Journal of the Korean Data & Information Science Society
2014, **25**(3), 493–499

http://dx.doi.org/10.7465/jkdi.2014.25.3.493
한국데이터정보과학회지

한국프로야구에서 피타고라스 지수의 추정[†]

이장택[1]

[1]단국대학교 응용통계학과

접수 2014년 2월 3일, 수정 2014년 2월 19일, 게재확정 2014년 3월 12일

요 약

야구의 승률은 총득점의 제곱을 총득점의 제곱과 총실점의 제곱의 합으로 나눈 것으로 추정된다는 야구의 피타고라스 정리에 대하여 많은 연구들이 활발하게 진행되고 있다. 본 연구에서는 피타고라스 정리에 사용되는 지수에 대한 새로운 추정방법을 제안하며 평균제곱오차의 제곱근 (root mean squared error; RMSE)을 이용하여 널리 알려진 추정방법들과 상대적 효율성을 비교하였다. 사용된 데이터는 1982년부터 2013년 사이의 모든 한국프로야구 기록이며, 그 결과 제안된 방법은 기존의 방법보다 RMSE 관점에서 바람직하다고 간주된다.

주요용어: 승률, 평균제곱오차의 제곱근, 피타고라스 정리.

1. 머리말

야구는 축구나 배구에 비하여 통계적인 접근이 용이한 스포츠이며, 우리는 통계를 통해 야구를 이해할 수 있다. 투수가 던지는 하나의 공에도 여러 가지 결과들이 기록되며, 타자도 마찬가지이다. 비록 한 경기의 결과가 시즌 전체에 미치는 영향은 작지만 수많은 시즌의 누적된 기록들은 통계적으로 신뢰할 수 있는 결과를 제공하는데 세이버메트릭스 (sabermetrics)는 이렇게 다년간 쌓인 통계데이터를 이용하여 야구에 대한 객관적 지식을 찾고자 하는 연구를 하는 분야이며, 세이버메트릭스 방법으로 데이터를 분석하는 사람들을 세이버메트리션 (sabermetrician)이라 부른다.

세이버메트리션들은 야구경기에서 득점이야말로 선수의 성취도를 평가하는 가장 좋은 방법이라고 인식하고 있다. 하지만 야구에 관심이 있는 통계학자들이 처음으로 1954년 득점과 팀 순위 사이의 상관관계를 논증했을 때는 통계학자들도 이미 결정된 팀 순위와 득점 사이의 상관관계를 인식하는데 그쳤으며, 득점 및 실점의 합으로 승률을 추정할 수 있다는 사실은 전혀 예상하지 못했다. 야구의 승률추정 문제가 팬들의 관심을 끌게 된 가장 큰 이유는 아마도 James (1980)의 역할이라고 할 수 있다. 그는 승률 ($Wpct$)은 다음과 같이 득점 (RS)의 제곱을 득점 (RS)의 제곱과 실점 (RA)의 제곱의 합으로 나눈 것으로 설명할 수 있으며, 이 식을 야구경기에 있어서 피타고라스 정리라고 불렀다.

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

또한 그는 실제 승률과 공식에 의한 승률의 차이를 보정하기 위하여 식 (1.1)을 일반화하여 득점과 실점의 γ승을 고려한 식 (1.2)와 같은 승률 추정식을 만들었으며, 이 식을 야구경기에 있어서 일반화 피타고라스의 정리라고 일컫는다. 일반적으로 지수 γ의 값은 RMSE와 같은 판정기준을 최소화하는 값으로

[†] 이 연구는 2014학년도 단국대학교 대학연구비 지원으로 연구되었음.
[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

정해지는데, James는 미국 메이저리그인 경우에 총득점과 총실점의 지수를 2에서 1.83으로 약간 낮추어 설명하는 것이 좀 더 바람직하다고 설명하였다.

