Journal of the Korean Data & Information Science Society
2015, **26**(3), 653–659

http://dx.doi.org/10.7465/jkdi.2015.26.3.653
한국데이터정보과학회지

한국프로야구에서의 피타고라스 정리의 정확도 측정[†]

이장택[1]

[1]단국대학교 응용통계학과

접수 2015년 3월 16일, 수정 2015년 4월 6일, 게재확정 2015년 4월 10일

요 약

야구의 피타고라스 정리는 야구의 승률을 추정하는 방법으로 오랜 기간 동안 타당성이 입증되고 또 활용되고 있다. 본 연구에서는 2005년부터 2014년 사이의 한국프로야구 팀대 팀 전체기록을 이용하여 실제승률과 피타고라스 정리에 의해 추정된 기대승률의 차이가 발생하는 원인을 회귀모형을 이용하여 살펴보았다. 기대승률과 실제승률의 차이가 큰 경우는 득점과 실점의 분포가 특이하다는 가정 아래에서 종속변수는 실제승률과 기대승률의 차이, 독립변수로는 게임당 득점 및 실점의 평균, 표준편차, 변동계수를 각각 이용하였다. 그 결과 실제승률과 기대승률의 차이에는 게임당 실점의 표준편차와 변동계수가 영향을 미치며 게임당 득점의 영향은 없는 것으로 나타났다.

주요용어: 득점, 승률, 실점, 피타고라스 정리, 한국프로야구.

1. 머리말

야구는 승점제인 축구와는 달리 각 팀의 순위를 승률로 결정한다. 따라서 각 팀의 입장에서 주어진 야구 환경을 이용하여 승률을 예측하는 문제는 매우 중요한 문제인데, 야구의 승률을 계산하는 공식 중에는 James (1982)가 제안한 야구의 피타고라스 정리라는 방법이 있다. 이것은 수학의 피타고라스 정리와 유사한 데가 있다고 해서 붙여진 이름으로 야구에서 수많은 시즌의 누적된 기록들을 이용하여 야구에 대한 객관적 지식을 찾고자 하는 세이버메트릭스 (sabermetrics)의 핵심적인 이론적 근거로 간주된다.

야구의 피타고라스 정리를 소개하면 승률 ($Wpct$)은 다음 식 (1.1)과 같이 득점 (RS)의 제곱을 득점 (RS)의 제곱과 실점 (RA)의 제곱의 합으로 나눈 것으로 정의되며, 승리한 게임의 수는 게임의 수 (G)에 승률을 곱해서 구할 수 있다.

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

James는 그 후 미국 메이저리그인 경우에 식 (1.2)와 같이 득점과 실점의 지수를 2에서 1.83으로 낮추어 설명하는 것이 좀 더 승률을 잘 예측할 수 있다고 설명하였다. 지수 1.83은 통계학에서 일반적으로 많이 사용되는 추정량의 선택기준인 제곱근 평균제곱오차 (root mean square error; RMSE)를 최소화하는 값이며, 적용기간의 선택에 따라 지수 추정치의 차이가 다소 있을 수 있다.

$$Wpct = \frac{RS^{1.83}}{RS^{1.83} + RA^{1.83}} \tag{1.2}$$

[†] 이 연구는 2015학년도 단국대학교 대학연구비 지원으로 연구되었음.
[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

한국프로야구에서도 야구의 피타고라스 정리는 미국 메이저리그의 경우와 비슷하게 적용되는데, Lee와 Kim (2006a)에 의하면 프로야구 원년부터 2005년까지의 데이터를 이용한 경우에는 지수가 1.87이 가장 적당하였으나, Lee (2014a)의 프로야구 원년부터 2013년까지의 데이터를 이용한 경우에는 지수가 1.82가 가장 적당하였다. 따라서 한국프로야구에도 데이터가 점점 더 누적될수록 미국 메이저리그의 피타고라스정리 지수 값과 유사해짐을 확인할 수 있다. 야구의 피타고라스 정리와 연관된 국외 연구들을 살펴보면 Miller (2006)는 몇 가지 가정과 와이블 분포를 이용하여 피타고라스 정리가 이론적으로 성립함을 보였으며, Davenport와 Woolner (1999), Cochran (2008)은 각각 피타고라스의 정리에 필요한 최적지수 추정문제를 다루었다. 또한 국내연구로는 Lee와 Kim (2006a, 2006b)은 한국프로야구에서 피타고라스 정리 지수 값 1.87을 사용하여 잘 적용된다고 설명하였으며, 한국여자프로농구와 프로축구에서도 지수 값을 각각 10.8과 1.378을 사용하여 승률을 잘 추정할 수 있다고 밝혔다. 또한 Lee (2014a)는 기존의 여러 가지 방법보다 우수한 피타고라스 지수를 추정하는 회귀직선을 새롭게 제안하였다. 그밖에 한국프로야구에 관한 일반연구를 몇 가지 소개하면 출루율과 장타율이 득점에 미치는 영향을 연구한 Kim (2012), 한국프로야구 타자들에 대한 세이버메트릭스 지수 값을 이용하여 선수들의 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012), 한국프로야구에서 출루율 계수의 추정문제를 다룬 Lee (2014b), 한국프로야구에서 투수평가지표를 제시한 Lee (2014c) 등이 있다.

