은 0.4357이 나왔으며 자유도의 차이는 5이다. 따라서 근사적 χ^2검정결과 유의확률은 0.9943이 나와 전체모형에 비해 이모형이 자료를 더 잘 설명하는 것으로 나타났다. 평균자책점이 마이너스의 추정치를 가지므로 평균자책점이 낮을수록 세이브의 개수가 많을수록 각 팀의 승리할 확률이 더 높다고 해석되어진다.

두 번째 최소의 AIC값을 갖는 평균자책점만 포함된 모형의 결과는 표 3.5와 같다. 평균자책점에 대한 유의확률은 0에 가까운 값이 나왔으며 이모형에 대한 AIC의 값은 364.44가 나왔다. 모든 변수를 포함한 모형과 이모형의 편차의 차이는 3.3389이며 자유도의 차이는 6이므로 근사적 χ^2검정결과 유의확률은 0.7653이 나와 전체모형에 비해 자료를 더 잘 설명하는 것으로 나타났다. 평균자책점이 낮을수록 각 팀의 승리할 확률이 높아진다. 이후 나타나는 모형들은 한국프로야구에서 관심의 대상이 될 수 있는 가설들을 설명하는 모형들로서 각 팀을 몇 가지 범주로 나누어 어떠한 것이 승패에 영향을 미칠 수 있는 범주인지 알아보고자 한다.

(2) 동군과 서군의 분석

올스타전에서 나뉘는 동군과 서군에 차이가 있는지 알아보기로 하였다. 동군의 경우 SK, 두산, 롯데, 삼성으로 이루어져 있고, 서군의 경우는 한화, 기아, 히어로즈, LG로 이루어져 있다. 분석에서 기준은 서군을 기준으로 하였다. 승패만 있는 모형의 경우 분석 결과는 표 3.6과 같다.

올스타전의 동군과 서군으로 분류를 기준으로 나누었을 때 동군의 추정치는 0.1182가 나왔고, 유의확률도 0.0006으로 나왔으며, 모형에 대한 유의확률도 0.1319가 나와서 모형도 적합하다고 할 수 있다. 그렇기 때문에 동군이 서군보다 승리할 확률이 높다고 할 수 있다. 이모형의 AIC값은 372.34가 나왔다.

(3) 골든글러브 수상자의 영향력 분석

한국 프로야구에서는 정규시즌과 플레이오프가 모두 끝난 후 매년 12월에 각 포지션별로 그해 가장 우수한 선수에게 상을 주게 되는 데 그 상을 골든글러브라고 한다. 그래서 각 팀에서 3년 동안 골든글러브를 탄 선수가 총 3명이상인 팀과 그렇지 못 한 팀으로 구분하여 분석을 하였다. 골든글러브가 있는 팀은 SK, 두산, 롯데, 삼성, 한화였고 골든글러브가 없는 팀은 기아, 히어로즈, LG였다. 분석에서 기준은 골든글러브 수상자가 3명 미만인 팀으로 하였다. 승패만 있는 경우 분석 결과는 표 3.7과 같다.

골든글러브 수상자가 3명 이상인 팀의 경우에 추정치에 대한 유의확률을 확인한 결과 유의하다는 결과

표 **3.7.** 승패만 있는 경우에 골든글러브 결과

분류	추정치	표준편차
3명 이상	0.1545	0.0359

표 **3.8.** 승패만 있는 경우에 다승왕을 보유한 팀들의 결과

분류	추정치	표준편차
다승왕	0.1453	0.0358

를 얻었고, 모형에 대한 검정 또한 0.3956으로 유의하다는 결과를 얻을 수 가 있었다. 즉, 골든글러브 수상자가 3명 이상인 팀들의 경우 그렇지 못한 팀들에 비해서 승리할 확률이 더 높다고 할 수 있다. 이 모형의 AIC의 값은 372.34로 나왔다.

(4) 투수의 영향력 분석

전체 팀에서 투수의 포지션에 있는 선수는 야수의 포지션별 선수에 비해 많이 있다. 하지만 투수의 경우 타자와 달리 매일 나오기는 힘들다. 특히 마무리를 전담하는 선수가 아닌 이상 더욱 힘들다. 그래서 투수들 중에서 다승왕을 차지한 선수가 있는 팀과 없는 팀으로 구분지어 분석을 하였다. 다승왕의 경우 우연한 경우를 제외하고는 매해 1명의 선수만이 차지할 수 있다. 조사된 3년의 자료에서는 매해 오직 1명의 선수만이 다승왕을 차지하였다. 다승왕을 보유하고 있었던 팀은 SK, 두산, 한화이다. 이 분석에서는 다승왕을 보유하지 않은 팀들을 기준으로 하여 분석을 하였다. 승패만 있는 경우 분석 결과는 표 3.8과 같다.

