Journal of the Korean Data & Information Science Society
2017, **28**(2), 309–316

http://dx.doi.org/10.7465/jkdi.2017.28.2.309
한국데이터정보과학회지

한국프로야구에서 쌍별 승률추정량의 효율성[†]

이장택[1]

[1]단국대학교 응용통계학과

접수 2017년 2월 10일, 수정 2017년 3월 5일, 게재확정 2017년 3월 13일

요 약

야구에서 승률 추정은 매우 중요한 문제이며 현재 이 분야에 대한 연구가 활발하게 진행되고 있다. 쌍별 승률추정은 팀대 팀의 경기결과를 이용하여 전체 승률을 추정하는 방법으로써 각 팀들의 추정된 승률의 합이 상수가 된다는 타당성을 가진다. 본 연구에서는 한국프로야구에서 피타고라스 승률과 선형 승률에 쌍별 추정을 적용하고 효율성을 RMSE와 MAD를 이용하여 살펴보았다. 사용된 데이터는 2013년부터 2016년 사이의 모든 한국프로야구 팀대 팀 기록이며, 그 결과 쌍별 피타고라스 추정이 기존의 방법들보다 RMSE와 MAD 측면에서 바람직하다고 간주되었다. 또한 쌍별 피타고라스 추정에 사용되는 바람직한 지수 값의 결정에 대하여 설명하였으며 추정에 사용된 지수 값의 변화에 따른 RMSE와 MAD의 차이는 크지 않음을 알 수 있었다.

주요용어: 선형, 승률, 쌍별, 피타고라스, MAD, RMSE.

1. 서론

프로야구경기의 궁극적인 목표는 단언컨대 우승이며, 시합에서 이긴 비율을 의미하는 승률 (winning percentage)이 가장 높은 팀이 프로야구 정규리그 우승을 한다. 따라서 야구팬들의 최우선 관심사가 이번 시즌에는 어느 팀이 우승할 것인가에 있으므로 승률의 정확한 추정은 매우 의미 있는 논쟁의 대상이다. 야구통계에서 승률추정 문제가 생활 통계학의 활용부분으로 야구팬 및 일반인들의 관심을 끌게 된 가장 큰 이유는 아마도 야구의 승률은 총득점의 제곱을 총득점의 제곱과 총실점의 제곱의 합으로 나눈 것으로 추정된다는 야구의 피타고라스 정리로 알려져 있는 James (1980)의 주장 때문이라고 할 수 있으며, 내용은 식 (1.1)과 같이 약술된다.

$$EW = \frac{RS^2}{RS^2 + RA^2} \tag{1.1}$$

여기서 EW는 기대승률 또는 피타고라스 승률 (Pythagorean winning percentage), RS는 팀의 총 득점, RA는 팀의 총 실점을 의미한다. 그리고 실제 승률의 추정값인 EW에 게임의 수 (G)를 곱하면 기대되는 이긴 게임의 수, 즉 기대승수를 구할 수 있으며 이를 피타고라스 승수 (Pythagorean wins)라고 한다. 보통 EW는 팀의 진짜 승률 값으로 간주하며 실제 승률 값과 EW의 차이를 이용하여 시즌에서 특정 팀의 행운측도로 사용하기도 한다.

식 (1.1)을 실제 상황에 적용할 때에는 보다 정확한 기대승률을 도출하기 위하여 지수 값으로 2 대신 미지수 γ로 두어서 추정하는 데, 미국 메이저리그인 경우 Davenport와 Woolner (1999)에 의하면

[†] 이 연구는 2017학년도 단국대학교 대학연구비 지원으로 연구되었음.
[1] (16890) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