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

야구의 피타고라스 정리에 대한 국외연구들을 살펴보면 Miller (2006)는 피타고라스 정리가 몇 가지 가정과 와이블 분포를 사용하여 성립함을 이론적으로 보였으며, Acharya (2006)은 메이저리그 2005년과 2006년 데이터를 이용하여 승률 자체보다 패배 대비 승리에 대한 비율을 추정하였다. Davenport와 Woolner (1999), Cochran (2008)은 각각 본 연구의 방향과 정확히 일치하는 피타고라스의 정리에 필요한 최적지수 γ의 추정문제를 다루었으며, γ의 값은 1.74부터 2.0 사이의 값으로 게임당 발생하는 득점의 수에 종속된다고 밝혔다. 한편 국내연구로는 Lee와 Kim (2005, 2006a, 2006b)은 한국프로야구에서도 야구경기의 피타고라스 정리가 지수 값을 2대신 1.87을 사용하여 잘 적용된다고 설명하였으며 또한 우리나라의 대표적인 프로스포츠인 여자프로농구와 프로축구에서도 지수 값을 각각 10.8과 1.378을 사용하여 승률을 잘 추정할 수 있다고 밝혔다. 한국프로야구에 관한 연구는 최근에도 활발한데 시계열모형을 이용하여 관중 수 예측을 다룬 Lee와 Bang (2010), 타자들에 대한 세이버메트릭스 지수 값을 이용하여 선수들의 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012), 출루율과 장타율이 득점에 미치는 연구를 한 Kim (2012) 등이 있다.

오늘날 야구의 피타고라스 정리는 미국의 스포츠전문 채널인 ESPN, 미국 메이저리그 공식 홈페이지인 mlb.com, 메이저 리그 야구의 모든 선수에 관한 정보를 제공하고 주요 언론 매체에서 자주 이용하는 웹사이트인 baseball-reference.com 등에서 모두 인용되는 유명한 야구관련 수식이다. 이와 같이 대중의 관심은 점점 늘어나지만 정작 야구의 피타고라스 정리에 관한 발전은 매우 더디다고 할 수 있는데, 앞에서 언급한 선행연구들은 RMSE 판정기준에서 거의 유사한 효율성을 보이며 상대적 효율성이 우수한 어떤 추정량을 제시하지는 못했다. 그리고 지수 γ는 단지 게임당 발생하는 득점의 수에 종속된다고 하였는데, 예를 들어 게임당 발생된 평균 득점의 수가 모두 9점인 경우에서 팀의 평균득점이 5점, 평균실점이 4점인 경우보다 평균득점이 8점, 평균실점이 1점인 경우가 승률이 높아질 가능성은 훨씬 크다고 할 수 있다. 이와 같은 이유로 본 연구에서는 승률의 표준편차를 고려하여 지수 γ의 추정문제를 좀 더 개선하고자 한다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 분석데이터 및 승률의 정의, 통계분석 및 모형평가 기준에 대하여 언급하였으며, 3절에서는 피타고라스 정리에 필요한 연도별 최적지수 값을 구하고, 새로운 지수 추정 방법을 제안하며, 기존 방법들과 비교하여 제안된 방법의 상대적인 효율성을 검증하였다. 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 연구방법

2.1. 분석데이터 및 승률의 정의

모형설정을 위하여 사용된 표본은 한국프로야구 기록실에 기록되어 있는 1982년부터 2013년 사이에 있었던 전 경기로 전부 244개 팀의 결과이며, W는 승리한 게임 수, L은 패배한 게임 수, T는 무승부 게임 수라고 표기할 때, 연구에서 사용된 승률의 의미는 한국야구위원회 (KBO)에서 1987시즌부터 1997시즌까지 사용한 승률 ($Wpct$)의 정의인 다음 식을 사용하였다.

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

본 연구에서 일반적으로 사용하는 승률의 정의인 무승부를 제외한 $W/(W + L)$을 사용하지 않은 이유는 한국프로야구에 대한 모든 공식적인 기록들은 모두 무승부인 경우도 포함하여 집계되었기 때문에 수십

년간의 데이터를 무승부를 제외하고 득점과 실점을 재집계하는 것은 매우 번거롭고 또한 무승부를 포함하지 않는 승률의 값과 매우 유사한 값을 제공하기 때문이다.