다승왕을 보유한 SK, 두산, 한화의 경우 추정치의 유의확률이 유의한 결과가 나왔고, 모형의 검정에서도 0.2956으로 유의한 결과가 나와서 다승왕을 보유한 팀들이 다승왕을 보유하지 못한 다른 팀들에 비하여 승리할 확률이 더 높다고 할 수 있다. 이모형의 AIC값은 374.46이다.

위에 나타난 연속형변수와 범주형변수를 결합한 모형이 관심의 대상이 될 수 있다. 최적의 모형으로 선택된 두 가지 모형(평균자책점만 있는 모형과 평균자책점과 세이브가 있는 모형)에서 그 가능성을 탐구해보았다. 그 결과 연속형변수와 범주형변수의 조합은 단순히 연속형변수를 포함한 모형들에 비해 항상 더 큰 AIC값을 주어 이들의 조합은 고려의 대상이 되지 않았다.

(5) 기타 모형들

위의 두 가지 모형들 이외에도 다른 몇 가지 모형들을 시도해 보았으나 몇몇의 경우 모형에 대한 적합도 검정의 결과 적절하지 않은 모형으로 나왔다.

예를 들면 각 팀별 홈구장의 위치를 보면 전국에 분포 되어 있는데 수도권에 있는 팀과 비수도권에 있는 팀으로 나뉘어 질수가 있다. 수도권에 있는 팀은 SK, 두산, 히어로즈, LG이고, 비수도권에 있는 팀은 롯데, 삼성, 한화, 기아로 분류 할 수 있다. 이 경우 승패만 있는 경우나 무승부가 포함된 경우 모두 모형이 설명력이 없는 것으로 나타났다.

어떤 팀의 평균적인 투수의 영향력을 분석하기위해 10승 투수가 3년 동안 몇 명이 있었는지를 분석해보았다. 3년 동안 10승 투수를 6명이상 보유한 팀은 SK, 롯데, 삼성, 한화, 히어로즈이고 5명이 하는 두산, 기아, LG로 분류되었다. 승패만 있는 모형에서 추정치는 유의한 결과가 나왔지만 모형은 유의하지 않은 결과를 보였다.

타자의 경우 여러 부분을 고려해 볼 수 있다. 최다홈런, 최다안타, 최다타점 등 여러 가지를 고려할 수가 있는데 여기서는 최고타율만을 고려하기로 하였다. 타율에는 홈런과 안타 등이 포함되는 영역이기 때문

표 **3.9.** 승패와 홈경기의 이점을 고려한 결과

팀	추정치	표준편차
SK	0.3003	0.0702
두산	0.2353	0.0699
롯데	0.0977	0.0698
삼성	0.2251	0.0699
한화	0.2091	0.0698
기아	0.0847	0.0695
히어로즈	0.1008	0.0696
LG	0	0
δ	0.1274	0.0524

에 최고타율을 분석에 이용하기로 하였다. 최고타율은 다승왕과 마찬가지로 한해 가장 타율이 좋은 선수를 선정하는 것으로 3년 치 자료에서는 매해 각각 1명씩만이 가장 높은 타율을 친 선수들이 선정되었다. 그래서 최고타율을 친 선수들을 보유한 3개의 팀 두산, 롯데, 기아와 나머지 팀들에 대한 분석을 하였다. 여기에서도 최고타율을 친 선수들을 보유하지 않은 팀들을 기준으로 하여서 분석을 실시하였다. 승패만 있는 모형에서 모형에 대한 적합도 검정의 결과 모형은 설명력이 없는 것으로 나타났다.

3.3.4. 홈경기의 이점을 고려한 모형 두 팀이 경기를 하는데 있어서 경기의 장소는 영향을 준다는 일반적인 생각이 한국 프로야구에 있어서도 경기 승패에 영향을 주는지 홈경기의 이점을 고려하여 분석을 하였다. 분석에서 사용되는 모형은 식 (2.10)이다. 이를 위한 자료는 앞에서 승패만을 고려한 자료를 기본으로 홈팀과 원정팀을 구분하여 경기결과를 세분화 시켰다.