1.83을 주로 사용하며, 한국프로야구인 경우도 Lee (2016b)에 의하면 1982년부터 2015년 전 경기를 이용한 γ의 최적 해는 메이저리그와 같게 1.83으로 나타났다. 피타고라스 정리의 최적지수 γ의 추정문제를 다룬 연구들은 메이저리그인 경우, Davenport와 Woolner (1999)와 Cochran (2008)의 결과가 있으며 결론은 γ의 값은 1.74부터 2.0사이의 값으로 게임당 발생하는 총득점의 값에 종속된다고 밝혔다. 한편 한국프로야구인 경우, 최적지수 γ의 추정문제를 다룬 연구로는 Lee (2014)가 있는데 선행연구들의 결과인 게임당 발생하는 총득점의 값과 승률의 표준편차를 같이 고려하여 지수 γ의 추정을 좀 더 효율적으로 하였으며, Lee (2015)는 한국프로야구 기록을 이용하여 실제승률과 피타고리안 기대승률의 차이가 발생하는 원인을 회귀모형을 통해 설명하였으며, Lee (2016a)는 한국프로야구에서 야구의 피타고라스 정리에 의한 기대승률의 수렴특성을 살펴보았다. 식 (1.1)은 비선형모형이라면 승률추정량으로 Jones와 Tappin (2005)에 의해 제안된 식 (1.2)로 기술되는 선형모형이 있다.

$$EW = 0.5 + \beta(RS - RA) \tag{1.2}$$

식 (1.2)의 기울기 β의 값은 미국 메이저리그인 경우에 1969년부터 2003년까지의 미국 메이저리그 데이터를 이용하여 β의 추정치가 0.00053부터 0.00078까지 나타나며 그 평균치는 0.00065로 나타나고 한국 프로야구인 경우에는 1982년부터 2015년까지 모든 팀 데이터를 이용하여 추정한 결과 0.00079로 나타났다 (Jones와 Tappin, 2005; Lee, 2016b). 한국프로야구 승률추정 문제를 다룬 연구들은 여러 가지 승률 결과들을 비교한 Kim과 Lee (2016), 피타고라스 및 선형모형을 비교한 Lee (2016b)가 있다.

한편 Heumann (2016)의 피타고라스 승수에 대한 결과는 선행연구들과 차별성이 있는데, 일반적으로 특정시즌 팀의 수가 N일 때 야구팀들의 추정된 피타고라스 승률의 합이 $N/2$이 되지 않는 점에 착안하여 팀대 팀의 결과인 쌍별 (pairwise) 표본을 사용하면 추정된 기대승률의 합이 $N/2$이 되는 이유를 설명하고, 메이저리그 30년간의 데이터를 이용하여 쌍별 피타고라스 승수 (pairwise Pythagorean wins)가 기존 피타고라스 승수보다 작은 평균제곱오차의 제곱근을 갖는다고 설명하였다. 본 연구는 Heumann (2016)의 후속연구로써 승수 대신 승률의 형태로 한국프로야구에서의 쌍별 피타고라스 승률의 효율성, 쌍별 선형추정량의 효율성, 쌍별 피타고라스 추정량에 적당한 지수 γ값에 대한 추론 및 최적지수를 사용한 피타고라스 승률과 쌍별 피타고라스 승률과의 비교를 시도하는 것이 주목적이다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 분석데이터와 승률의 정의, 통계분석 및 모형평가 기준에 대하여 각각 언급하였으며, 3절에서는 쌍별 피타고라스 정리의 설명 및 지수 γ값의 설정에 대해 설명하고 여러 가지 방법들의 상대적인 효율성을 검증하였다. 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 연구방법

2.1. 데이터 및 승률의 정의

연구에 사용된 데이터는 한국야구위원회 `www.koreabaseball.com`에 기록된 1982년부터 2016년 사이에 있었던 273개 팀의 결과와 2013년부터 2016년까지 프로야구 통계기록실 `http://www.kbreport.com`이 제공하는 모든 팀대 팀의 결과로 모두 324개 팀의 결과이며, 승률의 정의는 한국야구위원회 (KBO)에서 1987시즌부터 1997시즌까지 사용한 식 (2.1)을 사용하였다. 여기서 $Wpct$은 승률, W는 승리한 게임 수, L은 패배한 게임 수, T는 무승부 게임 수이며, 현재 사용되는 승률의 정의인 $W/(W+L)$을 사용하지 않은 이유는 한국프로야구에 대한 모든 공식적인 기록들은 모두 무승부인 경우도 포함하여 집계되었기 때문이다.

$$Wpct = \frac{W + 0.5 \times T}{W + T + L} \tag{2.1}$$

그리고 W^*와 L^*를 각각 $W^* = W + 0.5 \times T$, $L^* = L + 0.5 \times T$라고 할 때, W대신 W^*, L대신 L^*를 적용하여 피타고라스 지수 γ와 승률을 추정할 수 있다.