한편 W^*와 L^*를 각각 $W^* = W + 0.5 \times T$, $L^* = L + 0.5 \times T$로 두면, 위의 승률은 식 (2.2)와 같이 표시된다.

$$Wpct = \frac{W + 0.5 \times T}{W + T + L} = \frac{W + 0.5 \times T}{W + 0.5 \times T + L + 0.5 \times T} = \frac{1}{1 + L^*/W^*} \tag{2.2}$$

따라서 우리나라의 경우엔 W대신 W^*, L대신 L^*를 적용하여 승률 ($Wpct$)을 추정할 수 있다.

2.2. 통계분석 및 모형평가기준

통계패키지 SPSS 21K와 한국프로야구 전체 데이터를 이용하여 승률을 세이버메트리션들의 선행연구와 마찬가지로 기술통계와 회귀분석 기법을 사용하여 추정하였는데, 가설검정에서는 유의수준 5%와 1%에서 통계적 유의성 검정을 하였다. 또한 본 연구에서는 제안된 모형들의 효율성을 서로 비교하기 위하여 일반적으로 많이 사용되는 추정량 선택기준인 평균제곱오차의 제곱근 (root mean square error; RMSE)을 사용하였다. RMSE는 i번째 게임의 승률을 w_i, 승률추정량을 $\widehat{w}_i$, 총 게임 수를 n이라고 두면, 식 (2.3)과 같이 정의되며, 추정량들의 RMSE를 구해서 구한 값이 가장 작은 것이 제일 좋은 추정량이라 할 수 있다. 본 연구의 RMSE 값은 승률을 퍼센트로 나타낸 경우의 값이다.

$$RMSE = \sqrt{\sum_{i=1}^{n}(w_i - \widehat{w}_i)^2/n} \tag{2.3}$$

3. 연구결과 및 논의

3.1. 피타고라스 정리에 필요한 시즌별 최적지수 값

만일 일반화 피타고라스 정리가 성립한다면 다음 식이 성립함을 알 수 있다.

$$Wpct = \frac{1}{1 + (L^*/W^*)^\gamma} = \frac{1}{1 + (RA/RS)^\gamma} \tag{3.1}$$

그러므로 로그를 사용하여 다음과 같이 비선형 식을 선형 식으로 바꿀 수 있다.

$$\log(W^*/L^*) = \gamma \log(RS/RA) \tag{3.2}$$

따라서 우리는 주어진 데이터와 최소제곱법을 이용하여 절편이 없는 단순회귀모형을 고려하여 γ값을 추정할 수 있다.

Figure 3.1은 피타고라스 정리에 필요한 최적지수 값을 시즌별로 추정한 결과를 보여준다. 최적지수 값의 최솟값은 1983년 1.06, 최댓값은 1999년 2.34로 변화가 상당하다고 할 수 있다. 참조선은 1.82로 전체 데이터를 이용하여 추정한 최적지수 값을 의미한다.

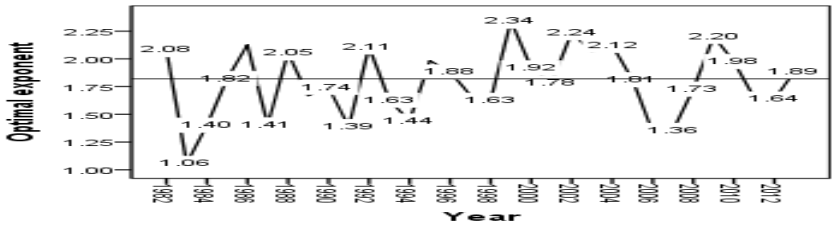

Figure 3.1 Korean Pro-baseball optimal exponents by year

선행연구에 의하면 실제로 최적지수 값은 경기당 발생득점 (runs per game; RPG)과 가장 연관성이 크다고 알려져 있는데, 구체적으로 G를 게임 수라고 할 때, RPG는 $(RS + RA)/G$로 나타낼 수 있다. Figure 3.2은 한국프로야구의 경우, RPG와 최적지수값을 표시한 산점도이다. 대략적으로 RPG가 증가할수록 최적지수 값도 증가함을 알 수 있으며, 실제로 두 변수 사이의 상관계수는 0.468, p값은 $p < 0.001$로 유의수준 1%에서 매우 유의한 것으로 나타났다.