승패에 홈경기의 이점을 고려한 결과는 표 3.9이다. 결과에서 홈경기의 이점에 대한 δ값이 0.1274로 나와서 홈팀이 경기에서 승리를 할 경우를 계산하는데 있어서 홈경기의 이점에 대한 δ의 특성값을 더한 값으로 계산한 확률이 승리를 할 확률로 계산할 수 있다. 그렇기 때문에 경기의 결과가 승패만 있는 경우에 경기를 하는 각 팀이 자신의 홈에서 경기를 하는 경우 승리를 할 확률이 더 높아진다고 할 수 있다. 유의확률은 0.015로 홈경기의 이점에 대한 δ값은 유의하다.

그리고 표 3.9의 모형은 자유도가 48인 카이제곱분포를 따르는데 유의확률이 0.7176으로 모형이 유의하므로 홈에서 경기를 하는 경우 홈팀이 이길 확률이 더 높다고 생각하는 일반적인 생각을 뒷받침 해줄 수 있을 것이다. 이 경우 AIC의 값은 664.09가 나왔다.

3.4. 한국시리즈 우승확률예측

지난 3년간의 각 팀별 페넌트레이스의 결과를 가지고 2008년도 플레이오프에 진출한 네 팀 (1위: SK, 2위: 두산, 3위: 롯데, 4위: 삼성) 각각의 우승할 확률을 예측해보고 이를 실제 결과와 비교하여 보았다. 가장 단순한 모형인 승패만 있는 경우의 모형을 가지고 우승확률을 계산해 보았다. 우승확률의 계산은 Searls (1963)에서 나타난 결과를 응용하여 계산하였다. 우선 j팀과 k팀이 경기를 하였을 시 j팀이 승리할 확률을 구하여야 하는데 이는 식 (2.1)에서 주어졌으며 표 3.2에서 주어진 각 팀별 가치모수, π_j를 이용하면 아래와 같이 확률들을 계산할 수 있다. 식 (2.1)에서 주어진 $p_{(jk)j}$를 기호의 편리상 p_{jk}로 표기하기로 하자. 표 3.10에 그 값을 계산하였다. 기호의 편리상 2008년도 순위로서 그 팀을 표기하기로 하자.

먼저 각각의 게임이 독립이며 각각의 게임에서 각 팀이 다른 팀에 승리할 확률은 표 3.10를 따른다고 가

표 **3.10.** 각행(j)팀이 각열(k)팀에 승리할 확률

j	k			
	SK(1)	두산(2)	롯데(3)	삼성(4)
SK		0.5328	0.5965	0.5377
두산	0.4672		0.5645	0.5049
롯데	0.4035	0.4355		0.4404
삼성	0.4623	0.4951	0.5596	

정하자. 이는 단기전에는 어느 정도 비현실적인 가정이지만 목적인 우승확률을 계산하고 이를 비교하는 데는 별다른 무리가 없으리라 생각되어진다. SK가 우승할 확률을 계산해보면 전확률의 정리에 따라 4가지 경우의 확률들의 합으로 구해진다. 첫 번째 확률은 준플레이오프에서 롯데가 승리하고 플레이오프에서 두산이 승리하고 그리고 한국시리즈에서 SK가 승리하는 경우이다. 두 번째 경우는 준플레이오프에서 롯데가 승리하고 플레이오프에서 롯데가 승리하고 그리고 한국시리즈에서 SK가 승리하는 경우이다. 세 번째 경우는 준플레이오프에서 삼성이 승리하고 플레이오프에서 두산이 승리하고 그리고 한국시리즈에서 SK가 승리하는 경우이다. 마지막은 준플레이오프에서 삼성이 승리하고 플레이오프에서 삼성이 승리하고 그리고 한국시리즈에서 SK가 승리하는 경우이다.

첫 번째 경우의 확률을 계산하면 5전 3승제이므로

$$\begin{aligned} P(\text{준플레이오프에서 롯데 승리}) &= P(3\text{연승}) + P(3\text{승1패}) + P(3\text{승2패}) \\ &= p_{34}^3 + \binom{3}{2} p_{34}^3(1-p_{34}) + \binom{4}{2} p_{34}^3(1-p_{34})^2, \end{aligned}$$

그리고 플레이오프에서 두산이 승리하는 확률은 7전 4승제이므로

$$\begin{aligned} &P(\text{플레이오프에서 두산승리} \mid \text{준플레이오프에서 롯데 승리}) \\ &= P(4\text{연승}) + P(4\text{승1패}) + P(4\text{승2패}) + P(4\text{승3패}) \\ &= p_{23}^4 + \binom{4}{3} p_{23}^4(1-p_{23}) + \binom{5}{3} p_{23}^4(1-p_{23})^2 + \binom{6}{3} p_{23}^4(1-p_{23})^3, \end{aligned}$$