2.2. 통계분석 및 모형평가기준

기술통계량과 회귀분석은 통계패키지 SPSS 23K를 사용하였다. 또한 제안된 추정량들의 효율성을 비교하기 위한 판단기준으로는 일반적으로 많이 사용되는 추정량 선택기준인 평균제곱오차의 제곱근 (root mean square error; RMSE)와 평균절대편차 (mean absolute deviation; MAD)를 사용하였다. RMSE와 MAD는 값이 작을수록 바람직하며, 정의는 식 (2.2)와 같다.

$$RMSE = \sqrt{\sum_{i=1}^{N}(\widehat{w}_i - w_i)^2/N}, \quad MAD = \sum_{i=1}^{N}|\widehat{w}_i - w_i|/N \tag{2.2}$$

여기서 $\widehat{w}_i$는 i번째 팀의 승률추정량, w_i는 i번째 승률, N은 총 팀의 수를 각각 의미한다. 본 연구의 RMSE와 MAD 값은 모두 승률을 퍼센트로 나타낸 경우의 값이다.

3. 쌍별 피타고라스 승률 및 분석

3.1. 쌍별 피타고라스 승률

Table 3.1 2016 KBO baseball season win loss records

Rank	Team	W	L	T	RS	RA	Wpct	Pyth(2)	Pyth(1.83)	Linear(0.097)
1	Doosan	93	50	1	935	682	.649	.653	.640	.700
2	NC	83	58	3	813	690	.587	.581	.574	.597
3	Nexen	77	66	1	852	757	.538	.559	.554	.575
4	LG	71	71	2	753	807	.500	.465	.468	.457
5	KIA	70	73	1	857	785	.490	.544	.540	.557
6	SK	69	75	0	786	784	.479	.501	.501	.502
7	Hanhwa	66	75	3	826	908	.469	.453	.457	.435
8	Lotte	66	78	0	777	865	.458	.447	.451	.430
9	Samsung	65	78	1	803	869	.455	.461	.464	.448
10	KT	53	89	2	672	927	.375	.344	.357	.299
Sum							5.000	5.008	5.007	5.000

Table 3.1은 2016년 한국프로야구 경기 최종결과로 10개 팀의 득점, 실점 및 승률을 보여주는데, 승률 ($Wpct$)의 합은 각 팀이 이길 기대확률이 0.5이고 팀의 개수가 10개이므로 5가 되나 지수 2를 사용한 피타고라스 승률인 Pyth(2), 지수 1.83을 사용한 피타고라스 승률인 Pyth(1.83)의 합은 5가 정확하게 되지 않는다. 왜냐하면 팀의 개수가 N일 때 추정된 피타고라스 승률의 합이 각각 $N/2$이 된다는 보장이 없기 때문이다. 하지만 1982년부터 2016년까지의 모든 데이터를 이용하여 추정한 기울기 β의 값 0.097을 사용한 선형추정량인 Linear(0.097)의 합은 5가 된다. 왜냐하면 기울기의 값과는 상관없이 선형추정량은 리그의 총득점이 총실점과 같기 때문이다. 만일 피타고라스 승률을 팀대 팀의 결과로 설명하면 앞에서 언급한 확률의 모순점을 극복할 수 있는 데, 그 이유는 팀 A와 팀 B가 게임을 하는 구조를 가정하면, RS_A는 RA_B, RS_B는 RA_A와 각각 같으므로 다음과 같이 피타고라스 추정량의 승률의 합은 1이 된다 (Heumann, 2016).

$$\frac{RS_A^\gamma}{RS_A^\gamma + RA_A^\gamma} + \frac{RS_B^\gamma}{RS_B^\gamma + RA_B^\gamma} = \frac{RS_A^\gamma}{RS_A^\gamma + RA_A^\gamma} + \frac{RA_A^\gamma}{RA_A^\gamma + RS_A^\gamma} = 1$$

팀대 팀의 결과인 쌍별 (pairwise)과 식 (1.1)로 기술되는 고전적인 피타고라스 정리와의 관계를 설명하면 RS_{ij}를 팀 i가 팀 j와 경기를 치러서 발생하는 득점이라고 하고, RS_i와 RA_i를 각각 팀 i의 시즌이 끝난 후의 총득점과 총실점이라고 할 때, 다음 식들이 성립한다.