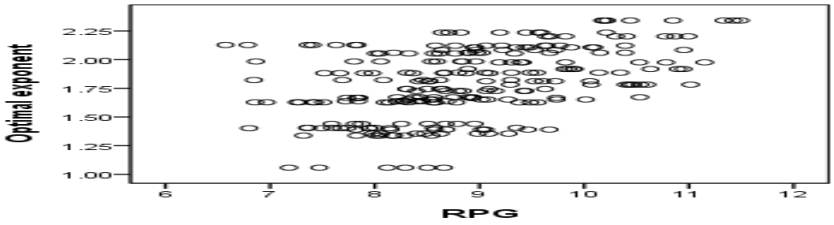

Figure 3.2 RPG - optimal exponent scatter diagram

Figure 3.2를 이용하면 최적지수 값을 RPG의 함수로 간주하여 추정하면 좀 더 효율적인 결과를 제공할 것으로 간주되는데, 이와 같은 아이디어를 이용하여 수많은 세이버메트리션들이 이 문제를 다루었으며, 지금도 인터넷 정보마당에는 이 토픽에 대한 여러 가지 답을 찾을 수가 있다. 하지만 온라인 백과사전인 위키피디아는 이 문제에 대한 유명한 두 가지 방법을 소개하고 있는데, 첫째는 Davenport와 Woolner (1999)의 방법으로 메이저리그 데이터를 이용하여 최적지수와 RPG 사이에는 다음 식 (3.3)이 성립한다고 주장하였다.

$$\gamma_D = 1.5 \log(RPG) + 0.45 \tag{3.3}$$

둘째는 Smyth와 Patriot (2004)의 결과로 최적지수와 RPG 사이의 관계는 식 (3.4)와 같다고 설명하였다.

$$\gamma_S = RPG^{0.287} \tag{3.4}$$

본 연구에서는 위의 두 가지 방법보다 RMSE의 측면에서 보다 효율적인 추정방법을 제안하려고 한다. Figure 3.3을 살펴보면 퍼센트로 표시된 연도별 승률의 표준편차와 최적지수 값은 서로 연관이 있음을 알 수 있는데, 승률의 표준편차가 크다는 것은 승률의 차이가 크다는 것을 의미하기 때문에 게임철학에서 시소게임과 다르다고 할 수 있겠다. 따라서 제안되는 새로운 지수 γ_N는 RPG와 승률의 표준편차 σ의 함수로 간주하는 것이 바람직할 것으로 여겨지며, 이 사실을 확인하기 위하여 독립변수로 RPG와 $\sqrt{\sigma}$를 고려하였는데, σ대신 $\sqrt{\sigma}$를 고려한 것은 Figure 3.3으로부터 σ의 제곱근 변환 관계가 타당하다고 생각할 수 있기 때문이다. 실제로 독립변수를 RPG와 σ로 간주하는 경우의 결정계수는 49.3%, RPG와 $\sqrt{\sigma}$로 간주하는 경우의 결정계수는 51.4%로 나타났으며, 또한 RPG와 $\sqrt{\sigma}$는 분산팽창계수는 1.01로 다중공선성 문제점이 없다고 할 수 있다.

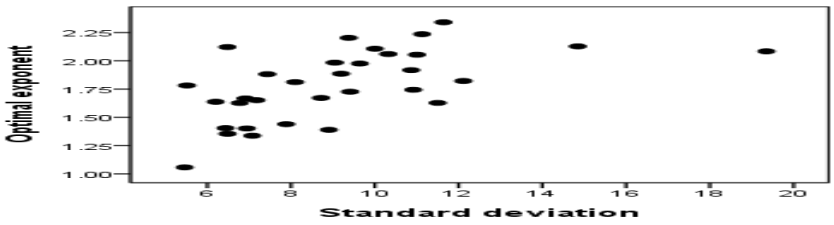

Figure 3.3 σ-optimal exponent sactter diagram

변수선택법으로 단계선택법을 사용하여 추정한 최종 모형은 식 (3.5)와 같이 표시되며, 첫번째 독립변수인 $\sqrt{\sigma}$의 계수에 대한 유의성 검정의 p값은 $p < 0.001$, 두번째 독립변수인 RPG의 계수에 대한 유의성 검정의 p값은 $p < 0.001$로 두 변수 모두 유의수준 1%에서 매우 유의한 것으로 나타났다.