오늘날 야구의 피타고라스 정리는 미국 메이저리그 공식 홈페이지인 mlb.com, 메이저 리그 야구선수에 관한 정보를 제공하는 baseball-reference.com, 미국의 스포츠전문 채널인 ESPN 등에서 모두 인용되는 유명한 야구관련 수식이다. 하지만 피타고라스 정리는 단지 팀 승률이 어떻게 진행되어 갈 것이라는 것을 예상해보는 이론이기 때문에 타당도의 정도는 무시되고 언급되는데, 본 연구에서는 한국프로야구 팀대 팀의 경기결과를 이용하여 실제승률과 피타고라스 정리에 의한 기대승률의 차이가 발생하는 원인을 살펴보았다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 승률의 정의, 지수추정, 분석데이터 및 통계분석에 대하여 언급하였으며, 3절에서는 실제승률과 피타고라스 정리에 의한 기대승률의 차이에 대한 기술통계량과 회귀추정식을 제안하였으며 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 연구방법

2.1. 승률의 정의 및 지수의 추정

한국프로야구에서 사용된 팀의 승률은 무승부제외 승률제, 무승부포함 승률제, 다승제가 있다 (Kim, 2011). 하지만 본 연구에서는 승률의 정의로 1987시즌부터 1997시즌까지 사용한 무승부포함 승률제인 식 (2.1)을 사용하였는데, $Wpct$는 승률, W는 승리한 게임 수, L은 패배한 게임 수, T는 무승부 게임 수를 각각 의미한다.

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

본 연구에서 무승부제외 승률을 사용하지 않은 이유는 한국프로야구에 대한 모든 공식기록들은 무승부인 경우도 포함하여 집계되었기 때문에 오랜 기간 동안의 데이터를 무승부를 제외하고 재집계하는 것은 불가능하고 무승부포함의 값과 유사한 값을 제공하기 때문이다. 또한 비록 야구의 피타고라스의 정리에 사용되는 승률이 $W/(W+L)$을 사용하지만 W^*와 L^*를 각각 $W^* = W + 0.5 \times T$, $L^* = L + 0.5 \times T$로 두면, 식 (2.1)은 $W^*/(W^* + L^*)$로 표현되어 James (1982)의 피타고라스 정리를 적용할 수 있다. 한편 지수의 추정은 다음과 같이 할 수 있는데, 만일 지수가 γ인 피타고라스 정리가 성립한

다면 다음 식 (2.2)가 성립함을 알 수 있다 (Lee, 2014a).

$$\log(W^*/L^*) = \gamma \log(RS/RA) \tag{2.2}$$

따라서 절편이 없는 단순회귀모형에서 주어진 데이터와 최소제곱법을 이용하여 γ 값을 추정할 수 있다.

2.2. 사용된 데이터와 통계분석

분석에 사용된 표본은 2005년부터 2014년 사이에 있었던 한국프로야구 팀대 팀 경기결과 전체기록을 이용하였는데, 자료의 개수는 모두 5,296개이며 자료의 출처는 롯데자이언츠 홈페이지 `http://www.giantsclub.com`이다. 통계패키지 SPSS 21K를 이용하여 연도별 각 구단의 승률과 야구의 피타고라스 정리에 의한 기대승률을 구하고 승률차이와 득점과 실점의 통계량과의 관계를 상관분석과 회귀분석을 통하여 규명하고 이를 기반으로 실제승률과 기대승률의 차이를 설명하는 통계모형을 구축하였다.