$$RS_i = \sum_{j\neq i}^{N} RS_{ij} \quad RA_i = \sum_{j\neq i}^{N} RS_{ji}$$

그리고 팀 i에 대한 피타고라스 승률을 RS_{ij}기호로 표기하면 식 (3.1)과 같다.

$$EW_i = \frac{\left(\sum_{j=1}^{N} RS_{ij}\right)^2}{\left(\sum_{j=1}^{N} RS_{ij}\right)^2 + \left(\sum_{j=1}^{N} RS_{ji}\right)^2},\ RS_{ii} = 0,\ i,j = 1,2,\cdots,N \tag{3.1}$$

또한 임의의 팀 i에 대하여 다음 피타고라스 승률비율을 P_{ij}, 팀 i와 팀 j가 갖는 게임의 수를 G_{ij}라고 하면,

$$P_{ij} = \frac{RS_{ij}^2}{RS_{ij}^2 + RS_{ji}^2}$$

다음 w_i는 팀 i의 쌍별 피타고라스 승수 (the pairwise Pythagorean win total)가 되며,

$$w_i = \sum_{i\neq j}^{N} P_{ij}G_{ij}$$

게임의 수에 대해서는 $G_{ij} = G_{ji}$가 성립하기 때문에 모든 팀에 대한 쌍별 피타고라스 승수의 합은 다음과 같이 모든 쌍별 팀들의 경기 수의 합과 같으며 이것은 원래의 피타고라스 정리에서는 성립하지 않는 성질이다.

$$P_{ij}G_{ij} + P_{ji}G_{ji} = G_{ij}(P_{ij} + P_{ji}) = G_{ij}$$

따라서 팀의 개수를 N, 양 팀 간의 동일게임의 수를 k라고 각각 하면, 총 게임수 G는 $k(N-1)$가 되며, 또한 $\sum_{i=1}^{N} w_i = kN(N-1)/2$가 된다. 예를 들면 2016년 한국프로야구는 모든 팀이 시즌동안 144게임을 하였고, 모든 팀대 팀의 경기 수는 같으므로 $k = 16$게임을 양 팀 간에 하였으며, $N = 10$이므로 $\sum_{i=1}^{N} w_i = 720$이 된다. 그리고 팀 i의 쌍별 피타고라스 승률 (PEW_i)는 다음과 같이 구할 수 있는데, 왜냐하면 $PEW_i = w_i/G$이고, $w_i = k\sum_{i\neq j}^{N} P_{ij}$과 $G = k(N-1)$가 성립하기 때문이다.

$$PEW_i = \frac{\sum_{i\neq j}^{N} P_{ij}}{N-1} \tag{3.2}$$

3.2. 여러 가지 피타고라스 승률들의 비교

Table 3.2는 2013년부터 2016년까지의 한국프로야구 팀대 팀 데이터를 이용하여 추정한 4개의 피타고라스 모형과 2개의 선형회귀모형에 대한 승률의 RMSE와 MAD 값을 보여준다. 고려된 모형은 지수 2를 사용한 피타고라스 승률 Pyth(2), 지수 1.72를 사용한 피타고라스 승률 Pyth(1.72), 지수 2를 사용한 쌍별 피타고라스 승률 p.Pyth(2), 지수 1.72를 사용한 쌍별 피타고라스 승률 p.Pyth(1.72), 기

울기 0.097을 사용한 선형회귀모형 Linear(0.097) 그리고 기울기 0.079를 사용한 선형회귀모형 Linear(0.079)와 같은 모두 6가지 모형인데, 지수 1.72은 2013년부터 2016년까지 KBO 데이터를 이용하여 추정된 단일 최적지수 값, 기울기 0.097은 1982년부터 2016년까지의 데이터를 이용하여 추정한 값, 0.079는 2013년부터 2016년까지의 데이터를 이용하여 추정한 값이다. 선형회귀모형을 사용할 때, 식 (1.2)에서 $RS-RA$대신 $(RS-RA)/G$를 사용하면 좀 더 승률과의 상관계수를 높일 수 있는 데, 1982년부터 2016년까지의 데이터를 사용하면 전자는 상관계수가 0.945, 후자는 0.948로 나타나서 본 연구에서는 독립변수로 $(RS-RA)/G$를 사용하였다.