$$\gamma_N = -0.468 + 0.388 \times \sqrt{\sigma} + 0.124 \times RPG \tag{3.5}$$

3.2. 여러 가지 지수 값에 대한 통계모형의 비교

Table 3.1은 1982년부터 2013년 한국프로야구 데이터와 여러 가지 피타고라스 지수 값들을 사용하여 구한 승률의 RMSE 값을 보여주는데, Method_O은 연도별로 추정한 최적지수값을 사용한 것이며, Method_T는 전체 데이터를 이용하여 추정한 단일최적지수 값인 1.823을 사용한 경우이며, Method_N은 본 연구에서 제안한 방법을 사용한 결과이다. 또 Method_D는 Davenport와 Woolner (1999)의 방법을 사용한 결과인데 한국프로야구의 경우, 식 (3.3)의 결과는 다음 식 (3.6)과 같은 결과로 나타났다.

$$\gamma_D = 1.212 \log(RPG) - 0.847 \tag{3.6}$$

그리고 Methods_S는 Smyth와 Patriot (2004)의 결과로 한국프로야구의 경우, 식 (3.4)의 결과는 다음 식 (3.7)와 같은 결과로 나타났다.

$$\gamma_S = RPG^{0.262} \tag{3.7}$$

Table 3.1 RMSEs of four optimal exponents

Method	RMSE
Method_O	2.7005
Method_T	2.9790
Method_D	2.9504*
Methods_S	2.9579
Method_N	2.8559**

그 결과 본 연구에서 제안한 γ_N를 이용한 방법인 Method_N가 상대적으로 효율성이 우수하다고 나타났고, 지수를 RPG만의 함수로 간주한 Davenport 등과 Smyth 등의 방법은 비슷한 결과를 제공하였으며, 단일지수 값의 경우가 가장 열등한 것으로 나타났다. Table 3.1에서 **는 가장 우수한 경우, *는 두 번째로 우수한 경우를 의미한다. 한편 지수 값의 선택이 승률의 추정에 어떤 의미가 있는 지를 확인하기 위하여 x를 $x = RS/(RS + RA)$로 두면, 승률은 식 (3.8)로 표현된다.

$$Wpct = \frac{x^\gamma}{x^\gamma + (1-x)^\gamma} \tag{3.8}$$

Figure 3.4는 실제로 한국프로야구의 경우, 최적지수 값의 최댓값인 1999년의 2.34와 최솟값인 1983년의 1.06, 전체 데이터를 이용하여 추정한 최적지수값 1.82를 이용한 세 가지 경우의 승률과 x의 산점도이다. 최적지수값 γ의 값이 클수록 승률에 x의 영향이 큰 것을 확인할 수 있다.