마지막으로 한국시리즈에서 SK가 승리하는 조건부 확률을 계산하면 7전 4승제이므로

$$\begin{aligned} &P(\text{한국시리즈에서 SK승리} \mid \text{플레이오프에서 두산승리, 준플레이오프에서 롯데 승리}) \\ &= P(4\text{연승}) + P(4\text{승1패}) + P(4\text{승2패}) + P(4\text{승3패}) \\ &= p_{12}^4 + \binom{4}{3} p_{12}^4(1-p_{12}) + \binom{5}{3} p_{12}^4(1-p_{12})^2 + \binom{6}{3} p_{12}^4(1-p_{12})^3. \end{aligned}$$

따라서 첫 번째 경우의 확률은 위 세 가지 경우의 곱으로 구해진다. 나머지 경우는 위와 비슷하게 계산하면 된다. 두산이 한국시리즈에서 우승하는 확률은 두 가지 경우의 확률의 합으로 구해지며 롯데와 삼성의 경우는 준플레이오프와 플레이오프 그리고 한국시리즈에서 연속으로 승리하는 한 가지 경우씩 있다. 이를 정리하면 표 3.11로 요약되어진다.

2008년도의 실제결과와 비교해보면 준플레이오프에서 삼성이 롯데를 3승으로 이기고 플레이오프에 진출하였으나 두산에 2승4패로 패하여 한국시리즈진출에는 실패하였다. SK와 두산이 치른 한국시리즈에서는 시리즈전적 4승1패로 SK가 한국시리즈 패권을 차지하였다. 이 결과는 표 3.11의 우승확률의 순서와 정확히 일치한다.

표 **3.11.** 각 팀별 우승확률

팀	우승확률
SK	0.2961
두산	0.1135
롯데	0.0166
삼성	0.0564

4. 결론

3절에서 분석한 결과를 종합해 보면 모든 경우의 모형에서 무승부가 포함된 모형의 AIC의 값이 크게 나오고 종종 모형의 적합도 검정결과 유의하지 않다는 결과를 보였다. 그러나 3절에서 언급되었듯이 비교되는 자료가 다르므로 AIC의 해석에 주의가 요구된다.

공변량으로 생각되어지는 여러 영역을 분석해본 결과 연속형 공변량을 사용한 모형에서 평균자책점과 세이브를 포함하는 모형이 최소의 AIC값을 가졌으며 이는 단순히 승패만 있는 모형에 비해 더 작은 AIC값을 가져 적절한 모형으로 판단되어진다. 몇 가지 관심 있는 가설에 대한 범주형 분석의 결과는 아래와 같다.

골든글로브를 수상한 선수를 포함한 팀이거나 다승왕을 이룬 선수를 포함한 팀 그리고 동군에 속한 팀들의 승리 확률이 그렇지 못한 팀들에 비해서 높다는 것을 알 수 있었다. 수도권과 비수도권으로 나눈 분류와 최고타율을 친 선수를 보유한 팀으로 나눈 경우는 승리에 영향을 미치는 조건이 될 수 없었다. 마지막으로 승패만 있는 경우 홈경기의 이점이 있었다.

이러한 분석결과를 종합하였을 때 승리를 하는데 있어서 영향을 미치는 요인은 평균자책점과 세이브의 개수가 자료를 잘 설명하는 것으로 드러났다. 관심의 대상이 되는 몇 가지 범주형 분류에 의한 분석결과 동군과 서군, 골든글러브 수상자와 다승왕, 그리고 홈경기가 승패에 영향을 미치는 것으로 드러났다.

최근에 이슈가 되었던 무승부를 패로 인정하는 제도에 대하여 통계적인 관점에서 이를 분석해 보았다. 본론에서 언급한 모형중 무승부를 포함한 4가지 모형(무승부를 포함한 모형, 동군과 서군, 골든글러브 수상자의 영향력 분석 그리고 다승왕에 의한 분류)과 해당 모형에서 무승부를 패로 인정한 모형들의 각각의 AIC를 비교하였다. 네 가지 모든 경우에서 무승부를 패로 인정한 모형들의 AIC값이 작게 나타났다. 이는 모형의 관점에서 무승부를 패로 인정한 모형이 무승부로 인정한 모형들보다 설명력이 좋다는 것이다.