Table 3.2 RMSE and MAD of six winning percentage models

Year	Pyth (2)	Pyth (1.72)	p.Pyth (2)	p.Pyth (1.72)	Linear (0.097)	Linear (0.079)
2013	2.163	2.065	2.112	2.023	2.097	2.675
	(1.710)	(1.765)	(1.696)	(1.626)	(1.506)	(2.381)
2014	3.339	2.877	2.813	2.690	4.086	2.982
	(2.616)	(2.295)	(2.208)	(2.276)	(3.075)	(2.234)
2015	2.049	1.820	1.698	1.582	2.180	1.752
	(1.673)	(1.656)	(1.452)	(1.482)	(1.842)	(1.590)
2016	2.539	2.174	1.841	1.590	3.080	2.250
	(2.044)	(1.807)	(1.504)	(1.309)	(2.574)	(1.894)
Total	2.540	2.277	2.120	2.024	2.957	2.437
	(2.016)	(1.860)	(1.686)	(1.676)	(2.247)	(2.010)

* MAD in parenthesis

2013년부터 2016년 전 데이터의 결과 (total)를 보면 p.Pyth(2), p.Pyth(1.72)의 RMSE와 MAD 모두 Pyth(2)와 Pyth(1.72)에 비해 작다. 심지어 피타고라스 모형인 경우는 최적지수 값 1.72를 사용하여도 RMSE와 MAD 모두 지수 2를 사용한 쌍별 피타고라스 모형보다 크다. 또한 선형회귀모형은 비록 추정된 승률의 합이 승률의 합과 같다는 쌍별 추정량의 좋은 성질을 가지고 있지만 피타고라스 모형보다 RMSE와 MAD 모두 크다. 이 사실은 쌍별 추정량이 좋은 추정치를 제공하는 수학적 보정의 역할을 하는 것이지만 좋은 추정량이 되기 위한 필수조건은 아닌 것을 알 수 있다.

Table 3.3은 쌍별 피타고라스 추정량 (p.Pyth)을 위한 지수 값의 선택 결과를 보여준다. 연도 밑 괄호 안의 지수 값은 연도별로 각각 추정한 최적지수 값을 의미하는데, 2013년부터 2016년까지 연도별 최적지수 값은 1.89, 1.57, 1.76, 1.68로 각각 나타났다. 또한 비교를 위해 사용한 지수 값으로 2013년부터 2016년까지의 단일 최적지수 값 1.72, 보통 알려져 있는 최적지수 값 1.83, 가장 보편화된 지수 2를 참고하여 비슷한 간격으로 1.63, 1.72, 1.83, 1.92, 2.00, 2.09와 같은 6개를 선택하였다. 지수 2와 1.72를 사용한 p.Pyth의 결과들은 다른 추정량들의 결과와 용이한 비교를 위해 Table 3.2의 결과를 다시 사용하였다. 그 결과 전체 데이터 (total)를 모두 사용하는 경우에는 지수 값이 1.83인 경우가 가장 RMSE와 MAD 값이 작았다. 이 사실은 해당 데이터인 경우에 대응되는 단일 최적지수 값 1.72인 경우에 가장 RMSE와 MAD 값이 작으리라는 상식과 일치하지 않는다. 이와 같은 결론은 연도 별로 살펴보아도 같은 결론으로 해석된다. 주목해야 할 또 하나의 사항은 쌍별 피타고라스 모형은 지수 값에 따라 RMSE와 MAD 값의 변화가 크지 않다는 점이다. 2013년부터 2016년까지 전체 데이터를 사용하는 경우에 RMSE 기준으로 보면 가장 작은 값과 큰 값의 차이는 0.239 정도다. 따라서 γ값을 일반적으로 보통 사용하는 범위인 1.74부터 2.0 사이의 값을 사용하면 쌍별 피타고라스 승률의 RMSE와 MAD의 값은 큰 차이가 나지 않는다고 할 수 있겠다.