3. 분석결과 및 논의

3.1. 승률차이에 대한 기술통계량

Table 3.1은 퍼센트로 계산한 승률차이 (DIF)와 DIF에 절대값을 취한 절대값 승률차이 (ADIF)에 대한 기술통계 값을 보여준다. 이 경우 DIF는 실제승률에서 피타고라스정리에 의해 추정된 기대승률을 뺀 값으로 정의하며, 기대승률은 지수값을 프로야구 원년부터 2014년까지의 팀별 승률, 득점, 실점 데이터를 이용하여 추정한 최적지수값 1.834를 사용하여 계산하였는데, 승률차이의 최대값은 6.24%, 최소값은 5.08%이다. 또한 ADIF의 평균값을 보면 야구의 피타고라스 정리를 이용한 추정값은 팀당 평균적으로 대략 1.95% 가량의 오차가 생기는 것을 알 수 있는데, 야구의 피타고라스 정리가 승률을 대단히 잘 예측한다고 할 수 있겠다. 또한 DIF와 ADIF의 왜도와 첨도의 값이 0에서 크게 벗어나지 않기 때문에 두 변수 모두 정규분포와 유사한 형태를 취한다고 할 수 있겠다.

Table 3.1 Descriptive statistics for baseball statistics

	Mean	Standard Deviation	Maximum	Minimum	Skewness	Kurtosis
DIF	−0.110	2.415	6.239	−5.077	0.237	−0.321
ADIF	+1.950	1.413	6.239	+0.029	0.720	+0.029

Table 3.2은 절대값 승률차이 (ADIF)가 큰 상위 10개 팀의 결과인데 ADIF의 크기와 부호를 주어진 득점과 실점의 수치로 판단하기는 쉽지 않은 듯하다.

Table 3.2 Team results of the top 10 ADIF scores

Rank	Year	Team	RS	RA	WPCT	EWPCT	DIF
01	2011	Hanhwa	568	727	45.113	38.873	+6.239
02	2009	Hanhwa	657	805	35.714	40.791	−5.077
03	2014	Hanhwa	619	889	39.063	33.987	+5.076
04	2008	Samsung	557	596	51.587	46.901	+4.686
05	2013	NC	512	551	42.188	46.639	−4.452
06	2006	Lotte	488	541	40.873	45.287	−4.414
07	2014	Nexen	841	716	61.719	57.325	+4.394
08	2014	Lotte	716	719	45.703	49.808	−4.105
09	2007	LG	532	600	48.413	44.507	+3.906
10	2008	Lotte	624	518	54.762	58.454	−3.692

Table 3.3 Correlations for statistics with DIF

	RS/G_M	RA/G_M	RS/G_{SD}	RA/G_{SD}	RS/G_{CV}	RA/G_{CV}
Correlation	0.013	0.040	−0.101	0.311**	−0.118	0.417**
p-value	0.905	0.721	+0.367	0.004	+0.292	0.000

*$p < 0.05$, **$p < 0.01$

야구의 피타고라스 정리가 득점과 실점의 함수이므로 득점과 실점의 기술통계량으로 본 연구에서는 평균, 표준편차 및 변동계수 (coefficient of variation)를 고려하였다. 변동계수는 득점과 실점의 팀별, 연도별 평균이 다르기 때문에 데이터의 산포를 비교하기 위해서 사용하였는데, 그 이유는 게임의 승부가 경기 초반전에 거의 확정지어지면 나머지 야구 횟수에서는 양 팀이 최선을 다하지 않은 경우가 빈번하게 벌어질 것이며 이런 경우가 DIF가 크게 나타날 것이라고 가정했기 때문이다. Table 3.3은 DIF와 여러 가지 통계량과의 피어슨 상관계수를 보여주는데, 이 경우 RS/G와 RA/G는 각각 게임당 득점과 실점을 의미하며, 또한 사용된 아래첨자의 의미는 M은 평균, SD는 표준편차, CV는 변동계수를 각각 의미한다. DIF와 상관이 큰 통계량은 유의수준 1%에서 게임당 실점의 표준편차 및 변동계수가 유의하였고, 게임당 득점에 관한 통계량은 유의수준 10%에서도 유의한 것이 전혀 없었다.