이 연구를 통해서 우리나라의 프로야구에 승리에 영향을 미치는 요인에 대하여 분석을 하면서 어려웠던 점은 프로리그라는 것이 한해마다 선수의 이동과 구단의 지원 등에 따라서 그 해의 성적이 엇갈리게 나오게 된다. 하지만 연구를 하면서 이런 부분들을 고려하기에는 모형과 분석자체가 너무 많은 양을 필요하고, 자료의 수집자체가 어렵다는 것을 느끼게 되었다. 그리고 최근 3년의 자료를 본 연구에서는 사용하여서 나온 결과치를 보면 2009년에 한국 프로야구리그에서 나온 결과와는 사뭇 다른 결과라는 것을 확인 할 수가 있었다. 그렇기 때문에 본 연구를 통하여 그 다음해의 성적을 예상하기에는 어려운 점이 있다. 본 연구로는 1년의 경기 결과를 예상하기 보다는 분석에서 사용된 가장 최근의 경기결과 다음에 해당되는 경기에 대한 결과를 예측하는데 도움이 될 수 있다. 그리고 본 연구에서 더 필요한 요인들을 고려하여 추가적인 연구를 한다면 연구를 한 다음의 경기에 대한 예측을 하기에 좋은 모형을 만들 수 있을 것이라고 생각된다. 한 예로 지난 3년간의 자료를 이용하여 2008년도의 한국시리즈 우승확률을 계산하고 이를 실제 결과와 비교하였다. 우승확률들의 순서는 2008년도의 결과와 정확히 일치하였다.

참고문헌

신상근, 박기철, 조영석, 최세현 (2007). 한국프로야구팀의 승패요인분석에 관한 연구: 삼성라이온즈를 중심으로, *Journal of the Korean Data Dnalysis Society*, **9**, 2071–2083.

이장택, 조현식 (2009). 로지스틱 회귀모형을 이용한 프로야구 홈경기의 이점에 관한 연구, *Journal of the Korean Data Dnalysis Society*, **11**, 533–543.

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition, John Wiley & Sons, New Jersey.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs Ⅰ: The method of paired comparisons, *Biometrika*, **39**, 324–345.

David, H. A. (1988). *The Method of Paired Comparisons*, 2nd edition, Oxford University Press, New York.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association*, **65**, 317–328.

Davidson, R. R. and Beaver, R. J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects, *Biometrics*, **33**, 393–702.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university ranking, *Applied Statistics*, **47**, 511–525.

Fienberg, S. E. and Larntz, K. (1976). Loglinear representation for paired and multiple comparison models, *Biometrika*, **63**, 245–254.

Francis, B., Green, M. and Payne, C. (1993). *The GLIM System: Release 4 Manual*, Clarendon Press, Oxford.

Glickman, M. E. (1993). Parameter estimation in large dynamic paired comparison experiments, *Applied Statistics*, **48**, 377–394.

Searls, D. T. (1963). On the probability of winning with different tournament procedures, *Journal of the American Statistical Association*, **34**, 1064–1081.

Sinclair, C. D. (1982). GLIM for preference, In: Gilchrist, R. (Eds.): GLIM 82., In *Proceedings of the International Conference on Generalized Linear Models, Springer Lecture Notes in Statistics*.

Springall, A. (1973). Response surface fitting using a generalization of the Bradley-Terry paired comparison model, *Applied Statistics*, **22**, 59–68.

A Statistical Study on Korean Baseball League Games

Young-Gun Choi[1] · Hyoung-Moon Kim[2]

[1]Department of Applied Statistics, Konkuk University
[2]Department of Applied Statistics, Konkuk University

(Received October 2010; accepted June 2011)

Abstract

There are a variety of methods to model game results and many methods exist for the case of paired comparison data. Among them, the Bradley-Terry model is the most widely used to derive a latent preference scale from paired comparison data. It has been applied in a variety of fields in psychology and related disciplines. We applied this model to the data of Korean Baseball League. It shows that the loglinear Bradley-Terry model of defensive rate and save is optimal in terms of AIC. Also some categorical characteristics, such as east team and west team, existence of golden glove winning players, team(s) with seasonal pitching leader, and team(s) with home advantage, influenced the game result significantly. As a result, the suggested models can be further utilized to predict future game results.

Keywords: Bradley-Terry model, loglinear Bradley-Terry model, Korean Baseball League.

This work was supported by the Konkuk University.
[2]Corresponding author: Professor, Department of Applied Statistics, Konkuk University, 1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Korea. E-mail: hmkim@konkuk.ac.kr