Table 3.3 RMSE and MAD of pairwise Pythagorasmodels

Year	p.Pyth. $\gamma = 1.63$	p.Pyth. $\gamma = 1.72$	p.Pyth. $\gamma = 1.83$	p.Pyth. $\gamma = 1.92$	p.Pyth. $\gamma = 2.00$	p.Pyth. $\gamma = 2.09$
2013	2.263	2.023	2.006	1.988	2.112	2.115
($\gamma = 1.89$)	(1.889)	(1.626)	(1.543)	(1.571)	(1.696)	(1.756)
2014	2.726	2.690	2.699	2.745	2.813	2.915
($\gamma = 1.57$)	(2.309)	(2.276)	(2.237)	(2.204)	(2.208)	(2.247)
2015	1.671	1.582	1.558	1.608	1.698	1.842
($\gamma = 1.76$)	(1.493)	(1.482)	(1.471)	(1.461)	(1.452)	(1.496)
2016	1.633	1.590	1.627	1.720	1.841	2.010
($\gamma = 1.68$)	(1.244)	(1.309)	(1.388)	(1.451)	(1.504)	(1.605)
Total	2.100	2.024	2.003	2.045	2.120	2.242
($\gamma = 1.72$)	(1.714)	(1.676)	(1.648)	(1.660)	(1.686)	(1.764)

* MAD in parenthesis

3.3. 최적지수를 사용한 피타고라스 승률과 쌍별 피타고라스 승률과의 비교

고전적 피타고라스 승률을 추정할 때 지수 γ를 사용하는 경우에는 최소제곱법을 이용하여 다음과 같은 단순회귀모형을 고려하여 γ값을 추정할 수 있다.

$$\log(W^*/L^*) = \gamma \log(RS/RA) \tag{3.3}$$

이렇게 최적지수 값을 구하면 2013년은 1.89, 2014년엔 1.57, 2015년엔 1.76, 2016년엔 1.68과 같음을 알 수 있으며 이 값들은 고전적 피타고라스 승률의 추정 시에 가장 작은 RMSE 값을 제공한다.

Table 3.4는 연도별로 최적지수 값을 사용한 고전적인 피타고라스 승률과 잘 알려진 지수 값 1.83을 선택한 쌍별 피타고라스 승률의 RMSE와 MAD 값을 보여준다. 결과적으로 모든 연도에 있어서 *로 표시된 쌍별 피타고라스 승률의 RMSE와 MAD 값이 더 작음을 알 수 있다. 이 사실로부터 알 수 있는 것은 지금까지 고전적인 피타고라스 승률의 최적지수 값은 시즌이 끝난 후에나 알 수 있었으며 많은 통계학자들이 최적지수의 추정을 중요한 연구대상으로 삼았는데, Table 3.4의 결과는 굳이 최적지수 값을 알 필요 없이 지수 1.83을 이용하여 쌍별 피타고라스 승률을 사용하면 좀 더 좋은 결과가 나타난다는 것을 알려준다. 하지만 1.83을 사용하는 것이 최선이라는 보장은 없으며 수많은 데이터를 적용시켜보면 좀 더 적당한 값을 찾을 수 있을 것이다. 그러나 그렇게 찾은 최적값과 1.83을 사용하여 구한 RMSE와 MAD의 값은 큰 차이가 나지 않을 것으로 간주되는데 이와 같은 결론이 Heumann (2016)이 향후과제로 제시한 지수 값의 선택에 관한 구체적인 설명이다.

Table 3.4 RMSE and MAD from two Pythagorasmodels (2013-2016)

Year	Method	RMSE	MAD
2013	Pyth(1.89)	2.023	1.577
	p.Pyth(1.83)	2.006*	1.543*
2014	Pyth(1.57)	2.811	2.349
	p.Pyth(1.83)	2.699*	2.237*
2015	Pyth(1.76)	1.817	1.656
	p.Pyth(1.83)	1.558*	1.471*
2016	Pyth(1.68)	2.164	1.782
	p.Pyth(1.83)	1.627*	1.388*

4. 결론

우리가 살고 있는 이 시대는 기업들이 스포츠마케팅에 눈을 뜨면서 스포츠의 과학화는 한층 더 가속이 붙고 있다. 야구는 다른 스포츠하고는 비교가 안 될 정도의 빅데이터를 가지고 있는 데 각종 데이터가 팀 단위는 물론 선수 개개인 수준까지도 매우 정밀하게 구비되어있다. 구단의 감독들은 데이터를 통해 상대 팀에 대한 전략을 세우고 야구통계학자들은 데이터를 바탕으로 야구에 대한 속설 또는 사실을 좀 더 명확하게 설명할 수 있는 공식을 만들기 위해 노력하고 있다. 야구에서는 득점이 많은 팀은 적은 팀보다 이길 확률이 높아지고 그 반대도 성립하는 데, 야구통계에서는 득점과 실점을 바탕으로 여러 가지 기대승률 지표를 개발하였으며 대표적인 것이 피타고라스 승률과 선형 승률이다. 하지만 피타고라스 승률에 의해 추정된 승률의 합은 기초적인 확률의 성질을 충족시키지 못하는 단점이 있으나 팀대 팀의 결과인 쌍별 추정으로 접근하면 이와 같은 문제점을 근본적으로 해결할 수 있다.