3.2. 승률차이에 대한 회귀추정식

실제승률과 피타고라스 기대승률의 차이를 설명하는 모형을 생성하기 위하여 종속변수를 DIF, Table 3.3의 득점과 실점의 6개 통계량을 독립변수로 사용하여 회귀모형을 설정하였다. 데이터는 2005년부터 2014년 사이의 10년간의 각 팀별 승률데이터 82개, 변수선택법은 단계선택법을 이용하였으며, 야구에도 예측과 전혀 다른 결과가 나타나고 이런 현상은 종속변수에 대한 이상치로 나타날 가능성이 크기 때문에 종속변수의 이상치를 제거하기 위하여 외적 스튜던트화 잔차 (externally studentized residual)의 절대값이 2보다 큰 데이터를 모두 삭제한 후에 모형을 구축하였으며 그 결과 모형은 76개의 데이터로 추정되었다. Table 3.4는 제안된 모형의 적합도를 보여주는데, 결정계수는 34.2%, 더빈-왓슨 통계량이 2.027이므로 1차자기상관은 없다고 할 수 있다.

Table 3.4 Model summary for regression analysis

R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
0.342	0.324	1.764	2.027

Table 3.5 Estimated regression model coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
	B	Std. Error	Beta			
Constant	−18.732	3.032		−6.178	0.000	
RA/G_{CV}	20.004	3.593	0.529	5.568	0.000	1.000
RA/G_{SD}	1.384	0.533	0.247	2.599	0.011	1.000

회귀분석 결과는 분산분석의 p값은 p <0.001로 유의수준 1%에서 매우 유의한 것으로 나타났으며, Table 3.5는 추정된 회귀식, 표준화 회귀계수 및 VIF를 보여주는데, 2개의 변수 모두 VIF의 값이 10보다 작아서 다중공선성의 문제는 없는 것으로 나타났다. 따라서 추정된 회귀식은 다음 식 (3.1)과 같이 기술할 수 있으며 표준화 회귀계수를 이용하면 중요도가 RA/G_{CV}가 2배 이상 RA/G_{SD}보다 의미가 있다고 나타났다. 결과적으로 Table 3.3의 유의한 2개의 변수가 모두 회귀모형 구축 시에 사용되어진

것으로 나타났다.

$$\widehat{DIF} = -18.73 + 20.00RA/G_{CV} + 1.38RA/G_{SD} \tag{3.1}$$

Figure 3.1은 DIF와 식 (3.1)을 이용한 추정된 DIF의 관계를 보여주는 산점도이며 두 변수의 수치는 모두 퍼센트를 단위로 하였다.

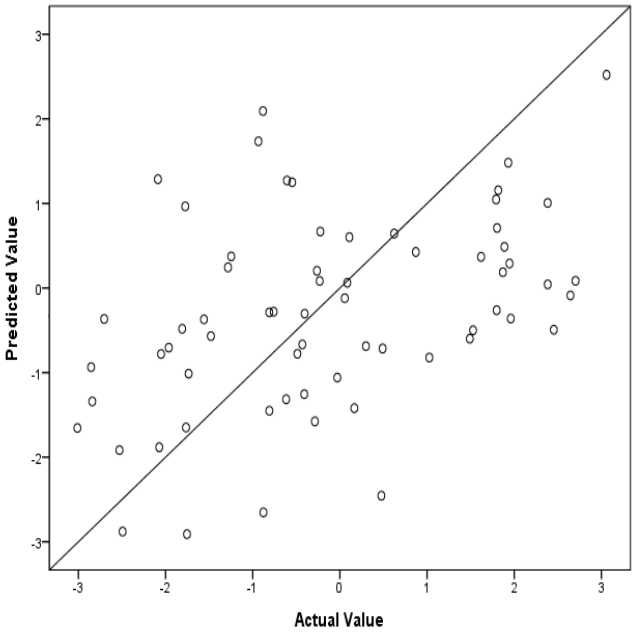

Figure 3.1 Scatterplot of predicted vs. actual values

지금까지 피타고라스 승률이 실제 승률보다 어떤 경우에 과대하게 추정하며, 또 어떤 경우에 과소하게 추정하는 지를 살펴보았는데, 식 (3.1)을 이용하여 DIF가 양수인 경우에 양수라고 판정하고, 음수인 경우에 음수라고 판정하는 분류정확률을 살펴보는 것은 매우 흥미로운 일이다. Table 3.6은 회귀분석에서 사용했던 이상치를 제외한 데이터 76개에 대한 교차표와 전체 데이터 82개에 대한 교차표를 보여준다. 이상치를 제외한 데이터는 분류정확률이 69.73%이나 전체 데이터인 경우에는 분류정확률이 67.07%로 약간 감소한다.