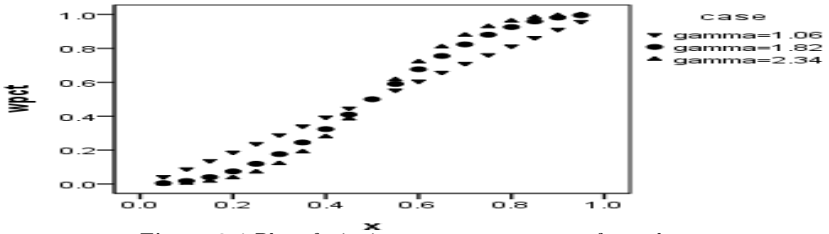

Figure 3.4 Plot of winning percentage versus x for each γ

4. 결론

오늘날 우리가 살고 있는 21세기는 모든 스포츠의 과학화가 급속도로 진행되고 있는 시대다. 야구에 있어서도 세이버메트리션들은 야구의 기록을 바탕으로 객관적인 분석을 하고 새로운 사실을 좀 더 명확하게 보여줄 수 있는 공식을 만들기 위해 노력하고 있다. 야구는 팀과 팀의 대결을 통해 승부를 결정지으며 승리는 상대 팀 보다 많은 점수를 얻는 것으로 결정된다. 득점이 많은 팀은 적은 팀보다 이길 확률이 높아지고 실점이 많은 팀은 적은 팀보다 패할 확률이 높아 질 것인데, 이를 이용하여 한 시즌 동안 얼마만큼의 승리를 거둘 수 있는지를 예측해 볼 수 있을 것이다. 이와 같은 철학을 바탕으로 완성된 공식이 야구의 승률을 계산하는 야구의 피타고라스 정리인데, 팀의 득점과 실점을 이용하여 승률을 계산함으로써 향후 팀의 경기 내용이 상승세 또는 하락세가 될지 예측할 수 있다. 본 연구에서는 세이버메트리션의 입장에서 야구의 피타고라스 정리에 필요한 최적지수 값을 정하는 문제를 한국프로야구에 적용하여 다루어 보았다. 제안된 방법은 기존의 연구들이 최적지수 값을 단지 RPG 만의 함수로 간주하였으나 승률의 표준편차가 추가로 고려됨으로써 RMSE의 관점에서 기존의 방법들보다 상대적으로 효율적이라고 할 수 있다.

본 연구의 결과와 관련된 제한점으로 기존의 연구들이 사용한 RMSE의 관점에서만 추정량들을 비교하여 좀 더 포괄적인 비교를 못한 점이 있으며, 한국프로야구 전 경기 데이터를 모두 사용하였기 때문에 이상치나 영향점을 제거하여 회귀분석을 적용하면 좀 더 나은 결과를 제공할 수 있을 것으로 간주되어진다. 또한 본 연구의 접근방법을 우리나라의 대표적인 프로스포츠에도 적용하여 피타고라스 지수 값을 추정하는 연구도 의미 있을 것으로 간주된다.

References

Acharya, R. A. (2006). Brief introduction to the Pythagorean theorem. *Harvard Sports Analysis Collective*, `http://www.hcs.harvard.edu/hsac/Resources/pythagorean.pdf`.

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`.

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

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. (2005). A study on runs evaluation measure for Korean pro-baseball players. *Journal of the Korean Data Analysis Society*, **7**, 2289-2302.

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. and Bang, S. Y. (2010). Forecasting attendance in the Korean professional baseball league using GARCH models. *Journal of the Korean Data & Information Science Society*, **21**, 1041-1049.

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.

Smyth, D. and Patriot. (2004). W% estimators, `http://gosu02.tripod.com/id69.html`.

Journal of the Korean Data & Information Science Society
2014, **25**(3), 493–499

http://dx.doi.org/10.7465/jkdi.2014.25.3.493
한국데이터정보과학회지

Estimation of exponent value for Pythagorean method in Korean pro-baseball[†]

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 3 February 2014, revised 19 February 2014, accepted 12 March 2014

Abstract

The Pythagorean won-loss formula postulated by James (1980) indicates the percentage of games as a function of runs scored and runs allowed. Several hundred articles have explored variations which improve RMSE by original formula and their fit to empirical data. This paper considers a variation on the formula which allows for variation of the Pythagorean exponent. We provide the most suitable optimal exponent in the Pythagorean method. We compare it with other methods, such as the Pythagenport by Davenport and Woolner, and the Pythagenpat by Smyth and Patriot. Finally, our results suggest that proposed method is superior to other tractable alternatives under criterion of RMSE.

Keywords: Pythagorean method, RMSE, winning percentage.

[†] The present research was conducted by the research fund of Dankook University in 2014.
[1] Professor, Department of Applied Statistics, Dankook University, Yongin 448-701, Korea. E-mail: jtlee@dankook.ac.kr