본 연구에서는 쌍별 피타고라스 추정이 한국프로야구에서도 바람직한지를 알아보기 위해 효율성을 구체적으로 살펴보았으며 쌍별 피타고라스 승률방법에 필요한 지수의 선택 문제를 논의했다. 그 결과 지수 값 1.83을 사용한 쌍별 피타고라스 추정은 기존의 어떤 피타고라스 방법보다도 더 바람직하다고 결론을 내릴 수 있었다. 연구의 결과와 관련된 제한점으로는 한국프로야구 경기 결과에 대한 데이터 축적이 오래 되어 있지 못해서 최근 몇 년 동안의 결과만을 사용하여 쌍별 피타고라스 승률 추정에 대한 결론을 내렸기 때문에 좀 더 포괄적인 비교를 못한 점이 있으며, 한국프로야구 전 경기 데이터를 모두 사용하였기 때문에 이상치나 영향점을 제거하면 좀 더 정밀한 결론을 내릴 수 있을 것으로 간주된다. 또한 쌍별 승률 추정방법을 축구, 농구 및 하키와 같은 스포츠에 적용하여도 기존 결과들보다 상대적으로 바람직할 것으로 예측되며 이 부분은 향후 연구과제로 남겨둔다.

References

Cochran, J. J. (2008). The optimal value and potential alternatives of Bill James Pythagorean method of baseball. *STAtOR*, **2** , 2008.

Davenport, C. and Woolner, K. (1999). Revisiting the Pythagorean theorem: Putting Bill James' Pythagorean theorem to the test. *The Baseball Prospectus*, `http://www.baseballprospectus.com/article.php?articleid=342`.

Heumann, J. (2016). An improvement to the baseball statistic "Pythagorean Wins". *Journal of Sports Analytics*, **2**, 49-59.

James, B. (1980). The Bill James abstract, self-published, Lawrence, KS.

Jones, M. and Tappin, L. (2005). The Pythagorean theorem of baseball and alternative models. *The UMAP Journal*, **26** .

Kim, S. K. and Lee, Y. H. (2016). The estimation of winning rate in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **27**, 653-661.

Lee, J. T. (2014). Estimation of exponent value for Pythagorean method in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 493-499.

Lee, J. T. (2015). Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **26**, 653-659.

Lee, J. T. (2016a). Convergence characteristics of Pythagorean winning percentage in baseball. *Journal of the Korean Data & Information Science Society*, **27**, 1477-1485.

Lee, J. T. (2016b). A comparison of formulas to predict a team's winning percentage in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **27**, 1585-1592.

Journal of the Korean Data & Information Science Society
2017, **28**(2), 309–316

http://dx.doi.org/10.7465/jkdi.2017.28.2.309
한국데이터정보과학회지

Efficiency of pairwise winning percentage estimators in Korean professional baseball[†]

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 10 February 2017, revised 5 March 2017, accepted 13 March 2017

Abstract

In baseball, estimation of winning percentage is critical and many studies for this topic have been actively performed. Pairwise winning percentage estimation using Pythagorean winning percentages of individual teams against other individual teams has the property that the sum of estimated winning percentage totals must be a constant. In this paper, we consider two types of pairwise estimation including linear formula and Pythagorean formula to the Korean baseball data of seasons from 2013 to 2016 under the criterions of RMSE and MAD. In conclusion, pairwise Pythagorean methods have the smaller RMSE and MAD than traditional Pythagorean methods. We suggest the optimal pairwise Pythagorean formula with a fixed exponent. Also we show that there are very little differences of RMSE and MAD between variation in exponent values.

Keywords: Linear, MAD, pairwise, Pythagorean, RMSE, winning percentage.

[†] The present research was conducted by the research fund of Dankook University in 2017.
[1] Professor, Department of Applied Statistics, Dankook University, Yongin 16890, Korea.
E-mail: jtlee@dankook.ac.kr