Table 3.6 Crosstabulation of DIF and predicted DIF for two cases

			Predicted DIF		Total
			< 0	> 0	
Regression	DIF	< 0	34	11	36
		> 0	12	19	34
	Total		46	30	76
Total	DIF	< 0	34	12	41
		> 0	15	21	41
	Total		49	33	82

4. 결론

본 연구에서는 한국프로야구에서 실제 승률과 피타고라스 기대승률의 차이가 어떤 요인에 의해 주로 연관되는 지를 살펴보았다. 그 결과 한국프로야구에서 실제 승률이 피타고라스 승률보다 커지는 경우는 게임당 실점의 표준편차와 변동계수가 큰 경우로 나타났다. 이 사실을 상식적인 선상에서 설명하면 실제 승률이 기대 승률보다 높은 현상은 대개 불펜이 강한 팀에서 나타날 가능성이 많은데, 일단 리드하면 득점이 많지 않아도 승리를 지킬 수 있기 때문이다. 즉 마음만 먹으면 실점을 많이 할 수도 있고 적게 할 수도 있는 능력을 보유한 팀들이 실제 승률이 기대승률보다 커질 수 있다고 설명이 가능하다.

야구통계 전문가 James가 메이저리그 팀들의 과거 성적을 정리하다가 발견한 야구의 피타고라스 정리에 의한 기대승률은 실제승률과 유사하지만 그 차이는 발생한다. 그 차이는 우리나라의 경우에 평균적으로 대략 1.95% 정도 나는데, 발생되는 이유가 설명 불가능한 랜덤한 현상인지 특정한 원인들이 있는지는 알 수가 없다. 이런 이유로 본 연구에서는 랜덤성을 통계적 규칙으로 설명하여 보려고 노력하였다. 하지만 고려된 독립변인들보다 더 많은 원인들이 있을 수도 있으며, 따라서 본 연구는 이런 노력의 초기단계 수준이라고 할 수 있다. 향후 연구과제로 좀 더 다양한 변수들을 고려하면 보다 나은 결과를 제공할 수 있을 것으로 간주되어지며 본 연구의 접근방법은 다른 프로스포츠에도 적용할 수 있을 것으로 간주된다.

References

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

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`.

James, B. (1982). *The Bill James baseball abstract*, Ballantine Books, New York.

Kim, H. J. (2011). Suggestion of a new method of computing percentage of victories for the Korean professional baseball. *The Korean Journal of Applied Statistics*, **24**, 1139-1148.

Kim, H. J. (2012). Effects of on-base and slugging ability on run productivity in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **23**, 1065-1074.

Lee, J. T. and Kim, Y. T. (2006a). A study on the estimation of winning percentage in Korean pro-baseball. *Journal of the Korean Data Analysis Society*, **8**, 857-869.

Lee, J. T. and Kim, Y. T. (2006b). Estimation of winning percentage in Korean pro-sports. *Journal of the Korean Data Analysis Society*, **8**, 2105-2116.

Lee, J. T. (2014a). 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. (2014b). Estimation of OBP coefficient in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **25**, 357-363.

Lee, J. T. (2014c). Pitching grade index in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 485-492.

Miller, S. J. (2006). A derivation of the pythagorean won-loss formula in baseball. *By the Numbers*, **16**, 40-48.

Seung, H. B. and Kang, K. H. (2012). A study on relationship between the performance of professional baseball players and annual salary. *Journal of the Korean Data & Information Science Society*, **23**, 285-298.

Journal of the Korean Data & Information Science Society
2015, **26**(3), 653–659

http://dx.doi.org/10.7465/jkdi.2015.26.3.653
한국데이터정보과학회지

Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball[†]

Jangtaek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 16 March 2015, revised 6 April 2015, accepted 10 April 2015

Abstract

The Pythagorean formula for baseball postulated by James (1982) indicates the winning percentage as a function of runs scored and runs allowed. However sometimes, the Pythagorean formula gives a less accurate estimate of winning percentage. We use the records of team vs team historic win loss records of Korean professional baseball clubs season from 2005 and 2014. Using assumption that the difference between winning percentage and pythagorean expectation are affected by unusual distribution of runs scored and allowed, we suppose that difference depends on mean, standard deviation, and coefficient of variation of runs scored per game and runs allowed per game, respectively. In conclusion, the discrepancy is mainly related to the coefficient of variation and standard deviation for run allowed per game regardless of run scored per game.

Keywords: Korean professional baseball, Pythagorean method, runs allowed, runs scored, winning percentage.

[†] The present research was conducted by the research fund of Dankook University in 2015.
[1] Professor, Department of Applied Statistics, Dankook University, Gyeonggi-do 448-701, Korea. E-mail: jtlee@dankook.ac.kr